Perception of Disfluency: Language Differences and Listener Bias

Catherine Lai, Kyle Gorman, Jiahong Yuan, Mark Liberman

Department of Linguistics, University of Pennsylvania, Philadelphia, USA

{laic, kgorman, jiahong, myl}@ling.upenn.edu

Abstract

This paper describes a crosslinguistic disfluency perception experiment. We tested the recognizability of pause fillers and partial words in English, German and Mandarin. Subjects were speakers of English with no knowledge of Mandarin or German. We found that subjects could identify disfluent from fluent utterances at a level above chance. Pause fillers were easier to identify than partial words. Accuracy rates were highest for English, followed by German and then Mandarin. Although German accuracy rates were higher than those for Mandarin, discriminability analysis suggests that this is due to conservative bias towards false negatives rather than non-recognition of the acoustic material. The fact that subjects could identify disfluent speech in languages they did not know shows that there are real phonetic crosslinguistic cues to disfluency.

Index Terms: crosslinguistic perception, disfluency, pause filler, partial words.

1. Introduction

Disfluencies are a fact of life when it comes to spontaneous speech. Repetitions and filled pauses are dealt with and discarded easily by human listeners. However, automatic detection and resolution of disfluencies appears much harder [1, 2]. Disfluencies are also a crosslinguistic fact of life. Comparative studies of disfluencies rates [3] do not suggest great interlanguage differences distributionally. So, understanding how disfluencies are recognized is important in developing language general speech models. Identification of crosslinguistic cues of disfluencies is clearly a useful step in unravelling the processing problem.

The natural place for crosslinguistic similarities is in in the acoustic properties of disfluencies. Earlier work has posited specific phonetic editing signals marking disfluencies for correction [4]. However, subsequent studies have not found such a unique acoustic cue. Instead phonetic features appear to change depending on the disfluency type and discourse. While the focus has mainly been on English, some acoustic cues seem to extend to other languages. For example, pause fillers tend to have low pitch with level or falling contour [5, 6, 7] along with extended duration. They also co-occur with lengthening of preceding syllables and silent pauses. These last features are also associated with repair disfluencies. These are often associated with laryngealization at the interruption point, changes in coarticulation patterns [8] or prosodic parallelism [9].

In order to accurately model the relationship between disfluencies and speech, we also need to understand how these acoustic realizations relate to speech perception. In particular, we would like to know whether disfluencies can be perceived in a language that the listener is not familiar with, so that top-down information is not available. If so, we are in a good position to identify robust cues to disfluency. More generally, this type of investigation can also give us a better insight into language differences and similarities, and what we might expect language universals to look like.

This paper presents experiments testing the perception of partial words and pause fillers in English, German, and Mandarin by native English speakers. German and Mandarin clearly differ in their similarity to English and to each other. In particular, Mandarin is a tonal language. Given that F0 is often cited as a cue for disfluency, the presence of tone in disfluency perception seemed worth investigating. We found that disfluent speech is crosslinguistically recognizable. However, subjects were biased against accepting utterances as disfluent. This bias was highest regarding the Mandarin data, lower for German and lowest for English. In order to set the stage, however, we first review some previous work on disfluency perception.

2. Perception of Disfluencies

A number of studies investigated the perception of disfluencies in English. Understanding of perception clearly contributes to an understanding of the cues to disfluency. One of the main questions that has been asked is whether or not disfluencies are perceived at all. Perception studies shed light on how disfluencies are processed with respect to the normal speech stream.

