Queen Margaret University
EDINBURGH

Papers presented at

THE UNIVERSITY OF EDINBURGH

DiSS 2015

The 7th Workshop on Disfluency in Spontaneous Speech

8–9 August 2015
Edinburgh, Scotland

Workshop organised by
Robin Lickley

Papers compiled by
Robert Eklund

Table of contents

PITCH CHARACTERISTICS OF FILLED PAUSES

Malte Belz[1], Uwe D. Reichel[2]

[1] Humboldt-Universität zu Berlin, [2] Ludwig-Maximilians-Universität München
malte.belz@hu-berlin.de, reichelu@phonetik.uni-muenchen.de

ABSTRACT

We investigate the pitch characteristics of filled pauses in order to distinguish between hesitational and floor-holding functions of filled pauses. A corpus of spontaneous dialogues is explored using a parametric bottom-up approach to extract intonation contours. We find that subjects tend to utter filled pauses more prominently when they cannot see each other, which indicates an increased floor-holding usage of filled pauses in this condition.

Keywords: Disfluencies, Filled Pauses, Intonation, Floor-holding

1. TURN-HOLDING AND INTONATION

Filled pauses such as *uh* or *um* have been assigned different functionalities. In this study, we explore the pitch characteristics of filled pauses in spontaneous dialogues and whether different f_0 contours distinguish between different functions.

Different hypotheses try to account for hesitations. The floor-holding hypothesis [7, 8] claims that filled pauses are used to hold the turn during speaking. This seems to be more challenging in heated debates. Lallgee and Cook [6] confront interlocutors with a high pressure situation by means of a quickly interrupting confederate. However, this setting does not effect the frequency of filled pauses as opposed to low pressure situations. The authors suggest that speakers use other means for signalling turn-holding, for example by raising their voice.

A factor that might have an additional effect on the frequency of filled pauses is visual contact. Kasl and Mahl [5] note that filled pauses occur more frequently in situations where interlocutors cannot see each other. This indicates that some compensation processes take place in remote communication sitations, as speakers in natural face to face dialogues mark turn-holding by non-verbal cues as well [4]. When bereft of this modality, they might compensate not only by increasing the filled pause frequency, but also by using other verbal cues more often, e.g., a high f_0 level [19] (for Swedish and English).

Nevertheless, the mere frequency information of filled pauses might distort the picture, as the distribution of filled pauses is widespread and other explanations such as planning problems are possible. Therefore, a closer look at the intonation of filled pauses might help discriminate the filled pauses that might enable, and account for, turn-holding. We will explore the link between a rising fundamental frequency (f_0) and filled pauses, two features of spontaneous dialogues that are used for turn management. Thus, intonation patterns of filled pauses may help with identifying the functional implementations of filled pauses.

Previous research on the intonation of filled pauses implies that they show gradual downsteps of f_0 [10], at least within clauses [17]. This is in accordance with Tseng [18] , who states that filled pauses usually show flat intonational contours. This notion may also be described as showing only little or no prominence. We explore whether there are prominent contours for filled pauses and whether speakers make use of them to compensate in remote dialogue situtations. We will use a bottom-up approach for extracting the relevant intonation contours.

2. METHOD AND DATA

2.1. f_0 Preprocessing

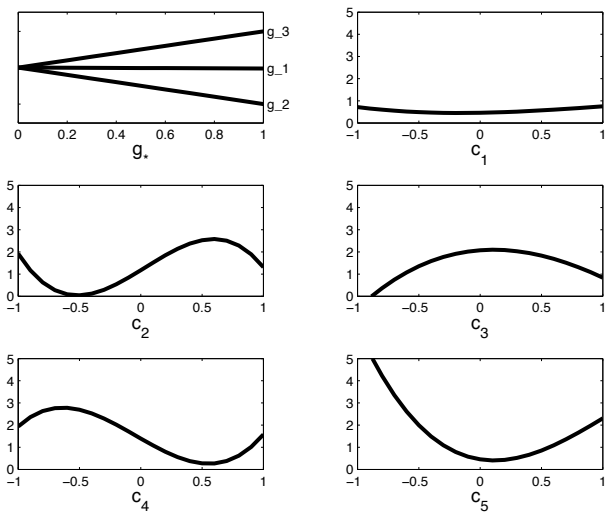
f_0 was extracted by autocorrelation (PRAAT 5.3.16 [3], sample rate 100 Hz). Voiceless utterance parts and f_0 outliers were bridged by linear interpolation. The contour was then smoothed by Savitzky-Golay filtering using third order polynomials in 5 sample windows and transformed to semitones relative to a base value [15]. This base value was set to the f_0 median below the 5th percentile of an utterance and serves to normalize f_0 with respect to its overall level.

2.2. Stylization

For intonation stylization we adopt the parametric CoPaSul approach of Reichel [13]. Within this framework, intonation is stylized as a superposition of linear global contours representing the f_0 baseline and third order polynomial local contours rep-

resenting the local pitch movements related to accent groups. The domain of global contours approximately related to intonation phrases is determined automatically by placing prosodic boundaries at speech pauses and punctuation in the aligned transcript. Within this domain, the baseline is fitted in a robust way through a sequence of f_0 medians of values below the 10th percentile [14]. The domain of local contours is determined by placing boundaries behind each content word determined by POS tagging [12]. Thus, these local contour domains roughly correspond to syntactic chunks [1] and generally contain at most one pitch accent. Additionally to this parametric analysis we carried out a kmeans clustering of the stylization coefficient vectors as in [9, 13] to describe the f_0 shapes also in categorical terms. The resulting contour classes are displayed in Fig. 1.

Figure 1: Global (g) and local (c_n) stylized intonation contour classes. The vertical axis represents semitones, the horizontal axis represents normalized time.



2.3. Data

Our database consists of GECO [16], a corpus of 48 German spontaneous mono- and multimodal dialogues (25 min each) of 13 women aged from 20 to 25 years who are strangers to each other. Only the multimodal condition allows subjects to see each other. Seven subjects participate in both conditions.

3. RESULTS

3.1. Frequencies

First, we test the hypothesis whether subjects produce more filled pauses in the monomodal condition (no visual cues) as predicted by the compensation hypothesis. The seven subjects that participate in both modality conditions, however, produce significantly less filled pauses in the monomodal condition, as computed by a one-sample t-test on the mean of the filled pauses frequencies ($\mu_0 = 0.5$, $t = -3.7$, $df = 6$, $p < 0.01$). Thus, frequencies do not confirm the compensation hypothesis. Nevertheless, the intonational features may still show such an effect.

3.2. Pitch characteristics

Table 1 shows that all subjects utter filled pauses with a higher proportion of steady intonation (global class g_1) in comparison to words with two to three characters ($t_{paired} = -16.1, df = 12, p < 0.001$). At first glance, it seems that the findings of [10] and [17] are corroborated, as 88% of filled pauses are uttered with a steady f_0 contour.

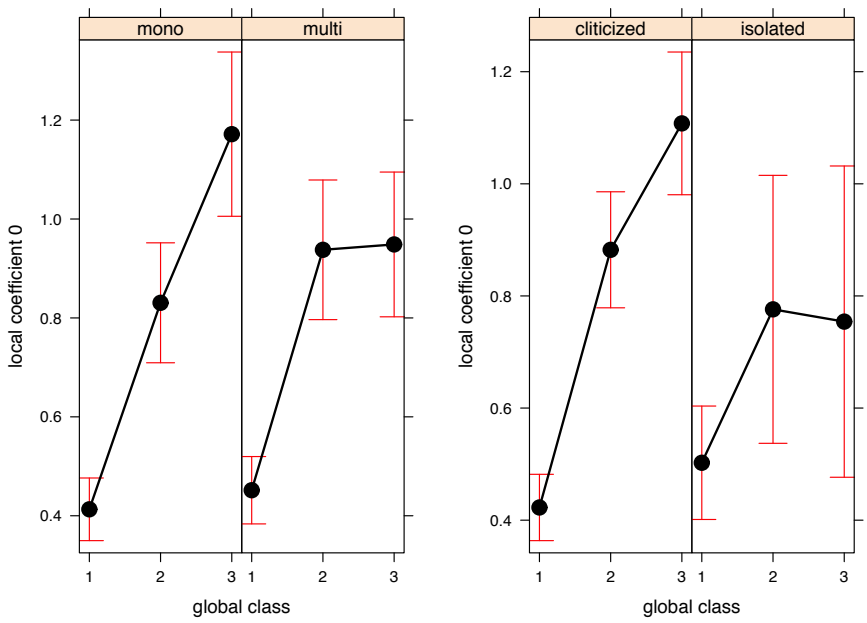
Table 1: Frequencies of global and local intonation classes, filled pauses and 2–3 character words.

Global class	1	2	3
Words	0.52	0.25	0.23
Filled pauses	0.88	0.07	0.05

Rising and falling f_0 baselines are reflected by global classes g_3 and g_2, respectively (cf. Fig 1). As every global class is described in more detail by overlaying local classes, we may now investigate prominence. Prominence is mainly modelled by the zeroth order polynomial coefficient of the local contour (coefficient 0). High values indicate a prominence-lending f_0 deviation from the baseline. To explore our data, we start with a full linear regression model of prominence, using the predictors modality (mono vs. multi), isolation (clitic vs. non-clitic), global class (steady vs. falling vs. rising) and their interactions as fixed effects and subjects as random effects in R [11] using the *lme4*-package (v.1.1-7). The isolation factor allows us to account for filled pauses in the context of a left and right silent pause, as opposed to cliticized filled pauses. We insert this factor as a fixed effect because its isolated position might lend the filled pauses more prominence *qua* isolation. We use the same cut-off for the left and right tail of the distribution, which removes 2.2% of the data. The model is reduced by a stepwise backward selection based on AIC [2].

Two two-way interactions remain in the model: global class interacts with modality ($\chi^2 = 6.8$, $df = 2, p = 0.03$) as well as with isolation ($\chi^2 = 9.0$, $df = 2, p = 0.01$), as confirmed by a log-likelihood

Figure 2: Effects of the multi- vs. monomodal and the cliticized vs. isolated predictor. The local coefficient 0 is used as a measure of prominence over global classes 1 (steady), 2 (falling) and 3 (rising).



test (cf. Fig. 2). The effect in the modality condition suggests that filled pauses are marked more prominently when interlocutors cannot see each other. The effect in the isolation condition does not hold for modality, but suggests that filled pauses in cliticized positions are marked more prominently than in isolated positions.

4. CONCLUSION

From these first exploratory findings we can conclude that filled pauses might in fact differ in their intonational appearances, and that these vary systematically. Instances in which filled pauses are uttered prominently—either with a rising or falling f_0 contour—indicate that these specific filled pauses can be used for holding the turn, especially so if interlocutors cannot see each other. Thus, our analysis implies that speakers are compensating the lack of visual contact in the monomodal condition with the help of intonation.

For further research, interesting insights to the functions and distributions of disfluencies may come from comparing the context of filled pauses showing a 'turn holding contour' to those of other instances. Additionally, annotating the data for turn-holding signals may reveal whether the detected instances are in fact perceived as holding the turn.

5. REFERENCES

[1] Abney, S. 1991. Parsing by chunks. In: Berwick, R., Abney, S., Tenny, C., (eds), *Principle-Based Parsing*. 257–278.

[2] Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-19, 716–723.

[3] Boersma, P., Weenink, D. 1999. PRAAT, a system for doing phonetics by computer. Technical report Institute of Phonetic Sciences of the University of Amsterdam. 132–182.

[4] Goodwin, C. 1981. *Conversational Organization*. Academic Press.

[5] Kasl, S. V., Mahl, G. F. 1965. Relationship of disturbances and hesitations in spontaneous speech to anxiety. *J Pers Soc Psychol* 1, 425–433.

[6] Lallgee, M. G., Cook, M. 1969. An experimental investigation of the function of filled pauses in speech. *Language and Speech* 12, 24–28.

[7] Maclay, H., Osgood, C. E. 1959. Hesitation phenomena in spontaneous English speech. *Word* 5, 19–44.

[8] Malandro, L. A., Barker, L., Barker, D. A. 1989. *Nonverbal Communication*. Random House.

[9] Möhler, G., Conkie, A. 1998. Parametric modeling of intonation using vector quantization. *Proc. 3rd ESCA Workshop on Speech Synthesis* 311–316.

[10] O'Shaughnessy, D. 1992. Recognition of hesitations in spontaneous speech. *Acoustics, Speech, and Signal Processing, ICASSP-92*. volume 1.

IEEE 521–524.
[11] R Core Team, 2012. R: A language and environment for statistical computing.
[12] Reichel, U. D. 2007. Improving data driven part-of-speech tagging by morphologic knowledge induction. *Proc. AST Workshop* 65–73.
[13] Reichel, U. D. 2014. Linking bottom-up intonation stylization to discourse structure. *Computer, Speech, and Language* 28, 1340–1365.
[14] Reichel, U. D., Mády, K. 2014. Comparing parameterizations of pitch register and its discontinuities at prosodic boundaries for Hungarian. *Proc. Interspeech* 111–115.
[15] Savitzky, A., Golay, M. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36(8), 1627–1639.
[16] Schweitzer, A., Lewandowski, N. 2013. Convergence of articulation rate in spontaneous speech. *Proc. Interspeech* 525–529.
[17] Shriberg, E. E., Lickley, R. J. 1993. Intonation of clause-internal filled pauses. *Phonetica* 50, 172–179.
[18] Tseng, S.-C. 1999. *Grammar, Prosody and Speech Disfluencies in Spoken Dialogues*. University of Bielefeld, Germany.
[19] Zellers, M. 2014. Duration and pitch in perception of turn transition by Swedish and English listeners. *Proc. from FONETIK* 41–46.

BOTH NATIVE AND NON-NATIVE DISFLUENCIES TRIGGER LISTENERS' ATTENTION

Hans Rutger Bosker[1], Jade Tjiong[2], Hugo Quené[2], Ted Sanders[2], and Nivja H. de Jong[2]

[1]Max Planck Institute for Psycholinguistics, Nijmegen; [2]UiL OTS, Utrecht University
HansRutger.Bosker@mpi.nl

ABSTRACT

Disfluencies, such as *uh* and *uhm*, are known to help the listener in speech comprehension. For instance, disfluencies may elicit prediction of less accessible referents and may trigger listeners' attention to the following word. However, recent work suggests differential processing of disfluencies in native and non-native speech. The current study investigated whether the beneficial effects of disfluencies on listeners' attention are modulated by the (non-)native identity of the speaker. Using the Change Detection Paradigm, we investigated listeners' recall accuracy for words presented in disfluent and fluent contexts, in native and non-native speech. We observed beneficial effects of both native and non-native disfluencies on listeners' recall accuracy, suggesting that native and non-native disfluencies trigger listeners' attention in a similar fashion.

Keywords: disfluencies, attention, non-native speech, Change Detection Paradigm.

1. INTRODUCTION

Disfluencies are "phenomena that interrupt the flow of speech and do not add propositional content to an utterance" [13], such as silent pauses, filled pauses (e.g., *uh* and *uhm*), slow speech, corrections, and repetitions. Despite their negative effects on listeners' impressions of the speaker's fluency level [5], disfluencies may have beneficial effects on the cognitive processes involved in speech comprehension, such as prediction and attention. For example, because disfluencies tend to occur before less accessible lexical items [19, 3, 18, 15], listeners may use disfluencies to predict more complex content to follow [1].

However, recent work suggests that there are differences in the way native and non-native fluency characteristics are processed (e.g., [7]). For instance, the beneficial effect of disfluencies on prediction may be attenuated when listening to a non-native speaker. That is, where native disfluencies may elicit prediction of low-frequency referents, disfluencies in non-native speech do not [6]. This attenuation has been argued to be due to variation in the distribution of disfluencies in native and non-native speech. Non-native speakers produce more disfluencies than native speakers and with a more irregular distribution [12, 22, 16]. As such, they are worse predictors of the word to follow, attenuating listeners' predictive strategies.

Apart from prediction effects, disfluencies are also known to trigger listeners' attention to the following word [9] as evidenced by higher recall accuracy of words heard in a disfluent context compared to a fluent context [11, 8, 14]. However, it is unknown whether the beneficial effects of disfluencies on listeners' attention are also modulated when listening to non-native speech. Using the Change Detection Paradigm [8], the current study compared how native and non-native disfluencies affect listeners' retention of words that were heard either in a fluent or a disfluent context (i.e., following a filled pause).

Given the finding that disfluency effects on prediction are attenuated in non-native speech [6], one may similarly hypothesize that effects of non-native disfluencies on attention will also be attenuated. Because native disfluencies introduce less accessible information, listeners may benefit from raising their attention as a precautionary measure to ensure timely comprehension of the unexpected information. However, non-native disfluencies follow a more irregular distribution, and, therefore, raised attention levels in response to non-native disfluencies may not prove advantageous to the listener. Therefore, the attentional effects of non-native disfluencies may be different from native disfluencies (e.g., attenuated).

Alternatively, the effects of disfluencies on attention have also been interpreted in terms of automatic cognitive consequences of temporal delay. The Temporal Delay Hypothesis [10] argues that temporal delay – inherent to disfluency – facilitates listeners' comprehension of the following content (e.g., better retention) by allowing more time to ori-

ent to the upcoming information. Following this hypothesis, native and non-native disfluencies would have similar effects on listeners' attention since they both delay the onset of the following word. The current study was designed to compare these two hypotheses.

2. METHOD

2.1. Participants

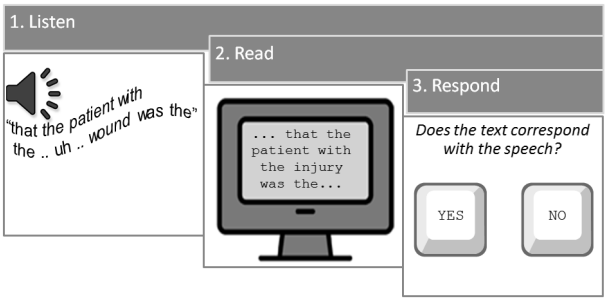
A sample of 80 native Dutch participants with normal hearing took part with implicit informed consent in accordance with local and national guidelines (M_{age}=23.3, SD_{age}=5.8, 11M/69F). Participants were randomly assigned to the native or non-native speaker condition.

2.2. Materials

A native speaker of Dutch was recorded (male, age=25) producing disfluent versions of 36 experimental story passages. These story passages were adopted from Collard [8] and consisted of three sentences. The passages were fluent except for a single filled pause (*uh*) preceding a target word. The speaker was instructed to speak as clearly as possible and to make the disfluencies sound as natural as possible. A highly proficient non-native speaker of Dutch (male, age=43, L1 Hebrew, LoR=13 years), reporting adequate knowledge of Dutch (self-reported CEFR level C1) and extensive experience with using Dutch in daily life, listened to the native recordings and subsequently imitated the native speech. Thus, matching native and non-native speech materials were obtained.

Fluent versions of the story passages were created by excising the filled pause from the disfluent version (at positive-going zero-crossings, using Praat [4]). If removing the disfluency led to an unnatural result, we instead inserted a disfluency into a fluent sentence, which was required for three native passages. Using this splicing method, fluent and disfluent versions of story passages were acoustically identical except for a filled pause appearing before a particular target word.

Figure 1: Schematical representation of the Change Detection Paradigm. Example of the Close Change condition.



2.3. Procedure

In our Change Detection Paradigm (schematically represented in Figure 1, adopted from [8]), participants listened to the fluent and disfluent passages, containing a particular target word, and then saw a written transcript of the passage. Their task was to indicate whether the transcript matched the spoken passage or not. The passages appeared in three conditions:

1. **No Change condition**:
   the transcript is identical to the spoken passage.
   (e.g., target word *wound* → *wound*)
2. **Distant Change condition**:
   the transcript contains one substitution involving a semantically unrelated noun.
   (e.g., *wound* → *handkerchief*)
3. **Close Change condition**:
   the transcript contains one substitution involving a semantically related noun.
   (e.g., *wound* → *injury*)

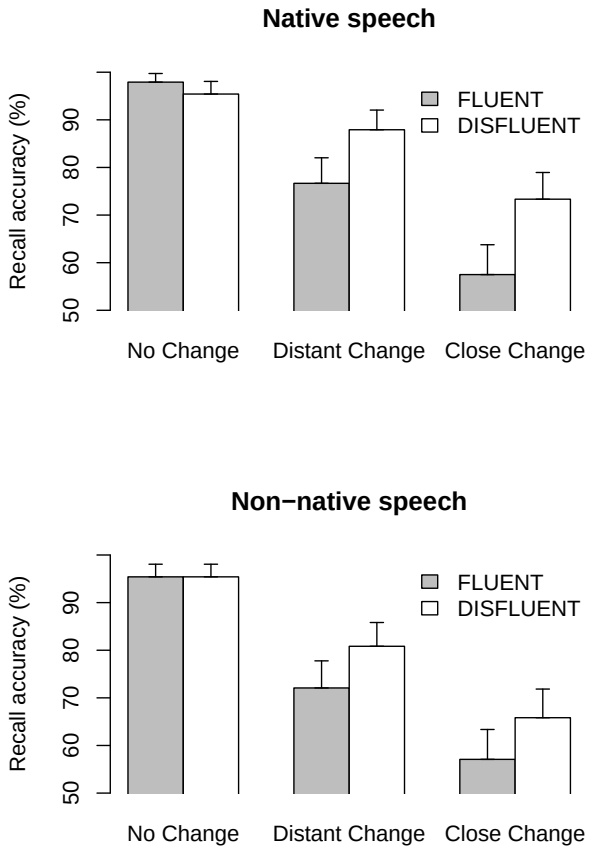
Target words from the three change conditions were matched in the log-transformed frequency of occurrence per million words, obtained from SUBTLEX-NL [17]. They were also matched in the number of characters. Target words always appeared halfway through the passage in a prepositional phrase that was out of focus.

To avoid the participants becoming accustomed to the co-occurrence of target words and disfluencies, 18 filler passages were included in the experiment which contained disfluencies without subsequent substitutions in other parts of the spoken passages. Trials were presented in pseudo-randomized order using a Latin-square design, such that all participants listened to all conditions without repeating passages. If the participant detected a substitution, he/she was asked to report the word from the audio passage that had been replaced. Finally, global accent ratings of both speakers were collected from participants using scales ranging from 1 (no accent) to 9 (very strong accent).

3. RESULTS

The accent ratings revealed a clear distinction between native and non-native speech (M_N=1.23, M_{NN}=8.08), indicating that participants clearly perceived a foreign accent in the non-native speech materials. For the comparison of recall accuracy for native vs. non-native speech, trials in which participants noticed a substitution but failed to provide the correct target word were coded as 'incorrect'. Overall recall accuracy is given in Figure 2 and was analyzed using a Generalized Linear Mixed Model (GLMM; [20]) as implemented in the `lme4` library [2] in R [21], with crossed random effects of Participants and Items. This GLMM included fixed effects of Nativeness (intercept: native speech), Disfluency (intercept: fluent speech), Change Condition (intercept: Close Change), and their interactions. This model revealed (1) an effect of Disfluency, showing a beneficial effect of disfluency on recall accuracy ($\beta = 0.86$, $z = 4.21$, $p < 0.001$), (2) effects of the different Change Conditions (No Change > Distant Change > Close Change), and (3) an interaction between Disfluency and No Change ($\beta = -1.81$, $z = -2.94$, $p = 0.003$), showing a smaller disfluency effect in the No Change condition, most likely due to a ceiling effect. However, no interactions were found between the factor Nativeness and any other predictor. The lack of interactions with the factor Nativeness indicates similar recall accuracy across native and non-native speech.

Figure 2: Mean recall accuracy in percentages. Error bars enclose 1.96 x SE, 95% CIs.



4. DISCUSSION

Results revealed that disfluencies have a beneficial effect on participants' recall accuracy. When our participants were presented with a transcript of an earlier spoken passage, they were more accurate in detecting a change in this text when the target word in the spoken passage had been preceded by a disfluency. This beneficial effect of disfluency was found for both native and non-native disfluencies, suggesting that both native and non-native disfluencies induce heightened attention to the following content.

These findings are in line with the Temporal Delay Hypothesis [10] arguing that the delay inherent to both native and non-native disfluencies allows listeners more time to orient to the upcoming information. However, one may consider an alternative explanation related to the perceived proficiency of our non-native speaker. Since our speech materials consisted of a variety of story passages with high lexical diversity and perfect grammatical accuracy, our non-native speaker produced relatively proficient Dutch speech. This may have indicated, to our listeners, a relatively high L2 proficiency. This, in turn, may have led listeners to treat non-native speech as similar to native speech. Future studies, manipulating perceived proficiency, may investigate how different (perceived) levels of L2 proficiency can affect the way non-native disfluencies are processed.

Alternatively, the absence of modulation of the disfluency effect for non-native speech may have been a result of our particular speech materials. Because we wanted to match the native and non-native speech as closely as possible, we used scripted passages (adopted from [8]). Listeners may have been aware that our speakers 'acted out' the story passages, thus preventing them from interpreting the non-native disfluencies as authentically different from the native disfluencies. Future experiments, involving spontaneous non-native speech materials and matched native counterparts, may shed light on the generalizability of the present findings.

Despite the fact that we cannot draw definitive conclusions about how non-native disfluencies affect listeners' perceptual mechanisms, our results, nonetheless, emphasize the role of attention in an account of disfluency processing. Disfluencies trigger listeners' attention with consequences for the retention of words following the disfluency.

5. REFERENCES

[1] Arnold, J. E., Hudson Kam, C. L., Tanenhaus, M. K. 2007. If you say -thee uh- you're describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33(5), 914–930.

[2] Bates, D., Maechler, M., Bolker, B. 2012. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-39.

[3] Beattie, G. W., Butterworth, B. L. 1979. Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech* 22(3), 201–211.

[4] Boersma, P., Weenink, D. 2012. Praat: doing phonetics by computer [computer program]. Version 5.3.18.

[5] Bosker, H. R., Quené, H., Sanders, T. J. M., De Jong, N. H. 2014. The perception of fluency in native and non-native speech. *Language Learning* 64, 579–614.

[6] Bosker, H. R., Quené, H., Sanders, T. J. M., DeJong, N. H. 2014. Native 'um's elicit prediction of low-frequency referents, but non-native 'um's do not. *Journal of Memory and Language* 75, 104–116.

[7] Bosker, H. R., Reinisch, E. 2015. Normalization for speechrate in native and nonnative speech. *Proceedings of the 18th International Congress of Phonetic Sciences 2015 [ICPhS XVIII], Glasgow*.

[8] Collard, P. 2009. *Disfluency and listeners' attention: An investigation of the immediate and lasting effects of hesitations in speech*. PhD thesis The University of Edinburgh.

[9] Collard, P., Corley, M., MacGregor, L. J., Donaldson, D. I. 2008. Attention orienting effects of hesitations in speech: Evidence from ERPs. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(3), 696–702.

[10] Corley, M., Hartsuiker, R. J. 2011. Why um helps auditory word recognition: The temporal delay hypothesis. *PLoS One* 6(5), e19792.

[11] Corley, M., MacGregor, L. J., Donaldson, D. I. 2007. It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition* 105(3), 658–668.

[12] Davies, A. 2003. *The native speaker: Myth and reality*. Clevedon: Multilingual Matters.

[13] Fox Tree, J. E. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language* 34(6), 709–738.

[14] Fraundorf, S. H., Watson, D. G. 2011. The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language* 65(2), 161–175.

[15] Hartsuiker, R. J., Notebaert, L. 2010. Lexical access problems lead to disfluencies in speech. *Experimental Psychology* 57(3), 169–177.

[16] Kahng, J. 2014. Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning* 64(4), 809–854.

[17] Keuleers, E., Brysbaert, M., New, B. 2010. SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods* 42(3), 643–650.

[18] Levelt, W. J. M. 1983. Monitoring and self-repair in speech. *Cognition* 14, 41–104.

[19] Maclay, H., Osgood, C. E. 1959. Hesitation phenomena in spontaneous English speech. *Word* 15, 19–44.

[20] Quené, H., Van den Bergh, H. 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language* 59(4), 413–425.

[21] R Development Core Team, 2012. R: A language and environment for statistical computing. ISBN 3-900051-07-0.

[22] Skehan, P. 2009. Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics* 30(4), 510–532.

DISFLUENCIES IN CHANGE DETECTION IN NATURAL, VOCODED AND SYNTHETIC SPEECH

Rasmus Dall[1], Mirjam Wester[1] & Martin Corley[2]

[1]The Centre for Speech Technology Research, The University of Edinburgh, UK

[2]School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, UK
r.dall@sms.ed.ac.uk, mwester@inf.ed.ac.uk, martin.corley@ed.ac.uk

ABSTRACT

In this paper, we investigate the effect of filled pauses, a discourse marker and silent pauses in a change detection experiment in natural, vocoded and synthetic speech. In natural speech change detection has been found to increase in the presence of filled pauses, we extend this work by replicating earlier findings and explore the effect of a discourse marker, like, and silent pauses. Furthermore we report how the use of "unnatural" speech, namely synthetic and vocoded, affects change detection rates. It was found that the filled pauses, the discourse marker and silent pauses all increase change detection rates in natural speech, however in neither synthetic nor vocoded speech did this effect appear. Rather, change detection rates decreased in both types of "unnatural" speech compared to natural speech. The natural results suggests that while each type of pause increase detection rates, the type of pause may have a further effect. The "unnatural" results suggest that it is not the full pipeline of synthetic speech that causes the degradation, but rather that something in the pre-processing, i.e. vocoding, of the speech database limits the resulting synthesis.

Keywords: change detection, filled pauses, speech synthesis

1. INTRODUCTION

Filled pauses (FPs) in naturally occurring spontaneous speech have received considerable attention and a variety of interesting phenomena have been found, such as faster reaction times [7, 8], faster word integration [3] and more accurate object identification [1].

This work explores the effect of filled pauses ('uh') in the context of "unnatural" speech, namely vocoded and synthetic speech, and compares it to the effects in natural speech. In other work we've explored the effects in various reaction time (RT) experiments [6, 14]. In these studies the same general tendency has been found. Vocoded speech generally mirrors natural speech effects, however no effects are found in synthetic speech except a generally slower RT in response to synthetic speech compared to the other types. While the reaction time experiments provide evidence that FPs affect people's on-line processing, FPs may have other, and longer term, effects. Change Detection [13] is a paradigm in which participants are asked to listen to short paragraphs of speech and are subsequently presented with the contents of the speech in writing. It is then the task of the participant to detect if a single change has occurred in the text as compared to the speech. This requires participants to not only process the speech as they hear it, but also to memorise it long enough to detect a change at a later point. Thus change detection, as opposed to reaction time, experiments provide a measure of the memorability of the speech in a slightly longer term context.

The basic effect reported by Collard [2] (Chapters 6 & 7), is that the presence of an FP prior to the changing word, as compared to fluent speech, increases the change detection rate by 10-15%. Collard [2] concludes that the acoustic quality of the FP is responsible for the effect (Chapter 7.6, pp. 128). His conclusion was based on manipulating silences around the FP but [12] has shown that a simple silent pause can make the same effect appear. Effects of silence are also found in related studies [4]. We therefore extend this work by including silent pauses and a discourse marker ('like') in natural speech to see if the effect is unique to FPs. Furthermore, as we are interested in the effects of "unnatural" speech types on listeners, we also perform the experiment using vocoded and synthetic speech. Vocoding, in speech synthesis, is the step of parametrising the speech in a manner suitable for statistical machine learning. This parametrised version can be re-formed directly by the vocoder, with some loss in quality, and this is what constitutes vocoded speech. Alternatively, a statistical model of the parametrised speech can be used to generate the speech, this is the method of synthesis applied in this paper.

The working hypothesis was that a similar pattern to the RT experiments would appear, in which the effect of disfluencies is present in natural and vocoded speech, but not in synthetic. This is motivated by the results of the prior experiments, but also by the assumption that current vocoding techniques do not degrade the quality of the speech in a way that would prevent the effect from appearing. It is possible however, that a different pattern will emerge due to the differences between the two paradigms. In RT experiments we are testing people's online monitoring and recognition of speech, whereas in change detection people are required to memorise the speech in order to detect the change at a later point. This means that even though participants may understand the speech, they may not be able to efficiently memorise it.

2. CHANGE DETECTION EXPERIMENTS

To perform the change detection experiments 43 short paragraphs, 20 critical, 20 filler and 3 practice, said by the same speaker in a spontaneous conversation were prepared. In each paragraph a target word was chosen and four alternative paragraphs were created. One where the target was preceded by an FP ('uh'), a silent pause (SP), the discourse marker 'like' (DM) or by nothing (i.e. fluent speech). The original paragraph was of one of these four cases, and the alternatives were made by altering the original by splicing out the segment immediately preceding the target word and splicing in the relevant replacement. The change word was a near-synonym or semantically related to the target word (i.e. the close-change condition of [2]). For the filler sentences no change existed, however a dummy target word was still chosen in front of which either an FP, SP or DM was placed. The paragraphs potentially included other FPs, DMs and SPs than the critical one so participants could not learn to use those as cues for the change. Two of the practice sentences contained no change and one a single change.

The vocoded versions were created taking the natural paragraphs and vocoding them using STRAIGHT [11], no further modifications to the audio was made. The synthetic utterances were made using HTS [15] and a good-quality state-of-the-art HMM-based voice trained on approximately 8 hours of speech. The transcripts of the paragraphs were used for the synthesis, and versions including a FP or DM was made by inserting these as words in the token stream, whereas the SP version was made in a similar way as in the RT experiments in [6], the length of the SP was thus similar to that of the FP.

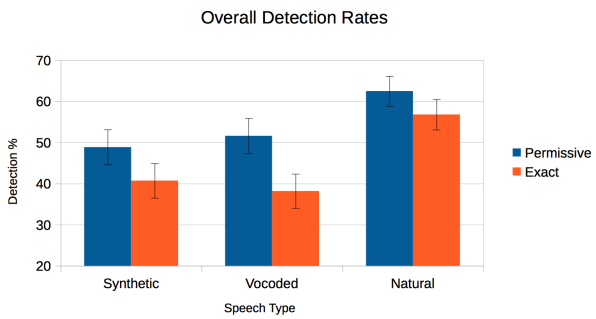


Figure 1: Detection rates per speech type. Permissive includes correct detection of change but incorrect identification. Exact does not.

2.1. Method

108 participants were recruited, 36 listened to natural speech only, 36 to vocoded and 36 to synthetic speech. Each participant only heard samples with either an FP, SP or DM such that for each type of speech and each type of pause there were 12 participants. Each participant listened to the practice sentences and then to each of the 40 paragraphs in a random order, of the 20 critical, half contained the appropriate form of pause, and the other half no pause (with 6 participants getting one set and other 6 the other set). In total this yielded 720 (36*20) critical evaluations per speech type and 240 (12*20) per condition (FP, SP or DM) within each speech style.

3. RESULTS

Due to an error in the experiment scripts 96 trials were invalid (4.4%) and were removed from the analysis. In 116 of the remaining trials (5.6%) participants correctly detected a change but incorrectly specified which change. In 16 of these the participant answered that the DM was the change which can arguably be considered correct. Therefore, two analysis were carried out - with (Exact) or without (Permissive) the exact specification of change. Notably however, the pattern of the results are identical. Please note that in the following analysis *disfluent* speech includes FPs, DMs and notably SPs, thus *fluent* speech is speech with none of these present.

A two-way ANOVA over the by-subject mean scores per condition was run. There was no overall effect of Disfluency Type (FP, DM, SP) or Disfluency Condition (Fluent or Disfluent), however a significant effect of Speech Type (Permissive: $F(2, 99)=5.917$, $p<0.005$, Exact: $F(2, 99)=10.377$, $p<0.0001$) was found and an interaction between Speech Type and Disfluency Condition for the Exact analysis ($F(2, 99)=5.180$, $p<0.01$) which was only marginal in the Permissive ($F(2, 99)=2.788$, $p=0.066$). Using Bonferroni correc-

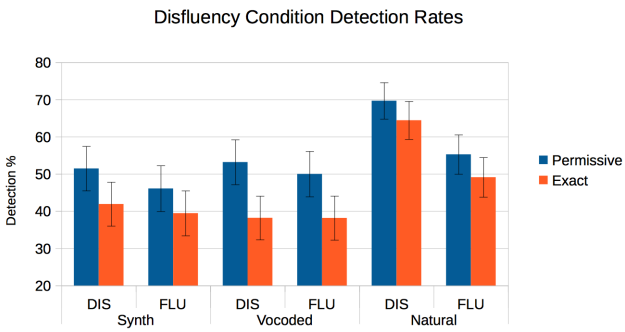


Figure 2: Detection rates divided by disfluency condition and speech type. DIS are disfluent conditions and FLU the fluent condition.

tion the effect of Speech Type is such that for the Natural Speech detection rates were significantly *higher* than Vocoded (Permissive: t(139)=2.692, p<0.05, Exact: t(140)=4.745, p<0.0001) and Synthetic (Permissive: t(142)=3.878, p<0.001, Exact: t(139)=4.699, p<0.0001), but no difference existed between Synthetic and Vocoded (Permissive: t(138)=0.870, p=1, Exact: t(133)=0.662, p=1), see Figure 1. That is, changes are generally detected better in natural speech than in synthetic (by 13.6% in the Permissive and 16.1% in the Exact case) and vocoded (by 10.9% in the Permissive and 18.6% in the Exact case).

The interaction effect (see Figure 2) was explored as it was significant in the Exact case and near significant in the Permissive. Using Bonferroni correction, there was no effect of disfluency condition in synthetic (Permissive: t(70)=1.374, p=0.521, Exact: t(70)=0.582, p=1) and vocoded speech (Permissive: t(70)=0.355, p=1, Exact: t(70)=0.075, p=1), however a significant effect was present in natural speech (Permissive: t(70)=3.326, p<0.005, Exact: t(70)=3.307, p<0.005). The presence of a disfluency did not have any effect on detection rates in synthetic and vocoded speech, however in natural they increased detection rates by 14.4% in the Permissive and 15.3% in the Exact case.

As disfluency had an effect in natural speech, individual tests for each disfluency type was run. Using Bonferroni correction a marginal effect of the FP was found in the permissive case (t(220)=2.356, p=0.058) which was significant in the exact (t(220)=2.468, p=0.043). For the DM a significant effect was found in the permissive case (t(223)=2.736, p=0.020) which was marginal in the exact (t(223)=2.3608, p=0.057). There was no effect of SP in the permissive case (t(236)=1.739, p=0.250) but a marginal effect in the exact (t(236)=2.234, p=0.079). See below discussion about this. Figures 3, 4 and 5 show individual detection rates for each disfluency type in each speech type.

4. DISCUSSION

Disfluent speech increases change detection rates in natural speech compared to fluent speech with no disruption. However, this is not the case in vocoded or synthetic speech (Figure 2).

In natural speech the FP and DM provides the larger and more significant benefit, while the contribution of a SP is less clear-cut (Figure 5). The results are, seemingly, in line with [2] who concludes that the acoustic quality of the FP is important in providing a benefit. While Collard investigated varying the length of SPs surrounding the FP he did not evaluate SPs on their own as done here. Our SP results are, however, different from those found by [12] in a very similar experimental setting. Their results are in line with the temporal delay hypothesis of [4] that it is simply the disruption which causes the increase in change detection rates. Something which, in a strict interpretation, is not supported by our results. While the tendency was for the SP to have lower detection rates than either FP or DM it did still increase detection rates (Permissive: 9%, Exact: 15%). It may be that the difference between the SP/DM and FP results is a consequence of our many tests and as such we have lost statistical power. That the effect appears with the DM can support both the hypothesis that it is the disruption which is the cause but also the idea that the use and purpose of DMs and FPs is similar (e.g. as seen in [9]). To determine which is more likely to be true using a non-speech condition as in [4] could be considered in future studies, besides replicating the SP experiment with a focus purely on natural speech.

Current Synthesis and Vocoding techniques do not produce speech for which the change detection results observed for natural speech are replicated (Figure 4 and 3). Where FPs, DMs and SPs increase the detection rate with 11-17% in natural speech there is no discernible pattern in synthetic and vocoded speech, rather, they tend to produce the same detection rates. Not only did the natural effect not appear, for both vocoded and synthetic speech the overall detection rate dropped as compared to natural speech by 11 to 18%. This is not just an effect of increased detections in the disfluency conditions of the natural speech, but rather an overall effect of the speech type. It is notable that this inability to replicate the effect occurs in *both* synthetic and vocoded, as the initial expectation was that current vocoding techniques were good enough to replicate the effect. That they do not suggests that it is not simply a matter of the speech prosody and

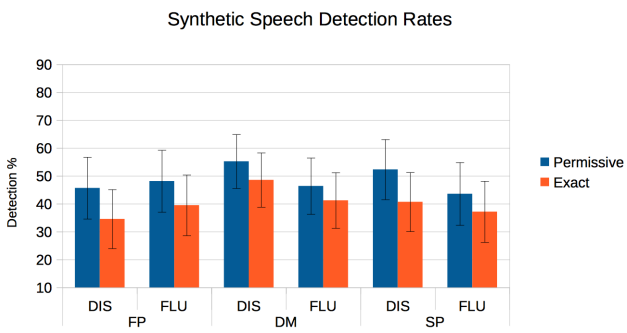


Figure 3: Detection rates per disfluency type for synthetic speech. FP = filled pause. DM = discourse marker. SP = silent pause. FLU = fluent. DIS = disfluent.

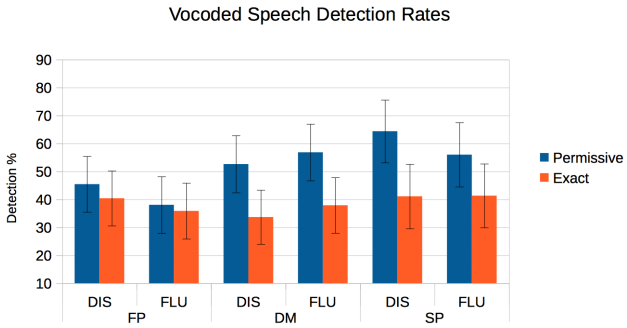


Figure 4: Detection rates per disfluency type for vocoded speech. FP = filled pause. DM = discourse marker. SP = silent pause.

general naturalness being poor, but rather that there is something about the inherent speech quality of the vocoder which limits synthetic speech in this regard.

In reaction time experiments we have found that vocoded speech [6, 14] elicits the same patterns as natural speech, which is in contrast to current results. Vocoding is known to introduce a buzzy character to the speech, while we are aware of the perceived naturalness of this [10], other possible psychological effects of this buzziness are unknown. It is possible that this demonstrates one of them. To detect a change the participant must necessarily be able to commit to (short term) memory what was being said in the paragraph in order to compare with the text later. Thus if the effect of vocoding decreases participants ability to memorize the salient elements of the paragraph, it should show an overall decrease in a participant's ability to detect changes, something which is the case. This decrease is likely due to an additional strain on the participant's cognitive resources and can also explain the lack of disruption/temporal delay effect. The participant must use so many resources to simply process the incoming speech stream that any potential benefit to be had from the disruption is lost. Following [2], the effect of disfluency found in natural speech is due to heightened attention to the target word, resulting

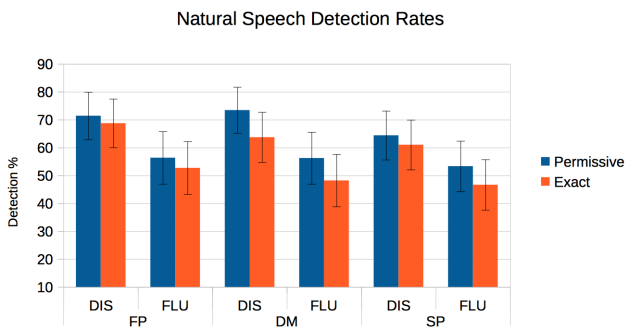


Figure 5: Detection rates per disfluency type for natural speech. FP = filled pause. DM = discourse marker. SP = silent pause.

in better recall and notice of changes. While durational and prosodic cues may still be present after vocoding, if the participant is already straining their cognitive resources to simply understand and commit the content to memory, it is likely that these cues do not result in an attentional shift. This is, however, speculative and further experimental evidence would be needed. Experiments explicitly manipulating the cognitive strain on participants, such as dual-attention tasks, could be used in combination with a change detection paradigm using natural speech, if this alters the results for natural speech to look similar to those of vocoded and synthetic speech it would provide evidence for a cognitive strain hypothesis.

5. CONCLUSION

We have shown that disfluent speech increase change detection rates in natural speech, but that this effect is not present in either vocoded or synthetic speech. Our vocoding results are in contrast to [6, 14] where the effect appears. The SP results seemingly support [2] and could be interpreted against the temporal delay hypothesis of [4]. As our results differ from [12] we have cautioned that this may be due to our high number of tests and given suggestions for further work which may resolve these tensions, including using a non-speech condition and a dual-attention paradigm. All research data associated with this paper can be found at Edinburgh DataShare [5] (http://hdl.handle.net/10283/808).

6. ACKNOWLEDGEMENTS

Thanks to Amelie Osterrieth and Anisa Jamal for help collecting the natural speech data. This work has been partially funded by the JSTCREST uDialogue project and EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). The NST research data collection may be accessed at http://datashare.is.ed.ac.uk/handle/10283/786.

7. REFERENCES

[1] Brennan, S. Feb. 2001. How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language* 44(2), 274–296.

[2] Collard, P. 2009. *Disfluency and listeners' attention: An investigation of the immediate and lasting effects of hesitations in speech.* PhD thesis University of Edinburgh.

[3] Corley, M., Hartsuiker, R. J. 2003. Hesitation in speech can. . . um. . . help a listener understand. *Proceedings 25th meeting of the Cognitive Science Society* Boston, USA.

[4] Corley, M., Hartsuiker, R. J. Jan. 2011. Why um helps auditory word recognition: the temporal delay hypothesis. *PloS one* 6(5), e19792.

[5] Dall, R., Wester, M., Corley, M. Experiment materials for "disfluencies in change detection in natural, vocoded and synthetic speech.", [dataset]. University of Edinburgh, School of Informatics, Centre for Speech Technology Research, http://dx.doi.org/10.7488/ds/274.

[6] Dall, R., Wester, M., Corley, M. 2014. The Effect of Filled Pauses and Speaking Rate on Speech Comprehension in Natural, Vocoded and Synthetic Speech. *Proceedings Interspeech* Singapore, Singapore.

[7] Fox Tree, J. E. 1995. The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language* 34(6), 709–738.

[8] Fox Tree, J. E. 2001. Listeners' uses of um and uh in speech comprehension. *Memory and Cognition* 29(2), 320–326.

[9] Fox Tree, J. E., Schrock, J. C. Feb. 1999. Discourse Markers in Spontaneous Speech: Oh What a Difference an Oh Makes. *Journal of Memory and Language* 40(2), 280–295.

[10] Henter, G. E., Merritt, T., Shannon, M., Mayo, C., King, S. September 2014. Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech. *Proceedings Interspeech* Singapore, Singapore. 1504–1508.

[11] Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A. 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication* 27(3-4), 187–207.

[12] Sanford, A. J. S., Molle, J. August 2006. Disfluencies and selective attention in speech processing. *Proceedings AMNLP* Turku, Finland. 183.

[13] Sturt, P., Sanford, A. J., Stewart, A., Dawydiak, E. 2004. Linguistic focus and good-enough representations: An application of the change-detection paradigm. *Psychonomic Bulletin and Review* 11(5), 882–88.

[14] Wester, M., Corley, M., Dall, R. 2015. The temporal delay hypothesis: natural, vocoded and synthetic speech. *Proceedings DiSS* Edinburgh, UK.

[15] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., Tokuda, K. 2007. The HMM-based Speech Synthesis System Version 2.0. *Proceedings SSW* Bonn, Germany. 294–299.

Uh – I forgot what I was going to say: How memory affects fluency

Stephanie Don and Robin Lickley

CASL Research Centre, Speech and Hearing Sciences
Queen Margaret University, Edinburgh, UK.

ABSTRACT

Disfluency rates vary considerably between individuals. Previous studies have considered gender, age and conversational roles amongst other factors that may affect fluency. Testing a non-clinical, gender-balanced population of young adults performing the same speaking tasks, this study explores how inter-speaker variations in working memory and in long-term (lexical) memory affect disfluency in two different ways. Working memory was tested by a forward digit span test; long-term lexical memory was tested by the Verbal Fluency Test, both semantic and phonological versions. In addition, each participant provided 3 one-minute samples of monologue speech. The speech samples were analysed for disfluencies. Speakers with lower working memory scores produced more error repairs in running speech. Speakers with lower lexical access scores produced a higher rate of hesitations. The two types of memory affected fluency in different ways.

Keywords: hesitation, error repair, working memory, long term lexical memory.

1. INTRODUCTION

Speech production involves the rapid interaction of several levels of planning processes and muscular movements. It is not surprising that things can and do go wrong, so that the flow of speech is interrupted. This interruption to the flow of speech we refer to as *disfluency*. Let us take two examples of things that can go wrong during production. First, at the lexical level, some words are harder to access than others and words that are less familiar to the speaker (for example, less frequent words) can take longer to access [1]; words with lower name-agreement and therefore more competitors for selection also present problems to the speaker [2]. As Hartsuiker and Notebaert showed, difficulties with lexical access can lead to disfluent hesitation [2]. Once words have been selected, speakers have to retain the phonological components of a word form in working memory until the word is articulated [3]. This process can lead to errors. If two word forms are active in working memory at the same time, there is potential for phonological errors, such as anticipations, perseverations or full exchanges, to appear when the words are articulated. Such a relationship has been observed in an experiment designed to elicit speech errors [4] as well as in unpublished experiments in our laboratory. Where there is error, there is usually repair, when the speaker detects and corrects the error. So when things go wrong during the highly complex production processes, disfluency often ensues, with hesitation-type disfluencies (silent or filled pauses, prolongations, repetitions) where there is uncertainty over the next word, for example, and with repair-type disfluencies where errors have occurred (see [5] for more detailed discussion).

Considerable inter-speaker variability in rates of disfluency can be observed in corpus studies of spontaneous speech (e.g., [6-8]). Some studies have speculated about male:female differences in fluency, suggesting, for example, that males tend to be more hesitant than females [6,7,9] (though see [10]). A weak relationship between fluency and the speaker's age has been shown by Bortfeld et al. [7], with older adults being slightly less fluent than younger. Aside from characteristics of the speaker, there is evidence that fluency clearly varies with the difficulty of the speaking task that the speaker is performing. When speakers are in a more dominant role in a dialogue, for example as instruction-givers in a map task scenario, then they are likely to have more strategic planning to do and their cognitive load is greater. As a result, they produce more disfluencies in general, with both more hesitations and more repair-type disfluencies [7,11]. So there are differences in fluency levels associated with characteristics of the speaker, like their sex and age, and differences associated with the difficulty of the tasks that speakers perform.

However, none of this accounts for the variability or the individual differences in fluency that we can observe between speakers of the same sex and age performing the same tasks. It is known that typical speakers vary considerably in their ability to access words in the lexicon [12] and in their performance in tests of working memory [13]. The questions in the present study are whether such differences in long term (lexical) memory and in working memory within a typical student population will relate to differences in fluency in running speech.

To this end, we designed a small study with three parts: A spontaneous speech task, which would allow us to assess fluency by measuring the frequency of hesitations and repairs; A task that assesses speakers' ability to access their lexicon; A task that assesses speakers' working memory.