Experiments in [10] found that participants robustly moved disfluencies inside constituents to constituent boundaries, when asked to reproduce recording of directed speech. Greater attention to disfluencies also degraded word recall accuracy. This suggests that listeners use strategies to discard disfluencies in order to process speech successfully. Disfluency displacement suggests these processes occur at a reasonably abstract level. Lickley and Bard [11] confirm that disfluency degrades normal word recognition processes. This leads them claim that listeners 'fail to recognize the acoustic material of disfluencies as words or they recognize it with so much delay that portions of the speech will be lost from memory as new input arrives'. This sits ill at ease with models where reparanda are signalled, recognized, and replaced. It is also somewhat at odds with both previous acoustic and perception studies. In similar experiments, it was found that disfluencies could be perceived in low-pass filtered speech [12]. That is, prosodic features appear to cue disfluencies.

If segments containing disfluencies are never actually recognized and processed then it is unclear what role the acoustic correlates of disfluency have to play in speech communication. It is also unclear how this deafness to disfluency should be resolved with respect to pause fillers. It has been claimed that pause fillers (e.g. English 'uh' and 'um') serve a discourse function with respect to turn taking. In fact, Clark and Fox Tree [13] argue that they are actually lexical items. In this case, their communicative function will be lost if they are simply not recognized. Moreover, if processing load cause disfluencies

to be discarded then listeners should not be able to recognize speech as disfluent in language they do not already know. Thus, crosslinguistic perceptibility has an interesting part to play in determining how disfluencies are processed.

There do appear to be language specific and universal acoustic properties of disfluencies. For example, work by Vasilescu et al. [14] suggest that universal prosodic components of pause filler are duration and F0. However, they also found listeners could distinguish isolated pause fillers as French or from another language (similarly, Portuguese). Moreover, using machine learning techniques, Chu et al [15] found that word fragments in Mandarin are not glottalized to the same extent they are in English. The following experiment adds a new perceptual perspective to the problem of crosslinguistic disfluency cues.

3. Experiment Design

Stimuli were drawn from the Callhome telephone speech corpora in English (LDC97S42), German (LDC97S43) and Mandarin (LDC96S34). For each language, 24 utterances (12 male/12 female) containing a pause filler, 24 containing a partial word, and 48 fluent fillers were selected. The pause fillers used were the ones most frequently used in the corpora: English 'uh', German 'äh' and Mandarin 'e'. The partial word experiment contained both false starts and repetitions. Utterances were initially randomly selected from the corpora and checked by a native speaker. 288 unique stimuli were used in the experiment. The experiment consisted of four sections (144 stimuli each). That is, subjects heard two repetitions of the entire data set and 576 utterances in total. Each section presented all utterances of one disfluency type with half the fluent utterances in random order.

12 University of Pennsylvania undergraduates (8 male, 4 female) participated in this experiment. The average age of the participants was 20 years. All subjects were speakers of English without any pedagogical exposure to Mandarin or German. Subjects were paid to participate in the experiment and a bonus was offered for top performers.

The stimuli were presented via a computer interface. For each section, the subject was prompted on screen with the question "Do you hear a partial word?" (resp. "pause filler") as they heard the utterance over high quality headphones. They responded by hitting on-screen buttons labelled "no partial word" or "partial word" (similarly for pause fillers) with a mouse or the keyboard. Subjects were asked to answer as accurately as possible. Each subject was presented with a short version of the experiment to help ensure they understood the task. New stimuli were presented only when a response for the current stimulus was registered. Subjects could only respond after the whole utterance had played and were not allowed to replay any stimulus. The results of one subject, who performed at chance on the English data, were removed from the data set. The rest of the results of the perception experiment are presented next.

4. Perception Experiment Results

4.1. Language Variation

The results show that disfluencies can be perceived crosslinguistically. However, language and disfluency type clearly had an effect on each subject's success rate. Pause fillers were more perceptible than partial words. This trend persists looking at the three language separately (c.f. Table 1) Subjects were able to detect disfluencies in English with over 80% accuracy. However, subjects were less accurate with German and Mandarin. Mandarin incomplete words appear to be the most difficult to detect.