In order to measure lexical access, we employ the so-called *Verbal Fluency Test* (VFT), which has for decades played an important role in neuropsychological testing, both in clinical situations and in research [14]. In VFT, participants are asked to name as many words as they can of a given category within a given time frame (usually 60 seconds). There are two general versions of this test. In the phonemic version, the participant is asked to name as many words as possible that begin with a certain letter or sound, typically F, A or S. In the semantic version, a semantic category is given, most commonly *animals*. Usually, the measure taken in the test is simply the number of correct members of the given category that the participant names within 60 seconds [14].

For working memory, we employed the very simple, frequently-used task of testing forward digit span, where participants are asked to repeat back to the experimenter strings of digits of increasing length.

Since we expect that people who have more trouble accessing their lexicon will also hesitate more in spontaneous speech [2], we hypothesised that a negative correlation would be found between VFT scores and hesitation rates. No association was expected between VFT and rates of repair-type disfluencies.

Since we expect that people with weaker working memory capacity will also produce more errors during the speech production process [4], we hypothesised that a negative correlation would be found between forward digit span scores and the frequency of repair-type disfluencies. No association was expected between forward digit span scores and rates of disfluent hesitations.

2. METHOD

2.1. Participants and data collection

Participants were 20 native speakers of British English, including 10 females and 10 males, aged between 20 and 25 years. So the group was balanced for sex and participants represented a narrow age range. Participants reported no hearing impairments and no visual impairment that could not be corrected by wearing glasses. In addition, participants had no history of speech and language therapy assessment or intervention. All participants were students in higher education.

Ethical permission was given by the Research Ethics Committee at Queen Margaret University, Edinburgh

There were three parts to this study: We tested (1) working memory, using a digit span task (DST); (2) lexical access via 3 versions of the Verbal Fluency Test (VFT); (3) disfluency in running speech from 3 minutes of monologue. Digital audio recordings were made of all parts of the experiment

In the Digit Span Task, participants were presented with audio recordings of strings of digits (1-9), over headphones. They were asked to repeat the string that they had heard, verbatim. Participants heard a string of 5 digits first. If this was successfully recalled, a string of 6 digits was presented. If a participant failed to repeat the latest string correctly, another string of the same length was played. If this was repeated successfully, a longer string was presented. If the participant failed twice on a given string length, the test was stopped. This process continued up to a maximum string length of 10 digits. The score for this test equalled the number of digits in longest string successfully recalled.

In the Verbal Fluency Test, participants were given three lexical access tasks. In each task, participants were asked to name as many words as they could of a given category, within 60 seconds. The tasks were timed with a digital stopwatch. In two of the tasks, semantic categories were given: *Animals* and *Things that are yellow*. In the third task, participants were asked to name words beginning with the letter *B*. The score for each test was simply the number of real English words produced that correctly belonged in the category produced within the 60 seconds allowed.

Finally, participants were asked to speak on a given topic for 60 seconds, and to try to speak continuously, avoiding long pauses. Three different topics were given, and each participant spoke on all three topics: *My best night out; My favourite place in the world (and why); My biggest achievement.* The recordings were orthographically transcribed. These transcriptions were then perceptually analysed for filled pauses (uh, um, eh, em), silent pauses (judged subjectively), prolongations (judged subjectively), repetitions and repair disfluencies: substitutions, insertions and deletions as defined by the HCRC Disfluency Coding Manual [15] and in [5]. The participants' scores on this task were the frequencies (per 100 words) of these disfluencies.

2.2. Statistical Analysis

Pearson product-moment correlation tests were used to test for associations between the two memory tasks (Digit Span, Verbal Fluency Test) and disfluency rates in running speech.

3. RESULTS

All 20 participants completed all parts of the experiment.

Forward DST scores varied from 5 to 9 (Mean = 6.45, SD = 1.27).

As expected, given previous findings [12], participants produced different numbers of words, in each of the 3 VFTs, with the most words for the semantic category *Animals* (Mean = 26.05, SD = 8.02), the fewest for *things that are yellow* (Mean = 11.55, SD = 5.16) and slightly more in the phonological test *Words beginning with B* (Mean = 15.3, SD = 5.25).

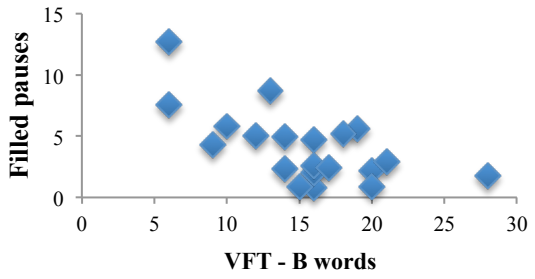
Speakers' disfluency rates varied from 7.6 to 29.4 per 100 words (Mean = 14.65, SD = 5.3) (note that including subjective judgments of silent pause and prolongation, rather than omitting them as most studies do, will lead to higher overall rates of disfluency).

As expected, DST scores did not correlate with VFT overall, but, unexpectedly, a strong positive correlation was observed between digit span and word count in the VFT for *things that are yellow* (*Pearson's r* = .607, n = 20, p = .002).

3.1. Verbal Fluency Test and Hesitation rates

We hypothesised that the VFT scores would correlate negatively with speakers' hesitation rates in running speech, and they did (Overall VFT compared to rate of Filled Pauses: *Pearson's r* = -.587, n = 20, p = .006). However, there was a difference between the three categories used in the VFT, and the strongest correlation was between results for the phonemic test (words beginning with B) and filled pause rate (r = -.683, n = 20, p = .002, Figure 1). The second strongest correlation was between the scores for the semantic category *animals* (r = -.513, n = 20, p = .021), and there was no significant correlation for the category *things that are yellow*.

Figure 1: Number of words beginning with "B" named in 60 seconds plotted against filled pauses per 100 words in 3 minutes of free speech.

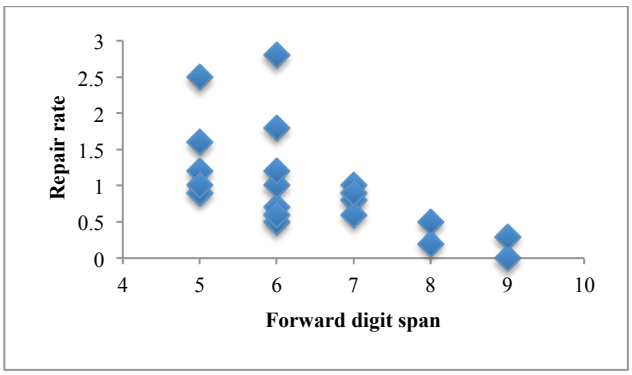


There was no significant correlation between VFT scores and repair-type disfluency rates.

3.2 Digit Span and Repair rates

We hypothesised that the DST scores would correlate negatively with speakers' repair-type disfluency rates in running speech, and they did. We assumed that substitution, insertion and deletion type disfluencies represented error repairs (following [5]). As hypothesised, a strong negative correlation was found between rates of substitutions, insertions and deletions combined (*r* = -.615, n = 20, p = .004, Figure 2). While substitutions (*r* = -.429, n = 20, p = .03) and deletions (*r* = -.425, n = 20, p = .031) both correlated negatively with DST, the correlation of DST with insertions was not significant (*r* = -.284).

Figure 2: As forward digit span increases, error repair rate falls



There was no significant correlation between forward digit span scores and rates of filled pauses (*r* = -.306, n = 20, p = .19).

4. DISCUSSION

This study provides initial evidence that individual differences in two aspects of disfluency may be attributed in part to differences in two aspects of memory. Speakers with more effective access to the long-term memory storage referred to as the lexicon produced fewer filled pauses in running speech. Speakers with stronger performance in the working memory task of digit span produced fewer error-repairs. There was no correlation between the scores for the lexical access test (except in one non-standard version) and the digit span task, so no association between the long-term memory task and the working memory task. Our initial hypotheses, based on previous research (e.g., [2,4]), are supported by this set of data.

Since the disfluent behaviours observed in our samples running speech are so strongly associated with results of tests typically used in clinic, even with a high functioning typical population, it is interesting to speculate over the extent to which changes in fluency in running speech may be used diagnostically to detect the onset of neurological conditions that affect memory and speech.

However, the study was relatively small, with a low number of participants and a small sample of running speech for each participant, consisting of just 3 minutes of monologue. So, despite the strength of the correlations and the significance levels in the results, the results need to be treated with caution. Despite the limitations, the results suggest that further study in this area is warranted. A current project employing larger number of participants, each producing a larger sample of running speech, and subject to a broader range of tests of memory will shed further light on the issues raised by this study.

NOTE

This work was completed as part of Stephanie Don's Honours Dissertation, contributing to her BSc Honours degree in Speech and Language Therapy at Queen Margaret University, Edinburgh. Address for correspondence is rlickley@qmu.ac.uk

5. REFERENCES

[1] Jescheniak JD, Levelt WJ. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. Journal of Experimental Psychology-Learning Memory and Cognition 1994;20(4):824-43.

[2] Hartsuiker RJ, Notebaert L. Lexical access problems lead to disfluencies in speech. Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie) 2010;57(3):169-77.

[3] Levelt WJM. Speaking: from intention to articulation. Cambridge, Massachusetts: The MIT Press; 1989.

[4] Saito S, Baddeley AD. Irrelevant sound disrupts speech production: Exploring the relationship between short-term memory and experimentally induced slips of the tongue. Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology 2004;57(7):1309-40.

[5] Lickley RJ. Fluency and Disfluency. In: Redford M, editor. The Handbook of Speech Production: Wiley-Blackwell; 2015, p. 445-469.

[6] Shriberg E. To 'errrr' is human: ecology and acoustics of speech disfluencies. Journal of the International Phonetic Association 2001;31(1):153-69.

[7] Bortfeld H, Leon SD, Bloom JE, Schober MF, Brennan SE. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. Lang.Speech 2001;44(2):123-47.

[8] Branigan H, Lickley RJ, McKelvie D. Non-linguistic influences on rates of disfluency in spontaneous speech. Proceedings of the 14th International Congress of Phonetic Sciences San Francisco, California; 1999.

[9] Binnenpoorte D, van Bael C, den Os E, Boves L. Gender in everyday speech and language: a corpus-based study. INTERSPEECH; 2005.

[10] Ian R. Finlayson. Testing the roles of disfluency and rate of speech in the coordination of conversation. PhD Thesis: Queen Margaret University, Edinburgh; 2014.

[11] Lickley R. Dialogue moves and disfluency rates. ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech; 2001.

[12] Gladsjo JA, Schuman CC, Evans JD, Peavy GM, Miller SW, Heaton RK. Norms for letter and category fluency: demographic corrections for age, education, and ethnicity. Assessment 1999 Jun;6(2):147-78.

[13] Jarrold C, Towse JN. Individual differences in working memory. Neuroscience 2006;139(1):39-50.

[14] Tombaugh TN, Kozak J, Rees L. Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. Archives of Clinical Neuropsychology 1999;14(2):167-77.

[15] Lickley RJ. HCRC disfluency coding manual. 1998;TR-100.

Neural correlates of the processing of unfilled and filled pauses

Robert Eklund[1], Peter Fransson[2], Martin Ingvar[2,3]

[1] Department of Culture and Communication, Linköping University, Linköping, Sweden
[2] Department of Clinical Neuroscience, Karolinska Institute, Stockholm, Sweden
[3] Osher Center for Integrative Medicine, Karolinska Institute, Stockholm, Sweden
robert.eklund@liu.se, peter.fransson@ki.se, martin.ingvar@ki.se

ABSTRACT

Spontaneously produced Unfilled Pauses (UPs) and Filled Pauses (FPs) were played to subjects in an fMRI experiment. While both stimuli resulted in increased activity in the Primary Auditory Cortex, FPs, unlike UPs, also elicited modulation in the Supplementary Motor Area, Brodmann Area 6. This observation provides neurocognitive confirmation of the oft-reported difference between FPs and other kinds of speech disfluency and also could provide a partial explanation for the previously reported beneficial effect of FPs on reaction times in speech perception. The results are discussed in the light of the suggested role of FPs as floor-holding devices in human polylogs.

Keywords: speech disfluency, filled pauses, unfilled pauses, speech perception, spontaneous speech, fMRI, Auditory Cortex, PAC, Supplementary Motor Area, SMA, Brodmann Area 6, BA6

1. INTRODUCTION

A characteristic of spontaneous spoken language is that almost no one is completely fluent, and the most common voiced disfluency is the filled pause, “eh”. The reported average frequency of filled pauses (FPs) ranges from 1.9 to 7.6% (Eklund, 2010).

While a commonly expressed view regards speech disfluency as errors in speech production, there are several studies that indicate that certain kinds of disfluencies can have beneficial effects on listener perception (Fraundorf & Watson, 2011; Barr & Seyfeddinipur, 2010; Ferreira & Bailey, 2004; Fox Tree, 2001, 1995; Reich, 1980).

While several behavioral studies of speech disfluency have been carried out over the years, only a few strictly neurocognitive studies have been performed, mostly using electrophysiological methods and/or scripted or enacted disfluencies. Moreover, most of these studies have focused on the effect of speech disfluency on higher-level speech processing, like syntactic parsing. The present study uses functional Magnetic Resonance Imaging (fMRI) to analyse the effect of authentic disfluencies proper to study the effect of unfilled and filler pauses on brain processing.

2. METHOD

2.1. Subjects

The subjects were 16 healthy adults (9 F/7M) ages 22–54 (mean age 40.3, standard deviation 9.5) with no reported hearing problems. All subjects were right-handed as determined by the Edinburgh Handedness Inventory (Oldfield, 1971). All subjects possessed higher education. After a description of the study, including a description of fMRI methodology, written and informed consent was obtained from all subjects. A small participation remuneration was also administered.

2.2. Equipment

The fMRI scanner used was the General Electric 1.5T Excite HD Twinspeed scanner at Karolinska Institute, MR-center, Stockholm, Sweden. The coil used was a General Electrics Standard bird-cage head-coil (1.5T).

2.3. Stimulus data

The stimulus data used were excerpts from the human–human dialog speech data described in Eklund (2004: 187–190). Subjects were asked to play the role of travel agents listening to customers making travel bookings over the telephone, following a task sheet (Eklund, 2004: 185)

From the original data set, four speakers were chosen (2M/2F) and a number of sentences were excised that were fluent except that they included a number of UPs and an approximately equal number of FPs. The resulting number of both UPs and FPs roughly corresponded to reported incidence of UPs and FPs in spontaneous speech.

Stimulus data are shown in Table 1.

Table 1. Stimulus data. Legend: UPs = Unfilled Pauses; FPs = Filled Pauses; MIT = Mean Interstimulus Time in seconds.

Stimulus File	No. UPs / MIT	No. FPs / MIT
1	17 / 11.9 s	23 / 7.1 s
2	9 / 9.7 s	8 / 13.8
3	10 / 5.5 s	9 / 8.7 s
4	22 / 7.2 s	15 / 10.7 s

2.4. Experimental design

The stimulus files described above were used in an event-related experiment. After initial localizer anatomical scanning sessions, the four stimulus files (M/F/M/F) were played. During the intermissions the subjects were briefed whether they were still awake/focused on the task. Interstimulus intervals were of sufficient duration so as to allow for reliable BOLD acquisition. FPs and UPs were modelled as events in SPM and were convolved using the Haemodynamic Response Function (HRF) in SPM.

2.5. Experimental setting

The subjects lay supine/head first in the scanner with earplugs to protect them from scanner noise and headphones with the sound data played to them. The perceived sound level was quite sufficient and no subjects reported having any problems hearing what was said. Head movement was constrained using foam wedges and/or tape.

2.6. Experimental instructions

The subjects were instructed to listen carefully to what was said, as if they were the addressed travel agent, but that they were not expected to react verbally to the utterances or say anything, only that they needed to pay attention to the information provided by the clients. All subjects understood the instructions without any confusion.

2.7. Post-experiment interview

After the scanning, all subjects were interviewed in order to confirm that they had been awake and focused during the experiment. A self-rating scale of how attentive the subjects felt they had been during the sessions was also used. All subjects reported that they had been attentive at a satisfactory level.

2.8. MRI scans

For each subject, a T1-weighted coronal spoiled gradient recalled (SPGR) image volume was obtained to serve as anatomical reference (TR/TE= 24.0/ 6.0 ms; flip angle 35°; voxel size = $1 \times 1 \times 1$ mm^3). Moreover, BOLD-sensitized fMRI was carried out by using echo-planar imaging EPI using 32 axial slices (TR/TE=2500/40 ms, flip = 90 deg, voxel size = $3.75 \times 3.75 \times 4$ mm^3).

In total, T2*-weighted images were collected four sessions: (3m30s/80 volumes; 2m22s/53 volumes; 1m33s/33 volumes; 3m05s/70 volumes).

2.9. Post-processing

The images were post-processed and analyzed using MatLab R2007a and SPM5 (Friston et al., 2007). The images were realigned, co-registered and normalized to the EPI template image in SPM5 and finally smoothed using a FWHM (Full-Width Half Maximum) of 6 mm. Regressors pertaining to subject head movement (3 translational and 3 rotational) were included as parameters of no-interest in the general linear model at the first level of analysis. No subjects were excluded due to head motion or for any other image acquisition related causes. Analyses were also carried out using the SPM Anatomy Toolbox (Amunts, Schleicher & Zilles, 2007; Eickhoff et al., 2007, 2006, 2005).

3. ANALYSES AND RESULTS

Using Fluent Speech (FS) as the baseline condition, the following three contrasts were analyzed:

(1) Filled Pauses > Fluent Speech

(2) Unfilled Pauses > Fluent Speech

(3) Filled Pauses > Unfilled Pauses

The results were calculated with a False Discovery Rate (FDR) at $p < 0.05$ (Genovese, Lazar & Nichols, 2002) with a cluster level threshold of 10 contiguous voxels.

First, no activation in BA22, associated with semantic processing, was observed.

For **FP > FS**, increased activation was found in Primary Auditory Cortex (PAC) (Morosan et al., 2001; Rademacher et al., 2001) bilaterally, and in subcortical areas (cerebellum, putamen) and most interestingly in the Supplementary Motor Area (SMA), Brodmann Area 6 (BA6). Activation was also observed in the Inferior Frontal Gyrus (IFG). Typical BA6 modulation is shown in Figure 1.

For **UP > FS**, increased activation was observed in PAC, bilaterally, and in Heschl's Gyrus, the Rolandic Operculum and BA44. We did not observe any activation of SMA.

For **FP > UP**, the activation was very similar to that of FPs over FS. This suggests that FPs and UPs equally affect attention in the listener, but while UPs modulates syntax processing areas, this is not the case for FPs that instead modulate motor areas in the perceiving brain. Also, from the point of view of FPs, fluent speech and UPs seem to constitute more or less the same phenomenon in that there is no observed difference between the two contrasts FPs > FS and FPs > UPs.

4. DISCUSSION

As has already been pointed out, we also focused our analyses on FPs/UPs proper instead of their effect on subsequent linguistic items (e.g. words). The effect on subsequent words following FPs has been studied with electrophysiology, e.g. N400 attenuation on the word following an FPs, as reported in Corley, McGregor and Donaldson (2007). Our main observations will be presented in the following sections.

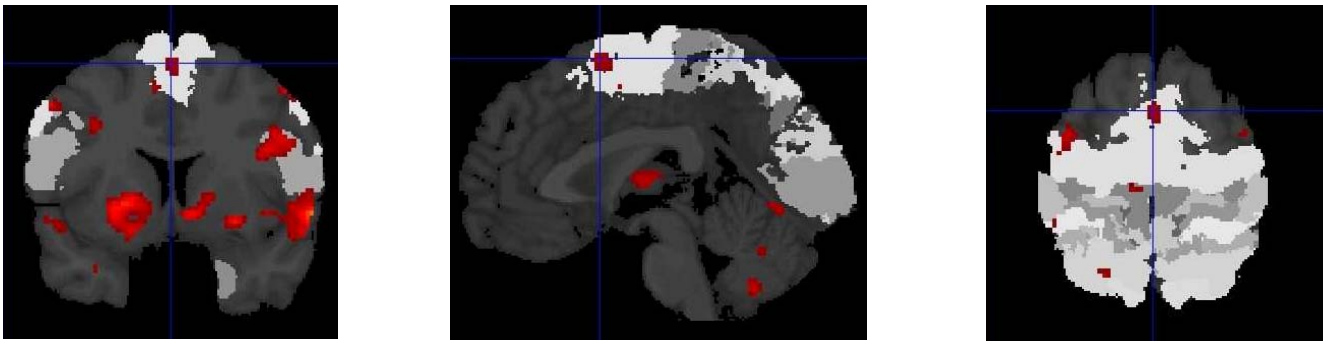


Figure 1. Observed modulation for the contrast Filled Pauses greater than Fluent Speech modulation. Cluster size = 63 voxels. Coronal, Sagittal and Axial views. 56.8% of cluster in right Area 6 (x = +8, y = –2, z = +55; MNI +8/+2/+50). 41.8% of cluster in left Area 6 (x = +0, y = +4, z = +65; MNI +0/+8/+60).

4.1. Activation of primary auditory cortex

Beginning with the strongest results, the **bilateral modulation of PAC**, it seems more than obvious that listeners' attention increases significantly when FPs/UPs appear in the speech. Top down regulation of primary cortices, e.g. the PAC, has previously been reported (Ghatan et al., 1998) and that heightened attention influences auditory perception has also been shown (Petkov et al. (2004). This attention-heightening function of FPs could possibly help explain the shorter reaction times to linguistic stimuli that follow FPs as reported by e.g. Fox Tree (2001, 1995). However, since UPs *also* modulated PAC in the listeners, conceivably *any* break in the speech signal might serve as a potential attention-heightener. Consequently, the shorter reaction times reported by Fox Tree might also be observed for unfilled pauses or other types of disfluency.

4.2. Activation of motor areas

Perhaps more interesting is the observation that **FPs activate BA6/SMA** in the listening brain. The most obvious explanation for this activation is that when hearing the speaker produce FPs, the listeners prepare to start speaking themselves, i.e. take over the floor. It has been known already since Brickner (1940) that SMA is active in the processing of speech, and several later studies have confirmed both SMA and pre-SMA play a role in both speech *production* (e.g. Goldberg, 1985; Alario et al., 2006) and speech *perception* (e.g. Iacoboni, 2008; Wilson et al., 2004). Moreover, an interesting result was reported in Wise et al. (1991) where subjects in a PET study were instructed to silently (i.e. non-vocalizing) generate verbs, which resulted in activation of the SMA, very much in accordance with our own observations. However, it could conceivably be the case that motor cortex activation during speech tasks at least partly occur as a part of motor planning of speech *breathing* (as distinct from baseline breathing), as is pointed out in Murphy et al. (1997).

4.3. Implications for two FP hypotheses

Our observed FP-induced activation of SMA could be seen in the light of the **"floor-holding" hypothesis** of FPs, as first proposed by Maclay and Osgood (1959). This hypothesis suggests that FPs are used by speakers who want to maintain the floor in conversation, as a (semi-deliberate) means to prevent interlocutors from breaking in. Although this might be true, our observations that FPs "kick-start" the speech production system in the listener would indicate that this use of FPs is counter-productive in that the effect on the listener is exactly the opposite of the suggested function to *prevent* interlocutors from speaking, not *preparing* them to do so.

An alternative hypothesis concerning the roles FPs might play in human dialog could be called the **"help-me-out" hypothesis**, as suggested in Clark and Wilkes-Gibbs (1986). This hypothesis suggests that FPs can be used as a (semi-deliberate) signal asking for interlocutor help in that they signal to the listener that the speaker is encountering problems in the production of speech, or simply put: when a speaker is looking for a word or term which is not immediately available to them, uttering "uh" signals to the listener that some help is desired. Our observation that motor areas are activated by FPs should mean that a helpful interlocutor would be faster to come to the rescue.

5. CONCLUSIONS

Three things make this study unique (we believe):

1. We used *fMRI to study disfluency perception*, unlike other perception studies that have relied on EEG and the concomitant focus on temporal aspects of speech perception.
2. We investigated perceptual modulation caused by *FPs proper*, not their effect on ensuing items (words, phrases) or general cognitive processing.
3. Unlike previous studies where the auditory stimuli often have been scripted laboratory speech, we used *ecologically valid stimulus data*.

Our results suggest that FPs – unlike FS and UPs – activate motor areas in the listening brain. However, both FPs and UPs activate PAC, which lends support to the attention-heightening hypothesis that has been forwarded in the literature. It would also seem clear that it is not the break in the speech stream per se that causes this activation, since UPs seemingly do not have this effect.

6. ETHICS APPROVAL

The study was approved by the Karolinska Institute ethics committee on April 4, 2007.

7. REFERENCES

Alario, F.-X., Chainay, H., Lehericy, S., Cohen, L. 2006. The role of the supplementary motor area (SMA) in word production. *Brain Research* 1076, 129–143.

Amunts, K., Schleicher, A., Zilles, K. 2007. Cytoarchitecture of the cerebral cortex—More than localization. *NeuroImage* 37, 1061–1063.

Barr, D. , Seyfeddinipur, M. (2010). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes* 25(4), 441–455.

Brickner, R.M. 1940. A Human Cortical Area Producing Repetitive Phenomena When Stimulated. *Journal of Neurophysiology* 3, 128–130.

Clark, H.H., Wilkes-Gibbs, D. 1986. Referring as a collaborative process. *Cognition* 22, 1–39.

Corley, M., MacGregor, L.J., Donaldson, D.I. 2007. It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition* 105, 658–668.

Eickhoff, S.B., Stephan, K.E., Mohlberg, H., Grefkes, C., Fink, G.R., Amunts, K., Zilles, K. 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage* 25, 1325–1335.

Eickhoff, S.B., Heim, S., Zilles, K., Amunts, K. 2006. Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps. *NeuroImage* 32, 570–582.

Eickhoff, S.B., Paus, T. Caspers, S., Grosbras, M.-H., Evans, A.C., Zilles, K., Amunts, K. 2007. Assignment of functional activations to probabilistic areas revisited. *NeuroImage* 36, 511–521.

Eklund, R. 2010. The Effect of Directed and Open Disambiguation Prompts in Authentic Call Center Data on the Frequency and Distribution of Filled Pauses and Possible Implications for Filled Pause Hypotheses and Data Collection Methodology. In: *Proceedings of DiSS-LPSS Joint Workshop 2010*, University of Tokyo, Japan, 25–26 September 2010, Tokyo, Japan, 23–26.

Eklund, R. 2004. *Disfluency in Swedish human–human and human–machine travel booking dialogues*. PhD thesis, Linköping University, Sweden.

Ferreira, F., Bailey, K.G.D. 2004. Disfluencies and human language comprehension. *TRENDS in Cognitive Sciences* 8(5), 231–237.

Fox Tree, J.E. 1995. The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language* 34, 709–738.

Fox Tree, J.E. 2001. Listeners' uses of *um* and *uh* in speech comprehension. *Memory and Cognition* 29(2), 320–236.

Fraundorf, S.H., Watson, D.G. 2011. The disfluent discourse: Effects of filled pauses on recall. *Journal of Memory and Language* 65, 161–175.

Friston, K., Ashburner, J.T., Kiebel, S.J., Nichols, T.E., Penny, W.D. (eds.). 2007 (second edition). *Statistical Parametric Mapping*. Amsterdam: Elsevier.

Genovese, C.R., Lazar, N.A., Nichols, T. 2002. Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate, *NeuroImage* 15, 870–878.

Ghatan, P.H., Hsieh, J.C., Petersson, K.-M., Stone-Elander, S., Ingvar, M. 1998. Coexistence of Attention-Based Facilitation and Inhibition in the Human Cortex. *Neuroimage* 7(1), 23–29.

Goldberg, G. 1985. Supplementary motor area structure and function: Review and hypotheses. The *Behavioral and Brain Sciences* 8, 567–616

Iacoboni, M. 2008. The role of premotor cortex in speech perception: Evidence from fMRI and rTMS. *Journal of Physiology – Paris* 102, 31–34.

Maclay, H., Osgood, C.E. 1959. Hesitation Phenomena in Spontaneous English Speech. *Word* 5, 19–44.

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., Zilles, K. 2001. Human Primary Auditory Cortex: Cytoarchitectonic Subdivisions and Mapping into a Spatial Reference System. *NeuroImage* 13, 684–701.

Murphy, K., Corfield, D.R., Guz, A., Fink, G.R., Wise, R.J.S., Harrison, J., Adams, L. 1997. Cerebral areas associated with motor control of speech in humans. *Journal of Applied Physiology* 83(5), 1438–1447.

Oldfield, R.C. 1971. The Assessment and Analysis of Handedness: The Edinburgh Inventory. *Neuropsychologia* 9, 97–113.

Petkov, C.I., Kang, X., Alho, K., Bertrand, O., E. Yund, E.W., Woods, D. L. 2004. Attentional modulation of human auditory cortex. *Nature Neuroscience* 7(6), 658–663.

Rademacher, J., Morosan, P., Schormann, T., Schleicher, A., Werner, C., Freund, H.-J., Zilles, K. 2001. Probabilistic Mapping and Volume Measurement of Human Primary Auditory Cortex. *NeuroImage* 13, 669–683.

Reich, S.S. 1980. Significance of Pauses for Speech Perception. *Journal of Psycholinguistic Research* 9(4), 379–389.

Wilson, S., Saygın, A.P., Sereno, M.I., Iacoboni, M. 2004. Listening to speech activates motor areas involved in speech production. *Nature Neuroscience* 7(7), 701–702.

Wise, R., Chollet, F., Hadar, U., Friston, K., Hoffner, E., Frackowiak, R. 1991. Distribution of cortical neural networks involved in word comprehension and word retrieval. *Brain 114*, 1803–1817.

A LONGITUDINAL STUDY OF FILLED PAUSES AND SILENT PAUSES IN SECOND LANGUAGE SPEECH

Lorenzo García-Amaya

University of Michigan
lgarciaa@umich.edu

ABSTRACT

This study provides a longitudinal analysis of speech rate and the use of filled pauses (FPs) and unfilled or silent pauses (SPs) in the oral production of L2 learners of Spanish in two learning contexts: a 6-week intensive overseas immersion program (OIM), and a 15-week US-based 'at-home' foreign language classroom (AH). Fifty-six native speakers of English performed two video-retell tasks at three different time points. A total of five measurements of oral production were calculated. The results show a significant increase in rate of speech over time in the OIM group compared to the AH group. Additionally, the OIM learners show greater use of "disfluencies" over time, namely FPs and short SPs. We suggest that OIM learners increase their use of hesitation phenomena over time as a speech processing and planning strategy and discuss this finding within the framework of L2 cognitive fluency.

Keywords: second language fluency, disfluencies, rate of speech, filled pauses, silent pauses, study abroad, Spanish

1. INTRODUCTION

Research on second language (L2) oral production, and specifically on L2 *fluency*, investigates factors that influence the way speech is perceived and judged by native speakers [5,13], including temporal variables of speech such as amount of spoken units produced, speed of delivery [8,10,13], and perturbations in the speech chain, also known as *disfluencies* or speech errors [20]. According to Levelt's [14] taxonomy for monitoring and error repair, disfluencies can be solved through a process of *repair*, which involves repeats, restarts, self-corrections, false starts, or *editing*, which entails the use of hesitation phenomena. Editing a speech error by means of a disfluency can be achieved by using silent pauses (SPs) or filled pauses (FPs). SPs can be analyzed by considering their different degrees of length. FPs may have a lexical status (e.g., the English word *like*) or non-lexical status (e.g., *uh, um*). Using the schematics of Levelt's model as a basis, Segalowitz [17] explains the use of disfluencies in L2 speech in his own model of L2 cognitive fluency. [19] shows that L2 learners of English use more SPs in phrase-medial position than native speakers of the same language, and that these SPs are associated with processes such as replacement, reformulation, and online planning (see [9] for Spanish).

As for the effect of context of learning on L2 acquisition, when comparing developmental patterns in the study abroad (SA) vs. the at-home (AH) contexts, SA learners are typically superior for aspects related to oral production [18]. The quantification of disfluencies in L2 speech has been a widely used assessment for determining oral fluency in the SA learners [6,13,15,21]. SA programs researched in the literature have varied in length and duration, with programs lasting an entire academic year or just a few weeks [12]. However, few studies have addressed the fluency abilities of L2 learners who participate a special type of SA program, namely an overseas immersion program (OIM), where learners pledge not to use their L1 for the duration of their stay. García-Amaya [7], using a cross-sectional design, shows that OIM learners use a significantly higher rate of speech than proficiency-matched AH and SA learners.

Two aspects of language learning for which SA/OIM programs are thought to be particularly efficient compared to the AH contexts are pronunciation and oral fluency, the fluidity or "smoothness" of language use [4]. Researchers in this area typically extract measures of fluency (e.g., rate of speech; total speech produced; mean length of run) as well as measures of disfluency or hesitation phenomena (e.g., FPs and SPs). However, few studies have conducted principled comparisons of L2 fluency in different learning contexts using longitudinal data (but see [18]). The current study addresses these gaps in our knowledge through an analysis of L2 oral production for two learner groups of Spanish, tracking learner rate of speech as well as hesitation phenomena (FPs and SPs) over time.

2. METHOD

2.1. Speakers

We elicited L2 speech data from a total of 56 native

speakers of English who were L2 learners of Spanish: 27 OIM and 29 AH learners. The OIM learners were in their junior year of high school, and participated in an overseas program in Spain. These learners signed a language commitment, pledging to use their L2 exclusively during their time abroad. The AH learners were Spanish majors or minors at a large Midwestern institution in the US. The OIM program lasted 6 weeks, whereas the AH program lasted 15 weeks.

For all learners proficiency was determined by means of a 45-item Proficiency test. At Time 1 the average score was 24.89 for OIM and 24.06 for AH, indicating comparable proficiency at the start of their respective programs. Within each group, learners who scored between 0-24 at Time 1 were classified as low proficiency, and learners who scored between 24-45 were classified as high proficiency. All learners completed the same proficiency test at the end of their programs. The average score was 32.11 for OIM and 24.76 for AH, indicating considerable proficiency development for OIM learners only.

2.2. Data collection

A video-retell task (Simon Tofield's 'Simon's Cat') was used to elicit speech. Participants performed the video-retell at three time points in their respective language learning experiences: beginning (Time 1), middle (Time 2), and end (Time 3). For both learner groups, Time 1 occurred over the first two days of classes, and Time 3 occurred over the last two days of classes. Time 2 occurred after the third week of classes for OIM learners, and after the seventh week of classes for AH learners. At each time participants saw two different videos (lasting approximately three minutes total), and were instructed to wait until each video ended to begin their retells in Spanish. There were no time limitations on the length of the retell, and participants were encouraged to retell each video with as much detail as possible.

2.3. Fluency analysis

For this paper, we calculated five measures of oral production. To account for oral fluency, we calculated *rate of speech* in syllables per seconds. To account for hesitation phenomena we established different metrics of FP and SP production. FPs include vocalizations such as *uh*, or vocalizations with nasalization such as *uhm* or *mmm*. SPs included stretches of time without linguistic content or vocalizations of any kind.

To account for FP production we calculated *seconds between FPs* [10]. As for SP production, although various authors [e.g., 2,3,11] have advocated for examining pauses shorter than 400 milliseconds, there is a large body of research that has discarded these shorter pauses since they are also common to native speaker speech [6,19]. We adopted the latter approach in this study. Wantanabe and Rose [22] also suggest that a more fine-grained fluency analysis could classify SPs into short, medium and long pauses in order to separate behaviour of articulatory-based from linguistically-based pauses in speech. We therefore decided to break silent pauses into three different lengths: silent pauses between 400 and 1000 ms (short SPs); silent pauses between 1000 and 2000 ms (intermediate SPs); and silent pauses between 2000 and 3000 ms (long SPs). We calculated the number of seconds between each of the three types of SPs.

2.4. Statistical analysis

We fitted five linear mixed-effects models (LMEMs), corresponding to the five dependent variables in the analysis (listed in Section 2.3). The predictor variables included in our models were TIME, GROUP, PROFICIENCY, and the two-way interactions between them. The LMEMs also included two random effects: the first was associated with each subject in general (to model correlations within a subject), and the second was associated with time blocks within a subject (to model additional correlations introduced by time blocks within a subject). All models were fitted using the MIXED procedure in SPSS, and degrees of freedom for approximate F-statistics for the fixed effects were computed using a Satterthwaite approximation. Post-hoc comparisons with Bonferroni adjustments were performed to compare selected means when fixed effects were found to be significant. In the following sections, we report results for statistically significant effects only.

3. RESULTS

Rate of speech increased for OIM learners across time, whereas it remained stable for AH learners (Figure 1). This general effect was constant for both proficiency levels. The LMEM showed significant effects for the fixed factors TIME ($F(2,108) = 16.782$, $p < 0.001$), PROFICIENCY ($F(2,108) = 16.782$, $p < 0.001$), GROUP ($F(1,53) = 17.790$, $p < 0.001$), and the GROUP X TIME interaction ($F(2,108) = 16.55$, $p < 0.001$). A Bonferroni adjustment used to determine the mean differences in the GROUP X TIME interaction returned significant effects for two of the three pairwise comparisons. The factor GROUP yielded a significant effect at Time 2 ($p < 0.001$) and at Time 3 ($p < 0.001$). This indicates that by the

second data collection time point the OIM learners had significantly increased their rate of speech in comparison to the AH learners.

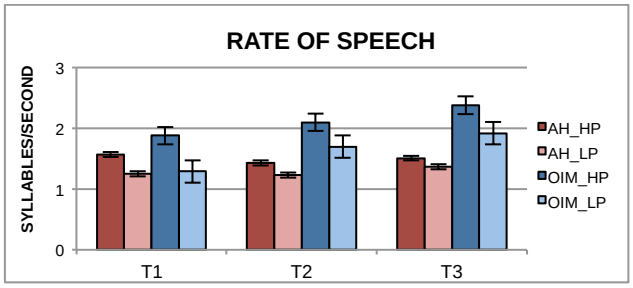


Figure 1: Rate of speech (syllables/second).

Regarding *seconds between FPs* (Figure 2), high proficiency AH and OIM learners remained constant in the time they took to produce FPs in Spanish. However, low proficiency learners showed different patterns based on Group. Whereas the low proficiency AH learners increased time between FPs, the low proficiency OIM learners decreased it. The LMEM yielded significant effects for the fixed factor GROUP ($F(1,49.104) = 16.310$, $p < 0.001$)) and the GROUP X PROFICIENCY interaction ($F(49.105) = 5.029$, $p=.029$). A Bonferroni adjustment revealed an effect of PROFICIENCY for the AH group ($p = .036$), indicating that the low proficiency AH learners used more FPs than high proficiency AH learners.

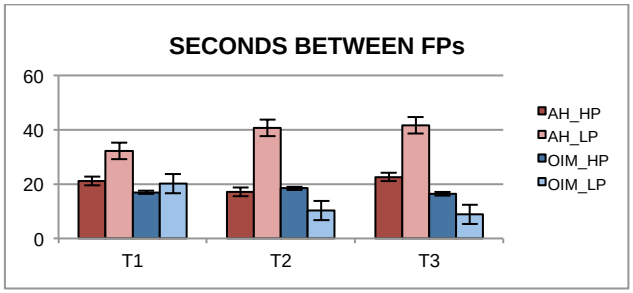


Figure 2: Seconds between FPs.

As for *seconds between short SPs* (Figure 3), AH learners increased their speaking time between short SPs across time, whereas OIM learners decreased it. The LMEM showed significant effects of GROUP ($F(1,50.657) = 21.542$, $p < 0.001$) and the GROUP X TIME interaction ($F(2,95.931) = 3.756$, $p=.027$). A Bonferroni adjustment showed that when we compared means for both groups at each time point, there was an effect at Time 1 ($p = .007$), Time 2 ($p < 0.001$), and Time 3 ($p < 0.001$). Another post-hoc analysis to compare the time effect within each GROUP indicated that there were significant differences for seconds between SPs 400-1000ms for the AH learners for Time 1 vs. Time 3 ($p = .016$).

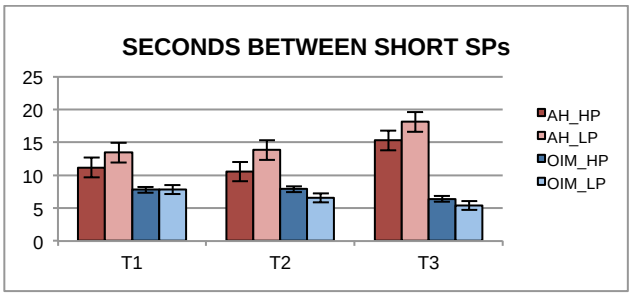


Figure 3: Seconds between short SPs.

Regarding *seconds between intermediate SPs* (Figure 4), both learner groups generally increased their speaking time between SPs of 1000-2000 ms. The results of the LMEM showed significant effects for GROUP ($F(1,49.086) = 4.599$, $p = .037$), TIME ($F(2,98.251) = 5.737$, $p = .004$), GROUP X TIME ($F(2,98.251) = 7.061$, $p = .001$), and GROUP X PROFICIENCY ($F(49.078) = 4.929$, $p = .031$). A Bonferroni adjustment indicated significant effects of GROUP at Time 2 only ($p < 0.001$) and of PROFICIENCY for the OIM group only ($p = .015$). The implication is that within the OIM group, those learners who started the program with high proficiency had significantly more time between intermediate SPs than low proficiency learners.

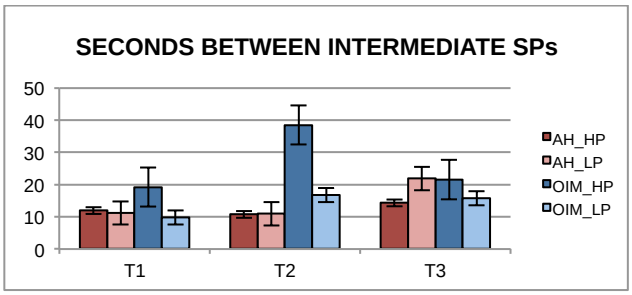


Figure 4: Seconds between intermediate SPs.

Finally, in terms of *seconds between long SPs* (Figure 5), AH and OIM learners increased their speaking time between long SPs with the exception of high proficiency OIM learners, who increased this amount by Time 2 but decreased it by Time 3. The LMEM returned significant effects for TIME ($F(2,91.311) = 10.241$, $p < 0.001$), GROUP X TIME ($F(2,91.311) = 4.277$, $p = .017$), and TIME X PROFICIENCY ($F(1,91.311) = 9.915$, $p < 0.001$). A Bonferroni adjustment indicated a significant effect of GROUP at Time 2 only ($p = .007$) and of PROFICIENCY at Time 1 only ($p = .014$).

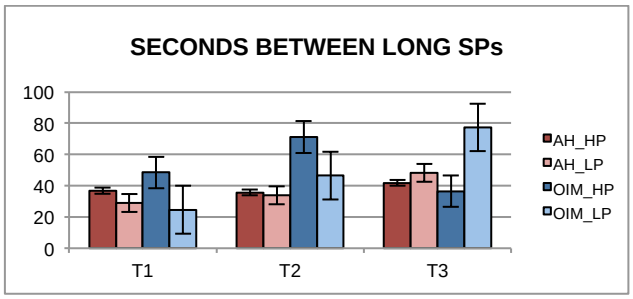


Figure 5: Seconds between long SPs.

To summarize the statistical results, we found significant effects of TIME for three measures (rate of speech, secs. between intermediate SPs, secs. between long SPs), GROUP for four measures (rate of speech, secs. between FPs, secs. between short SPs, secs. between intermediate SPs), and GROUP X TIME for four measures (rate of speech, secs. between short SPs, secs. between intermediate SPs, secs. between long SPs).

4. DISCUSSION

For rate of speech, the OIM learners clearly spoke faster by the end of their learning experience, compared to the AH learners. Although the OIM program was actually shorter in duration than the AH program, we attribute the fluency gain on part of the OIM learners to the increased opportunities for L2 interaction during their abroad experience. This interpretation is corroborated by the substantially greater improvement in proficiency test scores for OIM learners. This confirms previous research that compares speech rate based on learning context [7]. In addition, OIM learners decreased their seconds between FPs over time, implying that they increased their use of FPs between Time 1 and Time 3. Furthermore, OIM learners decreased their seconds between short SPs over time, whereas they increased their seconds between long SPs. The implication is that OIM learners produced more short pauses and fewer long pauses at Time 3 compared to Time 1.

Putting these data together, it becomes clear that we must account for why OIM learners showed faster rates of speech, but also used more FPs and short SPs over time. This combination of findings is not typically reported in the literature on L2 fluency. On the contrary, previous research reports that faster rates of speech proceed with a reduction of hesitation phenomena [18]. We would argue that the increase in use of FPs and short SPs is the result of multiple factors: first, since our OIM learners are communicating more linguistic content in less time by Time 3 (i.e., they have faster rates of speech), they may resort to a greater use of FPs and short SPs to facilitate lexical and syntactic planning [1,14]; second, since their speech by Time 3 is more complex from a syntactic standpoint (i.e., higher instances of subordination, etc.), their use of FPs and short SPs may serve as processing mechanisms to mitigate the greater demands of complex syntactic structure, etc. [19].

Recent research in L2 fluency research considers the relationship between hesitation phenomena and linguistic processes such as lexical access, syntactic processing or discourse planning [22]. Following a computational account, it can be posited that hesitation phenomena in the chain of speech are a by-product of processing difficulties at the cognitive level. Clark and Wasow [1] propose that these underlying processing difficulties may be associated with the degree of syntactic complexity of the utterance and the higher cognitive load necessary for processing it. Through their complexity hypothesis, these authors postulate that computing more complex syntax requires higher processing and may also cause longer delays. Based on the findings for our OIM learners, we would propose that short SPs (and not long SPs) provide a sufficient timing mechanism by which such processing and planning can occur. Although we did not assess syntactic complexity in this paper, previous research on similar OIM learners indicates that their speech contains a greater frequency of subordination and fewer sub-clause independent units than proficiency-matched AH learners [8].

5. CONCLUSION

To conclude, one important contribution of our study is that we have shown that OIM learners obtain faster rates by the end of their language learning program than AH learners. This development toward more “fluent” speech proceeds jointly with a development toward a greater use of disfluencies (i.e., FPs and short SPs). We suggest that these two strategies work in tandem as learners increase syntactic complexity [9]. In future work models of L2 cognitive fluency (e.g., [17]) will need to consider this interplay between fluency strategies and syntactic and grammatical development.

6. REFERENCES

[1] Clark, H., Wasow, T. 1998. Repeating words in spontaneous speech. *Cognitive Psychology* 37, 201-242.

[2] De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech (DiSS)*, 17–20. Stockholm, Sweden: Royal Institute of Technology.

[3] De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics 34*, 893–916.

[4] Freed, B. F. 1995. What makes us think that students who study abroad become fluent? In: Freed, B. (ed.), *Second language acquisition in a study abroad context.* Phildaelphia, PA: John Benjamins, 123-148.

[5] Freed, B. F. 2000. Is fluency, like beauty, in the eyes (and ears) on the beholder? In: Riggenbach, H. (ed.), *Perspectives on fluency*. Ann Arbor, MI: Univeristy of Michigan Press, 243-265.

[6] Freed, B. F., Segalowitz, N., Dewey, D. 2004. Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition 26*, 275-301.

[7] García-Amaya, L. 2009. New findings on fluency measures across three different learning contexts. In: Collentine, J., García, M., Lafford, B., Marcos-Marín, F. (eds.), *Selected Proceedings of the 11th Hispanic Linguistics Symposium*, Somerville, MA: Cascadilla Press, 68-80.

[8] García-Amaya, L. 2010. *The effect of clause type on filled pauses and discourse markers clause type in the oral speech production of second language learners across three different learning contexts.* Paper presented at Kentucky Foreign Language Conference, Lexington, KY. April 15-17, 2010.

[9] García-Amaya, L. 2011. *The effect of clause type on filled pauses and discourse markers in the oral production of second language learners across three different learning contexts.* Paper presented at American Association for Applied Linguistics Conference, Chicago, IL. March 26-29, 2011.

[10] García-Amaya, L. 2012. *Second Language Fluency and Cognition: The Study of Spanish Second Language Development in an Overseas Immersion Program and an At-Home Foreign Language Classroom*. PhD Dissertation, Indiana University.

[11] Kahng, J. (2014). Exploring Utterance and Cognitive Fluency of L1 and L2 English Speakers: Temporal Measures and Stimulated Recall. *Language Learning 64*, 809-854.

[12] Lafford, B. A. 2006. The effects of study abroad vs. classroom contexts on Spanish SLA: Old assumptions, new insights and future research directions. In: C. A. Klee & T. L. Face (eds.), *Selected Proceedings of the 7th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages*. Somerville, MA: Cascadilla Proceedings Project, 1-25.

[13] Lennon, P. 1990. Investigating fluency in EFL: A quantitative approach. *Language Learning 40*, 387-417.

[14] Levelt, W. 1989. *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

[15] Möhle, D. 1984. A comparison of the second language speech of different native speakers. In: Dechert, H. W., Möhle, D., Raupach, M. (eds.), *Second language productions*. Tubingen: Gunter Narr, 26-49.

[16] Mora, J., Valls-Ferrer, J. 2009. *Study Abroad Effects on the Oral Production of Advanced L2 Learners: The Interaction between Fluency, Accuracy and Complexity*. Presented at *AAAL 2009 Conference*, Denver, CO.

[17] Segalowitz, N. 2010. *The cognitive bases of second language fluency*. New York, NY: Routledge.

[18] Segalowtiz, N., Freed, B. F. 2004. Context, contact, and cognition in oral fluency acquisition: Learning Spanish at home and study abroad. *Studies in Second Language Acquisition 26*, 173-199.

[19] Tavakoli, P. 2011. Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal 65*, 71-79.

[20] Temple, L. 1992. Disfluencies in learner speech. *Australian Review of Applied Linguistics 15*, 29-44.

[21] Towell, R., Hawkins, R., Bazergui, N. 1996. The development of fluency in advanced learners of French. *Applied Linguistics 17*, 84-119.

[22] Watanabe, M., Rose, R. 2012. Pausology and hesitation phenomena in second language acquisition. In Robinson, P. (ed.), *The Routledge Encyclopedia of Second Language Acquisition.* New York/London: Routledge, 480-483.

DISFLUENCY IN MULTIPARTY SOCIAL TALK

Emer Gilmartin, Carl Vogel, Nick Campbell
Trinity College Dublin
{gilmare, vogel, nick}@tcd.ie

Keywords: disfluency, hesitation, repair, casual conversation, spoken interaction

1. INTRODUCTION

Much research on disfluencies in spontaneous spoken interaction has been carried out on corpora of task-based conversations, resulting in greater understanding of the role of several phenomena. Modern multimodal corpora allow the full spectrum of signals in face to face communication to be analysed. However, the 'unmarked' case of casual conversation or social talk with no obvious short-term instrumental goal has been less studied in this manner. Corpus-based work on social talk tends to deal with short dyadic interactions, although the norm for social conversation is for longer multiparty interaction. In this paper, we outline our programme of exploratory studies of disfluency in a longer multiparty conversation. We briefly describe the background to our research goals, and then report on the collection, transcription, and annotation of the data for our experiments. We present and discuss some of our early results.