	Pause Filler	Fluent Filler	Partial Word
Eng	0.89(0.71-1.00)	0.91(0.84-0.97)	0.83(0.5-0.94)
Ger	0.80(0.46-0.98)	0.83(0.64-0.95)	0.63(0.35-0.79)
Man	0.71(0.33-0.96)	0.89(0.75-0.99)	0.57(0.15-0.75)

Table 1: *Accuracy rates by language and disfluency type.*

	PF A'	PF B'	PW A'	PF B'
English	0.95	0.16	0.93	0.34
German	0.90	0.21	0.81	0.44
Mandarin	0.90	0.71	0.81	0.61

Table 2: *A' (disciminability) and B' (bias) values by language.* PF=*pause filler,* PW=*partial word.*

Table 2 gives results of a nonparametric discriminability analysis [16] based on signal detection theory. A' gives a measure of discriminability: scores near 1 indicate high discriminability, 0.5 indicates chance performance. B' is a measure of bias ranging over $[-1, 1]$. Positive scores indicate conservativity or bias towards false negatives. The A' scores confirm that subjects found fillers more discriminable than partial words. Interestingly, German and Mandarin obtained the same discriminability. However, subjects were more conservatively biased towards Mandarin. There was considerable subject variation (c.f. Table 2) in their ability to detect Mandarin disfluencies.

4.2. Subject Variation

The variation in bias is shown in Figures 1 and 2. This shows by-subject the hit/false alarm plots. The stronger bias against Mandarin pause fillers, relative to German, is shown with red data points generally closer to the left edge of the graph. However, the German and Mandarin data points are more similarly distributed in Figure 2. Subject variation may also be linked to variation in the prosody of stimuli. This is reflected in Figure 3 which shows the overall pattern of disfluency detection for each subject and stimulus.

Subjects appeared to use different criteria for the subcomponents of each task (c.f. Tables 3 and 4). Subject 4 appeared

	English		German		Mandarin	
Subject	A'	B'	A'	B'	A'	B'
1	0.95	0.36	0.88	-0.44	0.85	0.00
2	0.96	0.15	0.90	-0.30	0.92	0.77
3	0.98	-1.00	0.96	-0.53	0.95	0.00
4	0.95	0.00	0.93	-0.87	0.97	-0.35
5	0.85	0.35	0.78	0.78	0.83	0.82
6	0.93	-0.37	0.84	0.53	0.83	1.00
7	0.97	0.21	0.92	0.09	0.90	0.93
8	0.96	0.15	0.88	0.14	0.87	0.72
9	0.97	0.21	0.95	0.00	0.95	0.59
10	0.88	0.44	0.89	0.76	0.89	0.85
11	0.97	0.62	0.92	0.79	0.93	0.88

Table 3: *Pause fillers: A' and B' values by subject.*

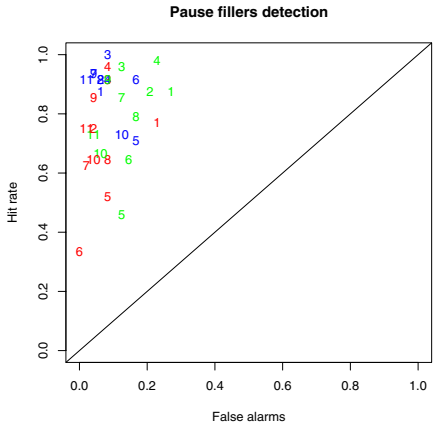

Figure 1: *Pause fillers by subject and language. Blue=English, red=Mandarin, green=German.*

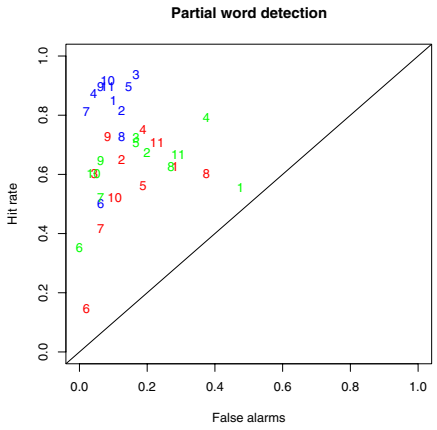

Figure 2: *Partial words by subject and language. Blue=English, red=Mandarin, green=German.*