2. SOCIAL TALK

Casual social conversation is defined as talk engaged in when 'talking just for the sake of talking'[7], and includes smalltalk, gossip, and conversational narrative. Malinowski described 'phatic communion' as an emergent activity of congregating people, comprising free aimless social conversation, which he viewed as the most basic use of speech [15]. For Malinowski, the purpose of such talk is not to exchange information or to express thought, but to avoid unfriendly silence and strengthen social bonding. This view is echoed in Jakobson's phatic component in his model of communication [11], Brown and Yule's distinction between interactional and instrumental language [4], and Dunbar's theory that language evolved to maintain social cohesion through verbal grooming as group size grew too large for physical grooming [6]. Laver focuses on the 'psychologically crucial margins of interaction', conversational openings and closings, postulating that small talk performs a lubricating or transitional function from silence to greetings to business and back to closing sequences and to leave taking [13]. Schneider analysed audio recordings of naturally occurring small talk, concentrating on the linguistic content of entire dialogues [21]. He described instances of small talk at several levels, from frames such as 'WEATHER' to sequences and adjacency pairs, and their constituent utterance types. He identifies idling sequences of repetitions of agreeing tails such as 'Yes, of course', 'MmHmm' as prevalent in social talk. Ventola viewed casual conversation as composed of several phases - with ritualised opening greetings, followed by approach segments of light uncontroversial small talk, which sometimes led to longer and more informative centre phases consisting of sequential but overlapping topics, and then back to ritualised leavetakings [23]. Thus a social conversation could range from a simple exchange of greetings, through a short exchange of small talk, to longer more varied stretches of spoken interaction covering several topics. Slade and Eggins state that through casual conversation people form and refine their social reality [7]. They cite gossip, where participants reaffirm their solidarity by jointly ascribing outsider status to another, and show examples of conversation between friends at a dinner party where greater intimacy allows differences of opinion. They identify story-telling as a frequent genre in conversation and highlight 'chat' (interactive exchanges involving short turns by all participants) and 'chunks' (longer uninterrupted contributions) elements of conversation. They report that casual conversation tends to involve multiple participants rather than the dyads normally found in instrumental interactions or examples from conversation analysis. Instrumental and interactional exchanges differ in duration; task-based conversations are bounded by task completion and tend to be short, while casual conversation can go on indefinitely. Indeed, early conversation analysts pointed out that these casual conversational situations or 'continuing state(s) of incipient talk' were not covered by the theories of (task-based) conversational structure being developed [19].

It seems likely that the distribution of disfluencies, both within turn pauses, hesitations and repairs, and phenomena such as recycled restarts and abandoned utterances, will vary between different types

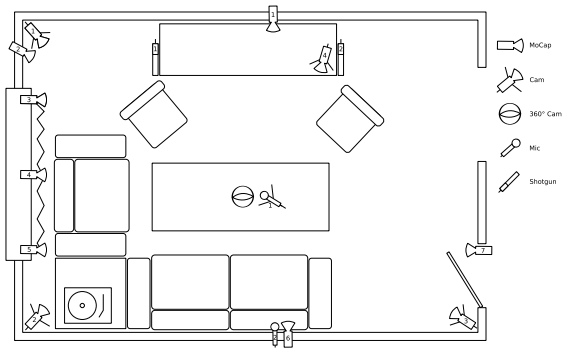
of interaction, and across different phases or subgenres of the same interaction. If, as is frequently claimed, disfluencies are a mark of planning difficulties or cognitive load, they should be less frequent in shorter more ritualised small talk or approach sequences. In more central sequences, they may appear more often in discussion sequences or at the beginning of narrative sequences as the speaker 'gets going', but less in idling sequences. Distributions may also vary between social and task-based conversations. In order to test these ideas, we have prepared a 70-minute sample of extended social talk data, on which we are currently experimenting, as described below.

3. DATA AND ANNOTATION

Corpora used for studies of disfluencies in human-human non-pathological adult spontaneous speech in English include the HCRC MapTask corpus of dyadic information gap task-based conversations [1], SWITCHBOARD corpus of dyadic telephonic speech [9], ICSI and AMI multiparty meeting corpora [12] [17], and resources such as recordings of televised political interviews [2]. However, the speech in these resources, while spontaneous and conversational, does not meet our need for longer multiparty social face to face conversation data. Therefore, for our preliminary studies, we have prepared a sample drawn from a longer conversation in the D64 corpus of spontaneous multiparty social talk.

The D64 corpus is a multimodal corpus of over 8 hours of informal conversational English recorded in an apartment living room, as shown in Fig. 1. Several streams of video, audio, and motion capture data were recorded for the corpus. There were between 2 and 5 people on camera at all times. There were no instructions to participants about what to talk about and care was taken to ensure that all participants understood that they were free to talk or not as the mood took them. Design and collection of the corpus is fully described in[18]. The audio recordings included near-field chest or adjacent microphone recordings for each speaker. These were found to be unsuitable for automatic segmentation as there were frequent overlaps and bleedover from other speakers. After a manual synchronisation, the audio files for each speaker were segmented manually into speech and silence intervals using Praat [3] on 10 and 4-second or smaller windows as necessary. The process was then repeated for the sound file recorded at the same time for each of the other speakers, resulting in annotations checked across five different sound files. Any remaining speech intervals not assigned to a particular speaker were resolved using Elan [24] to refer to the video recordings taken at the same time. There are concerns to note with the manual segmentation into speech and silence and indeed in manual annotation of disfluencies, as human hearing and comprehension is a filter rather than a simple sensor. Humans listening to speech can miss or imagine the existence of objectively measured silences of short duration, especially when there is elongation of previous or following syllables [16], and are known to have difficulty recalling disfluencies from audio they have heard [5]. However, in the current work, speech can be slowed down and replayed and, by zooming in on the waveform and spectrogram, annotators can clearly see silences and differences in amplitude on the speech waveform and spectrogram. This need to match the heard linguistic and non-linguistic content to the viewed waveform and spectrogram means that it was much more likely that pauses and disfluencies would be noticed.

Figure 1: Setup for Session 1 of D64 Recordings.



After segmentation the data for Session 1 of the corpus were manually transcribed and annotated, using a scheme largely derived from the TRAINS transcription scheme [10]. Words, hesitations, filled and unfilled pauses, unfinished words, laughs and coughs were transcribed and marked. The transcription was carried out at the intonational phrase (IP) level rather than the more commonly used interpausal unit (IPU) as IPs are a basic unit for intonation study and can easily be concatentated to the interpausal unit (IPU) and turn level as required. The transcriptions were then text-processed for automatic word alignment, which was carried by running the Penn Aligner [25] over a sound file and accompanying transcription for each intonational phrase annotated. Sections which could not be automat-

ically aligned, where there was significant overlap or cut off words, were manually aligned. In order to more fully investigate genre within casual talk, conversational sections were labeled as discussion, dominated, or idling. Idling was labelled orthogonally to discussion and dominated as it could occur within either modality. Discussion referred to stretches of talk shared more or less evenly among two or more participants throughout the bout, while dominated referred to bouts largely dominated by one participant. These often took the form of narratives or recounts of personal experiences, extended explanations or opinions. A total of 142 'bouts' were annotated, of which 14 were labelled as 'discussion' while the remaining 128 were classed as dominated.

Table 1: The annotation code used for disfluencies.

Symbol	Note
.	interruption point
-	unfinished word
tilde	unfinished utterance
caret	contracted word
r	repeated word
s	substituted word
d	deleted word
f	filled pause
x	pause
o	overlap

The word level transcription was then used with the sound files to annotate disfluencies using Praat. The scheme and procedures used were based largely those outlined in Shriberg's and Eklund's respective theses [22] [8], and in Lickley's annotation manual for the MapTask corpus [14], with extra labels and coventions for recycled turn beginnings [20], disfluencies in the presence of overlapping speech from another participant, and unfinished and abandoned utterances. The symbols used are outlined in Table. 1. Complex, or nested, disfluencies were labelled following Shriberg's method [22], and no indexing was used for substitutions or repetitions. Pauses within utterances were annotated with 'x' when they occurred within a larger disfluency or with '[.x]' when they occurred alone. The annotated data sample comprised 15,545 words across 6164 intonational phrase units, with 1505 annotated disfluencies. There were 653 lone pauses, which were removed from the dataset for the purposes of this analysis. Of the remaining 853 disfluencies, 117 were complex. Just over 15%, 128 disfluencies, occurred in the presence of overlap by another speaker

Below we describe preliminary results on the distribution of disfluencies in the corpus in general and particularly disfluencies in the presence or absence of overlapping speech in the dominated genre.

4. RESULTS AND CONCLUSIONS

We concentrate on the dominated genre of talk, comprising 777 disfluencies. For preliminary investigation, complex disfluencies were removed from the dataset, leaving 668 disfluencies, of which 13%, 87 disfluencies, were in the presence of overlap.

Table 2: Disfluency types in overlap in dominated genre (%).

Sp	Del	Rep	Sub	FP
All	52	34	3	11
Main	59	18	5	18
Other	67	25	2	6

In the overlap condition, the distribution of disfluency type over all disfluencies, and for the dominant or main speaker and other speakers in the dominated genre is shown in Table. 2 . It can be seen that the distributions are similar for both the main and other speakers with the bulk of disfluencies in overlap occurring as deletions of unfinished utterances. This finding is consistent across both dominant speakers and other speakers. However, speakers other than the main speaker were even more likely to abandon their utterance, while the main speaker used more filled pauses in overlap, possibly to indicate intention to continue.

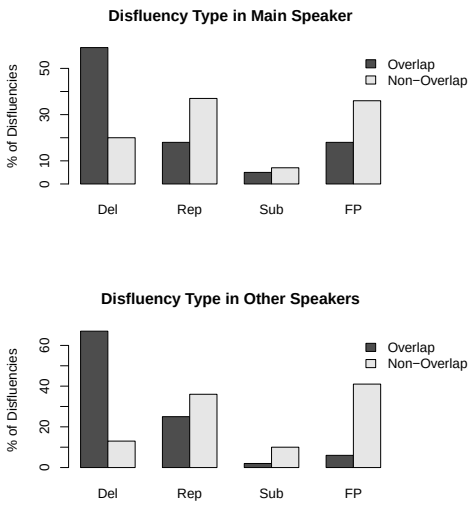
Table 3: Disfluency types in non-overlap in dominated genre (%).

Sp	Del	Rep	Sub	FP
All	18	36	8	38
Main	20	37	7	36
Other	13	36	10	41

In the non-overlap condition there were 581 disfluencies, distributions for all speakers, main and other speakers can be seen in Table, 3. In this condition, deletions and filled pauses are the most common types of disfluency and almost equally common, with the distribution showing consistency for main and other speakers. Figure. 2 contrasts the frequency of each disfluency type in main and other speakers in the overlap and non-overlap environments.

Our preliminary results show a strong tendency for speakers to stop in the presence of overlap,

Figure 2: Disfluency distributions in overlap and non-overlap environments.



although this does not always happen. It would be very interesting to analyse whether this tendency and indeed the occurrence of disfluency varies throughout the course of a bout of dominated conversation for main and other speakers. However, we cannot simply use distance of disfluencies from the bout start as the proportion of speech by different participants varies as each bout progresses. We are currently working on a measure of distance from the start of each bout which takes account of this variation. This will allow us to further explore the role of disfluency in casual multiparty conversation.

5. REFERENCES

[1] Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., others, 1991. The HCRC map task corpus. *Language and speech* 34(4), 351–366.

[2] Beattie, G. 1983. *Talk: An analysis of speech and non-verbal behaviour in conversation*. Open University Press.

[3] Boersma, P., Weenink, D. 2010. *Praat: doing phonetics by computer [Computer program], Version 5.1. 44*.

[4] Brown, G., Yule, G. 1983. *Teaching the spoken language* volume 2. Cambridge University Press.

[5] Deese, J. 1980. *Pauses, prosody, and the demands of production in language*. Mouton Publishers.

[6] Dunbar, R. 1998. *Grooming, gossip, and the evolution of language*. Harvard Univ Press.

[7] Eggins, S., Slade, D. 2004. *Analysing casual conversation*. Equinox Publishing Ltd.

[8] Eklund, R. 2004. *Disfluency in Swedish human-human and human-machine travel booking dialogues*. LinkkÌoping, Sweden: Dept. of Computer and Information Science, Linköping Studies in Science and Technology.

[9] Godfrey, J. J., Holliman, E. C., McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on* volume 1 517–520.

[10] Heeman, P. A., Allen, J. F. 1995. The TRAINS 93 Dialogues. Technical report DTIC Document.

[11] Jakobson, R. 1960. Closing statement: Linguistics and poetics. *Style in language* 350, 377.

[12] Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A. 2003. The ICSI meeting corpus. *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on* volume 1 I–364.

[13] Laver, J. 1975. Communicative functions of phatic communion. *Organization of behavior in face-to-face interaction* 215–238.

[14] Lickley, R. J. 1998. HCRC disfluency coding manual. *Human Communication Research Centre, University of Edinburgh*.

[15] Malinowski, B. 1923. The problem of meaning in primitive languages. *Supplementary in the Meaning of Meaning* 1–84.

[16] Martin, J. G. 1970. On judging pauses in spontaneous speech. *Journal of Verbal Learning and Verbal Behavior* 9(1), 75–78.

[17] McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V. 2005. The AMI meeting corpus. *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research* volume 88.

[18] Oertel, C., Cummins, F., Edlund, J., Wagner, P., Campbell, N. 2010. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces* 1–10.

[19] Schegloff, E., Sacks, H. 1973. Opening up closings. *Semiotica* 8(4), 289–327.

[20] Schegloff, E. A. 1987. Recycled turn beginnings: A precise repair mechanism in conversation's turn-taking organization. *Talk and social organization* 70–85.

[21] Schneider, K. P. 1988. *Small talk: Analysing phatic discourse* volume 1. Hitzeroth Marburg.

[22] Shriberg, E. E. 1994. *Preliminaries to a theory of speech disfluencies*. PhD thesis University of California.

[23] Ventola, E. 1979. The structure of casual conversation in English. *Journal of Pragmatics* 3(3), 267–298.

[24] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. 2006. Elan: a professional framework for multimodality research. *Proceedings of LREC* volume 2006.

[25] Yuan, J., Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America* 123(5), 3878.

FLUENCY IN ENL, ESL AND EFL: A CORPUS-BASED PILOT STUDY[1]

Sandra Götz

Justus Liebig University Giessen
Sandra.Goetz@anglistik.uni-giessen.de

ABSTRACT

Against the background of a 'cline model' of increasing fluency/decreasing disfluency from ENL to ESL to EFL forms of English, the present pilot study investigates (dis)fluency features in British English, Sri Lankan English and German Learner English. The analysis of selected variables of temporal fluency (viz. unfilled pauses, mean length of runs) and fluency-enhancement strategies (viz. discourse markers, smallwords and repeats) is based on the c. 40,000-word subcorpora of the British and the Sri Lankan components of the *International Corpus of English* (ICE-GB and ICE-SL) and the c. 80,000-word German component of the *Louvain International Database of Spoken English Interlanguage* (LINDSEI-GE). The study reveals that, while the EFL variant shows the lowest degree of temporal fluency (e.g. the highest number of unfilled pauses), the findings are mixed for ESL and ENL (e.g. the ESL speakers show a lower number of unfilled pauses, but the ENL speakers show a higher number of smallwords). Also, variant-specific preferences of using certain fluency-enhancement strategies become clearly visible.

Keywords: ENL vs. ESL vs. EFL, fluency, corpus-based (dis)fluency, fluency profiles.

1. INTRODUCTION

Kachru's [14] distinction between English as a native language (ENL), English as a second language (ESL) and English as a foreign language (EFL) has exerted an enormous influence on the modelling of Englishes worldwide. However, newer research shows that this distinction is far from clear-cut and the recent tendency has been to bridge this "paradigm gap" [22] and view EFL, ESL and ENL rather as a continuum (cf. [4, 5, 7, 13, 15, 17, 18]). Corpus-based studies in this vein have mainly focused on written English so far (very few exceptions being [1], [5] or [9]). A comparative approach to spoken fluency in inner-circle compared to outer-circle and expanding-circle varieties has not yet been undertaken; indeed, fluency has not played a major role so far in the description of post-colonial varieties of English. Thus, in the present pilot study, I put into perspective some all-pervading and typical features of spoken English, namely variables that are associated with fluency and/or disfluency. Against the background of a 'cline model' of ENL, ESL and EFL forms of English [7], I assume that the use of fluency variables such as filled or unfilled pauses by competent speakers of institutionalised ESL varieties shares some aspects with ENL usage while other aspects are shared with EFL usage. I also assume that there are correlations between the variety and frequencies of "fluency-enhancement strategies" [8], such as discourse markers on the one hand and the status of variant type of English (along the lines of [14]'s model from EFL over ESL to ENL) on the other.

2. FLUENCY IN ENL, ESL AND EFL

To a certain extent, differences between ENL, ESL and EFL are to be expected on all descriptive levels, caused by the different underlying acquisition processes, as L1 acquisition in an ENL context is held to be more holistic in nature while the acquisition of English as a non-native language in ESL and EFL contexts tends to be more analytic (see [23]). On the other hand, English is used much more widely and 'naturally' in both ENL and ESL contexts for intranational purposes, while in EFL contexts English is taught and learned primarily as an international means of communication without the EFL learners having much exposure to the target language. Therefore, an increase in fluency performance from EFL to ESL to ENL can be expected.

Against this background, this paper investigates a speaker's fluency level (1) on the one hand, by looking at his/her use of temporal fluency features (here, I count unfilled pauses [UPs] and measure the mean length of runs [MLR]), and (2) on the other hand, by seeing how a speaker employs "fluency-enhancement strategies", i.e. strategies of "dealing with speech planning phases and easing processing pressure" [8, 42]. In the pilot study, I included the following three categories that are used as fluency-enhancement strategies:

1. Discourse markers ([DMs], viz. *you know, you see*, *well*, *like*, *right*, *all right*, okay, *I don't know*);
2. Repeats that usually tend to occur with a high frequency at the beginning of an

utterance to ease the processing load [2] ([REPs], viz. pronouns, articles/determiners, possessive determiners, conjunctions, prepositions, high-frequency verbs and subject-verb-contractions); and

3. Smallwords, i.e. "small words and phrases, occurring with high frequency in the spoken language, that help to keep our speech flowing" [12] ([SWs], viz. *sort of*/*sorta*, *kind of*/*kinda*, *quite*).

In order to test if there are variant-specific preferences of using certain fluency-enhancing strategies within speaker groups, the use of these strategies will be set in relation to one another across the three variants. To this end, I investigated Great Britain from the inner circle, Sri Lanka from the outer circle and Germany from the expanding circle.

Thus, the aims of the present pilot study are

1. To test if temporal fluency increases from ENL to ESL to EFL and thus reflects the 'cline model';
2. To test whether there are variant-specific preferences in the use of fluency-enhancing strategies in ENL, ESL and EFL variants of English.

3. DATABASE AND METHODOLOGY

As the present pilot study investigates selected temporal fluency variables and fluency-enhancement strategies whose frequency and distribution are very prone to genre- and register differences, it is important to use comparable datasets for the analysis. However, since a spoken corpus including all three ENL, ESL and EFL variants is not (yet) available, the most practical solution for the time being is to use comparable corpus data from two different projects including similar and thus comparable genres and registers, namely the *International Corpus of English* (ICE; cf. [10]) and the *Louvain International Database of Spoken English Interlanguage* (LINDSEI; cf. [6]). I thus chose one inner circle and one outer circle 40,000-word subcorpus, compiled from the 'private conversations' and 'broadcast discussions' sections of the British and the Sri Lankan components of ICE (ICE-GB and ICE-SL), and the German component of LINDSEI (LINDSEI-GE; cf. [3]), a c. 86,000-word learner corpus compiled from 50 interviews with advanced German learners of English at the university level.

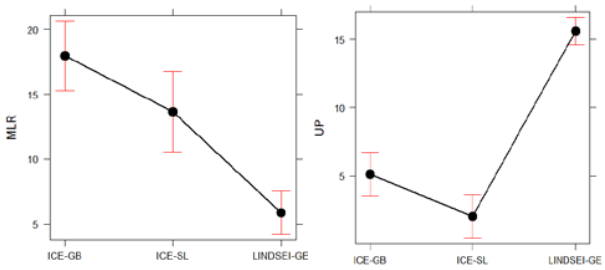
The analysis of (dis)fluency features was done in two steps: In a first step, all (dis)fluency features were automatically extracted from the corpora by using the corpus-analysis software *Wordsmith* v. 5.0; cf. [20] and then manually disambiguated (for discourse markers, repeats and smallwords) in a second step. The MLR was calculated semi-automatically. A run is defined as being a speaker's utterance until the occurrence of a disfluency. All the uninterrupted runs were measured manually and their mean value was then calculated automatically. Linear regression models were then fitted [11] to predict the effect of the variant type on the investigated fluency variables and fluency-enhancing strategies using the software package *R* [19].

4. FINDINGS

4.1. Temporal fluency

For the temporal variables, the hypothesized distinct cline in fluency from EFL to ESL to EFL is only visible in the variable MLR. A linear model revealed that the variant types differ significantly from each other (adjusted R^2=0.42, F=31.88, df=82, p<5.73e-11). As it becomes clearly visible from the left-hand panel in Figure 1, the model predicts the longest MLRs for ENL speakers of British English, followed by the ESL speakers of Sri Lankan English. The model predicts by far the shortest MLRs for the EFL speakers from Germany.

Figure 1: Effect plots of predicted MLR in ICE-GB, ICE-SL and LINDSEI-GE (left-hand panel) and number of UPs phw in ICE-GB, ICE-SL and LINDSEI-GE (right-hand panel).



For the frequency of UPs, the model revealed that, again, the variants differ significantly from each other. This model even accounts for 74% of the variance in the data (adjusted R^2=0.74, F=129.6, df=87 DF, p<2.2e-16). However, for UPs the model's predictions do not follow the hypothesized cline: The model predicts by far the highest frequency of UPs for German EFL this time, however, followed by British ENL, and the lowest frequency is predicted for Sri Lankan ESL. These findings are illustrated in the right-hand panel in Figure 1.

4.2. Fluency-enhancement strategies

The analysis of the functions of fluency-enhancement strategies across variants yields highly significant differences between the three variants (X^2=893.52, *df*=6, *p*<2.2e-16). These are illustrated in the association plot in Figure 2, where a significant overrepresentation is illustrated by a black upward block and a significant underrepresentation is illustrated by a red downward block. The frequency of a feature is accounted for in the width of the blocks (i.e. FPs are the most frequently used strategies and SWs the least frequently used ones).

Figure 2: Association plot of fluency-enhancement strategies (viz. FPs, SWs, REPs and DMs) in ICE-GB, ICE-SL and LINDSEI-GE.

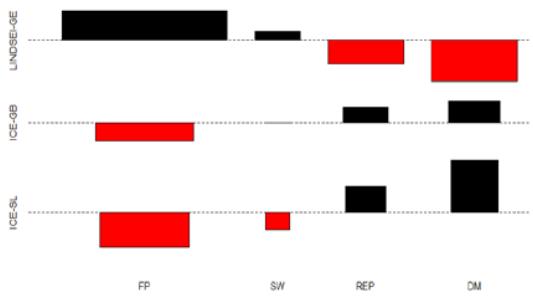
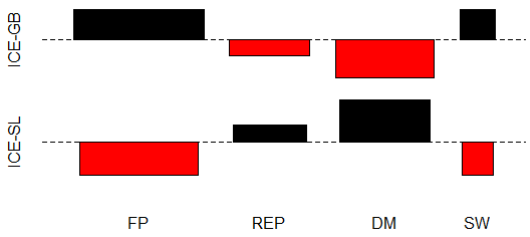


Figure 2 clearly illustrates that there is a large area of overlap in the preferred use of these strategies in ENL and ESL when compared to EFL (i.e. ENL and ESL speakers showing a preference of using more DMs and SWs than EFL speakers and a lower number of FPs). To test if possible differences between ENL and ESL went unnoticed by the highly different frequencies of all variables in the EFL data, I contrasted the ENL and ESL data separately from each other. This is reflected in the association plot in Figure 3.

Figure 3: Association plot of fluency-enhancement strategies in ICE-GB vs. ICE-SL.



There are indeed highly significant variant-specific preferences when we see which fluency-enhancement strategies are used in ICE-GB vs. ICE-SL. A comparison between British ENL and Sri Lankan ESL, again, reveals highly significant differences (X^2=70.60, *df*=3, *p*<3.182e-15). More specifically, the ESL variant clearly prefers to use REPs and DMs as fluency-enhancing strategies, whereas the British speakers have a marked preference for FPs and SWs.

5. CONCLUSION AND OUTLOOK

The present pilot study revealed that temporal fluency increases from EFL to ESL/ENL speakers. The German EFL speakers show by far the lowest temporal fluency reflected in their short mean length of runs and their high frequency of unfilled pauses. The same is reflected in the learners' immense overrepresentation of filled pauses as their main fluency-enhancement strategy. This is very much in line with the initial hypothesis that learners of English from the expanding circle who cannot take advantage of having much exposure to the target language in their everyday lives show a highly significantly lower temporal fluency (even though two thirds of the learners in the corpus spent time abroad in an English-speaking country).

However, things are somewhat different when we compare the British ENL speakers and the Sri Lankan ESL speakers, as the investigated fluency variables do not always follow the predicted cline of an increase in fluency from ESL to ENL. Speakers of ESL have a much lower number of unfilled pauses in their speech than ENL speakers, which is an indication of a high temporal fluency on the part of the ESL speakers. Also, when comparing the ESL speakers to the ENL speakers regarding the use of fluency-enhancement strategies, we find that ESL speakers have clear variant-specific preferences of establishing spoken fluency through strategies that are different from the British ENL speakers'. This seems to be one first indication that speakers of ESL variants of English can establish an even higher temporal fluency than ENL speakers regarding some variables. Also, ESL speakers establish spoken fluency by using different strategies than the other two variants. EFL speakers' fluency performance seems to be affected by the low degree of exposure to the target language, and by the same token, the ESL speakers' fluency performance seems to highly benefit from their exposure to English in their everyday lives.

However, the present study included one representative of each variant only and can thus only be seen as a preliminary pilot study that will need much more empirical back-up before it will be possible to make any valid claims across variant types. Future studies thus need to include more data of different variants from all three circles: 1) from

the inner circle, in order to determine whether there are differences within the circle of native speakers how they establish spoken fluency by using possibly different fluency-enhancement strategies; 2) from the outer circle that represent different "evolutionary stages" [21], in order to test if the trends found in the present pilot study can be extended to other variants at different levels of nativization, as such correlations have already been shown to exist at the lexicogrammatical level in previous studies (e.g. [16]); and 3) from the expanding circle that have different degrees of exposure to the target language, in order to determine if all learner variants show a lower degree of fluency than the representatives of the other two circles.

From the perspective of corpus comparability, it would be highly desirable to have a language database that includes spoken data of representatives of all three circles in order to be able to compare datasets that are maximally compatible.

7. REFERENCES

[1] Ballier, N., Martin, P. 2011. CV resyllabifications in French learner corpora and syllable-timed varieties of English. Paper presented at the Learner Corpus Research 2011. Université catholique de Louvain. Louvain-la-Neuve. 15-17 September 2011.

[2] Biber, D., Johannson, S., Leech, G. Conrad, S., Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

[3] Brand, C., Kämmerer, S. 2006. *The Louvain international Database of Spoken English Interlanguage*: Compiling the German component. In: Braun, S., Kohn, K., Mukherjee, J. (eds), *The Handbook of Phonetic Science*. Oxford: Blackwell, 462-506.

[4] Deshors, S. 2014. A case for a unified treatment of EFL and ESL: A multifactorial approach. *English World-Wide* 35(3): 279-307.

[5] Gilquin, G. 2015. At the interface of contact linguistics and second language acquisition research: New Englishes and Learner Englishes compared. *English World-Wide* 36(1): 91-124.

[6] Gilquin, G., De Cock, S., Granger, S. 2010. *The Louvain International Database of Spoken English Interlanguage*. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.

[7] Gilquin, G., Granger, S. 2011. From EFL to ESL: Evidence from the *International Corpus of Learner English*. In: [17], 55-78.

[8] Götz, S. 2013. *Fluency in Native and Nonnative English Speech*. Amsterdam: John Benjamins.

[9] Götz, S., Schilk, M. 2011. Formulaic sequences in spoken ENL, ESL and EFL: Focus on British English, Indian English and Learner English of advanced German learners. In: [17], 79-100.

[10] Greenbaum, S. 1991. ICE: the *International Corpus of English*. *English Today* 28(1): 3-7.

[11] Gries, S.T. 2013. *Statistics for linguistics with R*. 2nd rev. and ext. ed. Berlin & New York: De Gruyter Mouton.

[12] Hasselgren, A. 2002. Learner Corpora and language testing: Smallwords as markers of learner fluency. In: Granger, S., Hung, J., Petch-Tyson, S. (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins, 143-173.

[13] Housen, A., Schoonjans, E, Janssens, S., Welcomme, A., Schoonheere, E., Pierrard, M. 2011. Conceptualizing and measuring the impact of contextual factors in instructed SLA: The role of language prominence. *International Review of Applied Linguistics in Language Teaching* 49(2): 83-112.

[14] Kachru, B.B. 1985. Standards, codification and sociolinguistic realism: The English language in the outer circle. In: Quirk, R., Widdowson, H.G. (eds), *English in the World: Teaching and learning the language and literatures*. Cambridge: Cambridge University Press, 11-30.

[15] Laporte, S. 2012. Mind the gap! Bridge between World Englishes and Learner Englishes in the making. *English Text Construction* 5(2): 265-292.

[16] Mukherjee, J. & S.T. Gries. 2009. Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English. *English World-Wide* 30(1): 27-51.

[17] Mukherjee, J., Hundt, M. (eds) 2011. *Exploring Second-language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins.

[18] Nesselhauf, N. 2009. Co-selection phenomena across New Englishes: Parallels (and differences) to foreign learner varieties. *English World-Wide* 30(1): 1-26.

[19] R Development Core Ream. 2010. R: A Language and Environment for Statistical Computing. Foundation for statistical computing. Vienna, Austria. <http://www.R-project.org>.

[20] Scott, M. 2008. WordSmith Tools version 5. Liverpool: Lexical Analysis Software.

[21] Schneider, E.W. 2003. The dynamics of new Englishes: From identity construction to dialect birth. *Language* 79(2): 233-281.

[22] Sridhar, K.K., Sridhar, S.N. 1986. Bridging the paradigm gap: second language acquisition research and indigenized varieties of English. *World Englishes* 5(1): 3-14.

[23] Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

[1] I am extremely grateful to Justus Liebig University Giessen for their generous financial support of the project in which this pilot study emerged, to Joybrato Mukherjee for kindly sharing the spoken pilot corpus of ICE-SL, to the student assistants for analyzing ICE-GB and ICE-SL with great care and to Rosemary Bock for her wonderful editing and proofreading of this manuscript. Of course, all remaining errors are my responsibility alone.

COMPLEXITY CUES OR ATTENTION TRIGGERS? REPETITIONS AND EDITING TERMS FOR NATIVE SPEAKERS OF FRENCH

Iulia Grosman

Université catholique de Louvain
iulia.grosman@uclouvain.be

ABSTRACT

A growing stream of research shows evidence of the metalinguistic information that disfluencies (*silent and filled pauses, repetitions, false-starts, repairs, etc.*) can display to listeners. As a result, disfluencies may work as fluent devices. By means of a decision task latencies, this study investigates whether lexical repetition co-occurring with an editing term affects the perception of native speakers of French. There is a lack of consensus in the literature: do disfluencies trigger conceptual priming of complex entity or act simply as attention cues? Results from multiple analysis of variance and linear mixed-effect modelling show that the presence of a disfluency triggers a faster response from the participant, however complex the following noun-phrase might be, supporting the hypothesis that repetition and co-occurring editing terms act as cognitive signposts rather than as cues of complexity of an upcoming event.

Keywords: Disfluencies, reaction time, perception, prosody, repetitions, French.

1. INTRODUCTION

Literature qualifies oral speech production as notoriously disfluent (i.e. containing silent and filled pauses, false-start, prolongations, repetitions, substitutions, deletions, insertions, etc.): disfluency rate range from 1 up to 6% of speech in conversation [9], varying according to multiple parameters as cognitive, social and situational factors that may interact to affect speech production (age, gender, familiarity of the subject, etc.) [15, 4]. In perception, most of these occurrences do not lead to comprehension issue. Research agrees that disfluencies may work as fluent devices as they depict information about the speaker's confidence, his planning difficulties, and are even used as tools to coordinate conversational interaction [18].

The goal of a cognitive approach to disfluencies would be to determine on the one hand which factors are responsible for disfluencies' production in everyday speech and on the other hand to understand how speech containing disfluencies would affect the processes of the listeners in terms of comprehension, prediction, memory or attention [5]. Accordingly, in Segalowitz' terminology, the hypothesis underling this contribution works towards a better understanding of the *perceived fluency* by native French speakers or "the inferences hearers make about speakers' cognitive fluency based on their perceptions of the utterance fluency" [17]. Within this framework, the utterance fluency is accoustically measured on the generated audio stimuli while the experimental design connects it to the cognitive fluency (or the "efficiency of operation of the underlying processes responsible for the production of utterances") [17] by measuring latency of responses (RT) to a decision task with sentences including a disfluency, namely a lexical repetition and a co-occuring editing term.

1.1. The present study

Repetitions are the second most frequent marks of (dis)fluency, after filled and silent pauses: they represent an average 25% of all disfluency types in various discourse genres [12]. Disfluency have been sometimes depict as a tool to capture attention and/or a cue for conceptual priming of a complex entity [2, 19]. As priming facilitates perceptual processing, some disfluencies would make utterances more fluent. Literature already attests for the structuring function of silent and filled pauses, especially in the hearers' expectations about the complexity of an upcoming event [7, 19] and its relevance [2]. The field of linguistics tends to assume that all disfluencies types display metalinguistic information [6] but neurological work attested that "not all disfluencies are equal": they do not affect underlying cognitive processes of speech perception identically in terms of prediction, attention, comprehension or memorization [13]. However, apart from works by [9], [16] and [19] little experimental evidence exists on the hypothetical coherence and fluency function of disfluencies other than filled pauses (FP), such as repetitions, false starts or reformulations.

1.2. Hypothesis

According to the fact that, as the simplest exemplars of a target category, visually non compound stimuli (a sphere without *vs* with arrows) are inherently more fluent (i.e. easier to process) [1], we predict that RTs to simple shapes would be shorter than to compound ones. Secondly, we hypothesize that RTs to compound shapes would be shorter when preceded by a repetition, than when there is none, as they might give a cue to the speaker's planning difficulty and allow some time to predict the content of the upcoming utterance [19]. Thirdly, due to FPs' relative salience in the speech flow, we predict an additional result of disfluencies: repetitions containing an editing term will result in quicker response than simple repetitions.

2. METHOD

2.1. Participants and Design

Inspired by [19], the within repeated measures design invited 62 native speakers of French to listen once to 96 sentences describing both a simple and a compound shape (*complexity factor* – as verbally, linguistic complexity in surface generally "includes more words"). Their task was to listen once to each sentence and press the left or right key as soon as, from a displayed pair on a computer screen, they identified the shape corresponding to the description.

The paired shapes (Figure 1) were always of the same type and colour so that these features would not affect the task. The simple shape was always on the left side and the compound one on the right. The actual measurement of the exhaustion of the free capacity (log(RT)) is used as a concrete definition of "retrieval disfluency", the speed with which objects are accessed. They were automatically extracted by subtracting the word onset-time (disambiguating syllable) from the RT measured from the start of the sound files.

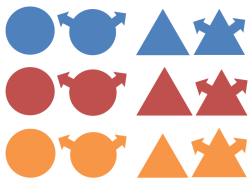


Figure 1: Example of 6 paired stimuli

2.2. Material

In summary, the experiment consisted of 96 items (24 similar items in 4 different conditions). The shapes in the sentences were either preceded by a repetition including (1) a FP (*original filler repetition*), (2) a silent pause of same duration (*repetition with silent pauses)*, (3) no pause (*contiguous repetition*) or (4) no repetition at all (*no repetition*) (*fluency factor*). In addition, each utterance contained a phrase describing a colour and a shape in this order (e.g. "Bring me the uhm the red circular paper from the living room" (simple shape with filler repetition), "Bring me the the orange circular paper with arrows from office" (compound shape with contiguous repetition)). All sentences were transcribed and phonetically aligned under Praat [3] using EasyAlign [11].

A thorough prosodic analysis of the 24 original sentences including a filler and a repetition were pronounced twice by a French native speaker, allowing the selection of stimuli with very little prosodic variation. The original utterances including a filler repetition were used to generate the 3 other sets of conditions by editing the appropriate audio segment. Using Prosogram v2.9 [14], the 24 stimuli showing the least prosodic variation were selected according to the smallest relative distance from the mean and median of : the speech rate, the pitch range in semitones, the top and bottom pitch, the pitch mean and median, the mean proportion of glissando, rises and falls of each audio-sample, the duration of the schwa in "le" (the first "le" in "Apporte-moi le euh le papier bleu triangulaire qui est dans le bureau") and the relative duration of the filler "euh".

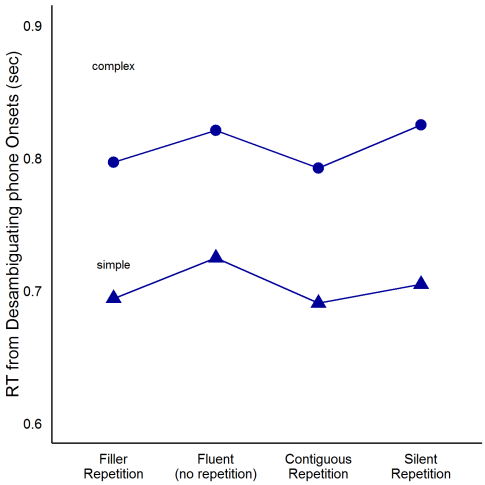
T-tests for the two levels of complexity were separately performed for each prosodic feature and none of them was proven significant apart from the absolute speech time of the stimuli between the two complexity conditions. On average compound shape stimuli are 600 ms. longer than simple stimuli. Filled pauses in the simple condition were not longer than in the compound shape condition. Schwas before the repetition were uttered with a homogeneous lengthening across conditions ($M = 350$ ms., $SD = 38$ ms.), in accordance with the tendency observed on the first part of repetitions and segments preceding a FP [18].

3. MAIN RESULTS

3.1. Error Rates and Types

Overall, the mean rate of correct responses was 98.147% ($SD = 1.73$), results comparable with [19]. Precocious responses (i.e. anticipated responses made before the disambiguating sound) were discarded and treated as errors. Error rates were significantly different in the simple and compound conditions ($t(4.92) = 4.92, p < 0.005$), supporting the fact that the compound shapes were indeed of higher complexity.

Figure 2: RT across conditions



Roughly, the participants pressed a key between 0.55 and 0.70 sec. after the onset of the disambiguating syllable (Fig. 2). As in [19], a two way repeated measures analysis of variance (ANOVA) revealed an expected main effect of the complexity factor $(F(1,61) = 26.76, p < 0.001)$. As the first hypothesis predicted, latency is indeed shorter for simple NPs than for complex ones. Moreover, contrary to the homologue Japanese study, we did find a significance of the main effect of the fluency factor as post-hoc comparisons using t-test with Bonferroni correction indicated that the mean scores within fluency conditions were statistically significant $(F(3,183) = 3.87, p < 0.02)$.

3.2. Linear mixed-effect Modeling

A linear-mixed effect model was selected based on AIC and log-likelihood $(pseudo - R^2 = 57\%)$. Aside from random items and subjects, the model included complexity and fluency factors as well as the position of the item in the experiment and the results of the participants to the Cognitive Reflection Test (CRT), a simple measure of cognitive ability associated with analytic thinking [10]. The latter parameter adds robustness to the model as it is highly correlated with the intercept $(r = 0.821)$.Table 1 shows the results for fixed effects (with default contrast treatment). The intercepts represent the original *filler repetition in the compound shape condition* $(M = (e^{660.82/1000}) = 1.94$ sec.). Increased knowledge about participants (age, gender, musical knowledge or bilingual aptitude) were not included in the model as they did not increase the explained variance.

Firstly, the negative estimates on the CRT reveals that the lower the participants score on the Cognitive Reflection Test (i.e. make assumption or anticipate on the upcoming content), the fastest participants react to the stimuli. Overall, RTs were indeed 1.65 sec. shorter for simple NPs than for complex ones. Moreover, the fluent condition elicited longer RTs than all other conditions, for both compound and simple shapes $(+1.01$ sec., $p < 0.05)$.

Table 1: Fixed effects

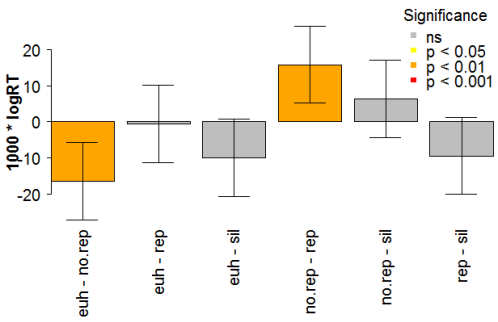
	Estimate (log)	Estimate (sec)	Df	t value	Sig. (p)
(Intercept)	660.82	1.94	66.13	20.62	<0.001***
Filler rep.- without arrows	-50.25	-1.65	90.52	-10.78	<0.001***
No rep. (fluent)	+13.95	+1.014	90.20	2.12	0.03*
Contiguous rep.	+0.60	1.000	90.23	0.09	0.93
Silent rep.	+10.79	+1.011	90.80	1.64	0.11
CRT scores	-1.79	-1.001	60.00	-4.01	<0.001***

Little variance is attested between other fluency levels. Post-hoc tests did show a significant difference among fluency conditions in the simple condition, $(F(3,58) = 2.79, p < 0.05)$, but no significant difference in the compound shape condition $(F(3,58) = 1.9, p = 0.131)$. These results contrast with [19] who reported and inverted significance. As a result, the second hypothesis according to which RTs to compound shapes would be shorter when there is a repetition than when there is none, and that RTs to simple shapes might be longer with a repetition than without can not be fully supported. In fact, due to the lack of interaction between fluency and complexity factors, we can solely conclude that, no matter the complexity condition, overall sentences with repetitions did induce shorter RTs than fluent ones. No matter the complexity condition, sentences with repetitions did induce shorter RTs than fluent ones.

The third hypothesis is partially confirmed. Overall, repetitions including a FP as an editing term show shorter RTs than their homologue silent version $(t(60) = 4.018, p < 0.001)$, thus leaning in favour of the interpretation of the filler and repetition additive weight. Nonetheless, there is no strong difference in latency between the filler condition and contiguous repetition condition, encouraging to consider that the decrease in latency may be due more to the presence of a repetition than to the presence of FP. Moreover, within the simple condition, replacing the FP by a silence doesn't significantly increase the latency, same is said when erasing the FP altogether $(+1.011$ and $+1.00$ sec. respectively$)$.

To summarize, Figure 3 illustrates the differences in least square means across fluency conditions and their confidence interval (based on Kenwar-Roger approximation). The least mean square differences were also found shorter for sentences including a

Figure 3: Differences in Least Square means for fluency factor with 95% Confidence Intervals



repetition with silence than for fluent ones, but not significantly. In decreasing order, fluent sentences, repetitions with silent pauses and contiguous repetitions show longer RTs than repetitions with filler. Furthermore, latency to contiguous repetitions are also significantly shorter than their fluent version ($p < 0.001$ for no.rep-sil).

4. DISCUSSION AND CONCLUSION

The study tested the hypothesis according to which hearers expect rather long or complex phrases when preceded by a disfluency. Our results do not entirely converge with previous studies: no interaction was found between fluency and complexity factors. Repetition and its editing terms do not seem to facilitate processing or bias hearers' expectations when complexity rises which does go in line with [13] and [8]. Significant difference in RTs was observed within the fluency condition for simple shapes but not for compound ones. Results show that native speakers of French are quicker to respond to compound and simple shapes when there is a contiguous repetition or a repetition with a FP. In decreasing order, fluent sentences, repetitions with silent pauses and contiguous repetitions show longer RTs than repetitions with filler. The findings seem to support that the combination of repetition with FP does not have an additive effect. According to the results, repetitions and FPs would cause hearers to react faster; they can thus be interpreted as a cue to catch their attention rather than as a complexity cue.

5. REFERENCES

[1] Alter, A. L., Oppenheimer, D. M. 2009. Uniting the tribes of fluency to form a metacognitive nation. *Pers. Soc. Psychol. Rev.* 13(3), 219–35.

[2] Arnold, J. E., Fagnano, M., Tanenhaus, M. K. 2003. Disfluencies signal theee, um, new information. *Journal of psycholinguistic research* 32(1), 25–36.

[3] Boersma, P., Weenink, D. 2015. Praat: doing phonetics by computer.

[4] Bortfeld, H., Leon, S., Bloom, J., Schober, M., Brennan, S. 2001. Disfluency Rates in Conversation : Effects of Age, Relationship, Topic, Role, and Gender. *Lang. Speech* 44(2), 123–147.

[5] Bosker, H. R., Quené, H., Sanders, T., de Jong, N. H. 2014. The Perception of Fluency in Native and Nonnative Speech: Native and Nonnative Fluency Perception. *Lang. Learn.* 64(3), 579–614.

[6] Brennan, S. E., Williams, M. 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *J. Mem. Lang.* 34, 383–398.

[7] Clark, H. H., Tree, J. E. F. 2002. Using uh and um in spontaneous speaking. *Cognition* 84(1), 73–111.

[8] Corley, M., Hartsuiker, R. J. 2011. Why Um Helps Auditory Word Recognition: The Temporal Delay Hypothesis. *PLoS ONE* 6(5), e19792.

[9] Fox Tree, J. E. 1995. The effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *J. Mem. Lang.* 34(6), 709–738.

[10] Frederick, S. 2005. Cognitive Reflection and Decision Making. *J. Econ. Perspect.* 19(4), 25–42.

[11] Goldman, J.-P. 2011. EasyAlign: an automatic phonetic alignment tool under Praat. *Proc. of Interspeech 2011.* Firenze, Italy. ISCA Archive. 3233–3236.

[12] Grosman, I. 2014. Dynamics of disfluencies within Basic Discourse Units. *NCLF* 31, 45–50.

[13] MacGregor, L. J., Corley, M., Donaldson, D. I. 2009. Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension. *Brain Lang.* 111(1), 36–45.

[14] Mertens, P. 2004. Un outil pour la transcription de la prosodie dans les corpus oraux. *Traitement Automatique des langues* 45(2), 109–130.

[15] Moniz, H., Batista, F., Mata, A. I., Trancoso, I. 2014. Speaking style effects in the production of disfluencies. *Speech Commu.* 65, 20–35.

[16] Moniz, H., Trancoso, I., Mata, A. I. 2009. Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts. Schuller, B., Steidl, S., Batliner, A., (eds), *Proc. of Interspeech 2009.* Brighton, U.K. ISCA Archive. 1719–1722.

[17] Segalowitz, N. 2010. *Cognitive Bases of Second Language Fluency.* New York: Routledge.

[18] Shriberg, E. 1995. Spontaneous Speech : How People Really Talk and Why Engineers Should Care. *Proc. of Eurospeech.* ISCA Archive. 1781–1784.

[19] Watanabe, M., Hirose, K., Den, Y., Minematsu, N. 2008. Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Commun.* 50(2), 81–94.

STUDYING THE DYNAMICS OF LEXICAL ACCESS USING DISFLUENCIES

Zara Harmon, Vsevolod Kapatsinski

Department of Linguistics, University of Oregon, Eugene, OR, USA
(zforough, vkapatsi)@uoregon.edu

ABSTRACT

Faced with planning problems related to lexical access, speakers take advantage of a major function of disfluencies: buying time. It is reasonable, then, to expect that the structure of disfluencies sheds light on the mechanisms underlying lexical access. Using data from the Switchboard Corpus, we investigated the effect of semantic competition during lexical access on repetition disfluencies. We hypothesized that the more time the speaker needs to access the following unit, the longer the repetition. We examined the repetitions preceding verbs and nouns and tested predictors influencing the accessibility of these items. Results suggest that speed of lexical access negatively correlates with the length of repetition and that the main determinants of lexical access speed differ for verbs and nouns. Longer disfluencies before verbs appear to be due to significant paradigmatic competition from semantically similar verbs. For nouns, they occur when the noun is relatively unpredictable given the preceding context.

Keywords: Repetition, lexical access, semantic competition, sentence planning, lexicalization.

1. INTRODUCTION

Several studies have suggested that disfluencies are used by speakers to buy time until the selection and planning of upcoming items is completed [9, 26]. This claim is supported by studies that show an increase in the rate of disfluneices as a result of utterance length [28] and difficulties in planning [21]. But more specifically, it is supported by the fact that disfluencies appear in production before words that are hard to access. Words may be hard to access because they are unpredictable [8, 2] or infrequent [17, 2], or because there are many semantically similar words that compete with them for selection [14, 25]. This difficulty in access is especially severe if the competitors are more frequent than the target word [27] or have been made highly accessible through priming [18]. The stronger the competition, the more time is needed for the lexical selection process. Disfluencies may therefore be longer in highly competitive contexts, as the speaker needs to buy more time to plan the upcoming material.

In the current paper, we examine this hypothesis for repetition disfluencies preceding nouns and verbs. Theories of sentence production differ on whether verbs and nouns are selected at the same time during planning. On the one hand, both are lexical rather than functional items and therefore might be expected to be selected at the same time following the building of the sentence's structural frame [6]. On the other hand, verbs have been suggested to project the sentence's argument structure into which nouns (or noun phrases) are then slotted or, in constructionist approaches to syntax, to be tightly fused with the sentence's argument structure construction, which again is selected early in sentence planning [10]. If verbs are selected before nouns, we expect to see predictors of accessibility vary with lexical category of the target, as nouns are selected in the context of verbs while verbs are selected in a more context-independent manner. Differences in predictors of accessibility for nouns vs. verbs, as reflected in lengths of repetition disfluencies preceding these items, may therefore shed light on the time course of sentence planning.

2. METHODS

2.1. Repetition disfluencies

Repetition disfluencies are interruptions in the flow of speech followed by repetition of one or more of the items preceding the interruption point. We limited our study to instances of one-word and two-word repetitions preceding main verbs and nouns in the Switchboard Corpus [7]. Some data were excluded. Repetitions that were within one word of the preceding clause boundary were excluded. This is due to the fact that repetitions never span clause boundaries, so in these cases, the speaker would be limited to produce a one-word repetition only. Cases in which another disfluency immediately preceded or followed the repetition disfluency were also excluded. Our data included 2858 verbs and 1899 nouns following one-word repetitions and 776 verbs and 452 nouns following two-word repetitions. Examples of one-word and two-word repetitions are shown in (1) and (2). The '+' indicates the

interruption point and the '[]' mark the disfluency boundary.

1) I think it gets them prepared [to, + to] learn how to volunteer as they get older.
2) Now I'm probably [going to, + going to] upset you.

Henceforth, we use the following templates to refer to the structure of repetitions in the preceding and following context of the interruption point. 'W' stands for 'word' and the subscriptions denote the distance from the interruption point. The word(s) within the brackets on either side of the interruption point is (are) the same words. W_1 is either a noun or a verb and immediately follows the repetition.

- One word repetitions
  $W_{-3}\ W_{-2}\ [W_{-1} + W_{-1}]\ W_1$
- Two-word repetitions
  $W_{-3}\ [W_{-2}\ W_{-1} + W_{-2}\ W_{-1}]\ W_1$

2.2. Cohesion and interruptibility

Besides accessibility, the length of a repetition disfluency may be influenced by characteristics of the preceding context. In particular, studies have shown that production tends not to restart from the middle of a cohesive unit [5]. Words, as the paradigm examples of cohesive units in language, are impervious to restarts: following an interruption, production always restarts from at least as far back as the preceding word boundary, in (3) and (4), and never from the middle of the word, hence the ungrammaticality of (5).

3) I [had a similar, + a similar] health plan.
4) I [had a similar, + similar] health plan.
5) * I [had a similar, + -milar] health plan.

But not all words are the same when it comes to interruptibility. Less frequent words are more susceptible to interruption than frequent ones. Logan [15] shows that in stop-signal tasks, it is more difficult to stop typing the word *the* compared to other less frequent words such as *thy*. There is also evidence that frequent words are less likely to be interrupted prior to completion than infrequent words [12, 23]. Likewise, cohesion of a unit bigger than a word may result from a high degree of co-occurrence between words rather than being completely determined by syntactic constituency. For example, sentence comprehension is sensitive to frequencies of compositional four-word strings that are not syntactic constituents (e.g. *in the middle of* vs. *in the side of*) even when frequencies of the component units are controlled [29]. This is in line with usage-based linguistic theory, which claims that "units used together fuse together" [4].

According to Levelt [13], speech production is restarted from the nearest major syntactic constituent boundary. However, the restart location is also influenced by co-occurrence, more specifically backward transitional probability: the probability of a word given the following word [11]. Speakers tend not to restart speech from transitions of high backward transitional probability. This observation is illustrated in (6) and (7). The probability of *of* before *the* is higher than the probability of *for* before *the*. While the speaker has an overall tendency to repeat as little as possible, thus interrupting the fairly cohesive *for the*, they do not interrupt the even more cohesive *of the* despite the nearest constituent boundary being before *the* in both cases.

6) That place is known for [the, + the] rudest waitresses.
7) The crime level is not as high as it is in other areas [of the, + of the] city.

High cohesion between words preceding the interruption point may therefore cause speakers to repeat more than one word, avoiding interrupting a cohesive unit. Thus, we need to take into account characteristics of the context preceding the disfluency before arguing that repetition length is affected by accessibility of the following word.

2.3. Independent measures

2.3.1. Accessibility

Influences on lexical accessibility include the frequency of the upcoming word [19], the number of semantic competitors (synonyms) of the word, and their frequencies [27]. Words that are infrequent relative to their competitors and that have many competitors are expected to be harder to access [16] and therefore be preceded by longer repetitions. A machine-readable version of Roget's Thesaurus [24] was used to retrieve and count the number of synonyms for verbs and nouns following disfluencies, but the synonyms were limited to the ones found in the Switchboard Corpus. Frequencies of verbs and nouns and their synonyms were then retrieved from the Switchboard Corpus. So, the predictors for measuring accessibility include: Frequency of the verb or noun following the disfluency; probability of the word following the disfluency given the word that precedes it (i.e., forward transitional probability or FTP) – $p\ (W_1|W_{-1})$; and competition index (the product of number of synonyms for the noun or verb and mean frequency of the synonyms).

2.3.2. Cohesion

Based on the findings of previous work on repetiion repair [11], backward transitional probability (BTP) was used in the model as the best index of cohesion

as it impacts repetition length. If BTP of W_{-2} or p (W_{-2} | W_{-1}) is high, sequence $W_{-2}W_{-1}$ is cohesive, so speakers should avoid restarting the speech from the middle of it, which would result in a longer (two-word) repetition. On the other hand, if BTP of W_{-3} or p (W_{-3} | W_{-2}) is high, sequence $W_{-3}W_{-2}$ is highly cohesive, which should prevent the repetition of W_{-2} because repetition of W_{-2} would require interruption of the cohesive sequence $W_{-3}W_{-2}$.

2.4. Analysis

Multimodel inference with logistic regression [3] was used to assess the predictors (MuMIn package in R [1]). In this method, models containing all possible subsets of predictors are built from a complex regression model. Coefficients are then derived by averaging across models, weighting each model in proportion to its predictiveness. There were no random effects in the model, as models with random effects (of speaker and following word) would not converge given the small number of observations per speaker and per following word. Note that to reduce skew, predictors were scaled using log, rank, and square root transformations. Therefore, the magnitudes of coefficients of the various predictors within a model cannot be directly compared.

3. RESULTS

The results of the analysis for verbs are summarized in Table 1. High BTP of W_{-2} resulted in longer repetitions, and high BTP of W_{-3} resulted in shorter repetitions (*p*s < .001). Thus, speech production tends not to restart from transitions with high backward transitional probability. In addition, there was an effect of accessibility of the upcoming verb: verbs with higher frequency tend to follow shorter (one-word) repetitions ($p < .001$) and verbs that face high competition tend to follow longer (two-word) repetitions ($p = 0.0139$).

Table 1: Model averaged coefficients for verbs.

Predictors	Estimate	z value	Pr(>|z|)
(Intercept)	2.986	6.794	< 2e-16
Verb Frequency	-0.1548	4.103	4.08e-05
BTP($W_{-2}W_{-1}$)	0.7702	15.90	< 2e-16
BTP($W_{-3}W_{-2}$)	-0.00066	8.655	< 2e-16
Competition	0.001935	2.459	0.0139
FTP(W_1|W_{-1})	-0.0128	0.470	0.6384

The results of the analysis for nouns are summarized in Table 2. Comparable to verbs, high BTP of W_{-2} resulted in longer repetitions, and high BTP of W_{-3} resulted in shorter repetitions (*p*s < .001). Thus, speech production tends not to restart from transitions with high backward transitional probability. In addition, there is an effect of accessibility: nouns with higher predictability (forward transitional probability) tend to follow shorter repetitions ($p = 0.0142$).

Table 2: Model averaged coefficients for nouns.

Predictors	Estimate	z value	Pr(>|z|)
(Intercept)	-0.351	0.535	0.5924
Noun Frequency	0.068	1.059	0.2898
BTP($W_{-2}W_{-1}$)	0.421	9.289	< 2e-16
BTP($W_{-3}W_{-2}$)	-0.002	8.231	< 2e-16
Competition	0.062	1.032	0.3020
FTP(W_1|W_{-1})	-0.129	2.451	0.0142

4. DISCUSSION

For both verbs and nouns, disfluency length was significantly influenced by both accessibility of the upcoming word and cohesion of the preceding string. Cohesive word sequences preceding hard-to-access nouns or verbs are likely to be repeated as a unit. However, the best predictors of accessibility were different for nouns and verbs. For verbs, accessibility was best captured by context-independent measures: frequency and the cumulative strength of semantic competitors. For nouns, accessibility was best captured by probability of noun given the preceding context. In other words, longer disfluencies before verbs happen when the verb is facing significant paradigmatic competition from semantically similar verbs. Longer disfluencies before nouns appear to occur when the noun is relatively unpredictable given the preceding context. This is consistent with the idea that sentence planning consists of selecting a grammatical frame / argument structure construction first, followed by filling in the content words during lexicalization [6] as long as verbs are considered to be tightly fused with the argument structure construction [10]. After verbs are specified during the selection of argument structure construction, nouns are selected and filled in during lexicalization, rendering the choice of noun a more context dependent process.

Why is backward transitional probability a reliable predictor for the length of repetition? Our current interpretation is that BTP is indexing cohesion in the sense of co-occurrence. Studies of statistical learning show learners to be sensitive to BTP in segmenting units out of a continuous speech stream [22]. However, BTP in pre-posing languages such as English is highly correlated with syntactic constituency [11, 20]. So, it is possible that BTP acts as an index of constituency rather than an independent influence on it.

5. CONCLUSION

We investigated factors that predict the lengths of repetition disfluencies based on the preceding and following context. We found that cohesive units preceding interruption point are less likely to be interrupted when restarting production. Additionally, speed of lexical access of verbs or nouns following the disfluency negatively correlates with the length of repetition. However, the best predictors of accessibility differ across lexical category. Frequency and semantic competition are best predictors of the length of repetition before upcoming verbs while the best predictor of repetition length for nouns is contextual predictability. This difference may be due to differences in the timecourse of planning, namely that accessing nouns occurs at a later stage than accessing verbs. The results also corroborate the hypothesis that repetition disfluencies are used to buy time for lexical access.

6. REFERENCES

[1] Barton, K. 2015. MuMIn: Multi-Model Inference. R package version 1.13.4. http://CRAN.R-project.org/package=MuMIn

[2] Beattie, G. W., Butterworth. B. L. 1979. Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech* 22, 201-211.

[3] Burnham, K. P., Anderson, D. R. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edition. New York: Springer.

[4] Bybee, J. 2002. Sequentiality as the basis of constituent structure. In: Givon, T., Malle, B. F. (eds), *The Evolution of Language out of Pre-Language*, Amsterdam: John Benjamins. 109-134.

[5] Clark, H. H., Wasow, T. 1998. Repeating words in spontaneous speech. *Cognitive Psychology* 37(3), 201-242.

[6] Garrett, M. F. 1980. Levels of processing in sentence production. In: Butterworth, B. (ed), *Language Production (Vol. 1: Speech and Talk*). London: Academic Press. 177-220.

[7] Godfrey, J. J., Holliman, E. C., McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *Proc. IEEE ICASSP '92* San Fransisco: 517-520.

[8] Goldman-Eisler, F. 1958. Speech production and the predictability of words in context. *Quart. J. Exp. Psychol.* 10, 96-106.

[9] Goldman-Eisler, F. 1968. *Psycholinguistics: Experiments in Spontaneous Speech,* London: Academic Press.

[10] Hampe, B., Schonefeld, D. 2006. Syntactic leaps or lexical variation? – More on "Creative Syntax". In: Gries, S. T., Stefanowitsch, A. (eds), *Corpora in Cognitive Linguistics. Corpus-based Approaches to Syntax and Lexis,* Berlin: Mouton de Gruyter. 127-157.

[11] Kapatsinski, V. 2005. Measuring the relationship of structure to use: Determinants of the extent of recycle in repetition repair. *BLS* 30, 481–492.

[12] Kapatsinski, V. 2010. Frequency of use leads to automaticity of production: Evidence from repair in conversation. *Language and Speech*, 53(1), 71-105.

[13] Levelt, W. J. M. 1989. *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

[14] Levelt, W. J. M., Roelofs, A., Meyer, A. S. 1999. A theory of lexical access in speech production. *Beh. and Brain Sciences*, 22(01), 1-38.

[15] Logan, G. D. 1982. On the ability to inhibit complex movements: A stop-signal study of typewriting. *J. Exp. Psychol: Human Perc. & Perf.* 8(6), 778-792.

[16] Luce, P. A., Pisoni, D. B. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19(1), 1-36.

[17] Maclay, H., Osgood, C. E. 1959. Hesitation phenomena in spontaneous English speech. *Word* 15(1), 19-44.

[18] Meyer, A. S. 1996. Lexical access in phrase and sentence production: results from picture-word interference experiments. *J. Mem. & Lg.* 35, 477-96.

[19] Oldfield, R. C., Wingfield, A. 1965. Response latencies in naming objects. *Quart. J. Exp. Psychol*. 17(4), 273-281.

[20] Onnis, L., Thiessen, E. 2013. Language experience changes subsequent learning. *Cognition* 126(2), 268-284.

[21] Oomen, C. C., Postma, A. 2001. Effects of divided attention on the production of filled pauses and repetitions. *J. Sp. Lg. & Hear. Res*. 44(5), 997-1004.

[22] Perruchet, P., Desaulty, S. 2008. A role for backward transitional probabilities in word segmentation? *Memory & Cognition* 36, 1299–1305.

[23] Plug, L., Carter, P. 2013. Prosodic marking, pitch and intensity in spontaneous lexical self-repair in Dutch. *Phonetica* 70(3), 155-181.

[24] Roget, P. 1852. *Roget's Thesaurus of English Words and Phrases.* Harlow UK: Longman.

[25] Schachter, S., Christenfeld, N., Ravina, B., Bilous, F. 1991. Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology* 60, 362-267.

[26] Schegloff, E. A., Jefferson, G., Sacks, H. 1977. The preference for self-correction in the organization of repair in conversation. *Language* 53, 361–382.

[27] Schnadt, M. J. 2009. Lexical influences on disfluency production. Unpublished doctoral dissertation. University of Edinburgh.

[28] Shriberg, E. E. 1996. Disfluencies in switchboard. *Proc. 4th ICSLP* Philadelphia, Addendum 11-14.

[29] Tremblay, A., Baayen, R. H. 2010. Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In: Wood, D. (ed), *Perspectives on Formulaic Language: Acquisition and Communication*, London: Continuum. 151-173.

Disfluency incidence in 6-year old Swedish boys and girls with typical language development

Clara Hedenqvist[1], Frida Persson[1], Robert Eklund[2]

[1] Department of Clinical and Experimental Medicine, Division of Speech and Language Pathology, Linköping University, Linköping, Sweden
[2] Department of Culture and Communication, Division of Language and Culture, Linköping University, Linköping, Sweden
clara.hedenqvist@rjl.se, frida.persson@edu.norrkoping.se, robert.eklund@liu.se

ABSTRACT

This paper reports the prevalence of disfluencies in a group of 55 (25F/30M) Swedish children with typical speech development, and within the age range 6;0 and 6;11. All children had Swedish as their mother tongue. Speech was elicited using an "event picture" which the children described in their own, spontaneously produced, words. The data were analysed with regard to sex differences and lexical ability, including size of vocabulary and word retrieval, which was assessed using the two tests *Peabody Picture Vocabulary Test* and *Ordracet*. Results showed that girls produced significantly more unfilled pauses, prolongations and sound repetitions, while boys produced more word repetitions. However, no correlation with lexical development was found. The results are of interest to speech pathologists who study early speech development in search for potential early predictors of speech pathologies.

Keywords: Speech disfluency, children, lexical development, sex differences

1. INTRODUCTION

Disfluency is a naturally occurring phenomenon that occurs in with some degree of individual variation in all speakers. While disfluency in adult speech has been devoted a numbers of studies over the past decades, the speech of children have not been subject to a similar number of studies, and in most cases research on non-pathological child speech mainly occurs in control groups in studies of stuttering and other pathologies. The development of speech disfluency in children is to a large extent unknown, and several theories have been forwarded, focusing on different aspects of language acquisition.

2. PREVIOUS RESEARCH

As was pointed out in the previous paragraph, disfluency in adult speech has been thoroughly studied over the past decades, partly from a speech technological perspective. Disfluency research on child speech looks a little different. While it could be argued that disfluency studies in the speech of children commenced already in the seminal papers by Davis (1939, 1940), later studies have appeared at intermittent intervals, using different test and control groups, and to complicate matters further present the reader with the problem of the fast and complicated language and speech development in the young child; comparing a three-year-old with a five-year-old does not amount to the same thing. Or, to use the words of Bornstein, Hahn and Haynes (2004: 268): "At virtually every age, children vary dramatically in terms of individual differences in their language abilities".

Moreover, many previous studies have focused on stuttering, and children with typical, or non-pathological, speech development have quite often been included mainly as control groups. Thus, it could be argued that even after 75 years of studies, the development of speech disfluency in children is still to a large extent unknown.

Studies of disfluency rates in adult speech have established that around 6% of spontaneously produced speech exhibits some kind of disfluency (Fox Tree, 1995; Oviatt, 1995; Brennan & Schober, 2001; Bortfeld et al., 2001; Eklund, 2004). Studies of general disfluency rates in children have presented similar figures, e.g. Guitar (2013) who reported that preschool children were observed to produce around seven disfluencies per one hundred words, although the percentage of disfluencies in children vary more. It has also been reported that disfluency rates are higher as a function of age, with younger children being more disfluent than older children (Gordon & Luper, 1989).

Levin and Silverman (1965) compared fluency and hesitation in 48 children who told two stories and two different situations: to an audience and to a microphone when no one was present. Using the schemata developed by Trager (1958) and Mahl (1956) they found that speech production was consistent over the different conditions, but that stressful hesitations were responsive to whether the children were speaking in public or privately.

Levin, Silverman and Ford (1967) compared speech disfluency in 24 children, six each from

kindergarten, second, fourth and sixth grade. The children were shown simple physical demonstrations and were then asked to provide descriptions and explanations of these demonstrations. For children of all ages, explanations exhibited more words, more pauses and hesitations, lower speech rate and also longer pauses.

Speech disfluency in children has also been examined from a syntax perspective. Westby (1974) found that highly "dysfluent" children made significantly more grammatical errors than fluent children, irrespective of whether or not the children exhibited stuttering or typical development. Likewise, Haynes and Hood (1978:79), studying 5year-olds, found "a significant relationship between linguistic complexity and disfluency in children". On the other hand, Muma (1971) reported no significant relation between disfluency and syntax in a group of highly fluent and highly disfluent 4-year-olds. Pearl and Bernthal (1980), studying 3- and 4-year-olds, reported significantly more disfluencies in passive sentences than in any other sentence type. Similar results, this time for 5-year-olds, were reported by McLaughlin and Cullinan (1989). It has also been shown that speech disfluency in children has a tendency to be subject to clustering (Sawyer & Yairi, 2010; Colburn, 1985) and especially frequent when language development begins and reaches a level where (rudimentary) sentences are being produced (Colburn, 1985).

As for lexical factors, the literature presents inconsistent findings. While some studies on have reported that disfluencies predominantly are related to content words (Juste et al., 2012), the opposite observation has also been made, relating early disfluency primarily to function words (Silverman, 1974). Moreover, Westby (1974) reported that children with atypical lexical development had a higher incidence of disfluency than children with normal lexical development, while e.g. DeJoy and Gregory (1983) reported no significant effect for either vocabulary or syntax.

While sex differences in adult speech disfluency (males being more disfluent), has been reported for e.g. American English (e.g. Shriberg, 1994), no sex differences between the sexes was found in studies of Swedish adult speech (Eklund, 2004; Bell, Gustafson & Eklund, 2000). In a study on speech disfluency in English 7-year-old children (Kools & Berryman, 1971), no differences between the sexes were observed. Likewise, Yairi (1981) found no sex differences in the speech of 2-year-olds.

Finally, looking at disfluency as a function of age, Haynes and Hood (1977) studied speech disfluency in 4- and 8-year-olds and reported only very small, non-significant, differences. DeJoy and Gregory (1978) studied 3.5- and 5-year-olds and reported both similarities (ungrammatical pauses, revisions) and differences (the younger children produced more phrase repetitions, dysrhythmic phonations etc; the older children produced more grammatical pauses). Wexler and Mysak (1982) reported similar disfluency patterns in 2-, 4- and 6-year-old nonstuttering children.

Summing up, it should be obvious from the previous passages that the situation is far from clear when it comes to speech (dis)fluency in children. Whatever factor one looks at, results seem to have a tendency to "point both ways", and individual variation seems to play a very large role, as pointed out by Bornstein, Hahn and Haynes (quote above). However, this does not mean that the results are "pointless". From a clinical perspective, any piece of information available to the clinician might be valuable in that it provides the basis to more careful diagnosis and makes the acting clinician more aware of what potential factors and the degree and type of variation present in the developing child. Consequently, all the previous as well as future studies help shed a little light on what apparently is a very complex phenomenon.

3. THE PRESENT STUDY

The present study focused on speech disfluency in a group of 6-year-old children without any diagnosed speech pathology, with focus on three parameters: (1) a general mapping of disfluency in 6-year-old girls and boys; (2) sex differences with regard to disfluency production; and (3) possible correlation with lexical development, including an analysis of vocabulary size and lexical retrieval.

4. METHOD AND ANALYSIS

The study included 55 children, 25 girls and 30 boys between 6;0–6;11 years old. All children had Swedish as their mother tongue. To elicit speech an event picture was used, taken from a neurolinguistic aphasia examination package described in Lindström & Werner (2006). Lexical ability was examined using the two tests *Peabody Picture Vocabulary Test* (PPTV) (Dunn & Dunn, 1997) and "Ordracet" (Eklund, 1996). The speech data were recorded using a discreet iPhone iOS7 and the sounds files were later converted into wave format and analysed in Praat (www.praat.org). Statistics were calculated using SPSS 22 (www.spss.com).

Interlabeller reliability was tested both between the two main investigators (CH, FP) and an outside speech therapist (MK) and was found to be .94.

Table 1. Summary statistics for all disfluencies broken down for type, number, proportion/percentage and position in sentence/word. All disfluency figures are given as disfluencies/words. Girls and boys are presented in separate columns. Statistical significance is calculated performing *Z* test-of-proportions, two-tailed. In cases where the difference is significant, the more disfluent group is specified using ♀ for girls and ♂ for boys.

	Girls *N* = 25	Boys *N* = 30	Statistical Significance
Total number of words	**3377**	**4600**	**–**
Total number of disfluencies	**575 (17.0 %)**	**670 (14.6 %)**	**–**
Unfilled Pauses	**328 (57%)**	**360 (53.7%)**	***p* < 0.01** (♀)
Filled Pauses – All Positions	**113 (19.6%)**	**153 (22.8%)**	**n.s.**
Filled Pauses – initial	58 (10%)	75 (11.2%)	n.s.
Filled Pauses – medial	55 (9.6%)	78 (11.6%)	n.s.
Filled Pauses – final	0 (0%)	0 (0%)	-
Segment Prolongations – All Positions	**81 (14.1%)**	**49 (7.3%)**	*p* < 0.01 (♀)
Segment Prolongations – initial	26 (4.5%)	12 (1.8%)	*p* < 0.01 (♀)
Segment Prolongations – medial	32 (5.6%)	13 (1.9%)	*p* < 0.05 (♀)
Segment Prolongations – final	23 (4%)	24 (3.6%)	n.s.
Sound Repetitions – All positions	**5 (0.9%)**	**6 (0.9%)**	**n.s.**
Sound Repetitions – initial	2 (0.4%)	6 (0.9%)	n.s.
Sound Repetitions – medial	3 (0.5%)	0 (0%)	*p* < 0.05 (♀)
Sound Repetitions – final	0 (0%)	0 (0%)	-
Syllable Repetitions – All positions	**6 (1%)**	**12 (1.8%)**	**n.s.**
Syllable Repetitions – initial	6 (1%)	12 (1.8%)	n.s.
Syllable Repetitions – medial	0 (0%)	0 (0%)	-
Syllable Repetitions – final	0 (0%)	0 (0%)	-
Word Repetitions	**23 (4%)**	**56 (8.4%)**	*p* < 0.05 (♂)
Truncations	**19 (3.4%)**	**34 (5.1%)**	**n.s.**

5. RESULTS

The results are summarized in Table 1 above.

As for general frequency, we found that the 6-year-olds in this study on average produced 15.6 disfluencies per 100 spoken words, a result that is different from what is reported in Guitar (2013), who presented adult-like figures in child disfluency. We also observed considerable individual variation was also observed, similar to what has previously been reported in the literature for adult speech (e.g. Oviatt, 1995; Bell, Eklund & Gustafson, 2000).

As for differences between the two sexes, we found that girls produced significantly more unfilled pauses, prolongations, sound repetitions in medial position and word repetitions, while boys exhibited more word repetitions. This observation replicates the results presented by Kools & Berryman (1971) for 7-year-olds, and Yairi (1981) for 2-year-olds.

We did not observe any correlation between the amount of disfluencies produces and lexical ability, including vocabulary size and word retrieval, which runs counter to Westby's (1974) observations.

6. DISCUSSION AND CONCLUSIONS

As has already been pointed out, there is an obvious lack of consistency in the results reported in the literature, across most variables. Although it must be borne in minds that given the very fast and complex development of language and speech in young children, comparisons between different and across studies are cumbersome, for obvious reasons. Comparing our results with previously reported studies highlights this phenomenon in that we both replicate results and present results that run counter to previous studies. It is our belief, and a limitation of the present study, that disfluency in young children must be studied in a way that is solidly based in general theories on child language acquisition. Similar to many previous studies, the present study included a small group of children with a given, and our results need to be corroborated or contested and future studies of children of the same or similar ages. We do hope, however, that the present study helps shed a little bit more light on disfluency incidence in young children.

7. REFERENCES

Bell, L., Eklund, R., Gustafson, J. 2000. A Comparison of Disfluency Distribution in a Unimodal and a Multimodal Human–Machine Interface. *Proceedings of ICSLP '00*, Beijing, 16–20 October 2000, 626–629.

Bornstein, M.H., Hahn, C-S., Haynes, O.M. 2004. Specific and general language performance across early childhood: Stability and gender considerations. *First Language* 24(3), 267–304.

Bortfeld, H., Leon, S.D., Bloom, J. E., Schober, M.F., Brennan, S.E. 2001. Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech* 44(2), 123–147.

Brennan, S.E., Schober, M.F. 2001. How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language* 44, 274–296.

Colburn, N. 1985. Clustering of Disfluency in Nonstuttering Children's Early Utterances. *Journal of Fluency Disorders* 10(1), 51–58.

Davis, D.M. 1939. The Relation of Repetitions in the Speech of Young Children to Certain Measures of Language Maturity and Situational Factors: Part I. *Journal of Speech Disorders* 4(4), 303–318.

Davis, D.M. 1940. The Relation of Repetitions in the Speech of Young Children to Certain Measures of Language Maturity and Situational Factors: Part I. *Journal of Speech Disorders* 5(3), 193–288.

DeJoy, D.A., Gregory, H.H. 1973. The relationship of children's disfluency to the syntax, length, and vocabulary of their sentences. Paper presented at the Annual Convention of the American Speech and Hearing Association, Washington, D.C., 1973. Cited (p. 152) in: Gordon, P.A., Luper, H.L., Peterson, H.A. 1986. The effects of syntactic complexity on the occurrence of disfluencies in 5 year old nonstutterers. *Journal of Fluency Disorders* 11, 151–164.

DeJoy, D.A., Gregory, H.H. 1985. The relationship between age and frequency of disfluency in preschool children. *Journal of Fluency Disorders* 10, 107–122.

Dunn, L. M., Dunn, D. M. 1997 (third edition). *Peabody Picture Vocabulary Test.* Circle Pines, Minnesota: American Guidance Service.

Eklund, H. 1996. *Ordracet.* Hargdata AB.

Eklund, R. 2004. Disfluency in Swedish human–human and human–machine travel booking dialogues. PhD thesis, Linköping Studies in Science and Technology, Dissertation No. 882, Department of Computer and Information Science, Linköping University, Sweden.

Fox Tree, J.E. 1995. The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language* 34, 709–738.

Gordon, P.A., Luper, H.L. 1989. Speech disfluencies in nonstutterers: Syntactic complexity and production task effects. *Journal of Fluency Disorders* 14(6), 429–445.

Guitar, B. (ed.). 2013. *Stuttering: An integrated approach to its nature and treatment* (4th Ed.). Lippincott, William & Wilkins.

Haynes, W.O., Hood, S.B. 1977. Language and Disfluency Variables in Normal Speaking Children from Discrete Chronological Age Groups. *Journal of Fluency Disorders* 2, 57–74.

Haynes, W.O., Hood, S.B. 1978. Disfluency changes in children as a function of the systematic modification of linguistic complexity. *Journal of Communication Disorders* 11, 79–93.

Juste, F.S., Sassi, F.C., de Andrade. C.R. 2012. Exchange of disfluency with age from function to content words in Brazilian Portuguese speakers who do and do not stutter. *Clinical Linguistics & Phonetics* 26, 11–12.

Kools, J.A., Berryman, J.D. 1971. Differences in disfluency behavior between male and female nonstuttering children. *Journal of Speech and Hearing Research* 14(1), 125–30.

Levin, H., Silverman, I., Ford, B.L. 1967. Hesitations in Children's Speech During Explanation and Description. *Journal of Verbal Learning and Verbal Behavior* 6, 560–564.

Levin, H, Silverman, I. 1965. Hesitation Phenomena in Children's Speech. *Language and Speech* 8(2), 67–85.

Lindström, E., Werner, C. 2006. *A-ning. Neurolingvistisk Afasiundersökning,* Standardisering. Stockholm: Ersta.

Mahl, G.F. 1956. Disturbances and silences in the patient's speech in psychotherapy. *Journal of Abnormal and Social Psychology* 53, 1–15.

McLaughlin, S.F., Cullinan, W.L. 1989. Disfluencies, utterance length, and linguistic complexity in nonstuttering children. *Journal of Fluency Disorders* 13, 17–36.

Muma, J.R. 1971. Syntax of Preschool Fluent and Disfluent Speech: A Transformational *Analysis.* Journal of Speech, Language, and Hearing Research 14(2), 428–441.

Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language* 9(1), 19.

Pearl, S.Z., Bernthal, J.E. 1980. The Effect of Grammatical Complexity Upon Disfluency Behavior of Nonstuttering Children. *Journal of Fluency Disorders* 5, 55–68.

Sawyer, J., Yairi, E. 2010. Characteristics of Disfluency Clusters Over Time in Preschool Children Who Stutter. *Journal of Speech, Language, and Hearing Research* 53, 1191–1205

Shriberg, E.E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California, Berkeley.

Silverman, E.M. 1974. Word Position and Grammatical Function in Relation to Preschoolers' Speech Disfluency. *Perceptual and Motor Skills* 39(1), 267–272.

Trager, G.L. 1958. Paralanguage: a first approximation. *Studies in Linguistics*, 13(1–2), 1–13.

Westby, C.E. 1974. Language performance of stuttering and nonstuttering children. *Journal of Communications Disorders* 12(2), 133–145.

Wexler, K.B., Mysak, E.D. 1982. Disfluency Characteristics of 2-, 4-, and 6-Yr-Old Males. *Journal of Disfluency Disorders* 7, 37–446.

Yairi, E. 1981. Disfluencies of normally speaking two-year-old children. *Journal of Speech, Language, and Hearing Research* 24, 490–495.

DISFLUENCY AND LAUGHTER ANNOTATION IN A LIGHT-WEIGHT DIALOGUE MARK-UP PROTOCOL

Julian Hough*, Laura de Ruiter**, Simon Betz** and David Schlangen*

Bielefeld University, *Dialogue Systems Group, **Workgroup Phonetics and Phonology
julian.hough@uni-bielefeld.de

ABSTRACT

Despite a great deal of research effort, disfluency and laughter annotation is still an unsolved problem, both in terms of consensus for a general applicable system, and in terms of annotation agreement metrics. In this paper we present a new annotation scheme within a light-weight mark-up for spontaneous speech. We show, despite the low overhead required for understanding the annotation protocol, it allows for good inter-annotator agreement and can be used to map onto existing disfluency categorization, with no loss of information.

Keywords: Disfluency annotation, laughter, German corpora, inter-annotator agreement, spontaneous speech

1. INTRODUCTION

Annotating spontaneous speech material is always constrained by a trade-off between time and effort on the one hand and coverage on the other. Here we develop a system for annotating spontaneous speech that reduces effort while increasing coverage of disfluency and laughter phenomena.

The low effort required is due to (1) a minimalistic and comprehensive vocabulary to be learned by the annotator and (2) the application directly on the transcription text *at transcription time*. This makes the scheme easy to learn and handle for non-experts who can facilitate the markup of spontaneous speech phenomena on the fly while transcribing. Further analysis and correction is left to interested experts who can focus on their own research questions more directly than in existing annotation systems where more expert work for identification and labeling of these phenomena might be necessary from the outset. Analyses of agreement confirm that non-experts indeed have little trouble in comprehension and application of this system.

Despite the existence of numerous annotation systems, some of which even focus on spontaneous speech phenomena (see §2 for an overview), these systems never cover all the phenomena potentially of interest to disfluency and dialogue researchers. Our system encompasses a light-weight way of covering all disfluency and laughter phenomena and the potential to be mapped onto existing annotation schemes.

2. EXISTING DISFLUENCY AND LAUGHTER ANNOTATION SCHEMES

There is a plethora of existing speech annotation schemes, some of which have been developed especially to capture disfluency phenomena. They vary in terms of practical implementation, the use of category labels and whether or not they mark non-verbal or paralinguistic events.

[15]'s thesis and the ensuing Switchboard disfluency annotation manual [9] are perhaps the most well known for disfluency mark-up. Other more general schemes, such as that used in the German Verbmobil corpus for task oriented dialogue annotation, have various conventions, covering a wide range of spontaneous speech phenomena, including, but not explicitly focusing on, disfluencies [2].

Technically, some schemes annotate disfluencies on a separate tier, for instance [4], [2] or [10]), while others use an in-text annotation on the transcription tier, for instance [9], [12], [13], with more recent schemes often doing this by means of an XML-style annotation [12], [1]. Several schemes require annotators to assign disfluency category labels like "repair", "insertion" or "stutter" from a pre-specified set (vocabulary) as they go through the data ([12], [2], [1], [4]), so training costs can be substantial.

3. A LIGHT-WEIGHT DISFLUENCY AND LAUGHTER ANNOTATION TECHNIQUE

In an effort to both improve the ease of annotation for non disfluency experts, and allow subsumption of existing categorical approaches, here we propose a more light-weight approach than those mentioned above which does not reduce the mark-up information. It combines transcription and annotation in one pass, so that turn segmentation and word transcription can be aided by directly observing their inter-

Table 1: Exemplary DF annotations and their equivalent labels in [1] and [8].

Annotation	[1]	[8]
(Ich + ich) will *(I + I) want*	<repetition> <rm> Ich </rm><rs> ich </rs> </repetition>	Covert repair
(nicht verwinkelt so dass +) und breit *(not contorted so that+) and wide*	<restart><rm> nicht verwinkelt so dass </rm><rs/></restart> und breit	Fresh start
die (<p="Küche">Krü-</p> + Küche) *the (kri- + kitchen)*	die <sot> Krü- </sot> Küche	Phonetic repair

action with disfluencies. Transcription and annotation can be conducted based on perception without lengthy training: transcribers do not have to learn categorical labels for the phenomena because such analyses can be carried out afterwards with the use of automatic search, dependent on the given interest of the researcher.

The advantage of avoiding category labels in the annotation process is that the annotated data is more versatile and more 'theory-neutral', allowing researchers to map it onto different existing classification schemes should they so wish, but not constraining them to do so. Labels like "repetition" or "replacement" (used for instance by [1]) can easily be derived from our bracketed annotation, which is similar to [15]. Other categories like those suggested by [8] can also be derived from our mark-up.

Table 1 provides some examples of how disfluencies annotated according to our scheme can be be mapped onto other classification schemes. While the bracketing of our mid-utterance repair disfluencies is based on [15], with reparandum, interregnum and repair phases available, with a strictly right-branching mark-up of chaining disfluencies such as in (4), we emphasize in our annotation manual the distinction between abandonment of turns (whereby annotators will simply transcribe an utterance final '-') as in (1) and (2) below, and turn-initial fresh starts marked with no repair phase such as in (3) below and in Table 1. Note that while not a focus of this paper, we account for phonetic variants of lexical items with <v=" . . ">. .</v> tags.

(1) Oder das <v="ist">is'</v> im -
Or that is in the -

(2) Ja eigentlich <v="wäre es"> wär's </v> cool in der Küche <v="einen">'n</v> kleinen Tisch zu haben wo man -
Yes actually it would be cool to have a small table in the kitchen where you -

We include filled pauses, marked simply by a {F } bracketing and other fillers simply use { }. We also include laughed speech with simple XML-style tags spanning the affected speech <laughter>...</laughter> and a <laughterOffset/> tag for the often audible deep inhalation of breath after laughed speech or a laughter bout marked <laughter/> (see (3)).

(3) (Und mit einem +) mit vielleicht Sachen die nicht <laughter> auseinander brechen </laughter> <laughterOffset/> -
(And with a +) with perhaps things that don't fall apart -

For partial words, we encourage transcribers to guess the complete form of the word where possible, again using a simple tag <p=" . . ">. . -</p>, as below:

(4) (<p="Wohnzimmer">Wohn-</p> + . {ja also} (die + (die + das)) {F äh} ... Wohnzimmer)
<p="living room">liv-</p> yes well the the the uh living room -

4. INTER-ANNOTATOR AGREEMENT

Category	Agreement	κ_{free}
reparandum	0.9477	0.8954
repair	0.9677	0.9353
filled pause	0.9968	0.9937
laughed speech	0.9558	0.9117

Table 2: Inter-annotator agreement scores for disfluent word types using κ_{free}

We test our annotation scheme on a corpus of dyadic interactions between German speakers, the Dream Apartment (DAP) corpus [7], which in contrast to existing corpora used to studies disfluencies

Class	SWBD % of words	DAP % of words
Reparandum words	5.16	5.51
Partial words*	0.75	1.10
Filled pauses*	1.12	1.81
Laughed words*	0.45	6.06

Table 3: The frequency of disfluent and laughed words in Switchboard (SWBD) and the Dream Apartment (DAP) corpora. Starred categories indicate a significantly different frequency between the two corpora.

like [6],[14] or [13], is a relatively domain-general corpus of face-to-face interactions. The DAP consists of 9 dialogues of 15 minutes in length. In the task, participant pairs were instructed to discuss their ideal apartment they could jointly design such that they could describe it to an architect. They are given a substantial budget of 500,000 Euros. The familiarity of the subjects varied with 2 of the pairs being strangers and the others varying in familiarity. All participants were students.

To test agreement we use one transcript and 3 annotators: one was the second author, while the other two were non experts. We compared the inter-annotator agreement of words being part of different disfluency and laughed speech elements using the marginal-free multi-rater metric κ_{free} [11]– we use this metric as other multi-rater agreement measures like Fleiss' κ suffer from an assumption annotators know a priori how many cases they should assign to each category, which is not the case here.

The results shown in Table 2 are both interesting and encouraging. Filled pauses and repair phase words have very good agreement, while the lower reparandum word agreement shows a deviation in the way annotators perceive the extent of repairs, and consequently their discourse effects– see [5] for a similar finding. The lower agreement for laughed speech segmentation is not detrimental, as it is still good enough to provide search terms for subsequent stand-off annotation.

5. USE CASE 1: DISFLUENCY AND LAUGHED SPEECH RATES IN PHONE AND FACE-TO-FACE CONVERSATIONS

One of the benefits of our scheme is that it is directly compatible with established schemes, including the Switchboard disfluency annotation mark up [9]. We can therefore directly compare the rates of the disfluency and laughter phenomena in the DAP with Switchboard.

In Switchboard we use the held-out data for disfluency detection (all files named sw4[5-9]* in the Penn TB III release: 52 transcripts, 6.5K utterances, 49K words) marked up according to the scheme in [9]. The DAP is smaller with fewer, but longer, dialogues (9 transcripts, 4.1K utterances, 20k words).

The proportion of reparandum words in each corpus was not significantly different ($\chi^2_{(1)}$=3.568,p=0.06) however the proportion of filled pauses, laughed words and partial words of the total word tokens was significantly lower in the Switchboard corpus.

The most striking difference is in the proportion of laughed words. We hypothesize this may have been due to the difference in topics between the DAP and Switchboard, and also due the familiarity of the participants. Upon inspection there were many opportunities for laughables based primarily on the incongruity of the situation of being students with a vast amounts of spending money.

6. USE CASE 2: DISFLUENCY AND LAUGHTER INTERACTION

A second use case we investigate for our annotation protocol is a qualitative analysis of the interaction of laughed speech, laughter and disfluency in the DAP corpus. In Figure 1 three extracts from the corpus with our mark-up scheme are shown with their English translations. In Example 1 we see some evidence to explain the high frequency of laughed speech described above being due to the topic of conversation. Here the subject of intimacy and privacy with one's partner in the dream apartment is not being fully addressed but is jointly laughed at. A filled pause is employed after the joint laughter.

In Example 2, disfluency and laughed speech interact again on the same topic as in Example 1, with a chaining replacement repair directly following the laughter. In Example 3, a laughable is taken up by both participants in response to a self-answered question by A who mocks her own lack of intelligibility in her explanation to B. Following this the turn is immediately held by A by use of a filled pause.

While these few examples do not provide a thorough analysis of the interactions, we hope they illustrate that Conversation Analysts and dialogue theorists may also benefit from our simple mark-up.

Example 1: Laughter and laughed speech (joint):

A	okay also (wenn wir + wenn wir) <v="eine">'ne</v> Küche haben ein {F ähm } Wohnzimmer zusammen haben und {F äh } ein [Badezimmer] das ist ja so wie in <v="einer">'ner</v> WG [ ] <breathing/> brauchen wir auf jeden Fall <breathing/> jeder so grosse Zimmer dass jeweils unsere <laughter> Partner die dann ja auch noch zu [Besuch kommen] </laughter> <breathing/> {F ähm } auch noch Platz finden
A-en	*okay so (when we + when we) have a kitchen, a {F uh } livingroom and {F uh } a [bathroom] together which is the way it is in a shared flat [ ], we must definitely have big rooms so when our <laughter> partners come [to visit] </laughter> <breathing/>{F um} they have a place to go.*
B	[mhm] / [ja] / [<laughter>sehr richtig</laughter>]
B-en	*[mhm] /[yes] / [<laughter>very true</laughter>]*

Example 2: Chaining substitution repair after laughed speech:

A	dann hat jeder genug Privatsphäre .. mit seinem <laughter> Partner </laughter> / (und die Küche + (und die + {F ähm } (und die + ... und das Wohnzimmer))) ist quasi so ... mittig
A-en	*then everyone has some privacy ... with their <laughter> partner </laughter> / (and the kitchen + (and the + {F um } (and the + ... and the living room))) is kind of ... central*

Example 3: Laughter at embarrassment of disagreement followed by turn hold filled pause:

A	und vom Wohnzimmer kannst du halt in die Küche gehen / <v="verstehst du">verstehste</v> das ? / [okay] nee / [gut] / {F ä:hm }
A-en	*and from the livingroom can you sort of go to the kitchen / do you understand ? / [okay] no / [good] / {F u:m }*
B	[nee] / [<laughter/>] <laughterOffset/>
B-en	*[no] / [<laughter/>] <laughterOffset/>*

Figure 1: Examples of the interaction between disfluency and laughter in the dream apartment corpus

7. CONCLUSION

We present a light-weight and reliable protocol for disfluency and laughter annotation which is currently being used in the DUEL (DisflUencies, Exclamations and Laughter in dialogue) project [3]. It is both fast and easy to use for non-experts, and subsumes existing schemes. Stand-off timing information for detailed investigation into phenomena of interest is derivable from this automatically using the MINT tools [7] software, among others. We have shown two use cases of the scheme, one which allows direct comparability to other schemes, and one which allows fast mark-up of dialogues at transcription time for quantitative and qualitative analysis.

8. REFERENCES

[1] Besser, J., Alexandersson, J. 2007. A comprehensive disfluency model for multi-party interaction. *SigDial 2007* volume 8 182–189.

[2] Burger, S., Weilhammer, K., Schiel, F., Tillmann, H. G. 2000. Verbmobil data collection and annotation. In: *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer 537–549.

[3] Ginzburg, J., Tian, Y., Amsili, P., Beyssade, C., Hemforth, B., Mathieu, Y., Saillard, C., Hough, J., Kousidis, S., Schlangen, D. 2014. The Disfluency, Exclamation and Laughter in Dialogue (DUEL) Project. *Proceedings of the 18th SemDial Workshop (DialWatt)* Herriot Watt University, Edinburgh. 176–178.

[4] Hedeland, H., Schmidt, T. 2012. Technological and methodological challenges in creating, annotating and sharing a learner corpus of spoken german. *Multilingual Corpora and Multilingual Corpus Analysis* 14, 25.

[5] Hough, J., Purver, M. Dec. 2013. Modelling expectation in the self-repair processing of annotat-, um, listeners. *Proceedings of the 17th SemDial Workshop (DialDam)* Amsterdam. 92–101.

[6] Kohler, K. J. 1996. Labelled data bank of spoken standard german: the kiel corpus of read/spontaneous speech. *ICSLP 96* volume 3. IEEE 1938–1941.

[7] Kousidis, S., Pfeiffer, T., Schlangen, D. 2013. Mint. tools: Tools and adaptors supporting acquisition, annotation and analysis of multimodal corpora. *Interspeech 2013*.

[8] Levelt, W. J. 1983. Monitoring and self-repair in speech. *Cognition* 14(1), 41–104.

[9] Meteer, M. W., Taylor, A. A., MacIntyre, R., Iyer, R. 1995. *Dysfluency annotation stylebook for the switchboard corpus*. University of Pennsylvania.

[10] Moniz, H., Batista, F., Mata, A. I., Trancoso, I. 2014. Speaking style effects in the production of disfluencies. *Speech Communication* 65, 20–35.

[11] Randolph, J. J. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online Submission*.

[12] Rodríguez, L. J., Torres, I., Varona, A. 2001. Annotation and analysis of disfluencies in a spontaneous speech corpus in spanish. *ISCA 2001*.

[13] Schiel, F., Heinrich, C., Barfüsser, S. 2012. Alcohol language corpus: the first public corpus of alcoholized german speech. *Language resources and evaluation* 46(3), 503–521.

[14] Schmidt, T., Hedeland, H., Lehmberg, T., Wörner, K. 2010. Hamatac–the hamburg maptask corpus.

[15] Shriberg, E. E. 1994. *Preliminaries to a theory of speech disfluencies*. PhD thesis.

INTERVENTION FOR CHILDREN WITH WORD-FINDING DIFFICULTY: IMPACT ON FLUENCY DURING SPONTANEOUS SPEECH FOR CHILDREN USING ENGLISH AS THEIR NATIVE OR AS AN ADDITIONAL LANGUAGE

Peter Howell

Psychology and Language Sciences, University College London
p.howell@ucl.ac.uk

ABSTRACT

Types of intervention that could be targeted when there are high rates of word-finding difficulty were examined for any impact they had on speech fluency (whole-word repetition rate in particular). Results are reported that are interpreted as showing that a semantic-based intervention has an impact on fluency as well as word-finding.

Keywords: EAL, word-finding, stuttering, intervention

1. INTRODUCTION

A procedure has been developed that identifies reception-class children who are disfluent [6, 8], allowing schools to refer them for speech language therapy (SLT). About 40% of the children in the schools we work with speak English as an additional language (EAL) and thus often experience word-finding difficulty (WFD) on some content words. They respond to WFD by repeating function words or pausing prior to the content word that they cannot retrieve. This interrupts the forward flow of speech and gives them time to seek an alternative way to communicate their message. Our screening procedure excluded whole-word repetitions (WWRs) as signs of disfluency [11]. Therefore, EAL children who had high rates of WWR because of WFD were not considered disfluent. However, other authors count WWR as disfluencies [10]. The speech-based assessment was appropriate as it was validated against a non-word repetition (NWR) test designed so that it applied to the majority of languages spoken by the EAL children [7]. Children with disfluencies (EAL or English only, EO) performed worse on the NWR test than did fluent controls, but children with WFD (EAL or EO) performed similarly to fluent controls. Whilst the children with disfluencies are considerations for referral to SLTs, schools regard WFD to be an issue that they should address themselves; improving vocabulary is an educational goal. A systematic review identified the types of intervention directed at EAL children, the majority of which involved literacy training (reading) in one form or another [9]. Based on the above observations, it would be expected that these interventions would also have an impact on fluency and reduce WWR rate in particular. This would also serve as a check as to whether children with disfluencies have been missed and classed as having WFD.

The literature on interventions used with patients with WFD, including that on WFD in second language learning, was examined to identify suitable possibilities for intervention. Semantic feature analysis (SFA) is one treatment used with aphasic patients. The semantic attributes of words are used as probes to activate semantic networks that encompass the target word to be retrieved [13]. SFA with aphasic patients leads to more accurate naming of words in isolation [1]) and in discourse [1,2]. Circumlocution-Induced Naming (CIN) is another method. CIN trains patients to talk round target words when they experience retrieval difficulty [3]. CIN improves word retrieval of test and novel words [3, 9]. Circumlocution has been successfully used with adults learning a second language [12]. The effects of circumlocution training depend on participants' language proficiency insofar as advanced learners use this, as well as language-based strategies, to avoid WFD more often than do beginning language learners [12].

Here a group-based procedure was designed for use with 4-5 year old children in schools who had WFD (not EAL children alone). However, more EAL children than EO children have WFD. Details of EAL children's language history were obtained. The standardized test of WFD [4] and fluency assessments [6, 11] were made before and after the intervention, and one week after the intervention finished. Children were assigned to treatment or non-treatment interventions in mixed language groups.

2. METHOD

2.1. Participants

Children's speech had been assessed as part of the ongoing screening study [7]. This procedure involves assessment of %WWR and % fragmentary disfluencies in a sample of speech which is at least 200 syllables long. The children who were selected had rates of WWR in the top 10% but had below

average disfluency counts. The selection was not based on language spoken nor gender, but 64% were EAL and 70% were male. Many languages were spoken (see [8]). Language history information was obtained from the children and verified by teachers. This included information where the child was born, language spoken at home and school etc. (space limitations preclude a full report). A protocol was used to score the questions, and these were used to match the EAL children for language level and to check EO children's status. Children's scores correlated well with parents' scores.

Sixteen children were selected for the study and were split into four groups of four. In each group of children, two were EO and two were EAL, and all were male. The two EAL children in each group spoke a different language (otherwise they may have used code-switching). Native language of the EAL groups was matched for two pairs of groups. Polish and Urdu were the languages in one pair and Akan African and Romanian in the other pair. One of these pairs of groups received the intervention and one received no intervention (see below).

2.2. Baseline performance

German's [4] Test of Word Finding in Discourse (TWFD) was used to measure WFD according to the instructions in the manual. The experimenter first conversed with the participants to put them at ease. A practice picture description task was given to children to familiarize them with the speech procedure. Participants then described a carnival scene. In this report only prevalence of word-finding characteristics (e.g. unnecessary repetition of words, empty words that add no content or specificity, speech delays etc.), are reported as indications of WFD.

A separate spontaneous speech sample was recorded and used to obtain %WWR [7]. Participants chose a topic and talked about it. Occasionally children did not speak long enough on the topic and the experimenter had to give short prompts. Each WWR was counted as one event irrespective of the number of times the word was repeated; this was expressed as the percentage of WWR out of all syllables reported. As a further check on fluency, fragmentary disfluencies were assessed, as described in [11] and it was confirmed that rates for these children were low.

2.3. Intervention and no-intervention conditions

There were two components to the intervention. The first checked that the children knew the name of each picture by pointing at one of the four pictures on a card – one picture was of the target word and the other three were fillers. Stimuli were selected (based on a pilot study) as likely to be known, but difficult to name by children with WFD. The words were low frequency and there were 30 nouns and 30 verbs in total. This component was done once, at the time of the TWFD assessment.

The second component was a picture-naming game that started one week after the first component and continued for three sessions, with one-week intervals between each session. Pairs of children were tested, and the pairings of members in the groups was counterbalanced such that each child had one session with each other member of the group. The procedure was identical in all three sessions, except that each child was tested on different sets of stimuli each time, so by the end of three weeks they had been tested on all stimuli). Thirty-three percent (10 nouns and 10 verbs) of the stimuli were assigned as the stimuli for that participant for the particular session. Different stimuli were given to each member of a pair. One child was selected for test at random. The other child helped the experimenter by showing the test child the picture.

The experimenter asked: 'What is this object?' when a noun was presented, and 'What action is happening?' when a verb was presented. When stimuli were named correctly, the next test item was presented. Two semantic prompts were prepared in case the child failed to name the word. Examples of probes for nouns were 'Where would you see this? and 'Who would use this?'. Examples for verbs would be 'Who would do this?' and 'Where would you see this?'. The test child was given an opportunity to respond after the first prompt. If they still could not name the object, they were given the second prompt. If the child still failed to name the object correctly, the first phone of the word was said by the experimenter and the child was given another opportunity to respond. If this last prompt did not help, they were told what the target word was. Scores for the amount of semantic and phonological help given ranged from 0 (no help), 2 (one semantic cue needed) 4 (two semantic cues needed) and 5 (two semantic cues and first phone). This scoring procedure weights semantic cues more than phonic cues. The roles of test child and helper were then reversed and the second child performed these tests. The order of who was tested first was arranged so all children were tested first and second. At the end of the three training sessions, the fluency and TWFD assessments were repeated. A week later these same tests were run again to establish whether any intervention effects on WFD and WWR had been retained. The non-treatment group received the same procedure, but the probes were given to words that the child had responded to correctly.