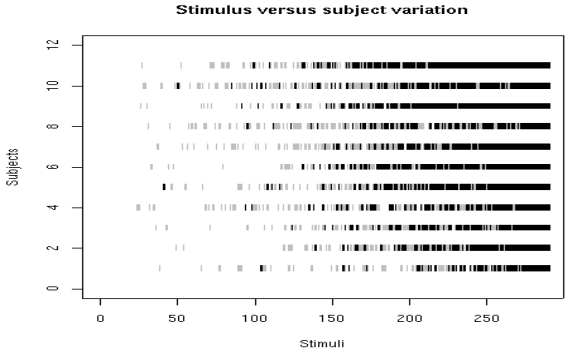

Figure 3: *Disfluency detection for all stimuli: Subjects and stimuli are sorted in ascending order of disfluency detection (i.e. all positive responses). Both fluent and disfluent stimuli are included here. Black=disfluency detected on all repetitions, grey=disfluency detected once, white=disfluency not detected.*

	English		German		Mandarin	
Subject	A'	B'	A'	B'	A'	B'
1	0.93	0.23	0.57	-0.06	0.76	0.20
2	0.91	0.23	0.82	0.32	0.85	0.58
3	0.94	-0.50	0.86	0.31	0.88	0.87
4	0.96	0.53	0.80	-0.39	0.86	0.18
5	0.93	-0.19	0.85	0.35	0.78	0.54
6	0.84	0.88	0.84	1.00	0.75	0.99
7	0.95	0.83	0.84	0.86	0.81	0.91
8	0.88	0.44	0.76	0.24	0.69	0.04
9	0.95	0.27	0.88	0.78	0.90	0.61
10	0.95	0.00	0.88	0.88	0.82	0.78
11	0.95	0.12	0.77	0.10	0.82	0.16

Table 4: *Partial words: A' and B' by subject.*

liberally biased with respect to Mandarin and German pauses fillers, but apparently unbiased towards the English data. However, subject 6 was extremely conservative with respect to Mandarin pause fillers, less conservative with respect to the German data, and liberally biased towards English pause fillers. This suggests that these subjects were using a different strategy for English than for foreign languages. In fact, subjects appear to be using their native language model in the perception task. This is visualized in the top-left English result cluster in Figure 2. This highlights how hard it is to control for the role of non-acoustic information when trying to study disfluencies. The fact that this information was not available for the German and Mandarin data gives us a good indication of how strong *acoustic* cues really are.

5. Discussion

The main finding of this experiment is that disfluencies can be detected by subject in a language they do not know. This strongly suggests that there are salient acoustic cues to disfluency that are valid crosslinguistically. This is somewhat incongruous with the theory that listeners are deaf to disfluencies (c.f. Section 2). Discriminability of German and Mandarin disfluency indicates that the acoustic material of disfluencies is perceived as such by the subject. The subjects had no access to lexical, semantic or syntactic information about these utterances. Repair phrases were sometimes reduplicated as part of the correction. However, this does not appear to have confounded the disfluency signal. Mandarin utterances with fluent reduplication were generally correctly perceived as fluent. So, we can conclude that prosodic factors did in fact signal these utterances as disfluent.

However, not all disfluencies are easily and equally recognizable in all languages. This study found pause fillers easier to detect than partial words across languages. This is consistent with the differing discourse status of these types of disfluency. On the one hand, partial words are errors and must be disposed of for normal speech comprehension to continue. On the other hand, Clark and Fox Tree [13] have argued that pause fillers are conventional words of English for floor holding/ceding, or to simply indicate that the speaker has not yet decided what to say. These results confirm that partial words have more language specific cues. Also, specific cues are not necessarily invoked at every disfluency.