3. RESULTS[1]

Raw scores for word-finding behaviours (WF) were calculated as %T-units with Word Finding characteristics specified in [4]. A mixed model ANOVA was conducted with two between group factors (factor one was intervention/no intervention and factor two was EO/EAL) and one within group factor (three levels, before, at the end and one week after the intervention) using WF as the dependent variable. The *means* for all levels of these factors and the associated *sds* (in brackets) are given in Table 1. Assessment (F (2, 24) = 39.02, p <.001, η^2_p = .765) was the only significant main effect and it interacted with intervention group (F (2, 24) = 5.08, p = .014, η^2_p = .297). Inspection of Table 1 shows that WF scores dropped over the intervention and stayed at the low level one week after the intervention. This occurred for both intervention and no-intervention groups, but the interaction of assessment with intervention indicated that the drop was more marked for the intervention group.

Table 1. Word-finding behaviours results

		Mean %T-units			
		Control		Intervention	
		EO	EAL	EO	EAL
Assessment Phase	Pre-treatment	11.20 (5.63)	13.85 (4.32)	11.35 (1.95)	14.18 (3.71)
	Post-treatment	10.23 (2.20)	11.70 (4.38)	7.70 (2.43)	8.05 (3.43)
	1 week follow-up	9.57 (3.33)	8.57(3.66)	5.60 (1.65)	7.70 (2.11)

A second analysis examined the same factors but used %WWR as the dependent variable. The *means* and *sds* are given in Table 2. Again assessment: F (2, 24) = 21.80, p <.001, η^2_p = .645) was the only significant main effect and it interacted with intervention group (F (2, 24) = 6.16, p = .007, η^2_p = .339). Table 2 shows that %WWR decreased for both groups, but the decrease was more marked for the intervention group.

Table 2. %WWR results

		Control		Intervention	
		EO	EAL	EO	EAL
Assessment phase	Pre-treatment	8.18 (2.11)	7.30 (3.55)	8.10 (2.78)	7.20 (3.27)
	Post-treatment	7.60 (1.91)	6.73 (3.63)	4.15 (0.89)	4.93 (1.27)
	1 week follow-up	6.90 (1.89)	5.78 (2.98)	3.93 (0.50)	4.18 (1.57)

4. DISCUSSION

The present study showed that an intervention for WFD that used a mixture of semantic and phonic cues resulted in improvements in word-finding behaviours [4]) and it reduced WWR rates. The improvements relative to baseline occurred immediately post-treatment and were sustained over at least a one-week follow-up period. Similar, but smaller changes occurred for the non-intervention group.

These results suggest that this intervention has promise for addressing WFD and associated fluency problems irrespective of what language a child speaks. Several further features need to be examined. First, only word-finding behaviours have been reported and other measures in [4] need to be examined too. Second, a novel non-word test could be conducted at the end of testing similar to what was done for the training material to see whether the results generalized to non-familiar material. Third, children were matched for WFD across EO and EAL groups. The data in all analyses were inspected for language group differences, but none were seen (as expected). Other ways of selecting children for tests may have revealed differences. For example, instead of matching samples so that EO and EAL children had similar %WWR and the number was constant across language groups, children could be selected at random and differences in sample size could have been allowed. Differences across language groups might then arise for all the dependent variables (WF, %WWR). Fourth, language history was used as a matching factor; it too merits investigation. Further analyses could use children with the same native language to see whether this resulted in higher rates of code-switching (although the current procedure mitigated against this by having different languages spoken by the EAL children). Fifth, a typology of circumlocutions is needed and results with this should be examined for differences across language groups. Finally, the results with this intervention have been explained in terms of the WFD intervention having an effect on fluency. The recent systematic review on interventions used with EAL children [9] shows that the majority of them address reading/literacy. An efficacious intervention for reading that would not be expected to have an impact on fluency should be run to verify whether or not the current intervention is the only one that addresses fluency.

5. REFERENCES

[1] Antonucci, S.. 2009. Use of semantic feature analysis in group aphasia treatment. *Aphasiology*, *23*, 854-866.

[2] Bryant, D., Goodwin, M., Bryant, B., Higgins, K. 2003. Vocabulary instruction for students with learning disabilities: A review of the research. *Learning Disability Quarterly*, *26*, 117-128.

[3] Francis, D.,Clark, N., Humphreys, G. 2002. Circumlocution-induced naming (CIN): A treatment for effecting generalisation in anomia?.*Aphasiology*, *16*(3), 243-259.

[4] German, D. 1991. *Test of Word Finding in Discourse TWFD: Administration, Scoring, Interpretation, and Technical Manual*. DLM Teaching Resources.
[5] Goodglass H., Kaplan E., Barresi B. 2001. Boston Diagnostic Aphasia Examination. 3rd ed. Lippincott Williams & Wilkins; Philadelphia.
[6] Howell, P. 2013. Screening school-aged children for risk of stuttering. *Journal of Fluency Disorders*. *38*, 102-123.
[7] Howell, P., Tang, K., Tuomainen, O., Chan, K., Beltran, K., Mirawdeli, A., Harris, J. Identifying fluency difficulty and word-finding difficulty in children with diverse language backgrounds. Submitted.
[8] Mirawdeli, A. In press. Identifying school-aged children who stutter or who have other difficulties in speech production in school reception classes. *Procedia, Social and Behavioral Sciences.*
[9] Murphy, V. Unthiah, A. 2015. A systematic review of intervention research examining English language and literacy development in children with English as an Additional Language (EAL). Doi: https://educationendowmentfoundation.org.uk/uploads/pdf/EAL_Systematic_review1.pdf
[10] Reilly, S., Onslow, M., Packman, A., Wake, M., Bavin, E.,. …Ukoumunne, O. 2009. Predicting stuttering onset by the age of 3 years: A prospective, community cohort study. *Pediatrics*, *123*, 270 – 277.
[11] Riley, G. 1994. *Stuttering severity instrument for children and adults (SSI-3)* (3rd ed.). Austin, TX: Pro Ed.
[12] Salomone, A., Marsal, F. 1997. How to avoid language breakdown? Circumlocution!. *Foreign Language Annals*, *30*, 473-484.
[13] Wambaugh, J., Ferguson, M. 2007. Application of semantic feature analysis to retrieval of action names in aphasia. *Journal of Rehabilitation Research and Development*, *44*, 381.

[1] The assistance of Kaho Yoshikawa, Yemi Marshall, Sofya Jolliff and Che Kwan Cheng on collecting data for pilot work and Li Ying Chuan in data analysis is gratefully acknowledged. This work was supported by the Dominic Barker Trust.

CAUSAL ANALYSIS OF ACOUSTIC AND LINGUISTIC FACTORS RELATED TO SPEECH PLANNING IN JAPANESE MONOLOGS

Hanae Koiso[1] and Yasuharu Den[2]

[1]The National Institute for Japanese Language and Linguistics, Japan

[2]Faculty of Letters, Chiba University, Japan

koiso@ninjal.ac.jp, den@chiba-u.jp

ABSTRACT

In this paper, we applied a general method of testing path models, investigating causal relationship between cognitive load in speech planning and four types of disfluencies in Japanese monologs. The four disfluencies examined were i) clause-initial fillers, ii) inter-clausal pauses, iii) clause-final lengthening, and iv) boundary pitch movements, which occurred at weak clause boundaries. The length of the constituents following weak clause boundaries was assumed to be a measure of the complexity affecting the cognitive load. By using a model selection technique based on the AIC, we found an optimal model with the smallest AIC, in which the constituent complexity had direct effects on all of the four disfluency variables. In addition, some of the disfluencies influenced one another; clause-final lengthening was enhanced by the presence of a boundary pitch movement and the occurrence of clause-initial fillers was affected by all the other three disfluency variables.

Keywords: path models, fillers, pauses, clause-final lengthening, boundary pitch movements

1. INTRODUCTION

Disfluencies such as fillers and word repetitions occur very frequently at initial positions of major constituents, where they help speakers gain time for speech planning[1, 2]. In Japanese, these linguistic devices for speech planning have also been found to occur at clause-initial positions within an utterance. For example, Watanabe [10] reported that fillers tend to occur more frequently at syntactically deeper, stronger clause boundaries than at shallower, weaker clause boundaries. She also found that the longer the following clause, the more frequently fillers occur at weak clause boundaries, but this tendency was not observed at strong clause boundaries. Watanabe and Den [11] also found that the duration of the filler *e* is positively correlated with the length of the following clause only at weak clause boundaries.

In order to examine acoustic and linguistic features related to speech planning at weak clause boundaries, Koiso and Den [5] focused on clause-initial fillers, clause-final lengthening, and boundary pitch movements (BPMs), showing that all these features were significantly correlated with the length of the following constituents, which was assumed to be a measure of the complexity affecting the cognitive load in speech planning. The study, however, conducted linear mixed-effects models in which the length of the following constituents was treated as a *response* variable, while disfluencies such as fillers as *explanatory* variables. This cause-effect relationship is reversed, if we suppose that disfluencies *are caused by* the complexity.

Two major problems arise if we use the complexity as an explanatory, rather than response, variable; i) there are two or more response variables, and ii) some of those response variables are correlated with each other. To overcome these problems, we apply path analysis, in which dependencies among variables are described by a directed acyclic graph. We show how the length of the following constituents directly affects disfluencies such as clause-initial fillers, inter-clausal pauses, clause-final lengthening, and BPMs at weak clause boundaries and how these disfluencies influence one another.

2. METHOD

2.1. Data

The *Corpus of Spontaneous Japanese* [6] was used in the present study. The CSJ is a large-scale corpus of spontaneous Japanese, consisting mainly of monologs. From among its *Core* data set (CSJ-Core), the simulated public speech subset (20 hours) was selected for the analysis, which consists of 107 casual 10- to 12-minutes narratives on everyday topics given by laypeople.

The CSJ-Core provides a variety of hand-corrected annotations, including clause units, *bunsetsu* phrases, words, phonetic segments, depen-

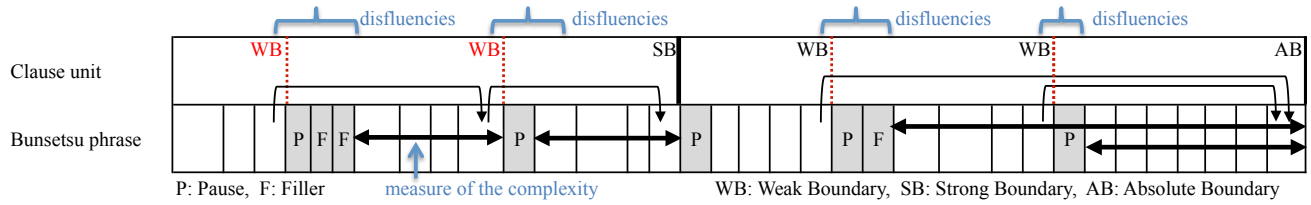


Figure 1: Measure of the complexity. Arcs indicate dependencies between clauses ending with WBs and the phrases they modify. The duration between the end of the current WB clause and the end of its depending phrase was measured as the complexity of constituents being planned. Pauses and fillers appearing at the beginning of this interval were excluded from the calculation.

dency structures, and prosodic information.

Clause boundaries are automatically detected and classified into one of the following categories based on the degree of completeness as a syntactic/semantic unit and on the degree of dependency on the subsequent clause. i) Absolute Boundary (AB): the sentence boundary in the usual sense; ii) Strong Boundary (SB): the boundary of a clause that is relatively independent of the subsequent clause; and iii) Weak Boundary (WB): the boundary of a clause that is relatively dependent on the subsequent clause.

Bunsetsu phrases are basic syntactic phrases in Japanese, which consist of one content word possibly followed by one or more function words. Dependency structures between *bunsetsu* phrases were labeled within clause units.

Prosodic labels were also provided based on the X–JToBI scheme [7]. An accentual phrase is always terminated with a low tone, which can be followed by an additional tone such as a simple rising tone and a rising-falling tone. These additional tones are called *boundary pitch movements* (BPMs).

2.2. Measure of the complexity

At clause boundaries, speakers need to plan the content and the structure of an upcoming syntactic unit. We assume that this unit corresponds to a region containing constituents between the boundary of the current clause and the phrase modified by that clause. Following this assumption, the duration between the end of the current WB clause and the end of its depending phrase was used as a measure of the complexity of constituents being planned, which affects the cognitive load in speech planning (Figure 1). Pauses and fillers appearing at the beginning of this interval were excluded from the calculation.

2.3. Statistical model

Since the corpus data is clustered by speakers and lexical items, the standard models for path analysis cannot be applied. In the field of ecology, Shipley [8] proposed a general method of testing path models, which is applicable to generalized mixed-effects models. In this method, the adequacy of a path model is tested in the following way.

1. Find a set of variable pairs such that the two variables in the pair are hypothesized to be independent of each other.
2. For each pair in the set, formulate a model with one variable in the pair, y_i, as a response variable and the other variable in the pair, x_i, together with other variables that are direct causes of y_i or x_i, as explanatory variables. Random effects may also be included in the model. Obtain the p value associated with the probability that the slope coefficient for x_i is zero.
3. Combine the obtained p values by the following formula:

   (1) $C = -2\sum_{i=1}^{n} \log p_i,$

   where n is the number of pairs tested and p_i is the p value for the ith pair. Compare the C value to a χ^2 distribution with $2n$ degrees of freedom to test the adequacy of the path model. The model is rejected if the C value is significant.

Shipley [9] also proposed a model selection technique for path models based on the AIC, where the AIC statistic is defined as follows:

(2) $AIC = -2\sum_{i=1}^{n} \log p_i + 2K = C + 2K,$

where C is the C statistic of the model, given the data, and K is the number of maximum-likelihood parameters that are estimated using the data.

2.4. Variables for statistical model

The duration of the constituents between the current clause and its depending phrase (`ConstDur`) was used as a variable that may cause four types of disfluencies at the current-clause's boundary, which

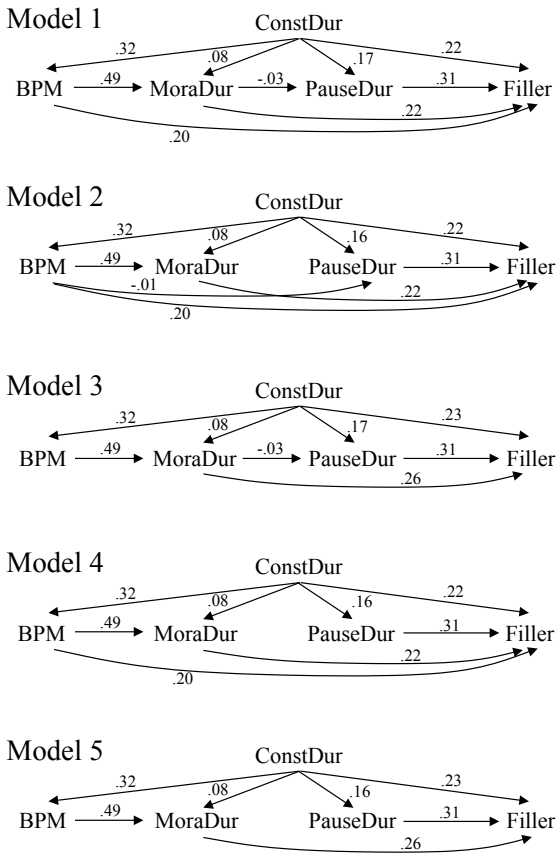


Figure 2: Path models that were not rejected at a significance level of .05. Arrows in each model represent hypothesized cause-effect relations. The numbers beside arrows indicate the estimated coefficients.

were measured by the following variables:

1. presence or absence of a filler at the beginning of the next clause (`Filler`)
2. duration of a pause between the current and the next clauses (`PauseDur`)
3. duration of the last mora of the current clause (`MoraDur`)
4. presence or absence of a BPM at the end of the current clause (`BPM`)

The duration variables were log-transformed and standardized before statistical analysis.

Models that represent all possible cause-effect relationships among these variables were hypothesized. (Some are shown in Figure 2.) In addition to the influence of the complexity to the four types of disfluencies, the influences of temporally preceding disfluencies on succeeding ones were also considered. Random intercepts and slopes for speakers and lexical items were used both in models for obtaining the C statistic and in models for obtaining the path coefficients. (Correlations among random parameters were not used.) Estimation of model parameters was conducted by using the `lme4` package of the R language, and the AIC statistic of each model was calculated by Equation (2).

Table 1: Statistics of models that were not rejected at a significance level of .05.

Model	C value	df	p value	AIC
Model 1	.31	2	.856	78.31
Model 2	3.25	2	.197	81.25
Model 3	8.13	4	.087	80.14
Model 4	3.54	4	.472	75.54
Model 5	11.36	6	.078	77.36

3. RESULTS

We excluded from the analysis clauses that did not modify any phrase or that did not coincide with accentual phrase boundaries. We also limited the data to clauses immediately followed by silent pauses, since clauses tightly connected to the following constituents without intervening pauses are less likely to serve as locations for speech planning. Of 11,244 weak clause boundaries in the data, 4,937 instances were retained for the analysis.

In each of the hypothesized path models, a set of variable pairs assumed to be independent of each other were identified. For example, in Model 4 in Figure 2, independent pairs, i.e., lacking an arrow between the paired variables, are (`BPM`,`PauseDur`) and (`MoraDur`,`PauseDur`). In the regression model of the second variable in each pair, the null probability was calculated for the slope coefficient for the first variable in the pair, considering other affecting variables (e.g., {`ConstDur`} for the pair (`BPM`,`PauseDur`) in Model 4) and random effects. The obtained null probabilities were combined based on Equation (1), getting the C statistic and the probability of observing this value by chance.

Table 1 shows the C statistics and the p values for five models that were not rejected at a significance level of .05, and Table 2 shows the claims of independence implied by these five models. Figure 2 represents the cause-effect relationships hypothesized in these models; the numbers associated with arrows indicate the estimated path coefficients, which were all highly significant. The AIC statistics of these models are also shown in Table 1.

4. DISCUSSION

According to the AIC statistics in Table 1, Model 4, with the smallest AIC, was optimal. In Model 4, the length of the following constituents, i.e., the

Table 2: Claims of independence implied by models in Table 1 with their null probabilities

Model	Claims of independence (Independent pair | Other affecting variables)	Variable whose regression slope is expected to be zero	Null probability
Model 1	(PauseDur, BPM)|{ConstDur, Mora}	BPM	.86
Model 2	(PauseDur, MoraDur)|{ConstDur, BPM}	MoraDur	.20
Model 3	(PauseDur, BPM)|{ConstDur, Mora}	BPM	.86
	(Filler, BPM)|{ConstDur, Mora, PauseDur}	BPM	.02
Model 4	(PauseDur, BPM)|{ConstDur}	BPM	.87
	(PauseDur, MoraDur)|{ConstDur, BPM}	MoraDur	.20
Model 5	(PauseDur, BPM)|{ConstDur}	BPM	.87
	(Filler, BPM)|{ConstDur, Mora, PauseDur}	BPM	.02
	(PauseDur, MoraDur)|{ConstDur, BPM}	MoraDur	.20

complexity, had direct effects on all of the four disfluency variables. When the complexity got higher, fillers and BPMs occurred more frequently and the durations of clause-final morae and inter-clausal pauses became longer. Fillers and segment lengthening have been reported to be affected by the complexity [10, 3]. Our results are consistent with these previous studies. Note that the occurrence of BPMs was directly affected by the complexity; when we removed the dependence between ConstDur and BPM from Model 4 in Figure 2, the resulting model was rejected ($\chi^2(6) = 40.89$, $p < .001$).

Some of the four disfluency variables also influenced one another. Clause-final lengthening was enhanced by the presence of a BPM and the occurrence of clause-initial fillers was affected by all the other three disfluency variables. Unexpectedly, BPMs had direct influence on the occurrence of fillers. In Model 5, which has excluded the dependence between Filler and BPM from Model 4, the slope coefficient for BPM in the regression model of Filler was significant ($p = .02$) in Table 2, suggesting that BPMs might have direct influence on the occurrence of fillers.

An alternative explanation for the direct cause-effect relationships between the complexity and BPMs and between BPMs and fillers might be that some unseen variable serves as a common cause of these three variables, i.e., BPM, ConstDur, and Filler. In this situation, we may observe spurious relationship between two variables. One candidate of such an unseen variable is the syntactic boundary depth, which is reported to be related with the rate of BPMs [4] as well as with the rate of fillers [10]. We will explore this possibility in future research.

5. ACKNOWLEDGMENTS

This work is supported by Grant-in-Aid for Scientific Research 26284055 led by Yasuharu Den and Grant-in-Aid for Scientific Research 25370505 led by Hanae Koiso.

6. REFERENCES

[1] Clark, H. H. 2002. Speaking in time. *Speech Communication* 36, 5–13.

[2] Clark, H. H., Wasow, T. 1998. Repeating words in spontaneous speech. *Cognitive Psychology* 37, 201–242.

[3] Den, Y. in press. Some phonological, syntactic, and cognitive factors behind phrase-final lengthening in spontaneous Japanese: A corpus-based study. *Laboratory Phonology* 6(3).

[4] Koiso, H. 2014. The relationship between boundary pitch movements and syntactic structures (in Japanese). *Journal of the Phonetic Society of Japan* 18, 57–69.

[5] Koiso, H., Den, Y. 2013. Acoustic and linguistic features related to speech planning appearing at weak clause boundaries in Japanese monologs. *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech* Stockholm. 37–40.

[6] Maekawa, K. 2003. Corpus of Spontaneous Japanese: Its design and evaluation. *Proceedings of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition* Tokyo. 7–12.

[7] Maekawa, K., Kikuchi, H., Igarashi, Y., Venditti, J. J. 2002. X–JToBI: An extended J_ToBI for spontaneous speech. *Proceedings of the 7th International Conference on Spoken Language Processing (INTERSPEECH 2002)* Denver, CO. 1545–1548.

[8] Shipley, B. 2009. Confirmatory path analysis in a generalized multilevel context. *Ecology* 90, 363–368.

[9] Shipley, B. 2013. The AIC model selection method applied to path analytic models compared using a d-separation test. *Ecology* 94, 560–564.

[10] Watanabe, M. 2009. *Features and roles of filled pauses in speech communication: A corpus-based study of spontaneous speech*. Tokyo: Hituzi Syobo.

[11] Watanabe, M., Den, Y. 2010. Utterance-initial elements in Japanese: A comparison among fillers, conjunctions, and topic phrases. *Proceedings of the DiSS-LPSS Joint Workshop 2010* Tokyo. 31–34.

VOICE-QUALITY ANALYSIS OF JAPANESE FILLED PAUSES: A PRELIMINARY REPORT

Kikuo Maekawa[†] and Hiroki Mori [††]

[†] Dept. Corpus Studies, National Institute for Japanese Language and Linguistics
[††]Graduate School of Engineering, Utsunomiya University

ABSTRACT

Using the Core of the *Corpus of Spontaneous Japanese*, acoustic analysis of F1, spectral tilt (TL), H1-H2, jitter and F0 was conducted to examine the voice-quality difference between the vowels in filled pauses and those in ordinary lexical items. It turned out by simple SVM analysis that the two classes of vowels could be discriminated with the mean accuracy of higher than 70%.

1. INTRODUCTION

Recently, analysis on the cognitive functions of filled pauses (FP hereafter) has made remarkable progress [1-3]. Speech production mechanisms of FP, on the other hand, remained largely unclarified even now. In this paper, results of preliminary voice-quality analyses of Japanese FP will be presented.

In Japanese, voice-quality of FP seems to be controlled by the speakers to transmit various "paralinguistic" information [4,5]. For example, FP /eH/ (prolonged /e/, the most frequent FP in Japanese) is often produced with noticeable creaky phonation, and the listeners tend to perceive the FP as transmitting strong 'hesitation' of the speaker. Similarly, /eH/ with breathy phonation tends to be perceived as implying the speaker's attitudes like 'politeness', 'weariness', or 'lack of self-confidence'.

Needless to say, similar control of voice-quality can be observed in ordinary lexical items (LX hereafter) like nouns and verbs [6]. But it seems that the control of voice-quality is more salient (and frequent) in FP than in LX. It is presumably because the number of possible word-form is considerably limited in FP than in LX.

The aim of the present study is to show that there is systematic acoustic difference related to voice-quality between the vowels in FP and LX in Standard (Tokyo) Japanese. In this study, the word "voice-quality" is used to refer to both laryngeal and supra-laryngeal features of speech. The phonation characteristics mentioned above belong to laryngeal voice-quality, while the variation in vowel articulation (measured in terms of vowel formant frequencies) belongs to supra-laryngeal, or segmental, voice-quality. Both of them are analysed in this paper.

2. DATA

Monologue talks in the CSJ-Core, i.e., the X-JToBI annotated part of the *Corpus of Spontaneous Japanese* was analysed [7]. 79 male and 58 female speakers were involved in this data. Among more than 30,000 FP recorded in the CSJ-Core, vocalic FP (namely, those FP consisting exclusively of monophthong vowels, /iH/, /eH/, /aH/, /oH/, and /uH/.) were analysed and compared to the corresponding LX long vowels. Only the LX vowels located in the word-initial positions like /eHgo/ "English", and /koHgi/ "lecture") were analyzed. Vowels that were estimated to have less than eight pitch cycles were omitted from the analysis, because jitter analysis (see below) required the duration of at least 5 pitch cycles.

Table 1 shows the number of samples analysed in the current study. Omitting the cases where numbers of samples were too small, male's /aH/, /eH/, and /oH/, and female's /aH/ and /eH/ were analysed.

Table 1: Numbers of analysed samples *

Speaker sex	Vowel	Filled Pause (FP)	Lexical Item (LX)
Male	/aH/	108	113
	/eH/	2411	764
	/iH/	16	281
	/oH/	66	2177
	/uH/	16	561
Female	/aH/	40	61
	/eH/	1049	529
	/iH/	2	248
	/oH/	10	1910
	/uH/	12	500

*Samples in the shaded cells are omitted.

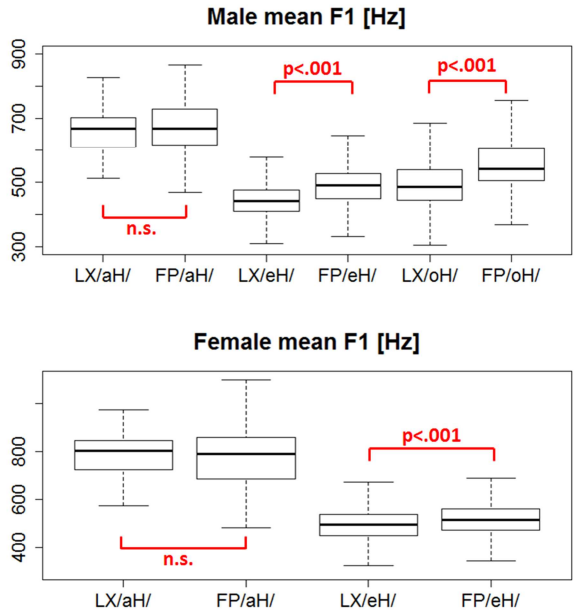
3. ANALYSIS

3.1. Formant frequency

Formant analysis was conducted using Praat [8]. Mean F1, F2, and F3 of vowels were computed by LPC method (number of poles was set to 12). An interesting result was obtained with respect to the first formant frequency (F1). F1 was significantly higher in FP than in LX in the male samples of /eH/

and /oH/ (See Fig.1). Female vowels showed significant difference in /eH/, but not in /aH/.

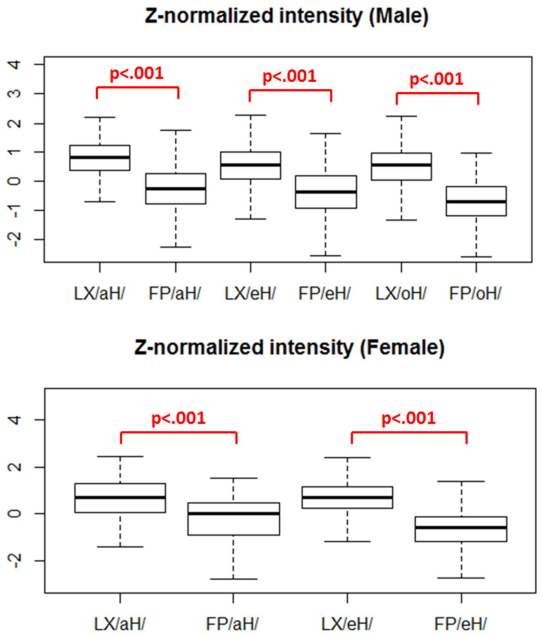
Figure 1: Mean F1[Hz]] in LX and FP.



3.2. Intensity

Intensity information of vowels was also analysed by Praat. Mean vowel intensities were z-normalized based upon the distribution of all vowel samples encompassing both FP and LX for each subject. Mean intensity was significantly smaller in FP than in LX in all vowels across male and female samples (Fig.2).

Figure 2: Mean z-transformed intensity in LX and FP.



3.3. Spectral tilt (TL)

Spectral tilt (TL) is a measure of phonation types. In order to estimate the TL of vowels in an efficient way, cepstrum-based method was used for analysis. Overall trend of a spectrum was approximated by the first cepstrum component, and the difference between the estimated amplitudes at 0 and 3000 Hz was used as the estimated TL [9]. Fig. 3 shows an example of the analysis where blue curve is the FFT spectrum, green curve represents the first cepstrum coefficient, and the red line stands for the TL. Note the TL estimated by this method is not the tilt of laryngeal source. Rather, it is the tilt of speech sound radiated from the mouth of speaker.

In male samples, TL was significantly larger in FP than in LX in all vowels (Fig. 4), while in female samples significant difference was not observed in any vowel.

Figure 3: Ceptstrum-based estimation of TL.

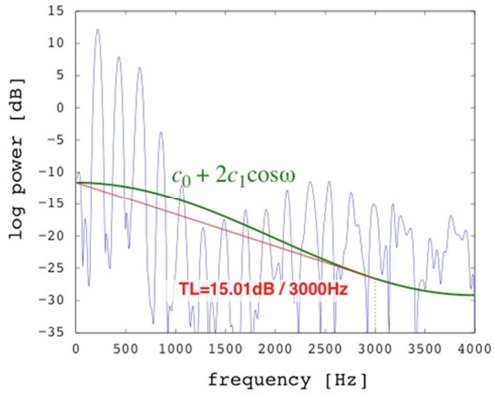
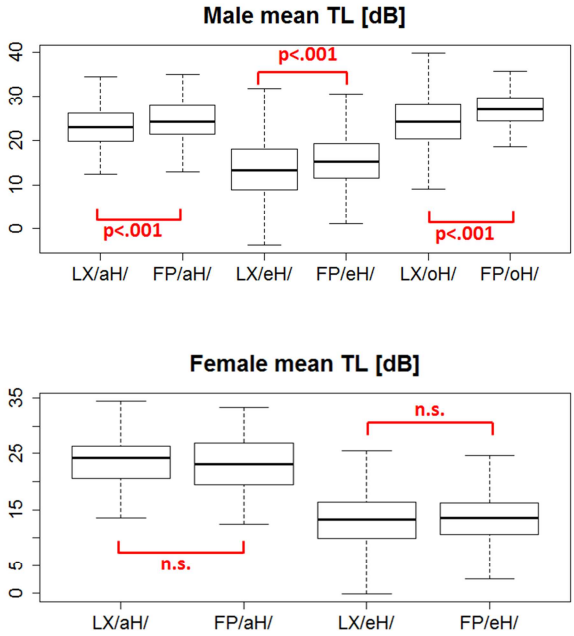


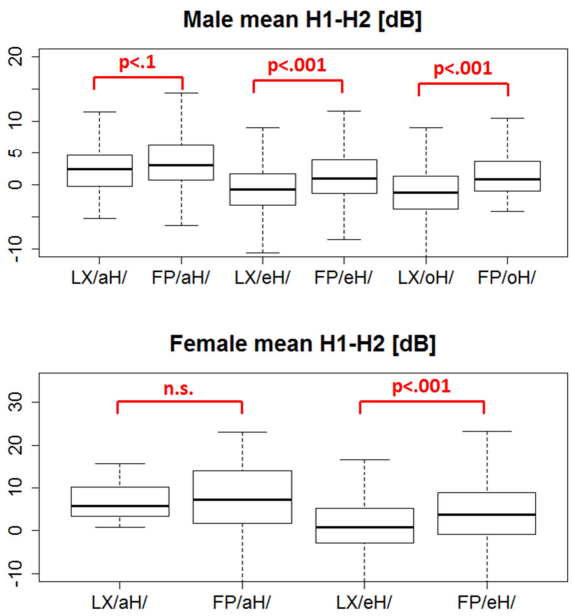
Figure 4: Mean TL [dB] in LX and FP



3.4. H1-H2

H1-H2 (the difference between the levels of the first and second harmonics) is another measure of phonation types, which is popular among phoneticians [10]. An open source script of Praat developed by Chad Vicenik was used for the computation [11]. Fig. 5 shows the mean H1-H2 values for male and female speech. In male speech, mean H1-H2 was significantly higher in FP than in LX (significance is marginal in the case of /aH/). In female speech, mean H1-H2 was significantly higher in /eH/, but not in /aH/.

Figure 5: Mean H1-H2 [dB] in LX and FP.



3.5. Jitter

Jitter is an index of the fluctuation of speech fundamental frequency (F0). In speech pathology, jitter is used frequently for the evaluation of pathological voices, and measured on the basis of the analysis of long sustained vowels. In this study, however, jitter analysis was applied for vowels in running spontaneous speech. Among various definitions of jitter, PPQ5 was computed using the voice report function of Praat. Mean jitter in logarithmic scale was significantly higher in FP than in LX in all vowels of male and female samples (Fig. 6). See the discussion below for the problem of this analysis.

3.6. F0

F0 contours of FPs is believed to be simpler compared to that of LX [12]. Moreover, the F0 values of FPs are supposed to be predictable, to some extent, from the surrounding phonological (tonal) context [13,14]. Here, mean F0 of FP vowels is compared to that of LX. As shown in Fig. 7, mean F0 was significantly lower in FP than in LX in all vowels across male and female samples.

Figure 6: Mean jitter in log scale in LX and FP.

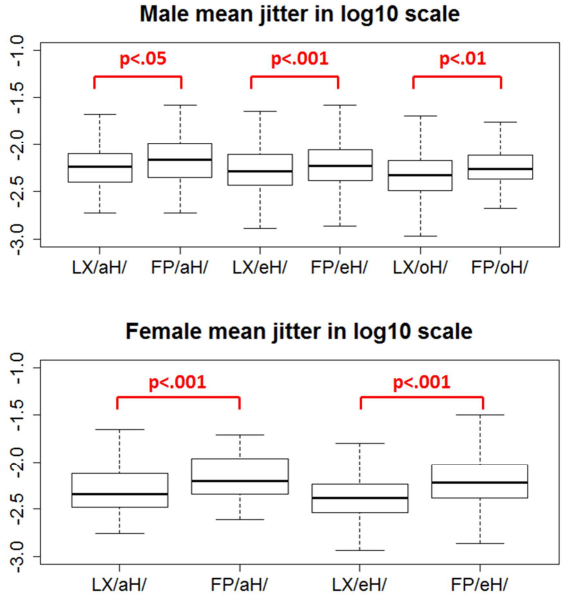
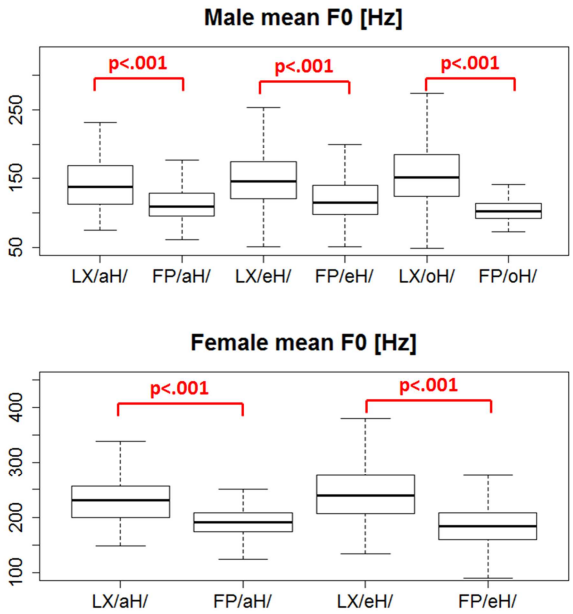


Figure 7: Mean F0 [Hz] in LX and FP.



3.7. Automatic classification

Acoustic analyses presented above suggested strongly the presence of voice-quality difference between the FP and LX. Automatic classification by means of SVM was carried out using the e1071 package of R language (version 3.1.3) in order to know how systematic the differences were.

Table 2 shows the results of ten-fold cross-validation. Note that SVM parameters were fixed to the same values (cost=3.0, gamma=0.5, and kernel is

radial) for all vowels. Note also that the samples of FP and LX vowels were resampled from the data set of Table 1 so that each class has the same number of samples. Therefore, the baseline accuracy was 50% in all cases. The number of samples thus resampled is shown in the third column ('N') of the table.

Mean accuracy of classification (%) was computed under four combinations of explanatory variables: 'All' stands for the case where all six acoustic features were used as explanatory variables. '-P' and '-J' stand respectively for the cases where pitch and jitter were removed, and '-PJ' mean the removal of both of them. The success rates are considerably higher than the baseline in all cases.

Table 2: Accuracy (%) of classification by SVM.

Sex	Vowel	N	All	-P	-J	-PJ
M	/aH/	200	72.0	65.0	70.5	60.5
	/eH/	400	76.0	70.3	76.3	69.5
	/oH/	100	71.0	61.0	74.0	70.0
F	/aH/	60	73.3	63.3	78.3	73.3
	/eH/	400	80.0	64.3	78.5	64.5

4. DISCUSSIONS

Systematic difference was observed between the voice-quality of vowels in FP and LX in spontaneous speech. As a general tendency, vowels in FP had relatively "softer" (as indicated by larger TL and H1-H2 values) and unstable (as indicated by larger jitter values) phonation. They were also marked by lower intensity, lower F0, and higher F1. Moreover, SVM analysis using these features as explanatory variable revealed that it was possible to make distinction between the FP and LX vowels with the mean accuracy of higher than 70%.

There are, however, two important cautions about the current analysis. First, computation of jitter values may be influenced seriously by F0 tracking error, which is inevitable in the current data that contains many samples with heavy creaky / breathy phonation. In this respect, it is noteworthy that in Table 2 the highest mean accuracy was often achieved when the jitter data was removed from the explanatory variables (i.e., the '-J' condition).

Second, the difference of F1 is not easy to interpret. There are at least two interpretations that seem to be equally plausible. For one, it can be the consequence of incomplete glottal closure of 'soft' phonation that makes vocal tract open at both ends [15]. For another, it can be the consequence of speakers' deliberate control over vowel articulation. Prior studies reported the case where speakers control their segmental articulation (in addition to the control of prosody) to transmit certain paralinguistic information [5,16].

As suggested in the introduction, it is the present authors' belief that the observed acoustic difference is related, in some way, to the transmission of paralinguistic information. But this remains as a mere conjecture at the current stage of inquiry. In the current analysis, samples having different paralinguistic information were analysed altogether. Because of this, it is highly probable that the observed acoustic characteristics of FP were biased strongly by the properties of the samples whose paralinguistic information have high occurrence frequencies. For example, higher values of TL and H1-H2 observed in FP samples could be the result of frequent use of FP to transmit 'politeness' and/or 'lack of self-confidence'.

Manual annotation of perceived phonation types and intended paralinguistic message is currently underway in view of automatic annotation. The classification data will help us to understand the production mechanism of FP more precisely.

5. REFERENCES

[1] Sadanobu, T. & Y. Takubo. "Danwa ni okeru monitā kinō." *Gengo Kenkyū*, 108, 74-93, 1995.

[2] Clark, H. & J. Fox Tree. "Using *uh* and *um* in spontaneous speaking." *Cognition*, 84, 73-111, 2002.

[3] Watanabe, M. *Features and roles of filled pauses in speech communication*. Tokyo: Hituzi, 2009.

[4] Ladd, D. R. *Simultaneous structure in phonology*. Oxford: Oxford Univ. Press, 2014.

[5] Mori, H, K. Maekawa, & H. Kasuya "*Onsei wa nani o tsutaete iru ka*". Tokyo: Korona-sha, 2014.

[6] Ishi, C., H. Ishiguro, & N. Hagita. "Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality." *Speech Communication*, 50, 531-543, 2008.

[7] Maekawa, K. "Corpus of Spontaneous Japanese: Its Design and Evaluation". *Proc. SSPR2003*, 7-12, 2003.

[8] http://www.fon.hum.uva.nl/praat/

[9] Maekawa, K. & H. Mori. "Filā no sēshitsujō no tokuchō ni kansuru yobiteki bunseki." *Proc. Spring Meeting of Acoust. Soc. Japan*, 3-2-9, 2015.

[10] Gordon, M. & P. Ladefoged. "Phonation types: a cross-linguistic overview." *J. Phonetics*, 29, 383-406, 2001.

[11]http://www.linguistics.ucla.edu/faciliti/facilities/acoustic/PraatVoiceSauceImitator.txt

[12] Shriberg, E. *Preliminaries to a theory of disfluencies*. Ph.D. Diss. Univ. California at Barkeley, 1994.

[13] Shriberg, E. & R. Lickley. "Intonation of clause-internal filled pauses". *Phonetica*, 50, 172-179, 1993.

[14] Maekawa, K. "Prediction of F0 height of filled pauses in spontaneous Japanese". *Proceedings of DiSS 2013*, Stockholm, 41-44, 2013.

[15] Honda, K. et al. "Visualisation of hypopharyngeal cavities and vocal-tract acoustic modelling." *Computer Methods in Biomechanics and Biomedical Engineering*. 13(4), 443–453, 2010.

[16] Maekawa. K. "Production and perception of 'paralinguistic' information." *Proc. Speech Prosody 2004*, Nara, 367-374 , 2004.

DISFLUENCY DETECTION ACROSS DOMAINS

Helena Moniz[1,2], Jaime Ferreira[1], Fernando Batista[1,3], Isabel Trancoso[1,4]

[1]Spoken Language Systems Lab - INESC-ID, Lisbon, Portugal

[2]FLUL/CLUL, Universidade de Lisboa, Portugal

[3]ISCTE - Instituto Universitário de Lisboa, Portugal

[4]Instituto Superior Técnico, Universidade de Lisboa, Portugal

helenam;jaime.ferreira;fmmb;isabel.trancoso@l2f.inesc-id.pt

ABSTRACT

This paper focuses on disfluency detection across distinct domains using a large set of openSMILE features, derived from the Interspeech 2013 Paralinguistic challenge. Amongst different machine learning methods being applied, SVMs achieved the best performance. Feature selection experiments revealed that the dimensionality of the larger set of features can be further reduced at the cost of a small degradation. Different models trained with one corpus were tested on the other corpus, revealing that models can be quite robust across corpora for this task, despite their distinct nature. We have conducted additional experiments aiming at disfluency prediction in the context of IVR systems, and results reveal that there is no substantial degradation on the performance, encouraging the use of the models in IVR domains.

Keywords: Disfluency detection, acoustic-prosodic features, cross-domain analysis, European Portuguese.

1. INTRODUCTION

Disfluencies are on-line editing strategies with several (para)linguistic functions. They account for a representative portion of our spoken interactions. Everyday we are analists of our own speech and of others, monitoring distinct linguistic and paralinguistic factors in our communications, using disfluencies to make speech a more error-free system, a more edited message, and a more structured system with coherent and cohesive mechanisms.

Disfluencies are an important research topic in several areas of knowledge, namely, Psycholinguistics, Linguistics, Automatic Speech Recognition, and more recently in Text-to-Speech conversion and even in Speech-to-Speech translation. Yet, whereas for several languages one can find much literature on disfluencies, for others, such as European Portuguese, the literature is quite scarce.

Detecting and filtering disfluencies is one of the hardest problems in rich transcription of spontaneous speech. Enriching speech transcripts with structural metadata [15] is of crucial importance for many speech and language processing tasks, and comprises several metadata extraction/annotation tasks besides dealing with disfluencies such as: speaker diarization (i.e. assigning the different parts of the speech to the corresponding speakers); sentence segmentation (also known as sentence boundary detection); punctuation and capitalization recovery; topic and story segmentation. Such metadata extraction/annotation technologies are recently receiving increasing attention [7, 6, 15], and demand multi-layered linguistic information to perform such tasks. A simple segmentation method, for instance, may rely only on information about pauses. More complex methods, however, may involve, e.g., lexical cues, dialog acts cues. In fact, the term *structural segmentation* encompasses all algorithms based on linguistic information that delimit *spoken sentences* (units that may not be isomorphic to written sentences), topics and stories.

Acoustic-prosodic features have been pervasive in disfluency prediction tasks, mostly due to the fact that ASR systems are still not mature enough to account for spontaneous speech in all its complexity. Therefore, the lexical output of a recognizer may not be reliable to train models to account for highly spontaneous data, as the datasets used in this work. Generally, the acoustic-prosodic set of features described for disfluency detection encompass: slower speech rate, lengthier adjacent silent pauses, higher values of spectral tilt, jitter, shimmer, pitch and energy differences relatively to the adjacent contexts (*e.g.*, [13, 18, 19]).

Previous experiments conducted for European Portuguese [12, 1, 9, 8, 11] reported that different structural regions of a disfluency can be predicted based exclusively on discriminative acoustic-prosodic features of disfluencies. Building on that, we now aim at a cross-domain disfluency prediction. Therefore, the aim of this work is threefold: to characterize the acoustic-prosodic features of disfluencies; to evaluate the impact of acoustic features within and across domains. In order to understand the realms of potential transversality in disfluency prediction tasks in dialogues in European Portuguese, domain-specific models will be created for each one of the different scenarios, and then those models will be reused across domains. The literature on cross-domain analysis of disfluencies in dialogues is, in general, quite scarce. It is, therefore, not clear, how one may transpose findings from one domain into another in human-human dialogues. Ultimately, this work may be seen as a contribution to the scarce studies on disfluencies in human-human dialogues, on the most salient features, and on the common vs. distinct properties of such phenomena.

2. DATASETS

This work focuses on two domains to predict disfluencies, namely: university lectures and dialogues. The choice of the corpora was influenced by the availability of large amounts of (highly spontaneous) transcribed data in European Portuguese for these two domains. Both corpora are available through ELRA.

LECTRA (ELRA-S0366) is a university lectures corpus, collected within the national project LECTRA (LECture TRAnscriptions in European Portuguese) [20], aiming at producing multimedia contents for e-learning applications, and also at enabling hearing-impaired students to have access to recorded lectures. It includes seven 1-semester courses, six of them recorded in the presence of students, and only one recorded in a quiet environment. Most classes are 60-90 minutes long. The initial set of 21 hours orthographically transcribed was recently extended to 32 hours [16]. The corpus was divided into train+development (89%) and test (11%). The sets include portions of each one of the courses and follow a temporal criterion, meaning the first classes of each course were included in the training and development sets, whereas the final ones were integrated into the test sets.

CORAL (ELRA-S0367) is a corpus of map-task dialogues [21]. One of the participants (giver) has a map with some landmarks and a route drawn between them; the other (follower) has also landmarks, but no route and consequently must reconstruct it. In order to elicit conversation, there are small differences between the two maps: one of the landmarks is duplicated in one map and single in the other; some landmarks are only present in one of the maps; and some are synonyms. The 32 speakers were divided into 8 quartets and in each quartet organized to take part in 8 dialogues, totaling 64 dialogues, which corresponds to 9h (46k words). The manual annotation of the last two quartets was finished very recently and could not be used in the scope of this work.

The corpora were multilayered annotated, as described in more detail in [11]. The disfluency annotation followed mostly [18, 2]. For the purposes of the present work, sentence-like units were transformed into "chunks" and sequences of distinct types of disfluencies were merged into "disfluent sequences". A binary classification will allow, at a first instance, to discriminate the most important classes. One of the goals is to provide a first evaluation of cross-domain characteristics in order to apply to an IVR system. The datasets comprise manual transcripts and force aligned transcripts, produced by the in-house ASR Audimus [14].

Overall statistics of the corpora are presented in Table 1. Chunks correspond to distinct types of sentence-like units (full stop, question mark, comma) and disfluent sequences encompass all the different types of disfluencies (e.g., filled pauses, complex sequences of disfluencies, repetitions, inter alia). The percentage of disfluent sequences in the university lectures corpus is of 28.5% in the train set and of 29.3% in the test set. The percentage of disfluent sequences in the dialogue corpus is relatively smaller, corresponding to 21%. These results, at a first inspection, point out to dialogues having fewer disfluencies than lectures. However, when considering more refined measures, e.g., the number of sentence-like units (either fluent or with disfluencies) uttered per minute, there are more sentences of both types per minute in dialogues than in lectures, clearly supported on the fact that dialogues have fewer words in both SUs, motivated by temporal constraints in the interaction between interlocutors. Further details on the characetrization of disfluencies in both corpora can be found in [11].

Table 1: Overall characteristics of the datasets.

	Lectra		Coral
	Train	Test	
Chunks	16569	4194	7968
Disfluent sequences	6619	1737	2120
% disfluent sequences	28.5%	29.3%	21.0%

Table 2: Performance for cross-validation and test set.

	C-value	Cross-validation	Test set	
		Acc.	Acc.	Kappa
OS-F	0.01	86.84%	84.60%	0.646
AS-F	1.0	85.51%	83.98%	0.629

3. EXPERIMENTS

The goal of this section is to find a set of features that allows us to discriminate disfluent segments from non-disfluent ones. The reported experiments are based on an automatic segmentation, provided by the in-house ASR system [10]. We have used the large set of openSMILE features from the Interspeech 2013 Paralinguistic challenge. The openSMILE toolkit [3] is capable of extracting a very wide range of speech features and has been applied with success in a number of paralinguistic classification tasks and for disfluency prediction **[17]**. It has been used in the scope of this study to extract a feature vector containing 6125 speech features (henceforth denoted as OS-F) by applying segment-level statistics (means, moments, distances) over a set of energy, spectral and voicing related frame-level features.

Different classification methods from the Weka toolkit [4] have been applied, including: Naïve Bayes, Logistic Regression, Decision trees, Classification and Regression trees, and Support Vector Machines (SVM). However, results reported in this section refer to SVM only, which have almost consistently achieved the best performances. The SVM has been setup to use Sequential Minimal Optimization with a Linear kernel as the training algorithm.

3.1. University lectures

Table 2 presents a summarized view of the results presented in this section. The cross-validation column shows the accuracy performance achieved using only the training data, with 10 folds, and offers useful insights about the model performance. The best experimented complexity of the model, expressed by the parameter C, was also

calculated based on training data only, and the resulting model was applied to the test set. The Kappa statistic is a chance-corrected measure of agreement between the classifications and the true classes. A value close to zero indicates that results could be achieved almost by chance, whereas a value close to 1.0 means an almost complete agreement.

The first row of Table 2 show results achieved with the initial set of features. The OS-F feature set comprises a very high number of features, most of them possibly not useful for our specific task. In this context, the need to search for a smaller subset, without lowering the classification performance, is crucial. Having a subset of these features, selected by their usefulness to distinguish the disfluent and non-disfluent classes is useful only if it allows to have simpler models that are equally discriminative. To this end, we automatically selected features using the WEKA's implementation on correlation-based feature subset selection [4], and used a best first search strategy. This approach evaluates the relevance of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Therefore, subsets of features that are highly correlated with the class, while having low intercorrelation, are preferred. Selecting over the training data gives us a total of 222 automatically selected features (AS-F) that best discriminate between the two classes of segments. Results achieved with this subset of features are presented in the second row of Table 2. Apart from a small degradation in accuracy performance, this subset allowed us to use a higher complexity model and significantly reduced the time required to train and test the model.

3.2. Cross-domain experiments

To test the robustness of our features in a cross domain scenario we created and applied models across corpora. We have started by creating an additional model using the CORAL corpus for training. These models were then applied to the LECTRA test set, and Table 3 shows the corresponding results. The achieved results are about 9% lower than the corresponding results achieved using the model created with the LECTRA training data (*vide* Table 2). Two main factors contribute to interpret these results, as shown in [11] on an analysis of speaking style effects in the production of disfluencies, considering the same corpora studied in the present work. Firstly, the distributional patterns of disfluencies evidenced that the selection of disfluency types is corpus dependent. Excluding filled pauses, the remaining disfluency categories have different distributional patterns. In dialogues, speakers produce more often repetitions and fragments than in lectures. In lectures, teachers prefer complex sequences of disfluencies (mostly repetitions and substitutions used for lexical search). Those strategies were associated with teachers having more time to edit their speech, displaying strategies associated with more careful word choice and speech planning, whereas dialogue participants had stricter time constraints. Secondly, the prosodic parameters analyzed showed that, although there is a cross corpora prosodic contrast marking between disfluency/fluency repair, there are significant differences in the degrees of contrast made in both corpora. Lectures exhibit the highest pitch maxima in all units of analysis, whereas dialogues exhibit the highest energy maxima. As shown in Tables 3 and 4, there is, in fact, an impact of speaking styles in automatic disfluency detection, however the features that best characterize the behavior of the disfluent sequences in the dialogue corpus fairly predict the disfluencies in the university lectures corpus.

Table 3: CORAL models applied to LECTRA test set.

	C	CORAL	LECTRA test set	
		Cross-val. Acc.	Acc.	Kappa
OS-F	0.01	82.00%	75.74%	0.495
AS-F	1.0	80.71%	74.85%	0.484

Table 4: LECTRA models applied to CORAL.

	Upperbound	CORAL	
	Acc.	Acc.	Kappa
OS-F	82.00%	74.25%	0.393
AS-F	80.71%	71.36%	0.353

The same parameters used in previous experiments were applied to the models used in Table 3. As we did not adjust the parameters, the performance reported for the cross-validation can be assumed to be an upper-bound for the CORAL data. Finally, we have applied our initial models, trained using the LECTRA training data, to the CORAL corpus. Table 4 shows the corresponding results, where the cross-validation performance achieved previously serve as an upperbound.

In the scope of an European Project and aiming at disfluency prediction in the context of IVR systems, the original disfluency detection models trained with 16kHz full bandwidth were downsample to 8kHz, with the use of the telephone simulator FaNT [5]. The main goal is to compare the performance between full and telephone bandwidth, replicating the same conditions of the previous experiments, using the openSMILE features, with the C parameter at 0.01, using the test set of LECTRA and the k-fold cross-validation (k=10) for CORAL. The results of the experiments with the telephone bandwith are displayed in Table 5, revealing no substantial degradation on the performance and encouraging the models' use in IVR domains.

The most informative features include the following: *mfcc_sma[6]_quartile3*, *mfcc_sma[14]_amean*, *mfcc_sma[12]_skewness*, *audspec_lengthL1norm_sma_lpc3*, and *pcm_Mag_harmonicity_sma_de_iqr1-3*. This selec-

Table 5: Telephone bandwith models.

	Acc.	Kappa
LECTRA	85.95%	0.64
CORAL	86.75%	0.54
LECTRA to CORAL	80.87%	0.45
CORAL to LECTRA	80.05%	0.55

tion points out to the importance of the MFCCs features and for the audiospectral differences. It is known that MFCCs highly contributes to distinct types of tasks, being quite transversal in a plethora of speech prediction tasks, what is specific of disfluencies are the audiospectral differences, in line with [19].

4. CONCLUSIONS AND FUTURE WORK

We have performed a disfluency detection task using openSMILE features, extensively used in Interspeech paralinguistic challenges. Amongst different machine learning methods being applied, Support Vector Machines achieved the best performance. Feature selection experiments revealed that the dimensionality of a large set of features can be further reduced at the cost of a small degradation (about 1% absolute). The robustness of the features across domains was investigated using university lectures and dialogues. Different models trained with one corpus were tested on the other, revealing that models can be quite robust across corpora for this task, despite their distinct nature. Models trained using university lectures achieved about 74% accuracy on the dialog corpus, and about 76% accuracy in the opposite direction. This later results is possibly due to the fact that CORAL contains more contrastive tempo characteristics, shares with LECTRA most of the pitch and energy patterns on disfluent sequences and therefore a model created with such data generalizes better. In the scope of an European Project, we have conducted additional experiments aiming at disfluency prediction in the context of IVR systems. Results show no substantial degradation on the performance, encouraging the use of the models in IVR domains.

Future work will encompass merging features now being extracted with openSMILE with knowledge-based features in order to their contribution in cross-domain disfluency prediction. Our work will also tackle other spontaneous domains and cross-language studies in human-machine interfaces.

5. ACKNOWLEDGEMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, under Post-doc grant SFRH/PBD/95849/2013, and by EU-IST FP7 project SpeDial, under contract 611396.

6. REFERENCES

[1] Batista, F., Moniz, H., Trancoso, I., Mamede, N., Mata, A. I. 2012. Extending automatic transcripts in a unified data representation towards a prosodic-based metadata annotation and evaluation. *Journal of Speech Sciences* (3), 115–138.

[2] Eklund, R. 2004. *Disfluency in Swedish Human-Human and Human-Machine Travel Booking Dialogues*. PhD thesis University of Linköpink.

[3] Eyben, F., Wollmer, M., Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. ACM, , (ed), *Proceedings of the international conference on Multimedia, MM '10* New York, NY, USA. 1459–1462.

[4] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. 2009. The weka data mining software: an update. *SIGKDD Explororations* 11(1), 10–18.

[5] Hirsch, H., Finster, H. 2005. The simulation of realistic acoustic input scenarios for speech recognition systems. *Proceedings of Interspeech 2005* 2697–2700.

[6] Jurafsky, D., Martin, J. H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR second edition.

[7] Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., Harper, M. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transaction on Audio, Speech and Language Processing* 14(5), 1526–1540.

[8] Medeiros, H., Batista, F., Moniz, H., Trancoso, I., Nunes, L. 2013. Comparing different machine learning approaches for disfluency structure detection in a corpus of university lectures. ACM, , (ed), *SLATE 2013* Oporto, Portugal. 259–269.

[9] Medeiros, H., Moniz, H., Batista, F., Nunes, L., Trancoso, I. 2013. Disfluency detection based on prosodic features for university lectures. *Interspeech 2013* Lyon, France. 2629–2633.

[10] Meinedo, H., Viveiros, M., Neto, J. 2008. Evaluation of a live broadcast news subtitling system for Portuguese. *Interspeech* Brisbane, Australia.

[11] Moniz, H., Batista, F., Mata, A. I., Trancoso, I. December 2014. Speaking style effects in the production of disfluencies. *Speech communication* 65(6), 20–35.

[12] Moniz, H., Batista, F., Trancoso, I., Mata, A. I. 2012. Prosodic context-based analysis of disfluencies. *Interspeech 2012* Portland, Oregon.

[13] Nakatani, C., Hirschberg, J. 1994. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America (JASA)* (95), 1603–1616.

[14] Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., Caseiro, D. 2008. Broadcast news subtitling system in Portuguese. *ICASSP 2008* 1561–1564.

[15] Ostendorf, M., Favre, B., Grishman, R., Hakkani-Tür, D., Harper, M., Hillard, D., Hirschberg, J., Ji, H., Kahn, J. G., Liu, Y., Maskey, S., Matusov, E., Ney, H., Rosenberg, A., Shriberg, E., Wang, W., Wooters, C. 2008. Speech segmentation and spoken document processing. *IEEE Signal Processing Magazine* 25(3), 59–69.

[16] Pellegrini, T., Moniz, H., Batista, F., Trancoso, I., Astudillo, R. 2012. Extension of the lectra corpus: classroom lecture transcriptions in european portuguese. *GSCP 2012* Brazil.

[17] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S. 2013. Paralinguistics in speech and language - state-of-the-art and the challange. *Computer Speech and Language* (27(1)), 4–139.

[18] Shriberg, E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis University of California.

[19] Shriberg, E. 2001. To "errrr" is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* 31, 153–169.

[20] Trancoso, I., Martins, R., Moniz, H., Mata, A. I., Viana, M. C. May 2008. The Lectra corpus - classroom lecture transcriptions in European Portuguese. *LREC 2008* Marrakesh, Morocco.

[21] Trancoso, I., Viana, M., Duarte, I., Matos, G. 1998. Corpus de dialogo CORAL. *PROPOR'98* Porto Alegre, Brasil.

THE WORD-ONSET EFFECT: SOME CONTRADICTORY FINDINGS

Sieb Nooteboom & Hugo Quené

Utrecht institute of Linguistics OTS
Utrecht University

ABSTRACT

In this paper we describe two experiments exploring possible for reasons for earlier conflicting results concerning the so-called word-onset effect in interactional segmental speech errors. Experiment 1 elicits errors in pairs of CVC real words with the SLIP technique. No word-onset effect is found. Experiment 2 is a tongue-twister experiment with lists of four disyllabic words. A significant word-onset effect is found. The conflicting results are not resolved. We also found that intervocalic consonants hardly ever interact with initial and final consonants, and that words sharing a stress pattern are a major factor in generating interactional errors.

1. INTRODUCTION

Recently, in an investigation on interactions speech errors in spontaneous Dutch [1], two things were demonstrated:

(1) Interactional consonant substitutions only rarely cross between initial, medial and final positions in the word (see Fig. 1).

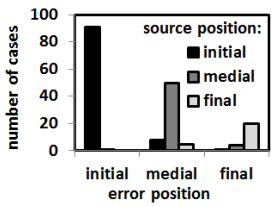


Fig. 1. Error frequencies as a function of source and error position in the word. Data from speech errors in spontaneous speech.

(2) Relative numbers of interactional segmental substitutions may be predicted rather accurately from the relative numbers of opportunities for phonotactically allowed interaction in different positions (see Fig. 2).

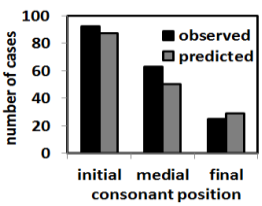


Fig. 2. Numbers. of consonant substitutions. in spontaneous Dutch in three positions. Predictions are made from relative numbers of opportunities for interaction.

Obviously, relatively many interactional consonant errors in spontaneous Dutch are in word-initial position, but this apparent word-onset effect (see Fig. 1) can be explained from the number of phonotactically allowed opportunities for interaction (Fig. 2): There are simply on average more onset consonants than other consonants in the immediate context.

However, tongue-twister experiments reported in the literature have shown a considerable and significant word-onset effect that, at least in the context of those experiments, cannot be explained from the relative numbers of opportunities for interaction for different positions, because those were kept equal. [2] found that in a tongue twister experiment focused on CVC real words no word-onset effect was found in sequences of four words, such as "leap note lap lute" but a considerable word-onset effect was found when such CVC words were embedded in phrases, as in "from the leap of the note to the nap of the lute". This result can perhaps be explained by the fact that in lists of CVC words there are equally many initial as final consonants, but in phrases there are many more initial than final consonants and therefore more opportunities for interaction between initial than between final consonants. But a result reported in [4] cannot be explained from numbers of opportunities. There, in a tongue twister experiment with lists of four CVC words, a considerable and highly significant word-onset effect was found for real words but no effect whatsoever for nonwords. The effect for real words conflicts with both the explanation from numbers of opportunities for interaction proposed in [1] and the data reported in [2]. In [3] some different tongue-twister experiments are described apparently demonstrating a considerable word-onset effect that cannot be easily explained from numbers of opportunities.

To explore possible causes for these contradictory findings, we have conducted two experiments eliciting interactional substitutions of single consonants. The first experiment was set up to elicit interactional substitutions of both the initial and the final consonant in CVC real words. The second experiment was a tongue-twister experiment with disyllabic words.

2. EXPERIMENT 1

In this experiment the classical SLIP technique was used, applying the phonological preparation of substitution errors by precursor word pairs. Targeted errors were either on the initial or on the final

consonant. In Table 1 we present examples of how interactional errors on initial and final consonants might be elicited using the SLIP technique.

Table 1. Examples of stimuli eliciting word initial or word final interactions.

	word pair (onset cons.)	**word pair (offset cons.)**
precursor	**k**uip **g**if	gie**t** poe**s**
precursor	**k**iem **g**oor	fei**t** ro**s**
precursor	**k**ies **g**ut	pie**t** geu**s**
precursor	**k**oos **g**ul	fa**t** be**s**
precursor	**k**aai **g**aar	waa**d** paa**s**
target	**g**aan **k**aan	gaa**s** gaa**t**
prompt	?????	?????
response	*kaan gaan*	*gaat gaas*

The precursor word pairs are visually presented to the speaker one by one in the middle of a screen, and have to be read silently. The target word pair is also to be read silently, but is then followed by a series of ?????. This a prompt for the speaker to speak the target word pair aloud. Every now and then in the response the two initial consonants are exchanged. Ninety-four speakers, all students of Utrecht University, participated in the experiment. This method was used to elicit exchanges, anticipations and perseverations of initial and final consonants in CVC real words. Ninety-four speakers, all students of Utrecht University, participated in the experiment. The main results are given in Fig. 3.

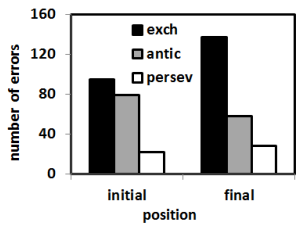


Fig. 3. Main results of Experiment 1. There is no significant difference between the numbers of interactional errors elicited in initial and final position.

The probability of an interactional error was analyzed by means of generalized linear mixed models (GLMM), using Markov Chain Monte Carlo simulations. The dependent variable was the binomial outcome of a response being an interactional substitution error or being a fluent and correct response. Other types of responses were ignored in the present analysis. Fixed effects were the position of the elicited error within the word, the direction of the elicited error, and the interaction of these two fixed effects. Subjects and items were included as random effects. For the current study, the interesting comparison is between the error rates for the initial and final positions. For elicited anticipations, there are fewer final errors than initial ones, but the difference is relatively small and not significant (p=.1854). For perseverations, the error rates are marginally *higher* for final than for initial consonants (p=.0557). Similarly, for elicited exchange errors, the error rate is significantly *higher* for final errors than for initial ones, i.e., fewer initial errors than final ones (p=.0141). This can perhaps be explained from the fact that final consonants tend to be more confusable than initial consonants [5], although it remains unclear why there are not more final than initial anticipations. In any case, these data do not show a clear and significant word-onset effect. They resemble more the results reported by [2] than those reported by [4]. These latter results, showing a considerable and highly significant word onset effect for real words but not for nonwords, in a tongue twister experiment, remain as yet unexplained.

3. EXPERIMENT 2

This experiment was a replication of an experiment reported by [3] but with a twist. Table 2 gives examples of stimuli in [3].

Table 2. Example of stimuli quartets used in [3]. Targeted consonants share word onset and pre-stress position in B, word-onset position in W, pre-stress position in S and neither in N. Numbers of elicited targeted interactional errors are given in the right-most column.

	[3]. (1992)	**NE**
B	*p*ack *f*ussy *f*ossil *p*ig	**253**
W	*p*ad *f*orsake *f*oresee *p*ot	**132**
S	*p*in su*ff*ice su*ff*use *p*et	**75**
N	*p*od so*f*a su*ff*er *p*eg	**26**

Note that in [3] in conditions S and N interactions are elicited between initial and medial consonants. This probably artificially reduces the numbers of errors (cf. Fig. 1 above). Therefore in our experiment we used two sets of Dutch stimuli, one comparable to the stimuli in Table 2, the other with disyllabic words only, thus avoiding the confound with crossing positions.

Table 3. Examples of stimuli used in the current tongue twister experiment.

	stimuli as in [3]: 1vs2 syllables	**stimuli with disyllables only: 2vs2 syllables**
B	**w**ok **r**ápper **r**óeper **w**al	*w*áter *r*ápper *r*óeper *w*állen
W	**w**ad **r**appórt **r**apíer **w**ol	*w*óeker *r*appórt *r*apíer *w*íkkel
S	**w**in par**r**ijs por**r**éus **w**el	be*w*ijs par*r*ijs por*r*éus ju*w*éel
N	**w**it píe**r**en pá**r**el **w**as	la*w*áai píe*r*en pá*r*el ge*w*ín

As exemplified here, stimulus word pairs of the "2vs2 syllables" type were derived from those of the "1vs2 syllables" type. We have decided that such related quartets should not be presented to the same participant because this might be confusing. Therefore we created two lists of stimuli each with 12 quartets of the "1vs2 syllables" type and 12 quartets of the "2vs2 syllables" type, in such a way that for each quartet of the "1vs2 syllables" type the

corresponding quartet of the "2vs2 syllables" type was in the other list and vice versa. Thus each list had 24 quartets and therefore 96 sequences of four words. There were 28 participants, 20 females and 8 males, all students of Utrecht University. Their age ranged from 18 to 26. Data from one participant (female, even-numbered) were lost due to technical malfunction. The analysis reported below is based on the remaining 27 participants.

Each speaker was tested individually in a sound-treated booth. He or she was instructed to repeat each sequence of words that appeared on the screen three times, then to push a button that made the stimulus disappear and to repeat the same sequence of words three more times from memory. The resulting speech from each participant was transcribed and coded separately for "1vs2 syllables" "stimuli and the "2vs2 syllables" stimuli. The source and error position of each interactional error was coded. There was quite some hysteresis in the sense that once a particular error was made, the participant tended to repeat that error during the six response utterances for that stimulus. Because of this we counted only the first of identical errors made to a stimulus. The errors were counted separately for the "visible" and "invisible" phase of the experiment. All interactional substitutions, targeted and not targeted, by exchange, anticipation and perseveration were counted.

Table 4 gives the main results for the targeted interactional substitutions only.

Table 4. Numbers of targeted single consonant substitutions separately for the "1vs2 syllables" and "2vs2 syllables" stimuli, for the visible and invisible part of the experiment, and for the four conditions B, W, S, N.

	1vs2 syllables				2vs2 syllables		
	vis	invis	sum		vis	invis	sum
B	32	85	117		47	80	127
W	18	35	53		19	32	51
S	8	14	22		20	38	58
N	4	4	8		3	12	15
sum	62	138	200		89	162	251

These data were fed into a mixed-effects logistic regression model (GLMM; Quené & Van den Bergh, 2008). Fixed effects were the condition (with N as baseline), 1vs2 syllables (baseline) versus 2vs2 syllables, and visible versus invisible (baseline). Random intercepts were included for participants and for stimulus quartets, and condition was included as a random slope between stimulus quartets (an extended model with random slope of condition between participants did not increase the model's performance). The three-way interaction between the fixed effects was not significant according to a likelihood ratio test (χ^2=6.9, df=3, p=.074), and it was therefore dropped from the model. The visibility factor does not interact with the condition factor and does not interact with the "1vs2 syllables" versus the "2vs2 syllables" factor (according to a likelihood ratio test, a simpler model from which these interactions were dropped performs equally well; χ^2=4.2, df=4, p=.38).

Two findings are remarkable: (1) There are more targeted interactional substitutions elicited by the "2vs2 syllables" stimuli than by the "1vs2 syllables" stimuli (p<.0001), but only in the B and S conditions. Consequently the distributions of interaction errors over conditions are significantly different for the "1vs2 syllables" stimuli and the "2vs2 syllables" stimuli. (2) There are significantly fewer targeted interactional substitutions in the visible phase of the experiment than in the invisible phase (p<.0001). When the participants do not see the four-word sequence on the screen they make more interaction errors. This suggests a memory problem. However, there is no interaction between visibility and condition.

Obviously, the distribution of the numbers over conditions is very different for the two sets of stimuli. This appears to be related to the fact that in the "1vs2 syllables" stimuli in two conditions targeted interactions involved two different positions in the word, whereas in the "2vs2 syllables" stimuli they did not. Therefore we refrain from further analysis of this data set and rather turn to the data set for the "2vs2 syllables" only. This time, however, we choose to include all valid interactional errors and not only the targeted ones.

First we want to see whether in this experiment interactional errors have the same resistance against crossing positions as found for errors in spontaneous speech by [1]. Fig. 4 gives the relevant data.

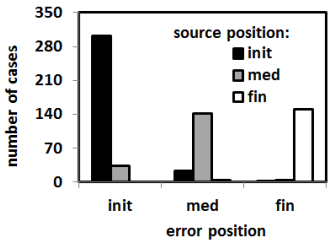


Fig. 4. Error frequencies as a function of the position of both error and source in the word. Data from Experiment 2.

Obviously, as in spontaneous speech, in this experiment also interactional errors have a strong resistance against crossing positions. Only 11 % of errors cross positions. We further analyze only cases for which source and error have the same position in the word.

Table 5 gives the breakdown over positions and conditions of all interactional substitutions for the "2vs2 syllables" stimuli.

Table 5. Numbers of single consonant substitutions for the "2vs2 syllables" stimuli, not crossing positions, for both targeted and not-targeted errors, collapsed over visible and invisible, and separately for the initial, medial and final consonant positions, and for the four conditions B, W, S, N. Positions targeted for interaction are printed in boldface. The highest number of each row is printed in italics.

	initial	medial	final	sum	fluent
B	***126***	55	73	254	456
W	***51***	13	11	75	481
S	*83*	**57**	28	168	462
N	*41*	**16**	38	95	476
sum	301	141	150	592	1875

Two things are immediately conspicuous:
(a) Overall numbers of errors differ considerably between conditions, also for consonants not targeted for interaction.
(b) The highest number of errors in each condition is in word-onset position, whether or not this is the targeted position.

In conditions B and S all 4 words in the tongue twisters share a stress pattern, viz. Sw in B and wS in S. This is not so in conditions W and N. Therefore we suspected that the enormous differences between conditions might have something to do with the sharing of stress patterns. We applied an analysis with a mixed-effects logistic regression model, and three different contrasts for the factor condition:.
(1) initial vs medial position (B & W vs S & N); (2) sharing vs not sharing a stress pattern (B & S vs W & N);
(3) sharing Sw vs sharing wS (B & N vs W & S). Contrast (1) was insignificant (p = .347), contrast (2) highly significant (p < .0001), contrast (3) significant (p = 0307). Contrast (1) being insignificant suggests that targeting specific pairs of consonants does little to generate interactions between those consonants. This may come as a surprise to those dedicated to doing tongue twister experiments. Contrast (2) being highly significant suggests that activation of all segments of words in each other's immediate context that share a stress pattern is increased, as if by stress-pattern-based priming. Thereby their probability for interactions is increased. The stronger effect of Sw than of wS probably is related to the fact that Sw is by far the most common of the two stress patterns in Dutch.

We found a similar effect of stress pattern in spontaneous speech: In a collection of single segment substitutions there were same and different stress patterns in 151 and 169 cases respectively . Expected values based on stress pattern frequencies were 98 and 222. The distributions are clearly different (Fisher's exact test: p<.0001).

The above analysis could not address the effect per se of position in the word, independent of other contrasts, because this was not an experimental variable. However, a post hoc logistic regression analysis with 1000 bootstrap replications over both speakers and matching stimulus sets. with position as main variable, clearly showed a significant word-onset effect in all 4 conditions (p<.05). This effect cannot be explained from relative numbers of opportunities for interaction.

4. CONCLUSION

We have not succeeded in explaining the absence of a word-onset effect in spontaneous speech, other than caused by number of opportunities for interaction, in [1], and also in lists of real CVC words in [2] and the current experiment 1, as compared to the presence of a clear word-onset effect in CVC words in phrases in [2] and in lists of real CVC words in [4], and also in lists of disyllabic words in the current experiment 2. We stumbled over two major effects that were previously hardly known: (1) Intervocalic consonants do not interact with initial or final consonants. This means that at the level where interactional substitutions arise, there is no resyllabification. (2) There is a major effect, confirmed in spontaneous speech, of words sharing a stress pattern on the frequency of their interactions.

5. REFERENCES

[1] Nooteboom, S.G. & Quené, H. 2015. Word onsets and speech errors. Explaining relative frequencies of segmental substitutions. *Journal of Memory and Language* 78, 33-46

[2] Shattuck-Hufnagel, S. 1987. The role of word-onset consonants in speech production planning: New evidence from speech error patterns. In: E. Keller & M. Gopnik (eds.) *Motor and sensory processes of language*. Hillsdale: Lawrence Erlbaum Associates, 17-52.

[3] Shattuck-Hufnagel, S. 1992. The role of word structure in segmental serial ordering. *Cognition, 42*, 213-259.

[4] Wilshire, C.E. 1998. Serial order in phonological encoding: an exploration of the 'word onset effect' using laboratory-induced errors. Cognition, 68, 143-166.

[5] Redford, M.A., Diehl, R.L. 1999. The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *Journal of the acoustical Society of America* ***106***, 1555—65.

UM AND UH AS DIFFERENTIAL DELAY MARKERS: THE ROLE OF CONTEXTUAL FACTORS

Ralph ROSE

Waseda University
rose@waseda.jp

ABSTRACT

The English filled pauses *uh* and *um* have been argued to correspond respectively to shorter and longer anticipated delays in speech production. This study looks at some contextual factors that might cause this difference by investigating filled pause instances in monologue and conversation speech corpora. Results are consistent with previously observed delay differences and further show that discourse-level processing may influence differential delay marking though monologue results are more conclusive than conversation results. However, no evidence was found that lexical factors (word type, frequency) correlate with filled pause choice. The findings suggest a limited view of how speakers use filled pauses as delay markers: Not all contextual factors may trigger differential delay marking.

Keywords: filled pause, delay, contextual factors

1. INTRODUCTION

According to the perceptual loop theory [1,2], speakers are constantly monitoring their speech production and, upon identifying a problem in their production, may initiate repair of the problem. In many cases, the repair may be covert and entirely unnoticeable to the listener. But in other cases, the repair takes some overt form. In this way, filled pauses (FPs)—like other hesitation phenomena (e.g., repeats, self-repairs, lengthenings, silent pauses)—can be said to mark such a repair sequence and thus constitute a delay in the communication of the speaker's message [cf., 3,4].

In English, the FP inventory is extremely limited, consisting almost exclusively of just an open syllable *uh* or a closed syllable *um*. Furthermore, the English FPs *uh* and *um* show some difference with respect to the length of the associated delay [3-7; but see 8-9 for counter-evidence]. *Uh* corresponds to a lower likelihood of an immediately following silent pause and an overall shorter delay (FP duration plus following silent pause duration) than *um*. Thus, it has been proposed [3-5] that when speakers detect a minor problem in their speech production and thus predict a minor delay to repair the problem, they are more likely to mark this delay with *uh* rather than *um*. Conversely, when speakers detect a major problem in their speech production, they are more likely to mark the associated delay with *um*.

Given the empirical observations of a differential delay between *uh* and *um*, then the next problem is to pinpoint the cause or causes of the respective minor and major delays that trigger the FPs. Several contextual factors have been observed to correspond with the occurrence of FPs. For example, much evidence exists to show that filled pauses are more likely to be used at major rather than minor discourse boundaries [7,10]. For instance, consider the hypothetical spoken discourse shown in (1). The beginning of the discourse {A} is a major discourse boundary requiring much planning effort as the storyteller plans the entire discourse to follow. The sentence boundary {B} is a less major boundary in the discourse as a whole, while the beginning of the subordinate clause {C} is a minor boundary and the clause-internal point {D} is a more minor boundary.

(1) {A}Yesterday I was walking down the street when I saw a surprising thing. There was this guy selling toys {E} in a small {F} stall and everyone was watching him because he was so unique. {B} He would balance several toys at once in one hand {C} while demonstrating a new toy {D} with the other hand. All the kids couldn't help but watch and so many parents had no choice but to buy something!

Major discourse boundaries like {A} and {B} incur greater speech production difficulty than minor boundaries like {C} or {D} and therefore increase the likelihood of a delay.

FPs have also been observed to occur more often before content words than before function words [11] and before words with low rather than high contextual frequency [12]. Hence, in (1), there is a greater probability of a FP at point {F} before the low-frequency word *stall*, than there is at point {E} before the high-frequency word *in*. More effort is required to retrieve low-frequency words from memory and thus the likelihood of a delay increases [13]. Since content words as a whole are lower frequency than function words, the same explanation applies, meaning a higher probability of FPs before content than function words.

Because such contextual factors as discourse boundary level and word frequency have been observed to incur speech delay as measured by the occurrence of FPs, a further hypothesis is that gradient differences in these contextual factors correspond to a speaker's anticipation of greater or lesser delay and thus greater or lesser choice of *um* or *uh*, respectively. This paper reports on a test of this hypothesis using speech corpora.

2. EXPERIMENT

2.1. Methods

In order to test the hypothesis in a broad range of speech contexts, this study used two English speech corpora: one of monologues and the other of conversations. The monologue corpus is the Corpus of Presentations in English (COPE: [14]) in which native English participants spoke in response to a prompt for about ten minutes, following minimal preparation time. Participants spoke in front of a small audience of peers. The conversation corpus is the Santa Barbara Corpus (SBC: [15]) which consists of unstructured, non-task-oriented conversations. Recordings were taken as speakers engaged in normal, everyday activities and conversations in non-laboratory settings (cf., BNC Spoken Corpus [16]).

A sample of FPs from each corpus were taken and the following delay measurements were made: FP duration, presence of immediately following silent pause, duration of following silent pause, and total delay (calculated as FP duration plus duration of any immediately following silent pause). Subsequently, the following contextual measurements were also made: syntactic location (clause boundary or clause-internal: a simplified view of discourse boundary level, corresponding to points A, B, and C versus D in (1) above), following word type (content or function), and following word frequency (using frequency counts from Brown Corpus [17]).

2.2. Results

The analysis is based on 163 FPs contained in 20 minutes of the COPE (monologue) corpus and 149 FPs contained in 165 minutes of the SBC (conversation) corpus. Turn-final FPs in the conversation corpus—where a following silent pause could arguably be attributed not to the speaker but rather to an interlocutor's latent uptake—were not included in this analysis. The results for the delay measures are shown in Tables 1-4 (with test statistics for main effects).

Table 1: Mean duration of FPs.

Speech type	*uh*	*um*	
monologue	328 ms	423 ms	F(1,308)=5.8
conversation	340 ms	363 ms	p<0.05
	F(1,308)=17.0, p<0.001		

Table 2: Proportion of FPs followed by silent pauses

Speech type	*uh*	*um*	
monologue	16.4%	38.2%	n.s.
conversation	16.9%	36.9%	
	p<0.001 (log. regr.)		

Table 3: Mean duration of silent pauses following FPs

Speech type	*uh*	*um*	
monologue	398 ms	825 ms	n.s.
conversation	744 ms	910 ms	
	F(1,133)=3.9, p<0.05		

Table 4: Mean total delay (duration of FP plus following silent pause).

Speech type	*uh*	*um*	
monologue	465 ms	876 ms	n.s.
conversation	592 ms	818 ms	
	F(1,308)=16.0, p<0.001		

The duration of FPs (Table 1) differs with respect to the type of FP (*uh*, *um*) and the speech type (monologue, conversation), and even the interaction is significant [F(1,308) = 6.1, p<0.05]. While *um* is longer on the whole, this difference is only significant in monologue speech [t(143)=4.6, p<0.001]. Although *um* consists of two phonemes compared to the one in *uh*, this does not necessarily mean that *um* consistently takes longer to articulate. The articulation of FPs seems to be partially modulated in the presence of interlocutors.

Results further show that there is more likely to be a silent pause (Table 2) after *um* than *uh*, but there is no difference between speech types in this regard. However, when looking at the length of the following silent pauses (Table 3), results show that despite the apparently very short pauses after *uh* in monologues, there is only a marginal difference between *um* and *uh* barely reaching significance at α =0.05 and no difference between speech types. This result therefore waters down differences between the FPs themselves (Table 1) when looking at the effect of the total delay (Table 4): There is a difference between *um* and *uh*, but no difference between the speech types and no interaction.

The results for the total delay show, therefore, that *um* marks a longer delay than *uh*. This replicates previous findings [3-7], and shows the results to be consistent across both monologue and conversational speech. However, the composition of that delay (i.e., the relative length of *uh/um* and its accompanying pause) differs between speech types.

As for the hypothesis that gradient differences in contextual factors corresponds to differential delays, the results for the contextual measures are shown in Tables 5-8.

Table 5: Proportion of clause boundary FPs that are closed FPs (*um*)

Speech type	boundary	internal	
monologue	74.7%	45.6%	n.s.
conversation	54.4%	58.6%	
	p<0.05 (log. regr.)		

Table 6: Mean total delay (duration of FP plus following silent pause) at clause locations

Speech type	*uh*	*um*	
monologue			Interaction: F(1,304)=7.3 p<0.01
boundary	433ms	1004ms	
internal	486ms	581ms	
conversation			
boundary	552ms	680ms	
internal	641ms	963ms	
	F(1,304)=16.4, p<0.001		

Table 7: Proportion of following words that are content words

Speech type	*uh*	*um*	
monologue	27.9%	27.4%	p=0.08 (log. regr.)
conversation	39.3%	35.5%	
	n.s.		

Table 8: Mean (log) frequency of following word

Speech type	*uh*	*um*	
monologue	7.52	7.81	F(1,296)=9.4 p<0.005
conversation	6.54	6.94	
	n.s.		

Overall, when a FP is used clause-internally, the probability of it being an *um* rather than *uh* seems to be near chance (Table 5). However, when a FP is used at a clause boundary, it is more likely that it will be an *um* than an *uh*. This result appears to be driven mostly by the monologue speech results, but in fact the difference between speech types is not significant. Hence, the overall result suggests that major discourse boundaries prompt greater use of *um*. However, when looking at the actual total delay results (Table 6), the situation is somewhat more complicated. While there is a main effect of FP with *um* longer than *uh* (i.e., same as shown in Table 4 above), there is an interaction between speech type and clause location. The monologue speech data shows that boundary *ums* are longer than others. Together with the results shown in Table 5, this follows the prediction that major discourse boundaries prompt longer delays which are marked by *um*. The conversation speech data seems to show something of an opposite trend where internal FPs have a longer duration than boundary FPs. However, this difference is not significant. Thus, only the monologue speech data is conclusive with regard to the contextual effect of discourse boundary on differential delay.

In contrast, as for the following word, choice of *um* and *uh* shows no connection with the type (Table 7) nor with the frequency (Table 8) of the following word. But the results do show differences between speech types with FPs in conversation (more so than in monologue) followed by more content words and by lower-frequency words. Results relating word type and frequency to mean total delay showed no relevant significant effects and are not shown here.

3. DISCUSSION

The aim of this study was to evaluate the hypothesis that differences in delays associated with FPs may be attributed to certain contextual factors. Results first show that *uh* and *um* do correspond with shorter and longer total delay, respectively (consistent with previous work). Results further show that *um* is more likely to be used at a clause boundary than clause-internally suggesting that processing associated with a major discourse boundary may be a factor that impels speakers to choose to use an *um* over an *uh*. Positive support for this conclusion comes only from the monologue speech data. The conversation speech data was inconclusive on this point. More discussion of this difference between the monologue and conversation data will be given below.

Despite the influence of discourse boundary level, comparable findings could not be obtained for the type or frequency of words following FPs, suggesting that lexical access effects are not sufficient to influence differential delay marking by FP choice.

These results suggest a limited view of FPs as delay-marking devices: While there may be many sorts of linguistic processing problems that speakers experience and which may cause an anticipation of delay such as the various contextual factors examined here, using a FP as a signal to mark the

degree of that expectation (i.e., short vs. long) is not a generic technique. Rather, it seems to be limited to cases where speakers recognize difficulty in processing major discourse constituents. Other delays might lead to different techniques.

Finally, the results here bear vaguely on a larger question regarding the use of FPs in spontaneous speech: the speaker's intent. As noted in the background, when speakers detect a minor delay in speech production, they mark it with *uh*, and mark a detected major delay with *um*. This can be unpacked into two hypotheses. First, there is the hypothesis that different FPs in English correspond to different delay lengths. This may be called the *differential delay hypothesis*. Previous studies as well as the present study confirm this hypothesis.

A second hypothesis, though, is that speakers intend to convey their anticipation of a delay differentially to their interlocutors. This can be called the *differential conveyance hypothesis*. None of the studies cited in the background provide clear evidence on this hypothesis: That is, there seem to be no tests of the speaker's intent to convey something different between *uh* and *um*.

Evaluating intent is surely a difficult task, but perhaps differences between the corpora used in this study are suggestive. In the monologue corpus, speakers were obliged to speak for a target amount of time, in order to accomplish the investigator's task. Furthermore, they were doing so in front of an audience. Here, speakers may feel more compulsion to communicate about their anticipated delays: Hence, their intent may be taken as a given.

On the other hand, in the conversation corpus, speakers were under no investigative time or task constraints: They could, so to speak, take their time freely within their conversation. In this context, they were under less compulsion to communicate about their anticipated delays to their interlocutors.

If this distinction is valid, then the prediction of the differential conveyance hypothesis would be that FPs would be used differently between the monologue and conversation corpora. The results here do in fact show this (cf., Table 6) where the differential use of *um* and *uh* at different discourse boundary levels occurs in the monologue but not the conversation speech data. Therefore the results provide support for the hypothesis. Of course, this is highly speculative and warrants much further examination.

4. FURTHER WORK

Although this work has looked at a broad sample of data with both monologue and conversational speech, the number of samples used was relatively limited and could be expanded to confirm the findings. Also, other factors that might lead to expectation of delay could be examined such as articulation, (co)reference processing, or syntactic and semantic effects. These could be investigated in corpus studies as performed here or in controlled production or perception experiments to see whether and how these various factors are related to differential delay as marked by *uh* and *um*.

5. REFERENCES

[1] Levelt, W.J.M. 1983. Monitoring and self-repair in speech. *Cognition* 14, 41-104.

[2] Levelt, W.J.M. 1989. *Speaking: From intention to articulation. Cambridge*. MA: MIT Press.

[3] Clark, H., Fox Tree, J.E. 2002. Using uh and um in spontaneous speaking. *Cognition* 84, 73-111.

[4] Clark, H., Fox Tree, J.E. 2014. On thee-yuh fillers uh and um. Language Log, November 11, http://languagelog.ldc.upenn.edu/nll/?p=15718

[5] Smith, V., Clark, H. 1993. On the course of answering questions, *Journal of Memory and Language* 32, 25-38.

[6] Kendall, T. 2013. *Speech Rate, Pause and Sociolinguistic Variation: Studies in Corpus Sociophonetics*. London: Palgrave Macmillan.

[7] Rose, R.L. 1998. *The Communicative Value of Filled Pauses in Spontaneous Speech*, Unpublished Master's Thesis, University of Birmingham.

[8] O'Connell, D., Kowal, S. 2005. Uh and Um Revisited: Are They Interjections for Signaling Delay? *Journal of Psycholinguistic Research* 34. 555-576.

[9] Corley, M., Stewart, O.W. 2008. Hesitation Disfluencies in Spontaneous Speech: The Meaning of um. *Language and Linguistics Compass* 2, 589-602.

[10] Swerts, M. 1998. Filled pauses as markers of discourse structure, *Journal of Pragmatics* 30, 485-496.

[11] Maclay, H., Osgood, C. 1959. Hesitation Phenomena in Spontaneous English Speech. *Word* 15, 19-44.

[12] Beattie, G.W., Butterworth, B.L. 1979. Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech* 22, 201-211.

[13] Oldfield R.C., Wingfield A. 1965. Response latencies in naming objects. *The Quarterly Journal of Experimental Psychology* 17, 273–281.

[14] Watanabe, M. Corpus of Oral Presentations in English (COPE). Unpublished corpus data.

[15] Du Bois, J.W., Chafe, W.L., Meyer, C., Thompson, S.A. 2000. *Santa Barbara corpus of spoken American English*, Part 1. Philadelphia: Linguistic Data Consortium.

[16] Crowdy, S. 1995 The BNC spoken corpus. In Leech, G., Myers, G. Thomas, J., eds. *Spoken English on computer: transcription, mark-up and application* Harlow: Longman, 224-235.

[17] Kucera, H., Francis, W.N. 1967. *Computational Analysis of Present-day American English*. Providence: Brown University press.

CAN YOU HEAR THESE MID-FRONT VOWELS? FORMANTS ANALYSIS OF HESITATION DISFLUENCIES IN SPONTANEOUS HEBREW

Vered Silber-Varod[1], Adva Weiss, Noam Amir[2]

[1]The Research Center for Innovation in Learning Technologies, The Open University of Israel
[2]Dept. of Communication Disorders, Sackler Faculty of Medicine, Tel Aviv University

vereds@openu.ac.il, advaweiss@gmail.com, noama@post.tau.ac.il

ABSTRACT

This study attempts to characterize the timbre of the default type of hesitation disfluency (HD) in Israeli Hebrew: the mid-front vowel /e/. For this purpose, we analysed the frequencies of the first three formants, F1, F2, and F3, of hundreds of HD pronunciations taken from The Corpus of Spoken Israeli Hebrew (COSIH). We also compared the formant values with two former studies that were carried out on the vowel /e/ in fluent speech. The findings show that, in general, elongated word-final syllables and appended [e]s are pronounced with the same amount of openness as fluent [e], while filled pauses tend to be more open (lower F1), and more frontal (higher F2). Following these results, we suggest to use different set of IPA symbols, and not the phonemic mid-front /e/, in order to better represent hesitation disfluencies.

Keywords: hesitation disfluency, filled pauses, LPC analysis, formants, spontaneous speech, Hebrew.

1. INTRODUCTION

Hesitation disfluencies (HDs) in spoken Hebrew are mostly produced by /e/, which is considered a mid-front vowel [4]. This vowel and four others, /a, i, o, u/, are the five vowels in the Israeli Hebrew (IH) vowel system [4]. This vowel system does not include a phonological difference of vowel length. Moreover, lax/tense differences are not phonemic in Hebrew, Nor does it includes the neutral Schwa [ə] as a phoneme. According to the International Phonetic Alphabet [9], diacritics indicating lowering [ę] or rising [ę̝], may be used to represent the quality of mid-front unrounded vowel (nevertheless, in the following, the symbol [e] will be used to represent this vowel). /e/ is the phonetically unmarked vowel in IH and is the vowel most likely to break up undesirable consonant clusters and to be affected by casual vowel deletion [3].

Fox [8] suggests that the 'hesitation vowel' used by speakers is different even in speakers of different varieties of English. When talking about filled pauses, he says that the 'neutral or rest position' is not always the same for different languages [8: 30]. Cruttenden [7] provides additional examples of the differences in the sound of filled pauses between dialects and languages: "In R.P. and in many other dialects of English the latter [the filled pause] typically involves the use of a central vowel [ə] ... in Scottish English a sound in the region of the vowel in *gate* and *play* is typical while in Russian an alveolar nasal is more common than a bilabial nasal" [7: 30]. In French, it is the closed-mid front vowel [Ø], commonly written as *euh* or [œ] but also as an [ɛ], while in American English, it is the open-mid back vowel [ʌ] or a vowel close to the schwa [ə], as mentioned in [15: 164]. For European Portuguese, the two most common phonetic forms for filled pauses are the near-open central vowel [ɐ] and the mid-central vowel [əə] [5].

This language dependent characteristic of filled pauses was already suggested by [16]: "Filled pauses have variants in many (perhaps all) languages, but their vowel quality varies predictably with the vowel inventory of the language" [16: 24-25]. Shriberg [15] demonstrates how disfluency is also associated with alternations in vowel quality – an elongated *the* is pronounced as the variant [ði] [15: 163].

In this paper we will use triple categorization of HDs in IH, i.e. three distinct manifestations with regard to word-level phonology [cf. 13, 14]:

A. Elongated word-final syllables (EWFS). Elongated syllables are mostly monosyllabic function words, such as /ve/ 'and'; /be/ 'in/at'; /ʃe/ 'that', and more.
B. Appended [e] vowels that are inserted after a word, but within the same intonation unit. This type will be further referred as Appended [e] (henceforth: AE).
C. HDs between silent pauses, known also as *filled pauses* (FP).

A similar distinction between the first and the third realizations (types A and C above) was described in [16: 16]: "lengthening of rhymes or syllables" and "hesitating much like they might display with a filled pause". It is important to note that although types B and C are only realized by the /e/ vowel, elongated syllables occurs in word final syllables, or monosyllable lexemes, and thus may be realized by each of the five vowels in IH vowel system. For example, the most common elongated

lexeme in spontaneous Hebrew is the definite article [ha] 'the' [14], which of course is realized by an elongated /a/.

This study focuses on two questions:

Q1: Will the three types of hesitation disfluencies show similar formant values?

Q2: Are formants of [e] when realized by hesitations more likely to be centralized in comparison with fluent [e]?

It is reasonable to hypothesize that formants of elongated word final [e] (type A above) will show closer, if not similar, values to fluent [e]. This question was investigated in other languages. For example, in Estonian it was found that hesitation formants are likely to be centralized or posterior and opened in comparison with related phonemes [11]. In European Portuguese a small tendency for higher F1 and lower F2 in vocalic fillers (termed FP in the present research) was found [12].

To summarize, in the present study, formant analysis of the three types of HDs with the vowel /e/ will be carried out and the results will be compared to two other studies on the acoustics of vowels in IH: The first is on the acoustic properties of Hebrew vowels as manifested in isolated utterances [10]; The second is on the phenomenon of vowel reduction in spontaneous speech [2], which is a study that was carried out in part on the same corpus as in the present research.

2. METHOD

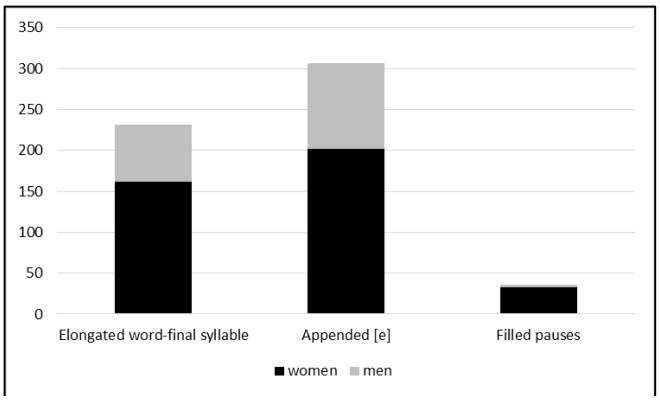
Speech samples were taken from the Corpus of Spoken Israeli Hebrew (CoSIH) [6]. The recordings, which were made during 2001-2002, are of authentic IH everyday conversations. Each dialogue consists of conversations between one core speaker and various interlocutors with whom the speaker interacted on that day.

The present research is based on recordings of 19 subjects who wore two unobtrusive microphones on their lapels, during an entire waking day. A total of 28 interlocutors were also recorded, therefore 47 different speakers are in the corpus, 19 of whom are men and 28 women. All speakers were native speakers of Hebrew, residing in Israel since birth. Mean age was 30.5 years (SD=14.6) for the men group, and 26.3 years (SD = 5.5) for the women. The research corpus consists of 31,760 word-tokens (over 5 hours of speech) of which 4,289 are word-types. 44% of the examined material is one-side telephone conversations; while 56% is face-to-face dialogues.

2.1. The database of HDs

Amongst 764 HDs that were annotated in the corpus, 575 HDs of types A, B, and C, as described above, were segmented and saved separately as sound files. The 575 HDs consist of [e] vowel, while the other 189 HDs consist of elongated vowels in varied syllables (mostly [a] in the Hebrew definite article /ha/ 'the'). This demonstrates that the [e] vowel is the default HD in Hebrew (75.2% HDs with [e] vs. 24.7% HDs with the other four vowels). Figure 1 demonstrates the distribution of the three types of HDs between women and men (with varied amount of productions between the 47 speakers): 70 elongated word-final syllables were collected from men's speech, and 162 from women's speech; 105 Appended [e]s were collected from men, and 202 from women; last, only 3 filled pauses, were collected from men and 33 from women.

Figure 1: Occurrences of the three realizations of hesitation disfluencies spoken by women and men.



2.2. Formant extraction method

A segment of 30-50ms, from the centre of each token was extracted, which was judged visually to be stationary. Linear Predictive Coding (LPC) analysis was applied to the signal. The first three formant frequencies were identified as the frequencies of the first two conjugate pole pairs obtained from this analysis. The results of this analysis may vary considerably, when applied automatically, depending on sampling rate and order of the LPC analysis. Therefore, we performed a manually supervised analysis of the first three formants of each token separately.

3. RESULTS

The results are divided into two parts: The first part describes intrinsic findings on the measurements of the first three formants in the three examined HDs – EWFS, AE, and FP. In the second part we bring

extrinsic findings which compare the current study results to the results in the two previous studies mentioned above ([2 and 10]), which will be termed "reduced /e/" study and "read /e/" study, respectively. A summary of the formant measurements (means and standard deviation values) is presented in Table 1 and in Figure 2.

3.1. Intrinsic comparison

The intrinsic comparison is a comparison between the three types of HDs – elongated word-final syllables (EWFS), Appended [e] (AE), and filled pauses (FP) – which were described above. First, results are presented for the 575 items in the database (Lines 1-6 in Table 1): A one way Analysis Of Variance (ANOVA) was conducted for men (M) and women (W) separately, for each of the first three formants, with HD type as a between-subject factor.

For women, all three ANOVA's (F1, F2, and F3) were significant ($p=0.012$, $p<0.001$, $p<0.001$, respectively). Post hoc comparisons revealed:

1. *F1*:
    a. EWFS is significantly different from FP ($p=0.01$)
    b. FP is significantly different from AE ($p=0.019$)
    c. EWFS is insignificantly different from AE.
2. *F2*:
    a. EWFS is significantly different from FP ($p<0.001$)
    b. FP is significantly different from AE ($p<0.001$)
    c. EWFS is insignificantly different from AE.
3. *F3*:
    a. EWFS is significantly different from FP ($p<0.001$)
    b. FP is significantly different from AE ($p<0.001$)
    c. EWFS is insignificantly different from AE.

For men, similar to the results for women, no statistical differences between formant values across the two HD types – EWFS and AE – were found (FPs are ignored here due to lack of tokens).

A similar analysis was performed next, which included only six main speakers, having 20 HDs or more (Lines 7-12 in Table 1). The motivation was to find if a more balanced token distribution will have an effect on the results. In this case, the number of tokens was reduced to 366 HDs. Results of this analysis were substantially the same as the analysis of all HDs. In general, results showed that EWFSs are closer in pronunciation to AEs and farthest from FPs.

3.2. Extrinsic comparison

The extrinsic comparison to previous works is presented in Figure 2. The main findings are the following:

1. F1 average values of the three types of HDs are within the standard deviations of the records in the two previous works ([2] and [10]), for both men and women. Yet, in comparison to "read /e/", the averages are higher in EWFS and AE (which reflects a more open manner of [e] in these two HDs) and lower in FP (which reflects a more closed [e] in FPs).
2. F2 of AE and EWFS in women's speech is beyond the standard deviations of the "read /e/" study [10]. It can be said that women pronounce AEs and EWFSs in a more back manner (i.e., lower frequencies of F2).
3. Women and men pronounce AE in a more frontal manner (higher F2), in comparison to the "reduced /e/" (as recorded in [2]).
4. Women and men pronounce EWFS in a similar frontal manner to "reduced /e/" (F2 within the standard deviation of [2] study).
5. F2 of FP in men's speech is beyond the standard deviations of F2 in IH [e] (in both [2] and [10] studies). Yet, since it reflects only three cases, men's pronunciation of FP remains for future research.

4. DISCUSSION

In this study, we attempted to characterize the timbre of the default type of hesitation disfluency (HD) in Israeli Hebrew: the mid-front vowel /e/. Of the three HD types, EWFS and AE formant measurements are closer to each other while FP formants show a significant deviation from the other two types.

As F1 values reflect the open-close vowel characteristics, we found that HDs are pronounced with the same amount of openness as fluent [e]. It is interesting to note that the deviations of F1 in elongated word-final syllables and Appended [e] cases, from the F1 in the "reduced /e/" are smaller than those in the "read /e/" and vice versa for the filled pauses. This can reflects discrimination between the attached types of HD (EWFSs and AEs), which have a reduced F1 nature, versus the isolated type of HDs (FPs), which has a more closed nature, as in read speech.

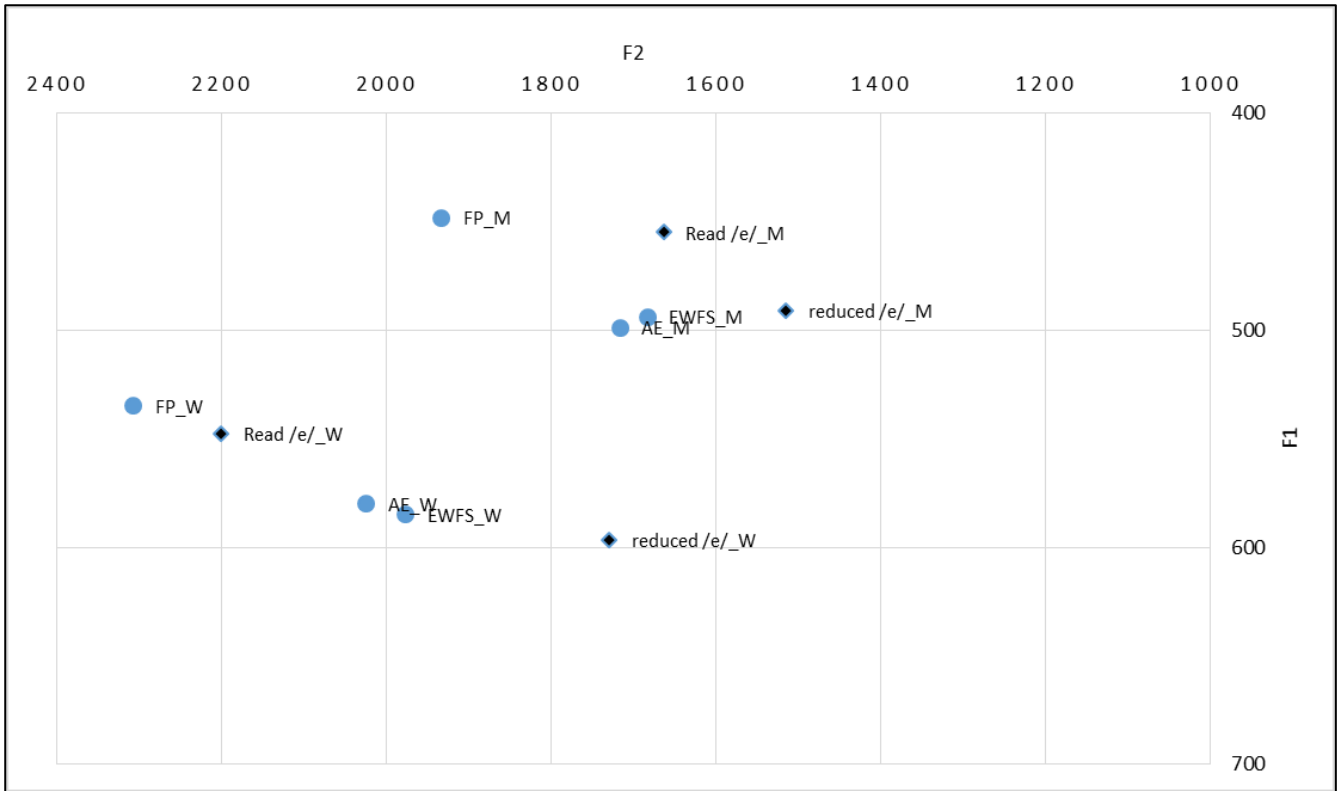
FPs in Hebrew are less central with comparison to reduced or fluent speech and to the two other types of HD (AE and EWFS), i.e., they are pronounced with a more closed manner (lower F1) and with more frontal manner (higher F2), thus becoming closer to the pronunciation of Hebrew [i] vowel. We can thus suggest that filled pauses may be indicated by the IPA symbol for closed-mid

vowel [e] while elongated syllables (EWFSs) and appended [e] (AEs) may be indicated by the open-mid symbol [ɛ], or to use diacritics indicating lowered [ę] (for FP) and raised [ę̞] (for EWFSs and AEs). Although these variations are still not investigated perceptually, in the present research we observe a potential for a diachronic shift from length to quality in at least part of the vowel system. By this, the present research contributes to the understanding of the historic rise and fall of distinctive vowels, as suggested by [1: 268].

Table 1: Intrinsic comparison of F1, F2, and F3 in the three types of HDs, for the whole corpus (lines 1-6) and for main speakers sub-corpus (lines 7-12)

	Formants	Gender	EWFS		AE		FP	
			Mean	SD	Mean	SD	Mean	SD
1	F1 (Hz)	M	494	85	499	88	448	51
2		W	585	88	580	88	535	85
3	F2 (Hz)	M	1683	246	1716	245	1933	284
4		W	1977	233	2024	283	2308	427
5	F3 (Hz)	M	2818	400	2853	383	2928	488
6		W	3025	286	3013	321	3307	543
7 (>20 HDs)	F1 (Hz)	M	466	53	492	92	452	72
8 (>20 HDs)		W	580	61	573	76	525	79
9 (>20 HDs)	F2 (Hz)	M	1655	279	1706	252	2020	342
10 (>20 HDs)		W	2005	243	2093	327	2365	433
11 (>20 HDs)	F3 (Hz)	M	2825	450	2905	383	-	-
12 (>20 HDs)		W	3037	245	3075	337	3299	514

Figure 2: Extrinsic comparison of F1 and F2 in the three types of HDs and in two previous works on fluent /e/ in lexical words ("reduced /e/" and "read /e/").



REFERENCES

[1] Abramson, A. S., Ren, N. 1990. Distinctive vowel length: duration vs. spectrum in Thai. *J. of Phonetics* 18(1), 79–92.

[2] Amir, N., Amir, O. 2007. Novel measures for vowel reduction. In *Proc. of the 16th International Congress of Phonetic Sciences*, 849–852.

[3] Bolozky, S. 1999. On the special status of the vowels *a* and *e* in Israeli Hebrew. *Hebrew Studies* 40, 233–250.

[4] Bolozky, S. 1997. Israeli Hebrew Phonology. In Kaye, A. S. & P. Daniels (eds.), *Phonologies of Asia and Africa 1*. 287–311. Winona lake: Eisenbauns.

[5] Candeias, S., Celorico, D., Proença, J., Veiga, A., Perdigão, F. 2013. HESITA(tions) in Portuguese: a database. *Proceedings of Disfluency in Spontaneous Speech, DiSS 2013, TMH-QPSR* 13, 13–16.

[6] CoSIH: The Corpus of Spoken Israeli Hebrew. Available at: http://humanities.tau.ac.il/~cosih.

[7] Cruttenden, A. 1997. *Intonation*. Cambridge: Cambridge University Press.

[8] Fox, A. 2000. *Prosodic features and prosodic structure: The phonology of suprasegmentals*. New York: Oxford University Press.

[9] IPA Chart. 2005. Available under a Creative Commons Attribution-Sharealike 3.0 Unported License. Copyright © 2005 International Phonetic Association. http://www.internationalphoneticassociation.org/content/ipa-vowels,

[10] Most, T., Amir, O., Tobin, Y. 2000. The Hebrew vowel system: Raw and normalized acoustic data. *Language and Speech* 43(3), 295–308.

[11] Nemoto, R. 2013. Which kind of hesitations can be found in Estonian spontaneous speech?. *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech, DiSS 2013, TMH-QPSR* 54(1), 53–54, Stockholm.

[12] Proença, J., Celorico, D., Veiga, A., Candeias, S., Perdigão, F. 2013. Acoustical characterization of vocalic fillers in European Portuguese. *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech, DiSS 2013, TMH-QPSR* 54(1), 63–66, Stockholm.

[13] Silber-Varod, V. 2010. Phonological aspects of hesitation disfluencies. *Speech Prosody 2010, The Fifth International Conference on Speech Prosody*, 11–14 May 2010, Chicago, Illinois.

[14] Silber-Varod, Vered. 2013. *The SpeeCHain Perspective: Form and Function of Prosodic Boundary Tones in Spontaneous Spoken Hebrew*. LAP Lambert Academic Publishing.

[15] Shriberg, E. 2001. To 'errrr'is human: ecology and acoustics of speech disfluencies. *J. of the International Phonetic Association* 31(01), 153–169.

[16] Shriberg, E. 1994. Preliminaries to a theory of speech disfluencies. Doctoral dissertation, Department of Psychology, University of California, Berkeley.

INVESTIGATING DISFLUENCY IN RECORDINGS OF LAST SPEAKERS OF ENDANGERED AUSTRONESIAN LANGUAGES IN TAIWAN

Jozsef Szakos[1] and Ulrike Glavitsch[2]

[1]The Hong Kong Polytechnic University
[2]EMPA, Swiss Federal Laboratories for Materials Science and Technology, Switzerland

ABSTRACT

The nearly three decades spent in Formosan language documentation produced hundreds of hours of recorded speech. In this paper, we show how the use of *SpeechIndexer* for transcribing and indexing the data visualises the problem of disfluency in the spontaneous narratives and dialogues. The semi-automatic alignment of speech and transcription needs to be adjusted manually each time when unpredictable pauses occur which are disfluencies, rather than markers of phrasal units. It is illustrated how the combination of *SpeechIndexer*'s pause finder with pitch measurements can help to pinpoint the difference of phrasal boundaries and pauses of disfluency.

Keywords: Austronesian, lesser-documented unwritten language, SpeechIndexer, pause finder.

1. INTRODUCTION

The languages of Tsou, Saaroa, Kanakanavu are the oldest vestiges of Austronesian languages in Taiwan. The natural speech of elderly is not influenced by any writing system.

We developed *SpeechIndexer* to help document the spontaneous narratives, dialogues of the last remaining fluent speakers [1]. *SpeechIndexer* is a semi-automatic indexing software, which suggests a segmentation of speech data on the basis of pauses/breaks, then links the phonetic units with the transcribed data [2]. *SpeechIndexer*'s built-in pause finder automatically partitions the voice recording into speech and pause segments. The pause finder is an extension of a basic utterance endpoint detection algorithm [3]. Most recently, a fundamental frequency (F0) estimation algorithm has been included in *SpeechIndexer* as well. In addition, a parallel text window is opened to enter a translation or comments.

The assumption in developing *SpeechIndexer* was that natural pauses appear between phrases and sentences. So it is easy to follow the suggested segmentation and link it to the transcription.

In reality, there are many more interruptions of the speech flow, and mostly these are the disfluencies not tackled by the software.

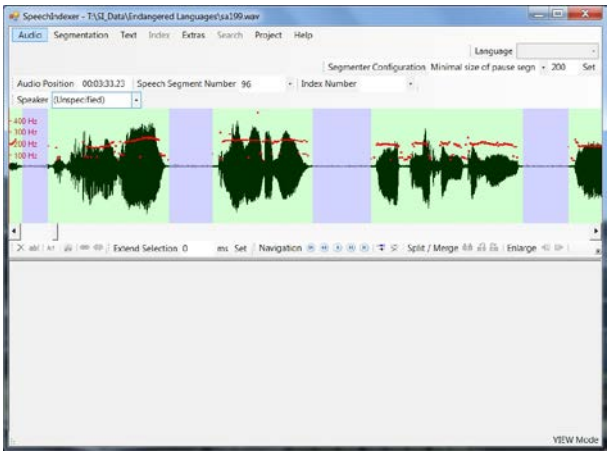
The data for this study have been selected out of 1'800 hours of recordings completed over the past three decades. The speakers were elderly male, representing the oldest branches of the Austronesian language family in the central mountain range of Taiwan. The languages are critically endangered, but important for the understanding of the other 1260 members of this language family extending over the Pacific.

This version of *SpeechIndexer* has been especially programmed to cope with the needs of the present investigation.

2. SPEECHINDEXER WITH F0 DETECTION

The F0 curve is computed on the spoken segments of the speech recordings only. As mentioned in Sec. 1, *SpeechIndexer* finds speech and pause segments of a voice recording automatically. The F0 detection algorithm is a simplified version of a cognition-oriented fundamental frequency estimation algorithm [4]. The algorithm has been rewritten to run on multiple processors for efficiency reasons. Fig. 1 shows the main window of *SpeechIndexer* with the signal window in the upper part. The speech segments are marked as light green sections with the F0 curve as dotted red line and the pause segments in bluish colour. (The speaker is an elderly male person, he speaks Saaroa.)

Figure 1: Spoken segments (green), pause segments (bluish) and F0 curve (red dotted line) of a voice recording shown with *SpeechIndexer*.



3. TYPES OF DISFLUENCIES

The indexed speech recordings were analysed for disfluencies in the flow of speech. A label was inserted into the text and linked with the recording as soon as a natural break was found.

We noticed different types of disfluencies that are listed in Table 1. They either occur alone or in combination with one another. For instance, a repetition often occurs together with a self-correction.

Table 1: Types of disfluencies encountered in narratives of the Tsou and Saaroa languages.

disfluency type	label
hesitation	H
repetition	R
pause	P
self-correction	SC, C
insertion	INSERTION

There are different ways to insert these markings into the text and link it with the corresponding phonetic unit. The easiest cases are voiced hesitations and pauses. Voiced hesitations are sounds sometimes similar to a "schwa". Such hesitations may occur at the beginning, in the middle or at the end of phrases but rarely in the middle of a word. In this case, the letter H can be inserted at the corresponding position in the text and appropriately linked with the phonetic unit. Pauses are treated similarly. Self-corrections and repetitions may occur in the middle of a word. In these cases, the labels are inserted either at the exact text position where the disfluency occurs or at the beginning or end of the word that is corrected or repeated. This is illustrated in the following example.

A hesitation combined with a self-correction occurs in the word "maahlüvürahla" in the narrative of an elderly Saaroa speaker after the sixth character position. Thus, this disfluency is labelled either as "maahlüCRvürahla" or "maahlüvürahla CR". The former has the advantage that the exact text position of the break is clearly recognizable, however, the original spelling of the word is lost and, hence, the word is not found by *SpeechIndexer*'s built-in search function. In the latter, the spelling remains intact, but it is not clear where the disfluency actually occurs.

Moreover, there are two different ways in which the word/phrase with the disfluency labels is linked to the corresponding speech segment. One possibility is to link the phrase and the disfluency labels as a whole with the spoken segment or only link the labels with the speech section. We leave it up to the researcher which type of labelling and which way of linking is preferred. In Sec. 4, we show examples with different types of labelling and linking.

In addition, researchers may create different text files and link them to the same speech recording with *SpeechIndexer*. The original text is linked with the voice recording in a first step. The result is a text file and an index file. Then, the researcher inserts the labels for the different types of disfluencies in the original text file and saves it under a new name. The new text file is linked with the original speech recording resulting in a new index file. This way, different text and index files are created to describe different types of linguistic phenomena occurring in the same speech recording.

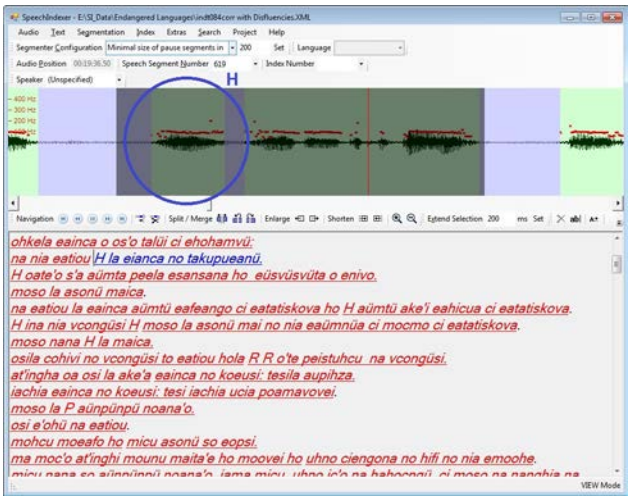
In the following, we will show concrete examples of the various disfluency types selected from a large collection of voice recordings of both Tsou and Saaroa.

4. EXAMPLES

4.1. Hesitations

Hesitations frequently occur at the beginning or end of a phrase if the speaker is not yet warmed up with talking. Figure 2 shows such a hesitation at the beginning of a narrative told by an elderly male Tsou speaker. The fundamental frequency (F0) curve in the hesitation is quite flat, i.e. the speech flow is not disturbed in any way. Hesitations are also characterized by interrupted words, thus short periods of silence. Less energetic speech segments usually appear combined with self-corrections, but not yet identifiable from the segments.

Figure 2: Hesitation at the beginning of a phrase, the speech segment is highlighted and the phonetic unit for the hesitation (H) circled in blue. The phrase in the text is highlighted in dark blue as well.

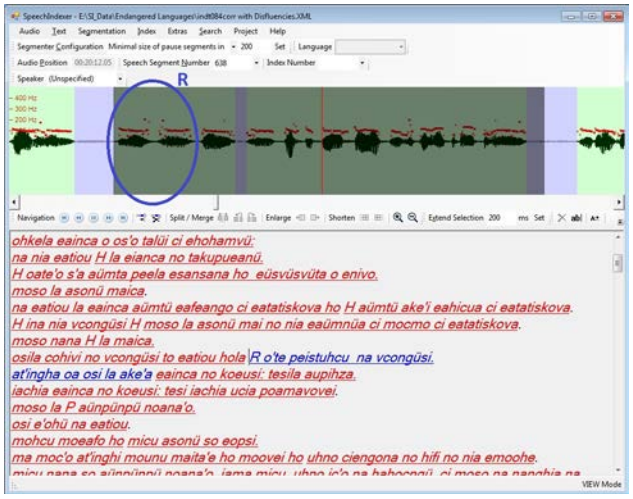


4.2. Repetitions

Repetitions occur quite frequently when the speaker wants to assure himself of something. Fig. 3 shows a double repetition of the last word of the previous phrase at the beginning of the current highlighted phrase. It is the same narrative as in Fig. 2 spoken by a male person in the Tsou language. The F0 curve in the circled repeated section of speech is flat again.

By the way, the most frequently used filler is *yainca* "that is to say" in the Tsou language. Speakers may insert it almost any point into the sentence, preceding verbs, nouns, etc. There is a significant personal style and variation in this.

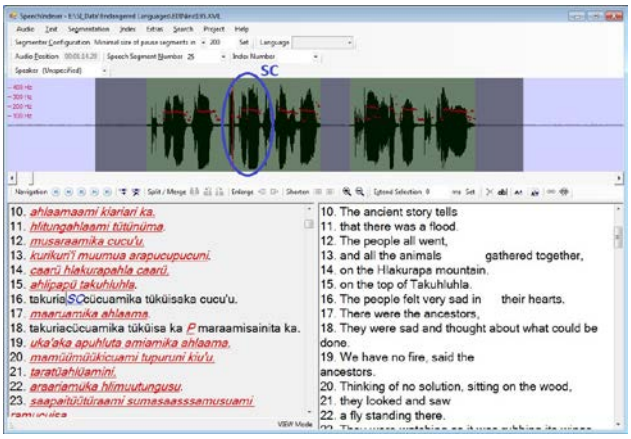
Figure 3: Repetition (R) at beginning of phrase. The word "hola" from the end of the previous word is repeated twice in the circled speech segment.



4.3. Self-corrections

Self-corrections take place on morpheme boundaries. Sentence-initial particles, tense and aspect markers (TAM particles) are especially prone to hesitation, correction and repetition. The speaker will always try to pronounce something complete, be they syllables or intonation units (curves on the screen). Self-corrections can be recognized by a sudden pause, followed by a more energetic word. Fig. 4 shows such a self-correction in the narrative of an elderly Saaroa speaker. Here, only the label SC is linked to the spoken phrase in which the self-correction occurs. The pause before the self-corrected part is clearly identifiable and also the following word with a high energy. The F0 curve in the self-corrected portion of speech shows more movements up and down than the sections before and after. The parallel text window shows the translation of the narrative.

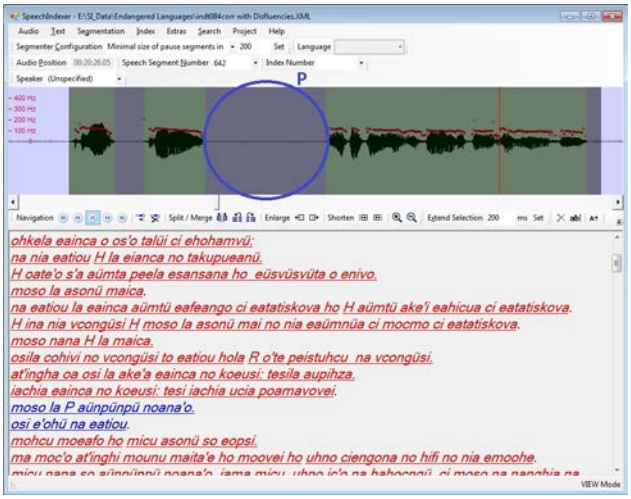
Figure 4: Self-correction (SC), the speaker breaks in the first part of the speech segment to start again with a highly energetic word.



4.4. Rhetorical break in phonation (pause)

Occasionally, speakers pause for some time before they continue talking. Most often, these are intentional interruptions of important words, however the 200 to 600 ms of silence is only used to emphasize the speaker's emotion. Fig. 5 shows a roughly 300 ms pause in the middle of a phrase of a Tsou narrative. The F0 curve clearly falls before the pause.

Figure 5: Significant pause (more than 300 ms) in the middle of a phrase.

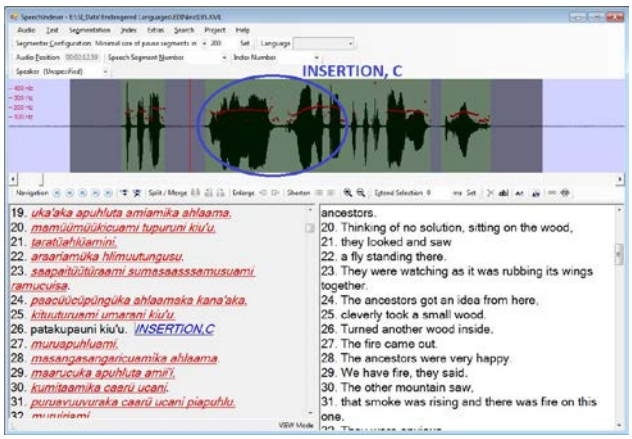


4.5. Spontaneous insertion

It occasionally happens that speakers make a spontaneous insertion. Fig. 6 shows such a case where the insertion is combined with a self-correction. The insertion is preceded by a longer

pause. It is the same Saaroa narrative as in Sec. 4.3. The F0 curve in the inserted portion of speech shows some movement upwards before the regular speech flow continues.

Figure 6: Insertion combined with self-correction in a narrative of an elderly Saaroa speaker.



5. CONCLUSIONS

Normal speech in narratives is characterized by falling intonation at the end of phrases. It is a combination of falling intonation and pause, which shows regular speech. Indications of disfluencies may be an abrupt change in F0, usually going high or showing up and down directions. We can categorize the visible disfluencies in this way, however, there remain other cases which require more data and semi-automatic ways of analysis to identify and differentiate them from normal speech flow. *SpeechIndexer* can in many cases show up disfluencies, but we need to find other factors, too.

6. ACKNOWLEDGEMENTS

This research was partially funded by Grant Number GRS-027/13 of the Gebert Rüf Stiftung in Switzerland.

7. REFERENCES

[1] Szakos, J., Glavitsch, U. 2004. Seamless Speech Indexing and Retrieval: Developing a New Technology for the Documentation and Teaching of Endangered Formosan Aboriginal Languages. Proc. EISTA'14 – International Conference on Education and Information Systems: Technology and Applications.

[2] Glavitsch, U., Simon, K., Szakos J. 2011. SpeechIndexer – A Flexible Software for Audio-Visual Language Learning. Proc. icEIC - International Conference on Education, Informatics and Cybernetics, 79 – 82.

[3] Rabiner, L. M., Sambur, M. R. 1975. An Algorithm for Detecting the Endpoints of Isolated Utterances. The Bell System Technical Journal, vol. 54, 297 – 315.

[4] Glavitsch, U., Simon, K. 2013. A Cognition-Oriented Approach to Fundamental Frequency Estimation. Proc. NLPCS, 105 – 116.

RECOGNIZING EMOTIONS IN DIALOGUES WITH DISFLUENCIES AND NON-VERBAL VOCALISATIONS

Leimin Tian, Catherine Lai, Johanna D. Moore

School of Informatics, University of Edinburgh
s1219694@sms.ed.ac.uk, clai@inf.ed.ac.uk, j.moore@ed.ac.uk

ABSTRACT

We investigate the usefulness of DISfluencies and Non-verbal Vocalisations (DIS-NV) for recognizing human emotions in dialogues. The proposed features measure filled pauses, fillers, stutters, laughter, and breath in utterances. The predictiveness of DIS-NV features is compared with lexical features and state-of-the-art low-level acoustic features.

Our experimental results show that using DIS-NV features alone is not as predictive as using lexical or acoustic features. However, adding them to lexical or acoustic feature set yields improvement compared to using lexical or acoustic features alone. This indicates that disfluencies and non-verbal vocalisations provide useful information overlooked by the other two types of features for emotion recognition.

Keywords: emotion recognition, dialogue, disfluency, speech processing, HCI

1. INTRODUCTION

Emotions are vital in human cognitive processes. Emotion recognition has long been a focus in human-computer interaction research. State-of-the-art approaches for improving performance of emotion recognition often focus on identifying better feature representations. In this work, our goal is to identify knowledge-driven features that can improve recognition performance.

Psycholinguistic studies have shown that emotions can influence the neural mechanisms in the brain, and thus influence sensory processing and attention [9]. This in turn influences speech processing and production, which may result in disfluencies and non-verbal vocalisations. Therefore, we would like to investigate the usefulness of DISfluencies and Non-verbal Vocalisations (DIS-NV) for recognizing emotions in dialogues.

One of the most predictive feature sets identified for emotion recognition is the set of acoustic features based on low-level descriptors (LLD). However, in our previous work [7] on the AVEC2012 database [8] of spontaneous dialogues, DIS-NV features were more predictive than acoustic or lexical features for recognizing emotions. We would like to study whether our DIS-NV features remain predictive when the data contains both non-scripted and scripted dialogues. Therefore, we compare our DIS-NV features with LLD acoustic features and lexical features on the IEMOCAP database [1]. Our results show that although DIS-NV features are less predictive than acoustic or lexical features when used alone, they improve performance when combined with existing models.

2. METHOD

2.1. The IEMOCAP Database

The IEMOCAP database contains approximately 12 hours of audio-visual recordings from 5 mixed gender pairs of actors. Each conversation was about 5 minutes long. There are 10037 utterances in total, of which 4782 utterances were not scripted. When collecting the non-scripted dialogues, the actors were instructed to act out emotionally intense scenarios, e.g., telling a best friend that (s)he has been accepted into his/her most desired university.

Emotions were annotated at the utterance-level with a 1 to 5 integer score of the Arousal (activeness), Power (domination), and Valence (positive or negative) emotion dimensions. The mean score over all the annotations was used when the annotators disagreed with each other. We categorized the scores into three classes (<3, =3, >3) to have a clearer view of the relation between emotions and features, and to reduce the influence of imbalanced classes.

2.2. Features

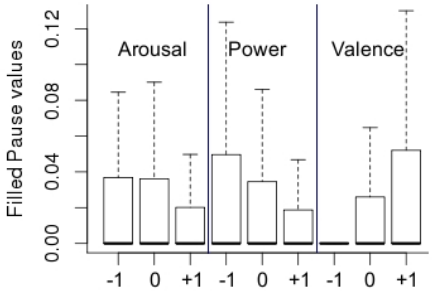
2.2.1. The DIS-NV Features

We studied 5 types of disfluencies and non-verbal vocalisations (DIS-NV): filled pauses (non-verbal insertions, e.g., “eh”), fillers (verbal insertions, e.g., “you know”), stutters, laughter, and breath. We choose them because they are the most common in the data, and they are relatively easy to extract from

transcripts. Disfluencies here refer to interruptions in the flow of speech production. Fluency of speech production may not always be the same with listener's perception of fluency [6]: Minor disfluencies may be ignored by the listener; In some cases, these tokens could also be perceived as part of a "fluent" utterance (e.g., using a filler at the beginning of an utterance while organizing sentences).

Feature values are calculated as the ratio between the sum duration of each type of DIS-NV and the total duration of the utterance, resulting in 5 DIS-NV features for each utterance. Descriptive statistics of filled pause features are shown in Figure 1 as an example. Utterances containing DIS-NVs are not very frequent in the IEMOCAP database (47.28% in the non-scripted utterances, 24.74% in the scripted utterances). To get a clearer view of value distributions, the statistics shown were computed on a subset of the data which contains all the utterances with disfluencies or non-verbal vocalisations (the DIS-NV subset).

Figure 1: Statistics of filled pause features.



2.2.2. The Lexical Features

The lexical features we extracted are 6 Point-wise Mutual Information (PMI) based features. PMI is a widely used measurement for the relation of words and emotions. It is based on the frequency of a word w having class label c, as shown: $PMI(c,w) = log_2(\frac{P(c|w)}{P(c)})$

To calculate PMI values, we first binarized all three emotion dimensions (<3, $\geq$3). PMI values of the scripted and non-scripted data are computed separately. The lexical features we proposed are calculated as the total PMI values of all the words in an utterance for each binarized emotion dimension, resulting in 6 lexical features for each utterance.

Example words with top PMI values are shown in Table 1. In the first column, "A-" represents unaroused, "A+" is excited, "P-" is dominated, "P+" is dominating, "V-" and "V+" represent negativeness and positiveness of emotion.

Table 1: Words with top PMI values.

	Non-scripted data
A-	Academy, Banking, Loan, Numb, Sleep
A+	Anger, Bloody, Flowers, Freak, Ruined
P-	Afraid, Beer, Error, Insane, Quit
P+	Bar, Chick, Duty, F*ck, Mad
V-	Abuse, B*tch, Die, Iraq, Unfair
V+	Australia, Cash, Dog, Snow, Tour
	Scripted data
A-	Lose, Non, Pets, Skip, Topic
A+	Bully, Cry, Gods, Jesus, Santa
P-	Bad, Cliff, Sacrifice, Sneak, Surprise
P+	Cry, Damn, Lose, Mad, Shut
V-	Ashamed, Crap, Hell, Sucker, Vile
V+	Delight, Eating, Gold, Loves, Wish

2.2.3. The LLD Acoustic Features

Our LLD acoustic features were the same as those used in the INTERSPEECH 2010 Paralinguistic Challenge extracted with OpenSMILE [3]. It represents a state-of-the-art feature set for emotion recognition. This feature set has been widely used as a reference for comparing emotion recognition feature sets and classification approaches.

There are 1582 LLD acoustic features, including those extracted by applying functionals (e.g., position of max) to low-level descriptors (e.g., MFCCs, F0, PCM loudness) and their corresponding delta coefficients, the number of pitch onsets, and the total duration of the utterance. Values are computed at the frame-level, with a window size of 60ms and a step of 10ms. Compared to DIS-NV and lexical features, LLD acoustic features overlook global characteristics of the utterance.

2.3. Experimental Settings

Our emotion recognition models were built with the LibSVM [2] classifier using WEKA [4]. We used the C-SVC approach with RBF kernel, and 10-fold cross validation. All features were normalized to [-1,1] before classification. Because of the imbalanced classes, we use weighted F-measure as the evaluation metric.

3. RESULTS AND DISCUSSION

The performance of different feature sets is shown in Table 2. "Mean" in the first row is the un-weighted average of the three emotion dimensions. In the first column, "DN" is the DIS-NV model, "PMI" is the lexical model, "LLD" is the LLD acoustic model.

Our results show that adding DIS-NV features to

Table 2: Performance on the full database.

Models	Arousal	Power	Valence	Mean
DN	0.363	0.407	0.328	0.366
PMI	0.483	0.483	0.332	**0.433**
PMI+DN	0.489	0.486	0.406	**0.460**
LLD	0.652	0.538	0.535	0.575

lexical feature set yields improvement on all emotion dimensions. This verified that DIS-NV features capture information neglected by the lexical content, thus helping with emotion recognition.

When used alone, DIS-NV features are less predictive than lexical or LLD acoustic features, which is different from our previous work. This may be caused by the different nature of the AVEC2012 and IEMOCAP database. Compared to the AVEC2012 database of spontaneous dialogues, disfluencies and non-verbal vocalisations are less frequent in the IEMOCAP database of acted data. To reduce such influence, we also performed experiments on the DIS-NV subset, as shown in Table 3.

Table 3: Performance on the DIS-NV subset.

Models	Arousal	Power	Valence	Mean
DN	0.470	0.453	0.329	0.417
PMI	0.500	0.467	0.316	**0.428**
PMI+DN	0.522	0.475	0.325	**0.441**
LLD	0.644	0.523	0.532	0.566
LLD+DN	0.645	0.525	0.533	0.568

Compared to using the full IEMOCAP database, when using this subset instead, performance of lexical features and LLD acoustic features has a small decrease, while performance of DIS-NV features increases greatly on all emotion dimensions. This verified the negative influence of infrequency of disfluencies and non-verbal vocalisations.

Adding DIS-NV features to lexical feature set remains helpful for all emotion dimensions. Adding DIS-NV features to LLD acoustic features only yields a small gain. The reason may be the great difference between the size of these two feature sets.

We further compared performance of individual DIS-NV features and LLD features with the CFS [5] method, which ranks features based on their individual predictiveness and their correlations with other features. DIS-NV features are always ranked among the top features, especially filled pauses, fillers, and laughter. This indicates that with a better fusion strategy, DIS-NV features may improve performance of LLD features greatly, by highlighting emotionally interesting segments.

Note that DIS-NV and lexical features describe data at the utterance-level, while LLD features describe data at the frame-level. In the future, with advanced fusion strategy that can combine feature sets at different levels with flexible weights, we may be able to combine information contained in these feature sets more efficiently and further boost performance of current emotion recognition models.

4. CONCLUSION

We proposed DIS-NV features measuring disfluencies and non-verbal vocalisations for recognizing emotions in dialogues. We compared their performance with lexical features and state-of-the-art LLD acoustic features. Our experiments on the IEMOCAP database show that using DIS-NV features alone is not enough for building a highly predictive emotion recognition model. However, these features contain information neglected by the lexical or LLD acoustic features. Thus, when fused properly, DIS-NV features may improve performance of current emotion recognition models greatly.

5. REFERENCES

[1] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42(4), 335–359.

[2] Chang, C.-C., Lin, C.-J. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3), 27.

[3] Eyben, F., Wöllmer, M., Schuller, B. 2010. OpenSMILE: the munich versatile and fast open-source audio feature extractor. *Proceedings of the international conference on Multimedia*. ACM 1459–1462.

[4] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter* 11(1), 10–18.

[5] Hall, M. A. 1998. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis University of Waikato Hamilton, New Zealand.

[6] Lickley, R. 2015. Fluency and disfluency. to appear.

[7] Moore, J., Tian, L., Lai, C. 2014. Word-level emotion recognition using high-level features. *Computational Linguistics and Intelligent Text Processing*. Springer 17–31.

[8] Schuller, B., Valster, M., Eyben, F., Cowie, R., Pantic, M. 2012. AVEC 2012: the continuous audio/visual emotion challenge. *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM 449–456.

[9] Vuilleumier, P. 2005. How brains beware: neural mechanisms of emotional attention. *Trends in cognitive sciences* 9(12), 585–594.

A LATTICE-BASED APPROACH TO AUTOMATIC FILLED PAUSE INSERTION

Marcus Tomalin[1], Mirjam Wester[2], Rasmus Dall[2], Bill Byrne[1], & Simon King[2]

[1]Cambridge University Engineering Department, University of Cambridge, UK

[2]The Centre for Speech Technology Research, University of Edinburgh, UK
mt126@cam.ac.uk, r.dall@sms.ed.ac.uk, mwester@inf.ed.ac.uk, wjb31@cam.ac.uk, Simon.King@ed.ac.uk

ABSTRACT

This paper describes a novel method for automatically inserting filled pauses (e.g., *UM*) into fluent texts. Although filled pauses are known to serve a wide range of psychological and structural functions in conversational speech, they have not traditionally been modelled overtly by state-of-the-art speech synthesis systems. However, several recent systems have started to model disfluencies specifically, and so there is an increasing need to create disfluent speech synthesis input by automatically inserting filled pauses into otherwise fluent text. The approach presented here interpolates Ngrams and Full-Output Recurrent Neural Network Language Models (f-RNNLMs) in a lattice-rescoring framework. It is shown that the interpolated system outperforms separate Ngram and f-RNNLM systems, where performance is analysed using the Precision, Recall, and F-score metrics.

Keywords: Disfluency, Filled Pauses, f-RNNLMs, Ngrams, Lattices

1. INTRODUCTION

In recent years, disfluent speech synthesis has started to receive more attention [1, 2, 3, 13]. The aim is to develop systems that produce convincing disfluencies such as filled pauses (FPs), discourse markers, repetitions, and restarts. It is well-known that such phenomena serve a wide range of important functions in conversational discourse. They can indicate psychological states [11], structure discourses [9], facilitate word recall [14], and improve word recognition [15, 10, 12]. Given this, it is desirable to model them overtly in automatic speech synthesis systems which seek to approximate a human-like conversational style.

The broad motivations underlying research into disfluent synthesis are closely related to those that have prompted the development of emotional or expressive speech synthesis systems [20, 19, 17, 4, 5]. Both endeavours ultimately seek to create synthetic speech that is able to convey a wider range of emotional or psychological states, thereby producing synthetic voices that can simulate certain character and personality types more convincingly. The main difference, however, is that while emotional or expressive speech synthesis concentrates primarily on modifying prosodic phenomena (such as pitch, speech rate, voice quality, inter-lexical pause duration) [20, 19, 17, 4, 5], disfluent speech synthesis additionally requires the augmentation of the (fluent) input token sequence [13].

This paper contributes to this ongoing endeavour by extending the basic approach to automatic FP-insertion introduced in [13]. That paper focused on the relatively simple task of inserting a single FP (*UH*) into a fluent token sequence at an appropriate Insertion Point (IP). By contrast, the current paper describes a system that can insert *multiple* FPs in *multiple* IPs. Therefore the sentence 'I NEVER LIKED GAMES' could be modified automatically to become '**UM** I NEVER LIKED **UH** GAMES'. In addition, the new system has a Disfluency Parameter (DP) that determines the degree of disfluency in the output text. The DP takes a value in the range $[0,1]$, where 0 = maximally fluent and 1 = maximally disfluent. Finally, while [13] used simple linear interpolation of word-level Ngram and RNNLM probabilities to rerank the potential sentences, a more robust lattice-based rescoring method is introduced here. As a modelling technique, it has clear advantages since simple re-ranking strategies become computationally inefficient when multiple FPs can be inserted in multiple IPs.

The structure of this paper is as follows. Section 2 describes the lattice-based modelling framework, and provides information about the training and test data used. Section 3 gives the results for the various FP-insertion systems compared using the Precision, Recall, and F-score metrics. Scores are given at the sentence level for output containing all the inserted FPs, along with breakdowns for each separate

FP subtype. The main conclusions and directions for future research are summarised in 4.

2. LATTICE-BASED LM INTERPOLATION

The lattice-based FP-insertion system developed here is similar to those recently implemented for Automatic Speech Recognition (ASR) tasks in [7, 18]. In the context of ASR language models (LMs), RNNLMs have become increasingly popular in recent years due to their inherently strong generalization performance. Specifically, Chen et al 2014 [7] has shown that f-RNNLMs facilitate an efficient parallisation of training in a Graphics Processing Unit (GPU) implementation. In addition, when used in a lattice-rescoring framework, they give both Perplexity and Word Error Rate improvements over standard RNNLMs. This is due in part to their use of an unclustered 'Full-Output' architecture.

This framework can be adapted for the FP-insertion task. There are five main stages in the modified process:

1. Create initial lattices in which each FP is accessible from each word token (Figure 1)
2. Expand the initial lattices using an Ngram (6g)
3. Rescore the expanded lattices using an interpolated LM with weighted Ngram and f-RNNLM sub-components
4. Output an *n*-best list for each sentence (where $n = 10000$)
5. Specify the desired degree of disfluency using the DP and generate final 1-best disfluent output

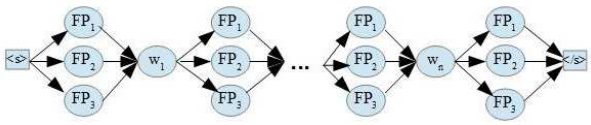


Figure 1: Example Initial Lattice for *n* words and 3 FPs

After FP-insertion, the versions of sentence S in the n-best list will have varying token counts since they will contain different numbers of automatically inserted FPs. All versions of S with p tokens are rank-ordered using the sentence-level interpolated LM score, and the 1-best version is output. The closed interval $[0,1]$ is divided equally between the various 1-best outputs for different p values. This provides the DP that determines the degree of disfluency. The impact of varying the DP parameter is shown in Table 1. This example provides a concrete instance of the impact that the DP value has on the resulting token sequences, and it illustrates the graded nature of the different DIS-insertion outputs. In particular, it shows how the perceived psychological state of the (synthetic) speaker can be altered as a fluent lexical sequence becomes increasingly disfluent.

DP	Output Sentence
0.00	WELL I GUESS THEY WERE SAYING
0.25	WELL I GUESS THEY WERE SAYING **UM**
0.50	**UM** WELL **UH** I GUESS THEY WERE SAYING **UM**
0.75	**UM** WELL **UH** I GUESS **HM** THEY WERE SAYING **UM**
1.00	**UM** WELL **UH** I GUESS **HM** THEY **UH** WERE SAYING **UM**

Table 1: An example of the impact of DP values on output disfluent token sequence

The LMs used in the experiments were trained on 20M words (1M sentences) of data from the Switchboard, Fisher, and AMI corpora, as well as an unreleased corpus of British conversational telephone speech [16, 8, 6]. Dev and Test sets were extracted from different subsets of the same corpora, and they comprised 7,365 sentences (145k words) and 6,910 sentences (139K words) respectively. Each sentence in the scoring reference contained at least one FP, and these FPs were removed to create the 'fluent' version of the test sets that were processed by the FP-insertion systems. The purpose of the experiments was to see whether the systems could insert the same FPs into the same IPs as those found in the scoring reference files. Seven FPs in total were modelled overtly by the various FP-insertion systems: *UH*, *UM*, *OH*, *UHUM*, *UHU*, *HM*, and *AH*. Information about the occerrence of these FPs in the training data is given in Table 2.

	#occs [%]
UH	213,924 [1.09%]
UM	200,499 [1.02%]
OH	123,028 [0.63%]
AH	69,288 [0.35%]
UHUM	29,515 [0.15%]
UHU	16,180 [0.08%]
HM	3,456 [0.01%]

Table 2: FP occurrence counts for the training data (and % of training data)

As the counts in Table 2 indicate, the FPs *UH* and *UM* occur most frequently in the training data. The fact that some of the other FPs have relatively low counts (<30,000) facilitates the exploration of the impact of data sparcity on the modelling of speech disfluencies.

3. EXPERIMENTS AND RESULTS

Three FP-insertion systems were compared:

1. **Ngram**: a standard 6g built using the training data; SRILM toolkit [21]; K-N discounting
2. **f-RNNLM**: a non-class-based f-RNNLM with

512 hidden layer nodes

3. **Ngram+f-RNNLM**: the 6g and f-RNNLM are interpolated with a 50%-50% weighting in the lattice-based framework described in section 2

The initial lattices (Figure 1) were expanded and rescored using the Ngram, the f-RNNLM, and the interpolated Ngram+f-RNNLM LMs. System performance was evaluated using standard Precision, Recall, and F-score metrics. The full range of sub-component weightings was explored for the Ngram+f-RNNLM system (e.g., 40%-60%, 60%-40%), but the 50%-50% weighting gave the optimal performance (as determined by the three metrics). Consequently, the 50%-50% weighting was adopted for all the experiments reported in this paper. The metric scores were also used to determine the optimal DP value for the Dev data and Figure 2 shows the metric scores for the Ngram+f-RNNLM system. The inverse relationship between Precision and Recall/F-score is apparent, and a DP value of 0.5 achieves a desirable balance between these extremes. Similar patterns were obtained for all three systems, so the DP was set to 0.5 for all subsequent experiments.

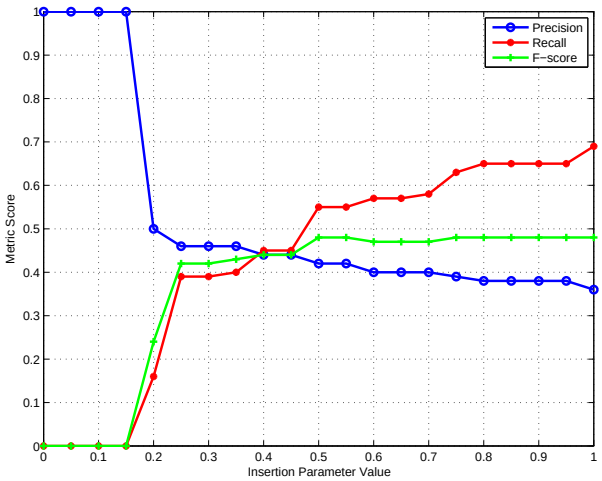


Figure 2: Precison, Recall, and F-score for Ngram+f-RNNLM for Different DP Values

	Precision (Dev/Test)		Recall (Dev/Test)		F-score (Dev/Test)	
Ngram	0.41	0.44	**0.55**	**0.60**	0.47	0.51
f-RNNLM	0.42	**0.47**	0.52	0.57	0.47	0.51
Ngram+f-RNNLM	**0.43**	**0.47**	**0.55**	0.59	**0.48**	**0.52**

Table 3: Dev and Test Sentence-level results for the Ngram, f-RNNLM, and Ngram+f-RNNLM systems

Table 3 shows that the Ngram+f-RNNLM system obtained the best (sometimes joint best) sentence-level Precision and Recall performance for every case except the Recall results for the Test set. Notably, the Ngram+f-RNNLM system obtained the best F-score results for both the Dev and Test sets. This suggests that the interpolated system combines the complementary properties of the two component LMs. Consequently, the Ngram+f-RNNLM system is comparatively more robust than either the Ngram or f-RNNLM systems, and the latter two are beneficially interpolated in the lattice-based framework.

	Dev #occs [%] (ref/hyp)		Test #occs [%] (ref/hyp)	
UH	3660 [2.29%]	6359 [3.97%]	3658 [2.44%]	6311 [4.10%]
UM	3331 [2.09%]	4711 [2.94%]	3392 [2.26%]	4201 [2.73%]
OH	2035 [1.27%]	3685 [2.30%]	2083 [1.39%]	3538 [2.30%]
AH	1053 [0.66%]	192 [0.12%]	348 [0.23%]	209 [0.14%]
UHUM	432 [0.27%]	73 [0.46%]	423 [0.28%]	92 [0.06%]
UHU	222 [0.14%]	8 [0.00%]	228 [0.15%]	23 [0.01%]
HM	61 [0.04%]	2 [0.00%]	55 [0.04%]	0 [0.00%]

Table 4: FP #occs for the Dev/Test reference files (ref) and the Ngram+f-RNNLM system output (hyp)

Table 4 gives the occurrence counts for both the scoring reference files and the Ngram+f-RNNLM system output hypotheses. These counts show that the Ngram+f-RNNLM models the various FP subtypes rather differently. There is a tendency to overgenerate the three most frequently occurring FPs (i.e., *UH*, *UM*, *OH*). The overgeneration ranges from 17.4% to 76.9%. By contrast, the system undergenerates the less frequently occurring subtypes (e.g., *AH*, *HM*). Presumably this is a consequence of the occurrence counts in the training data, which ensure that the LMs associate higher likelihoods with frequently occurring FPs. The patterns for all FP subtypes are similar for the Dev and Test sets. Table 5 further illuminates this by giving the Precision, Recall, and F-score scores for the distinct FP subtypes.

	Precision (Dev/Test)		Recall (Dev/Test)		F-score (Dev/Test)	
UH	0.42	0.46	0.72	0.74	0.53	0.57
UM	0.42	0.45	0.56	0.54	0.48	0.49
OH	0.48	0.53	0.70	0.70	0.57	0.60
UHUM	0.35	0.58	0.04	0.09	0.08	0.16
UHU	0.25	0.81	0.01	0.06	0.02	0.11
HM	0.50	0.00	0.02	0.00	0.04	0.00
AH	0.14	0.12	0.03	0.08	0.05	0.10

Table 5: Individual FP Results for the Ngram+f-RNNLM system

The scores in Table 5 show a fair amount of variation between the different FP subtypes. The scores for the three most frequently occurring FPs are relatively stable across the Dev and Test sets, achieving F-scores in the range 0.48-0.60. By contrast, the scores for the less common FPs sometimes fluctuate considerably (e.g., the Dev Precision for *HM* is 0.50, while the Test Precision is 0.00). Once again, this quantifies the impact of the data sparcity manifest in Table 2.

4. CONCLUSION

In recent years, interest in emotional or expressive speech synthesis has burgeoned. Dominant traits such as extraversion, conscientiousness, agreeableness, and openness are often considered to be essential to the creation of artificial personalities – and FPs are commonly occurring phenomena in natural conversational speech which convey important information about such traits. Consequently, this paper has described a novel approach to the task of inserting FPs into otherwise fluent token sequences to create disfluent input texts for speech synthesis systems.

A lattice-based rescoring framework has been presented which enables Ngram and f-RNNLM LMs to be interpolated. This framework enables multiple FPs to be inserted into multiple IPs. The experiments involving seven FPs show that, using standard metrics, the Ngram+f-RNNLM system is more robust than its constituent Ngram and f-RNNLM sub-components since it combines ther complementary tendencies.

Future research will focus on the modelling of other (more structurally complex) disfluency types, such as discourse markers, repetitions and restarts. It is also important to improve the way speech synthesis systems cope with disfluent input texts, and, to this end, data mixing, better outlier detection, and improved alignment methods will be explored.

5. ACKNOWLEDGEMENTS

This research was supported in part by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

6. REFERENCES

[1] Adell, J., Bonafonte, A., Escudero, D. 2007. Filled pauses in speech synthesis: Towards conversational speech. *Proc. 10th International Conference on Text, Speech and Dialogue* volume 1. Springer 358–365.

[2] Adell, J., Bonafonte, A., Escudero, D. 2007. Statistical Analysis of Filled Pauses Rhytm for Disfluent Speech Synthesis. *SSW6* Bonn, Germany. 223–227.

[3] Adell, J., Bonafonte, A., Escudero-Mancebo, D. 2010. Modelling Filled Pauses Prosody to Synthesise Disfluent Speech. *Speech Prosody* Chicago, USA.

[4] Aylett, M., Potard, B., Pidcock, C. 2013. Expressive Speech Synthesis: Synthesising Ambiguity. *In Proceedings of 8th ISCA Speech Synthesis Workshop*.

[5] Burkhardt, F., Campbell, N. 2014. Emotional Speech Synthesis. *The Oxford Handbook of Affective Computing*.

[6] Carletta, J. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation Journal* 41(2), 181–190.

[7] Chen, X., Wang, Y., Liu, X., Gales, M., Woodland, P. 2014. Efficient GPU-based Training of Neural Network Language Models Using Spliced Sentence Bunch. *In Proceedings of Interspeech* Singapore.

[8] Cieri, C., Miller, D., Walker, K. 2004. The Fisher Corpus: A Resource for the Next Generation of Speech-to-Text Fisher. *In Proceedings of LREC* Lisbon, Portugal.

[9] Clark, H. H., Tree, J. E. F. 2002. Using uh and um in spontaneous speech. *Cognition 84* 73–111.

[10] Collard, P., Corley, M., MacGregor, L. J., Donaldson, D. I. May 2008. Attention orienting effects of hesitations in speech: evidence from ERPs. *Journal of experimental psychology. Learning, memory, and cognition* 34(3), 696–702.

[11] Corley, M., Hartsuiker, R. J. 2003. Hesitation in speech can. . . um. . . help a listener understand. *Proceedings of the twenty-fifth meeting of the Cognitive Science Society* Boston, USA.

[12] Corley, M., Hartsuiker, R. J. Jan. 2011. Why um helps auditory word recognition: the temporal delay hypothesis. *PloS one* 6(5), e19792.

[13] Dall, R., Tomalin, M., Wester, M., Byrne, B., King, S. 2014. Investigating Automatic and Human Filled Pause Insertion for Speech Synthesis. *In Proceedings of Interspeech* Singapore.

[14] Dall, R., Wester, M., Corley, M. The Effect of Filled Pauses and Speaking Rate on Speech Comprehension in Natural, Vocoded and Synthetic Speech. *In Proceedings of Interspeech*.

[15] Fox Tree, J. E. 2001. Listeners' uses of um and uh in speech comprehension. *Memory and Cognition* 29(2), 320–326.

[16] Goodfrey, J. J., Holliman, E. C., McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *ICASSP* San Francisco, CA, USA. 517–520.

[17] He, L., Hyang, H., Lech, M. 2013. Emotional Speech Synthesis Based on Prosodic Feature Modification. *In Proceedings of the 8th International Conference on Bioinformatics and Biomedical Engineering*.

[18] Liu, X., Wang, Y., Chen, X., Gales, M., Woodland, P. 2014. Efficient Lattice Rescoring using Recurrent Neural Network Language Models. *In Proceedings of Interspeech* Singapore.

[19] Nass, C., Brave, S. 2005. Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship. The MIT Press.

[20] Schröder, M., Cowie, R., Cowie, E. 2001. Emotional Speech Synthesis: A Review. *In Proceedings of Eurospeech* volume 1 561–564.

[21] Stolke, A., Zheng, J., Wang, W., Abrash, V. 2011. SRILM at Sixteen: Update and Outlook. *In Proceedings of ASRU* Hawaii, USA.

THE RELATIONSHIP BETWEEN PRECEDING CLAUSE TYPE, SUBSEQUENT CLAUSE LENGTH AND THE DURATION OF SILENT AND FILLED PAUSES AT CLAUSE BOUNDARIES IN JAPANESE MONOLOGUES

Michiko Watanabe[a], Yosuke Kashiwagi[b], Kikuo Maekawa[a]

[a] National Institute for Japanese Language and Linguistics, [b] The University of Tokyo
[a] watanabem@ninjal.ac.jp, [b] kashiwagi@gavo.t.u-tokyo.ac.jp, [a] kikuo@ninjal.ac.jp

ABSTRACT

Filled pauses (FPs) are claimed to occur when speakers have some difficulties and need extra time in speech production. This study investigated whether the following two factors affect silent pause (SP) and FP durations at clause boundaries, using a spontaneous speech corpus: 1) boundary strength and 2) subsequent clause length. First, whether SP and FP durations increase with syntactic boundary strength was examined. Second, whether subsequent clause length affects SP and FP durations at the boundaries was investigated. Results show SP duration increased with boundary strength and subsequent clause length, but FP duration did not, suggesting only SP duration is affected by the two factors.

Keywords: silent pause, filled pause, clause boundary, speech planning, disfluency.

1. INTRODUCTION

Speech sounds such as “um” and “uh” in English and “eto” and “ano” in Japanese are rare in read speech, but common in spontaneous speech. These sounds are believed to be relevant to speech planning difficulties [1, 4]. We call such sounds “fillers”. The list of fillers in Japanese differs depending on researchers. In this study we employ the inventory of fillers in “The Corpus of Spontaneous Japanese (CSJ)” [6]. When the following sounds are used to fill in a gap, they are regarded as fillers:

- -ano, sono (originating from demonstratives, meaning “that” and “the”, respectively)
- -ato, eto, nto, to (some vowels or a nasal flowed by “to,” which is probably a quotation particle)
- -ma (originating from an adverb)
- -a, i, u, e, o, n (Japanese vowels and a nasal)

The items in the first three groups followed by “ne” (an interjectory particle) or “desune” (a copula + ne) are also included in the list.

Fillers are commonly observed at sentence and clause boundaries in spontaneous monologues in CSJ. They are often preceded and occasionally followed by silence at such locations. We call silences immediately after the end of clauses “silent pauses (SPs)”. If there is a filler between two consecutive clauses, we call the sound including the subsequent silence, if any, “filled pauses (FPs)” in this study, as shown below.

filler

[clause] silence um silence [clause]

SP FP

(silent pause) (filled pause)

Speakers are likely to be engaged in conceptualizing the message to be conveyed and encoding some part of it into linguistic forms at deep syntactic boundaries. It is conceivable that speakers use FPs when they need some more time to resume speech after a certain period of silence and want to inform the listener of their current situation. In addition to FPs, there are several means to do so. Repeating words, prolonging a part of words, and making false starts are among others. FPs seem one of the most common devices in English and Japanese.

Previous research indicates that the probability of fillers appearing at clause boundaries is affected by the boundary strength and the subsequent clause length [7]. It tested two hypotheses, “the boundary hypothesis” and “the complexity hypothesis”. The boundary hypothesis is that speakers are more likely to use fillers at deeper syntactic boundaries because they need to plan larger units of information than at shallower boundaries. The complexity hypothesis is that speakers are more likely to use fillers when the subsequent clause is more complex. The number of words in a clause was used as its index of complexity, because the number of words in a linguistic unit is reported to be highly correlated with its complexity [2]. Fillers appeared more frequently at stronger boundaries than at weaker boundaries, supporting the boundary hypothesis. Longer clauses were more often preceded by fillers

than shorter clauses, also supporting the complexity hypothesis. As previous research investigated only the rate of fillers appearing at clause boundaries, in this study we considered the durations of SPs and FPs at clause boundaries. We adjusted the two hypotheses to SP durations and FP durations as follows:

1) The stronger the boundary, the longer SPs and FPs at the boundary (the boundary hypothesis).
2) SP and FP durations increase with the subsequent clause length (the complexity hypothesis).

We regarded clause length as the amount of information expressed by the clause. We tested the two hypotheses by measuring SP and FP durations as a function of clause boundary strength and subsequent clause length.

Adverbial clauses in Japanese

We focus on adverbial clause boundaries in this study. Adverbial clauses are grouped into three types according to connective particles or certain forms of verbs, adjectives or auxiliary verbs at the end of the clause [3, 5]. Japanese is a head-final language and adverbial clauses always precede the main clause as schematically shown below.

(adverbial clause <connective particle>), (main clause).

Table 1 shows the classification of adverbial clauses. It is known that type C clauses are syntactically more independent from the main clauses than the other types because type C clauses can have their own topics as well as their own subjects. Type A clauses are the most dependent on the main clauses because the topics and the subjects of the main clauses are automatically those of type A clauses. That is, they cannot exist without the subsequent main clause. The degree of dependency of type B clauses is in between those of type A and type C clauses. Type B clauses can have their own subjects, but not their own topics. It is also known that the three types have a hierarchical structure. Type C clauses can contain any of the three types. However, type B clauses can contain type A and type B, but not type C. Type A clauses can contain only type A clauses. Based on the degree of dependency and the structure, we regarded boundaries after type C clauses as the strongest, boundaries after type A clauses the weakest, and those after type B clauses to be in between the two types. We also included sentence boundaries in the analysis, but regarded them as stronger than any clause boundary type. The ends of sentences are marked with sentence-final particles or sentence-final forms of verbs, adjectives or auxiliary verbs. The boundary hypothesis predicts the order of SP and FP durations at the boundaries as follows:

sentence boundaries > type C > type B > type A

Table 1: Classification of adverbial clauses

Boundary type	Connective	Meaning, usage
A	~tari, ~tari ~nagara, tsutsu	~ing and ~ing (listing actions) ~ing (expressing accompanying actions)
B	~to,ba,tara,nara	when ~, if ~, in case
	~te, de	~ and
	~te kara	since ~, after~
	~te mo	even if ~
	~yoo ni	so that ~
	adverbial forms of verbs and adjectives	~ and
C	~kara, node	as ~ (reason)
	~noni, ke(re)do	though ~
	~ga	although ~, ~but
	~si	~ and (listing actions or features)
	~desite, masite	~ auxiliary verb for politeness + and
sentence	~yo, ne, to	sentence-final particles
	~desu, masu, ta, n	sentence-final forms of auxiliary verbs

2. METHOD

2.1. Data

106 presentation speeches from the "*Core*" part of *The Corpus of Spontaneous Japanese* (CSJ) were used for analysis [6]. The presentations were given by paid volunteers to a small audience in an informal setting. In the corpus they are referred to as simulated public speaking (SPS). 53 of them were given by female speakers and 53 by male speakers. All the speakers spoke Tokyo dialect. They talked about general topics such as "the happiest memory in my life" or "my town" for about 10 minutes. The topics were given to the speakers beforehand. They were instructed to prepare notes for their speeches, but to not read from manuscripts. All the speeches were transcribed, and detailed linguistic information was given to the transcription.

2.2. Procedures

First, we identified clause boundaries and grouped them into three types according to the connectives. Sentence boundaries were also identified depending on sentence final forms or particles as shown in Table 1. We referred to a set of phrases between two consecutive boundaries a clause. Second, we counted the number of *Bunsetu*-phrases (*Bunsetu*, hereafter) in each clause. *Bunsetu* is composed of one content word with or without function words. We regarded the number of *Bunsetu* as an index of the amount of information in the clause. Third, we grouped boundaries according to the boundary type and the number of *Bunsetu* in the subsequent clause in each presentation. We measured SP and FP durations at each boundary as a function of the boundary type and the number of *Bunsetu* in the subsequent clause. An example of clause boundary grouping is shown below. The number of *Bunsetu* is given at the beginning of each clause. A single letter paired with a number at the end of each clause indicates a category of boundaries. For example, B2 means a type B boundary followed by a two-*Bunsetu* clause.

6: (F ano) ato kekkoo uchi-no ryooshin-wa
um also quite our parents-TOPIC
petto-ga shinu-<u>to</u> <conditional> B2
pet-SUB die-CONDITIONAL
2: hekomu hito-na-<u>node</u> /reason/ C3
depressed people-are-REASON
3: (F ano) zettai dame-to iwarete-i-te <te> B2
um absolutely no-QUATATION had-been-told-AND
2: medaka-sura kae-masen-deshita [sentence final] D1
killifish-even were-not-able-to-keep
1: sore-na-node /reason/ C2
that-is-REASON
(CSJ: S01F0006)

We took the mean value of SP durations, durations of SPs with subsequent FPs, where SPs were followed by FPs (SP + FP), and FP durations for each boundary group in each presentation, and averaged these durations over presentations.

3. RESULTS

The mean number of each boundary type per presentation was 5 for type A, 73 for type B, 39 for type C, and 42 for sentence boundaries. Type A boundaries were excluded from further analysis because of insufficient number of samples. Type A clauses were treated as parts of the larger clauses which contained them. We discuss SPs and FPs at type B, type C and sentence boundaries, hereafter.

Fig. 1 shows SP durations at the three boundary types as a function of subsequent clause length in the number of *Bunsetu*. First, the figure shows that SP durations are closely related to boundary strengths. SPs at sentence boundaries are longer than those at clause boundaries. Likewise SPs at type C boundaries are longer than those at type B boundaries. Fig. 1 also illustrates SP durations tend to increase with the subsequent clause length until the length reaches 8 *Bunsetu* at all boundary types.

Fig. 2 illustrates durations of SPs including following FPs (SP + FP) at the three boundary types as a function of subsequent clause length. Although SP + FP durations are naturally longer than SP only durations, the pattern of the duration change in Fig. 2 is very similar to that of Fig. 1. SP + FP durations are the longest at sentence boundaries, the shortest at type B clause boundaries, and in between at type C boundaries. SP + FP durations increase with the subsequent clause length until the length reaches 6 *Bunsetu* at all boundary types.

Fig. 3 shows FP durations at the three boundary types as a function of subsequent clause length. FPs at type C boundaries tend to be longer than those at type B and sentence boundaries. FP durations do not

Figure 1: SP durations at three boundary types as a function of subsequent clause length. Error bars indicate standard errors in all the figures.

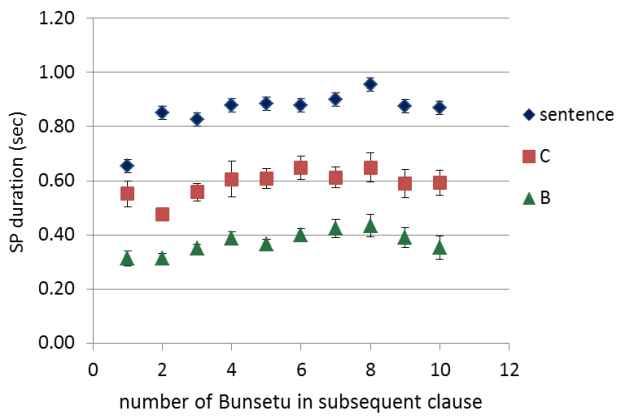
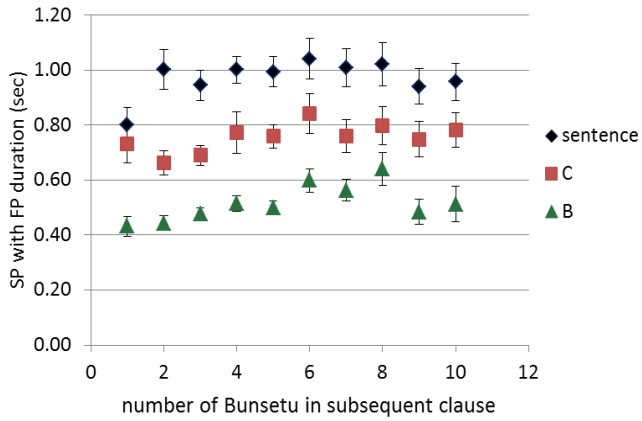
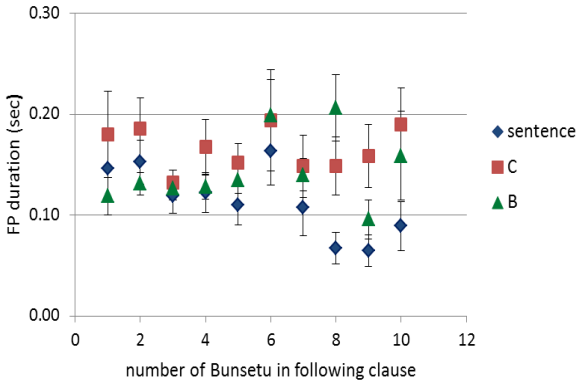


Figure 2: Durations SPs with following FPs at three boundary types as a function of subsequent clause length.



increase with the following clause length at any boundary type.

Figure 3: FP durations at three boundary types as a function of subsequent clause length.



4. DISCUSSION

Both SP durations and SP + FP durations were the longest at sentence boundaries, the second longest at type C boundaries, and the shortest at type B boundaries, supporting the boundary hypothesis. FPs alone did not support the boundary hypothesis, because FPs were the longest at type C boundaries, not at sentence boundaries. When FPs at type B and type C clause boundaries are compared, the hypothesis is supported because FPs at type C boundaries are longer than those at type B. SP durations are closely related to the boundary strength, whereas FP durations are related to the boundary strength only at clause boundaries.

As a general trend, both SP durations and SP + FP durations increased with the number of *Bunsetu* in the subsequent clause, supporting the complexity hypothesis. However, there were upper limits for both SP durations and SP + FP durations at all boundary types. SP durations were longest when the following clauses contain 8 *Bunsetu*, whereas SP + FP durations were longest when the subsequent clauses contained 6 *Bunsetu*. It is possible that speakers have the threshold value of appropriate pause duration for each boundary type in mind and try to avoid pausing longer than the value. Regarding SPs, the threshold value is likely to be about 900ms for sentence boundaries, about 600ms for type C, and about 400ms for type B clause boundaries. As for SP + FP durations, the value can reach up to about 1000ms for sentence boundaries, about 800ms for type C, and about 600ms for type B clause boundaries.

Another interpretation of the results is that speakers' maximum encoding span is about 8 *Bunsetu* at clause boundaries in casual presentations like these. When the message to be conveyed is too rich to be expressed within 8 *Bunsetu*, or within the threshold pausing time, speakers are likely to encode a part of the message at the beginning of the clause and encode further later. It is interesting to note that SP + FP durations at sentence boundaries are almost constant, around one second, regardless of the following clause length, except when the subsequent clause is composed of one *Bunsetu* (Fig. 2). The complexity hypothesis is hardly supported at sentence boundaries. It is possible that pause durations at sentence boundaries reflect cognitive load of macro level message planning, which cannot be measured by the number of *Bunsetu* in the immediately following clause.

A conjunction "de" (an abbreviation for "sorede", meaning "and") is frequent at the beginning of sentences, often followed by SPs and/or FPs. It is possible that some conjunctions are used to gain time for planning at sentence and clause boundaries. As a next step, we plan to investigate the distribution of conjunctions and their relationship with SPs, FPs and the subsequent clause length at sentence and clause boundaries.

5. ACKNOWLEDGEMENTS

This research was conducted as a part of the project, "Creation of Corpus Japanese Linguistics" at National Institute for Japanese Language and Linguistics, and supported by JSPS KAKENHI Grant Numbers 15K02553 and 26284062.

6. REFERENCES

[1] Clark, H. H. 2002. Speaking in time. *Speech Communication* 36, 5-13.

[2] Clark, H. H., Wasow, T. 1998. Repeating words in spontaneous speech. *Cognitive Psychology* 37, 201-242.

[3] Minami, F. 1974. *Gendai nihongo no kouzou* (The structure of modern Japanese) (in Japanese). Tokyo: Taisyukan syoten.

[4] Sadanobu, T., Takubo, Y. 1995. The monitoring devices of mental operations in discourse – a case of 'eeto' and 'ano (o)' – (in Japanese), *Gengo Kenkyu* 108, 74-93.

[5] Takubo, Y. 1987. Syntactic structure and contextual information (in Japanese), *Nihongogaku, 6* (5), 37-48.

[6] The National Institute for Japanese Language. http://pj.ninjal.ac.jp/corpus_center/csj/misc/preliminary/index_e.html

[7] Watanabe, M. 2009. *Features and Roles of Filled Pauses in Speech Communication -A corpus-based study of spontaneous speech* (Hituzi Linguistics in English No.14). Tokyo: Hituzi Syobo Publishing.

THE TEMPORAL DELAY HYPOTHESIS: NATURAL, VOCODED AND SYNTHETIC SPEECH

Mirjam Wester[1], Martin Corley[2], Rasmus Dall[1]

[1]The Centre for Speech Technology Research, The University of Edinburgh, UK
[2]PPLS, The University of Edinburgh, UK
m.wester@inf.ed.ac.uk, martin.corley@ed.ac.uk, r.dall@sms.ed.ac.uk

ABSTRACT

Including disfluencies in synthetic speech is being explored as a way of making synthetic speech sound more natural and conversational. How to measure whether the resulting speech is actually more natural, however, is not straightforward. Conventional approaches to synthetic speech evaluation fall short as a listener is either primed to prefer stimuli with filled pauses or, when they aren't primed they prefer more fluent speech. Psycholinguistic reaction time experiments may circumvent this issue. In this paper, we revisit one such reaction time experiment. For natural speech, delays in word onset were found to facilitate word recognition regardless of the type of delay; be they a filled pause (*um*), silence or a tone. We expand these experiments by examining the effect of using vocoded and synthetic speech. Our results partially replicate previous findings. For natural and vocoded speech, if the delay is a silent pause, significant increases in the speed of word recognition are found. If the delay comprises a filled pause there is a significant increase in reaction time for vocoded speech but not for natural speech. For synthetic speech, no clear effects of delay on word recognition are found. We hypothesise this is because it takes longer (requires more cognitive resources) to process synthetic speech than natural or vocoded speech.

Keywords: delay hypothesis, disfluency

1. INTRODUCTION

Various studies have shown that speech understanding can sometimes benefit from the presence of filled pauses (e.g., *um* and *uh*) and that words following a filled pause are recognised more quickly [8, 4, 3]. A study by Corley and Hartsuiker [5] showed that not just filled pauses but delays of any kind help auditory word processing. This study investigates whether synthetic speech understanding also benefits from delays in the form of either filled pauses or silence.

The end objective of this work is to produce synthetic conversational speech (interesting for e.g., artificial personalities, more natural speech synthesis) and including disfluencies in the synthetic speech is a possible way of achieving this. Evaluating a synthetic system which includes disfluencies however is not straightforward. The standard preference tests used in the synthetic speech field result in listeners either being primed to prefer sentences with fillers [1] or when they are not primed they prefer stimuli without fillers [6].

Experimental paradigms (e.g., reaction time and change detection experiments) borrowed from the field of psycholinguistics may be a way of circumventing this issue. First of all, in these paradigms listeners are not primed regarding the presence or absence of disfluencies. Secondly, listeners are not asked to judge the quality of the synthetic speech, they are asked to react to the speech they have processed. The idea is that if listeners respond to filled pauses in synthetic speech in the same way as they do to filled pauses in natural speech, we will have an indirect measure of the quality of synthetic speech and the validity of including disfluencies in synthetic speech will be strengthened.

A previous reaction time study [7], including filled pauses in synthetic and vocoded speech, showed that processes observed for natural speech were also observed for vocoded speech but not synthetic speech. The lack of effect for synthetic speech was hypothesised to be due to the poor quality of the synthetic filled pauses. In Corley and Hartsuiker [5] it was shown that delays in word onset facilitate word recognition regardless of the type of delay, whether they were filled with *um*, silence or even non-speech sounds (a tone). Corley and Hartsuiker refer to this as "the temporal delay hypothesis", i.e., it is the temporal delay that facilitates word recognition rather than that speech understanding benefits from the presence of filled pauses such as *um*, *uh*, or similar. If this temporal delay hypothesis applies to

synthetic speech then the quality of the filled pause should be of less importance. The Corley and Hartsuiker study, a reaction time experiment, was replicated here for natural, vocoded and synthetic speech.

2. METHOD

The experiment consists of participants viewing pairs of images on a computer screen and following instructions to press a button corresponding to one of the images as quickly as possible. Details of the materials, speech types and experimental procedure are given below.

2.1. Materials

The same experimental materials were used as in [5]. The materials consisted of both auditory and visual stimuli. The auditory stimuli were instructions to press a button corresponding to one of the pictures in a pair. In the delay conditions, listeners heard an instruction with a delay directly preceding the target word. In the control conditions, the delay was earlier on in the sentence. The instructions were either:

1. Now press the button for the <delay> <target>, please.
2. Now press the <delay> button for the <target>, please.

The delay was either a filled pause *um* or a silent pause of the same length. In addition to the delay there was also a task difficulty manipulation. In the difficult condition the words were low-frequency (LF) words and visually blurred. In the easy condition the target words were high-frequency (HF) words and visually intact. Two sets of 16 pictures were used (examples of LF words: kite, snail, vase, etc., HF examples: bed, foot, tree, etc.). For details of how the frequency category of the words was determined see [5]. Each LF picture was paired with four HF pictures (never in the same combination) resulting in 64 picture pairs. Each picture was shown twice on the left, twice on the right and was a target twice: once in an instruction with an early delay and once with a late delay. The delay was either a filled pause (*um*) or a silent pause. Three picture pairs with mid-frequency items (lamp-cake, clock-knife, wheel-cow) were used for practice trials at the start of the experiment. Figure 1 shows an example of the picture pair snail/tree.

2.2. Speech types

The above described experiment was run using natural, vocoded and synthetic speech. The natural speech recordings have been described in detail in [4, 5]. To summarise, a female native speaker of English was recorded reading the list of target words embedded in the above carrier sentence. Target words, together with the word *please*, were removed from their original contexts and spliced into one version of the carrier sentence that had not originally included any of the target items. The delay was created by asking the speaker to insert an *um* "as naturally as possible" when reading a list of low-frequency items in carrier sentences. A single *um* that was judged most natural was selected and spliced in before the target word (delay condition) and before the word *button* (control condition). All targets start at 2297 *ms*, the *um* or silent pause is 1078 *ms* long.

Figure 1: Example of picture pair snail/tree.

The vocoded and synthetic speech were generated as in [7]. The vocoded speech was created by taking the natural speech and vocoding the stimuli using STRAIGHT [10]. The durations output by STRAIGHT are not exactly the same as the natural speech durations, due to the way silence is dealt with. The target onset for vocoded speech with *um* matches the natural speech at 2297 *ms*. For the silence condition, the target onset time is at 2270 *ms* (the length of the pauses is 1078 *ms* in both conditions).

The HMM-based synthetic speech was generated using HTS 2 [18] in a system newer than but roughly similar to [17]. All the target words were synthesised in the carrier sentence. The *um* was generated by the system but some additional padding was added by hand to make the pause the same length as the pause in natural speech. The silence delay was spliced in by hand. Target onsets were measured by hand and vary from 2413 *ms* to 2507 *ms*.

2.3. Procedure

The experiment was run using OpenSesame [12]. The auditory stimuli were presented to native British English speakers with no hearing problems over Beyerdynamic DT770 headphones in individual sound-treated booths. In total 120 subjects took part, twenty per speech type. The participants were in-

Table 1: Reaction time results for the six experiments: Natural, Vocoded and Synthetic Speech including either *um* or silence delays. Participant mean correct reaction time (ms) relative to target onset. Standard error in brackets.

(a) Experiment 1– Natural Speech; *um*

Instructions	Target Type	
	clear HF	blurred LF
control (*um* early)	376 (10.4)	429 (10.9)
delay (*um* late)	369 (10.0)	417 (12.0)

(b) Experiment 2 – Natural Speech; silence

Instructions	Target Type	
	clear HF	blurred LF
control (silence early)	422 (12.4)	461 (12.5)
delay (silence late)	382 (11.4)	432 (12.0)

(c) Experiment 3– Vocoded Speech; *um*

	clear HF	blurred LF
control (*um* early)	398 (10.5)	444 (10.6)
delay (*um* late)	368 (11.0)	389 (11.2)

(d) Experiment 4 – Vocoded Speech; silence

	clear HF	blurred LF
control (silence early)	440 (10.1)	474 (10.2)
delay (silence late)	389 (9.0)	451 (11.1)

(e) Experiment 5 – Synthetic Speech; *um*

	clear HF	blurred LF
control (*um* early)	557 (9.2)	603 (10.0)
delay (*um* late)	564 (9.7)	592 (10.1)

(f) Experiment 6 – Synthetic Speech; silence

	clear HF	blurred LF
control (silence early)	553 (11.5)	582 (11.4)
delay (silence late)	561 (10.5)	586 (11.4)

formed that the study was about sentence comprehension and that the aim of the study was to follow instructions given in stressful situations. This minor deception was necessary to justify the disfluencies in the study. Ethical approval was obtained from the Ethics Committee of PPLS, University of Edinburgh. The participants were explicitly told to be as fast and accurate as they could. Prior to the experiment starting, the subjects were given the three practice trials to familiarise themselves with the procedure. Following this, the 64 items were presented in a random order. The experiment took just over 5 minutes to complete.

3. RESULTS

For each of the experiments, there are 1280 responses (20*64). Before the data was analysed some of the responses had to be removed: One participant in Experiment 3 did not complete the task and so was disregarded, reducing the number of responses for that experiment to 1216. Furthermore, all incorrect responses (e.g., a subject clicking left when it should have been right) were removed, as well as all responses with a reaction time (RT) smaller than 0 *ms* and all RTs larger than 1100 *ms*. RTs < 0 indicate a participant responded before the target started, RTs > 1100 correspond to button pushes well after the end of the utterance. The number of discarded responses and the total responses included in the analyses per experiment are given in Table 2. Analyses were carried out by fitting Generalized Linear Mixed-Effects models, as implemented in the lme4 library in R [14, 2].

Table 1 shows mean correct reaction times (RTs) relative to target onset with standard error between brackets. Experiments 1 and 2 here are the same as Experiments 1 and 2 in [5].

In Experiment 1, which included natural speech and *um* as the local delay, we found that the delay led to only very small decreases in RT (7 *ms*) for the clear HF words, and marginally larger decreases for LF blurred words (12 *ms*). This effect was not significant, in contrast to what was reported in [5]. On the other hand, the effect of task difficulty was found to be significant with participants taking 48 *ms* longer to react to blurred LF images ($p = .003$) which is more in line with the results reported in [5].

In Experiment 2, which again included natural speech but this time with a silent pause as the delay, a significant effect of delay was found with participants faster by 36 *ms* in the delay condition ($p = .03$). The effect of task difficulty was also significant with participants 50 *ms* slower to respond to blurred images ($p = .003$).

Experiments 3 and 4 show results for vocoded speech. In the *um* condition (Experiment 3), a significant decrease of 42 *ms* ($p = .008$) in RT was found due to the local delay. The effect of task diffi-

Table 2: Number of discarded trials per experiment, and the total number of included trials.

Experiment #	1	2	3	4	5	6
# Incorrect	38	26	39	17	22	42
#RT<0	20	0	3	2	1	0
#RT > 1100	19	59	14	21	23	92
Total responses	1203	1195	1160	1240	1234	1146

culty was significant ($p = .03$) with participants taking 33 *ms* longer to respond to blurred LF pictures. In the silence condition (Experiment 4), significant effects of both delay and task difficulty were found, participants were respectively 36 *ms* faster after a silence delay ($p = .003$) and 50 *ms* slower in the blurred LF condition ($p = .003$).

Experiments 5 and 6 show the results for synthetic speech. Overall the RTs are slower for synthetic speech than for natural and vocoded speech. There is no significant effect of delay for synthetic speech. In both *um* and silence conditions there is a main effect of task difficulty with blurred LF words processed less quickly than HF clear words 42 *ms* ($p = .0003$) and 33 *ms* ($p = .02$), respectively.

Cross-experiment comparisons in which we incorporate an additional "experiment" factor show significant effects of speech type ($p < 0.0001$) and of frequency ($p = 0.0002$). This frequency, or task difficulty, effect corresponds to the findings reported above per experiment. Regarding speech type, synthetic speech is 146 *ms* slower in the silence condition and 181 *ms* slower in the *um* condition than natural speech, and 131 *ms* and 179 *ms* slower than vocoded speech. There is no significant difference in RT between natural and vocoded speech (p=0.96).

4. CONCLUSIONS

The only robust result across all three types of speech (natural, vocoded and synthetic) is that it takes approximately $30 - 50$ *ms* longer to react to blurred images than to visually intact images.

Natural and vocoded speech show a similar picture to the findings in Corley & Hartsuiker's paper [5]. Experiments 2, 3 and 4 support their conclusion "... any delay in word onset can help word recognition". There was a main effect of delay after a silent pause in both natural and vocoded speech and there was a main effect of the *um* delay for vocoded speech. However, the results for Experiment 1 only show a slight increase in the speed of word recognition after *um*.

No effect of delays was observed for synthetic speech. Neither the *um* nor the silence conditions led to increases in RT. This is in line with our previous RT experiments [7], which followed Fox Tree's method [9, 8], and showed that filled pauses (*uh*) led to faster reaction times in natural and vocoded speech but slower reaction times in synthetic speech. At the time this was hypothesised to be due to the poor quality of the filled pauses and prosody. In the current study, not only did we consider filled pauses but also silent pauses.

We found that listeners are significantly slower in experiments 5 & 6 (synthetic speech) compared to the same experiments using natural and vocoded speech. Our findings give support to the theory that processing synthetic speech requires a listener to apply more cognitive resources than when processing natural speech. In [16], Pisoni and colleagues illustrate that the perception of synthetic speech requires more cognitive resources citing studies from the eighties using formant synthesis. For instance, listeners took more time to process synthetic stimuli than natural stimuli in a speeded lexical decision task [13] and they needed to hear more of synthetic speech before reliably identifying whole words [11]. Our RT studies suggest that this also holds for modern statistical parametric speech synthesis (SPSS). Future work revisiting some of the lexical decision, word recognition and sentence verification tasks comparing SPSS and natural speech should further inform how the acoustic-phonetic characteristics of SPSS influence speech perception and how speech synthesis has evolved psycholinguistically compared to earlier approaches.

For now, there is no evidence that including disfluencies is beneficial in synthetic speech as the disfluencies do not seem to be processed in the same way as in natural speech. We hypothesise that no clear effects of delay on word recognition are found because synthetic speech takes so much longer to process than natural or vocoded speech.

All research data associated with this paper can be found at Edinburgh DataShare [15] (http://hdl.handle.net/10283/806).

Acknowledgements This research was jointly supported by the EPSRC under Programme Grant EP/I031022/1 (Natural Speech Technology) and the JST Crest uDialogue Project.

5. REFERENCES

[1] Adell, J., Bonafonte, A., Escudero, D. 2007. Filled pauses in speech synthesis: towards conversational speech. *Proceedings of 10th International Conference on Text, Speech and Dialogue* volume 1. Springer 358–365.

[2] Bates, D., Maechler, M., Bolker, B. M., Walker, S. 2014. lme4: Linear mixed-effects models using Eigen and S4. Available: http://CRAN.R-project.org/package=lme4. R package version 1.1-7.

[3] Brennan, S., Schober, M. F. 2001. How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language* 44(2), 274–296.

[4] Corley, M., Hartsuiker, R. J. 2003. Hesitation in speech can... um... help a listener understand. *Proceedings of the 25th Meeting of the Cognitive Science Society* 276–281.

[5] Corley, M., Hartsuiker, R. J. 2011. Why um helps auditory word recognition: The temporal delay hypothesis. *PloS one* 6(5), e19792.

[6] Dall, R., Tomalin, M., Wester, M., Byrne, W., King, S. 2014. Investigating automatic & human filled pause insertion for speech synthesis. *Proceedings Interspeech* Singapore.

[7] Dall, R., Wester, M., Corley, M. 2014. The effect of filled pauses and speaking rate on speech comprehension in natural, vocoded and synthetic speech. *Proceedings Interspeech* Singapore.

[8] Fox Tree, J. E. 2001. Listeners' uses of um and uh in speech comprehension. *Memory and Cognition* 29(2), 320–326.

[9] Fox Tree, J. E., Schrock, J. C. Feb. 1999. Discourse Markers in Spontaneous Speech: Oh What a Difference an Oh Makes. *Journal of Memory and Language* 40(2), 280–295.

[10] Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A. 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication* 27(3-4), 187–207.

[11] Manous, L., Pisoni, D. 1984. Effects of signal duration on the perception of natural and synthetic speech. *Research on Speech Perception Progress Report No* 10.

[12] Mathôt, S., Schreij, D., Theeuwes, J. 2012. Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods* 44(2), 314–324.

[13] Pisoni, D. 1981. Speeded classification of natural and synthetic speech in a lexical decision task. *The Journal of the Acoustical Society of America* 70(S1), S98–S98.

[14] R Development Core Team, 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. ISBN 3-900051-07-0.

[15] Wester, M., Dall, R., Corley, M. 2015. Experiment materials for "The temporal delay hypothesis: Natural, vocoded and synthetic speech.", [dataset]. University of Edinburgh, School of Informatics, Centre for Speech Technology Research. http://dx.doi.org/10.7488/ds/272.

[16] Winters, S., Pisoni, D. 2005. Speech synthesis: Perception and comprehension. In: Brown, K., (ed), *Encyclopedia of Language and Linguistics* volume 12. 31–49.

[17] Yamagishi, J., Watts, O. 2010. The CSTR/EMIME HTS System for Blizzard Challenge 2010. *Blizzard Challenge Workshop*.

[18] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., Tokuda, K. 2007. The HMM-based Speech Synthesis System Version 2.0. *Proceedings of SSW6* Bonn, Germany. 294–299.

THE EFFECT OF STUDY ABROAD EXPERIENCE ON L2 MANDARIN DISFLUENCY IN DIFFERENT TYPES OF TASKS

Clare Wright[1] and Cong Zhang[2]

[1] University of Reading
[2] University of Oxford

ABSTRACT

Disfluency is a common phenomenon in L2 speech, especially in beginners' speech. Whether studying abroad can help with reducing their disfluency or not remains debated [8]. We examined longitudinal data from 10 adult English instructed learners of Mandarin measured before and after ten months of studying abroad (SA) in this paper. We used two speaking tasks comparing pre-planned vs. unplanned spontaneous speech to compare differences over time and between tasks, using eight linguistic and temporal fluency measures (analysed using CLAN and PRAAT). Overall mean linguistic and temporal fluency scores improved significantly ($p < .05$), especially speech rate ($p < .01$), supporting the general claim that SA favours oral development, particularly fluency [2]. Further analysis revealed task differences at both times of measurement, but with greater improvement in the spontaneous task.

Keywords: fluency; L2 Mandarin; Study Abroad

1. INTRODUCTION

Disfluency is often regarded as one of the first signs in non-native-like L2 speech. In order to improve students' L2 proficiency, many language programmes in the UK send students to the target countries to study for a period of time – usually around a year. Studies on oral proficiency generally show that L2 learners do benefit from the immersion in the target language countries [2]. Improvement on fluency is one of the most noticeable effects. The key factors that influence the improvement include quantity of input[4], quality of input [10] and higher levels of proficiency level at the start of the study abroad(SA)[3]. However, whether study abroad can really be a good way of improving the fluency for all types of L2 speech is not completely agreed [7] [11][12].

L2 fluency include various aspects of fluencies, such as cognitive fluency, utterance fluency and perceived fluency [13]. Skehan [14] further divided utterance fluency into breakdown fluency, speed fluency, and repair fluency. Our study examined all three aspects of utterance fluency using two non-temporal and six temporal measures, with some of the results reflecting cognitive fluency.

Although L2 (dis)fluency studies are not rare, most of them are on European languages such as English and Spanish. Mandarin, as a typologically different language, has received extremely little attention in L2 fluency research. A few studies touched upon L2 Mandarin fluency as a part of their proficiency studies (e.g. [6][16]). One of the first longitudinal Mandarin fluency studies was Du [4], which looked at the development of fluency during the period of one study-abroad semester (approximately four months). The data collected in this study are valuable. Task-wise, [4] used both recorded Chinese speaking classes for planned instructed output and Labovian-style individual interviews to elicit spontaneous output. The time range of the data collection and number of participants are significant enough – once a month over the course of roughly four months on 29 participants in different contexts on and off campus. This study showed significant fluency progress over time, but the results were taken from specific 2-minute segments chosen to highlight productive speech, limiting the generalizability of the findings.

Moreover, because Mandarin is typologically different from European languages, research methods from existing European L2 (dis)fluency studies cannot directly be applied in L2 Mandarin. This creates challenges for our study in terms of annotating data and choosing the right measures.

In this paper, we aim to contribute to the discussion of fluency development by presenting data from a set of 10 adult English university learners of Mandarin measured before and after ten months of studying abroad, to illustrate the complex nature of L2 speech and its development across a range of measures and different tasks.

2. METHODS

2.1. Participants

10 adult English university learners of Mandarin from a UK university participated voluntarily in the study. None had any contextual or long-term exposure to Mandarin. Data were collected individually in the context of the students' end of second year exams, and at the start of Year 4 using the same examination procedures.

2.2. Task design

We used two speaking tasks in monologic form to track changes in disfluency. The first task was a pre-prepared planned talk on participants' daily life in China – talking about what they expected before their study abroad at Time 1, or reflecting back after their return at Time 2. Participants were given at least 48 hours before the test to prepare the talk outside class. The second task was an unplanned spontaneous description of a photograph of a scene involving people doing typical daily-routine activities. The tasks, commonly used in class discussions and in practice oral assessments before the first time of recording, were both based on the theme of daily routines. All the tasks were transcribed using CHAT transcription method developed together with CHILDES [9].

2.3. Variable measurement

Eight measures were used to evaluate L2 Mandarin disfluency before and after study abroad, including two non-temporal measures (linguistic fluency) and six temporal measures (temporal fluency).

Linguistic fluency was measured as output as total characters and hesitation rate (number of retracing and filled pauses per word count), using CLAN analyses found in other SLA studies (e.g. [15]). The measures in CLAN, which was originally based on CHILDES [9] produce produces useful linguistic evidence of underlying grammatical and lexical competence.

Temporal fluency, using PRAAT analyses, was measured by a range of measures tapping both cognitive and articulatory fluency [4][7][11], and including composite and non-composite measures. Composite measures were speech rate of characters per second, articulation rate (speech rate during phonation – i.e. minus pausing time), phonation time ratio (time spent speaking divided by total time taken); non-composite measures were mean length of run, number and mean length of silent pauses. Silent pause data are shown separately here from the other hesitation phenomena in line with other studies, suggesting that pausing can be a specific indication of speech planning processes [9]. Silent pauses in this study are ≥250ms. Previous studies defined silent pauses as between 200ms and 400ms, but we adopt 250 ms as the cut of in line with other L2 studies (e.g. in [8]).

2.4. Annotation issues

2.4.1 Zi ("characters") or ci ("words")

Different from European languages which mostly have orthographic words, Mandarin does not have an intuitive division for "words". Especially in modern Mandarin, a word can be in either short form, which consists of one character, or long form which consists of two or more characters [5]. In disfluency studies, it is especially important to clarify the difference since the position of pauses, repetitions or retracing can make a big difference [1]. In our study, we made use of *zi* instead of *ci*, which is more commonly used in past studies in European languages, since it provides a good way for calculating the total output, annotating in-word fixes (including pauses, repetitions and retracing) in Mandarin Chinese. Past studies such as [4][8] also made use of this method.

2.4.2 Tonal errors

Some of the repetitions and retraces are due to tonal errors. For example, the following is an utterance from our CHAT transcriptions

*ABC: uh 在 中 国 <买> [/?] 买(卖?) 东 西.
Uh in China buy buy(sell?) stuff.
"buying(selling?) stuff in China"

"Buy" and "sell" are both legitimate in the context of this utterance. The only difference between these two characters is their lexical tones, which do not exist in previous studies in European languages. It is hard to differentiate the differences between repetition or retracing in such cases. Since prosodic aspects including lexical tones are among the easiest and earliest acquired features in L1 Mandarin, no norm has been formed for tonal errors in L1 Mandarin annotation either. How to annotate tonal errors therefore also becomes an issue that needs careful thinking in L2 Mandarin.

2.4.3 Other issues

In the process of annotating the data, many other problems occurred due to the exploratory nature of the current study. One problem was that some repeated morphemes were legitimate for intensification in the context but from the flow of the speech, they sounded like repetitions. For example, for the morpheme *zui* 最 ("most") which when repeated means "very", when repeated, it can be either repetition or intensification.

Another problem is whether filled pauses should be annotated as "um, er" as in the western convention. Given that different types of filled pauses may have different psycholinguistic discourse implications [1], target-like fillers may be worth differentiating from the source language fillers.

3. RESULTS

Linguistic fluency scores were analysed using CLAN; we measured output as total characters and hesitation rate (number of retracings and filled pauses per total characters).

The scores showed a clear general improvement as shown in Figure 1 and Figure 2. However, using paired-sample t-tests, in Task 1, only output showed a significant increase (t = -3.552, df = 9, p = .006). By comparison, in Task 2, output and hesitation rate both improved significantly (output: t = -3.432, df = 9, p = 007; hesitation rate: t=2.529, df = 9, p = .032).

Temporal scores were tabulated using PRAAT to measure speech rate, articulation rate, both measured as characters per second; also phonation time ratio, mean length of run (MLR), number and mean length of silent pauses (MLP).

These temporal measures showed consistent improvement across both tasks, with more marked development on measures in Task 2. In Task 1, only Speech rate showed a significant increase (t = -4.567, df = 9, p = .001), and Phonation time ratio (t = =2.286, df = 9, p = .048). By contrast, in Task 2, almost all improvements were highly significant. Speech rate improved very significantly (t = -9.739, df = 9, p = .000), as did Phonation time ratio (t = -3.953, df = 9, p = .003). Articulation rate improved (t = -2.656, df = 9, p = 0.26), MLR improved (t = -2.946, df = 9, p = .016), total number of pauses reduced (t = 2.353, df = 9, p = .043).

Figure 1: Average total output in Task 1 and Task 2 over time (number of characters).

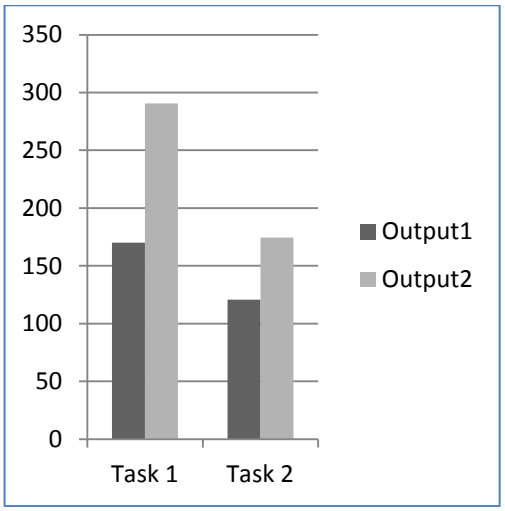
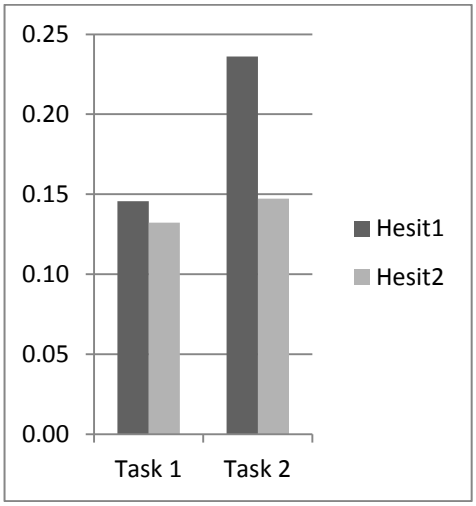


Figure 2: Average hesitation rate in Task 1 and Task 2 over time.



4. DISUCSSION

This study of the reduction of L2 Mandarin disfluency in adult learners after a period of study abroad showed clear improvement on many measures of linguistic and temporal fluency. Results showed a significant improvement in

- total ouput: how much overall participants produced in both pre-planned and unplanned spontaneous speaking tasks;
- speech rate: how fast they could articulate this output
- phonation time ratio: how much more of the time in the task was spent on speaking than pausing.

In general, these data reached the same conclusion as many previous studies did -- study abroad has a positive effect on improving L2 fluency.

Moreover, we found clear evidence of task differences, supporting predictions that planning can have a significant effect on speaking performance. In Task 1, although hesitation rate went down, it was not statistically significant. In Task 2 hesitation rate went down significantly. Moreover, the temporal fluency of Task 2 significantly improved while in Task 1 only speech rate and phonation time ratio were significantly improved.

The difference across tasks may be due to the different natures of the tasks. Task 1 used a pre-planned speech. In this task, participants tend to use simple shorter phrases, easily memorised and recited. The use of pre-planned verbal material reduces cognitive load, by accessing pre-created lexemes at the formulation stage; these sequences are rehearsed prior to the required time of utterance, which further boosts utterance fluency.

In contrast, Task 2 required unplanned spontaneous speech. The participants had to construct new speech materials, which created extra cognitive loads. It therefore revealed a predictable trade-off against utterance fluency in terms of more hesitation and pausing; this created a disadvantage compared to Task 1 at Time 1, but significantly less so by Time 2. Further analysis of location of pause is needed, but preliminary inspection suggested more shifting towards early-clause pausing rather than mid-clause pausing, i.e. general distribution of pauses throughout the utterance.

5. CONCLUSION

In summary, the results showed different levels of reduction on disfluency in pre-planned Task 1 and unplanned spontaneous Task 2. This suggests that disfluency in an online task is reduced more significantly than a pre-planned task after long-term immersion in target language, since online task involves much more cognitive processing than pre-planned task. We therefore argue that studying abroad helps with improving L2 fluency in different ways according to the nature of the speech tasks -- unplanned spontaneous tasks shows more improvements in terms of fluency than pre-planned tasks.

6. REFERENCES

[1] Clark, H.H. and Tree, J.E.F. 2002. Using uh and uh in spontaneous speaking. *Cognition, 84*, 73–111.

[2] Collentine, J., Freed, B. F. 2004. Learning context and its effects on second language acquisition: Introduction. *Studies in Second Language Acquisition, 26*(02), 153–171.

[3] Davidson, D.E. 2010. Study abroad: When, how long, and with what results? New data from the Russian front. *Foreign Language Annals 43*, 1, 6–26.

[4] Du, H. 2013.The Development of Chinese Fluency During Study Abroad in China. *The Modern Language Journal 97*, 1, 131–143.

[5] Duanmu, S. 1998. Wordhood in Chinese. In J.L. Packard, ed., *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese*. Walter de Gruyter, 135–196.

[6] Liu, J.J. 2009. Assessing Students' Language Proficiency: A New Model of Study Abroad Program in China. *Journal of Studies in International Education 14*, 5, 528–544.

[7] De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., Hulstijn, J. H. 2012. Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(01), 5–34.

[8] Kahng, J. 2014). Exploring Utterance and Cognitive Fluency of L1 and L2 English Speakers: Temporal Measures and Stimulated Recall. *Language Learning, 64*(4), 809–854.

[9] MacWhinney, B. 2000. *The CHILDES project: The database.* Psychology Press.

[10] Moyer, A. 2013. An Investigation of Experience in L2 Phonology : Does Quality Matter More than Quantity ? An Investigation of Experience in L2 Phonology : Does Quality Matter More than Quantity ? *67*, 2, 191–216.

[11] Pallotti, G. 2009. CAF : defining , refining and differentiating constructs Defining constructs, 590 − 601.

[12] Segalowitz, N., Freed, B. F. 2004. Context, Contact, and Cognition in Oral Fluency Acquisition: Learning Spanish in At Home and Study Abroad Contexts. *Studies in Second Language Acquisition, 26*, 173–199.

[13] Segalowitz, N. 2010. *Cognitive bases of second language fluency*. Routledge.

[14] Skehan, P. 2003. Task-based instruction. *Language teaching 36*, 01, 1–14.

[15] Wright C. 2012. An investigation of working memory effects on oral grammatical accuracy and fluency in producing questions in English. *TESOL Quarterly*. 47(2): 352-374.

[16] Yang, B. 2012. How Do Pauses Reveal Linguistic Plans by L2 Learners of Chinese ? *Proceedings of the 6th International Conference on Speech Prosody (SP2012)*, 390–393.