Overall accuracy scores suggest that Mandarin disfluencies are simply harder to recognize than those in German. However,

discriminability analysis suggest that lower accuracy is more to do with subject bias than with the ability to hear the acoustic realization of disfluencies. Note, there was a smaller, but still conservative bias, towards the German data. This suggests the foreignness of the language causes subjects to set a higher criterion for recognition of disfluency. It may be the case that there is less overlap with English disfluency features with German fluent speech than with fluent speech in Mandarin. The prime suspect seems to be the presence of lexical tone in Mandarin.

Zhao and Jurafsky [7] find that the Mandarin 'uh' type pause fillers is realized with a low and flat or falling pitch. Moreover, they claim this is the Mandarin lexical low tone. In fact, much like English, our experiment found that the most easily identifiable (100% accuracy) pause filler in Mandarin had low flat tone of extended duration, associated with lengthening of previous syllable or pauses. The case was similar for German. However, not all cues are always present. When cues such as duration, pauses or laryngealization are missing the listener may need to put greater reliance on pitch to identify the disfluency. The presence of lexical low tones may reduce the use of pitch as a disfluency cue. Conservative bias may also be caused by use of tone in general rather than tonal features local to the disfluency. It is clear that more investigation needs to be done to confirm this.

Finally, variation in subject bias seems to generalize findings in [17]. That study showed that stutterers are more likely to rate speech as disfluent than non-stutterers. This was independent of whether the speech was from a stutterer or not. This suggested that the monitoring is higher in people who stutter. It would be interesting to find out if these biases translated to monitoring rates in our subjects' speech production.

6. Conclusion

This paper presented the results of a crosslinguistic study of disfluency perception. We found that subjects could identify disfluencies in languages they did not know. Subjects did not have recourse to lexical, syntactic or semantic information in identifying disfluencies. This indicates that listeners do use phonetic cues to identify disfluencies. Our English speaking subjects were less accurate in recognizing Mandarin disfluency than German. However, discriminibility analysis from signal detection theory indicates that subjects could discriminate the acoustic material of disfluencies in Mandarin as well as German. The A' (discriminability) and B' (bias) measures also give a clearer indication than pure accuracy rates of what type of variation matters with respect to individual subjects and languages. The difference in the accuracy rates seems to come from the subjects conservative bias towards languages less like their native one. Further acoustic and perceptual studies are required to get to the root of this bias. However, it seems likely that the presence of lexical tone in Mandarin may cause listeners to be less confident about acoustic cues involving pitch.

Perception of speech disfluencies does not appear to have been as intensely studied as the acoustic component. However, perception data is crucial in understanding what acoustic cues are actually useful to human speech processing. We hope to replicate these results with wider subject pools and with subjects who have native languages other than English, along with the acoustic analysis. The differences in perceptual biases among subjects also suggests further avenues of research. This could lead to a better generalized models of speech perception and processing in general. Crosslinguistic studies provide a fruitful line of inquiry for finding robust cues of disfluency.

7. Acknowledgements

Thanks to Tatjana Scheffler for help with the German data.

8. References

[1] C. Nakatani and J. Hirschberg, "A corpus-based study of repair cues in spontaneous speech," *J. Acoust. Soc. Amer.*, vol. 95, no. 3, pp. 1603–1616, 1994.

[2] E. Shriberg, "To 'errrr' is human: ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, vol. 31, p. 1, 2001.

[3] R. Eklund, "Crosslinguistic Disfluency Modeling: A Comparative Analysis of Swedish and Tok Pisin Human–Human ATIS Dialogues," in *ICSLP-2000*, 2000, pp. 991–994.

[4] D. Hindle, "Deterministic parsing of syntactic nonfluencies," *Proceedings of the 21st conference on Association for Computational Linguistics*, pp. 123–128, 1983.

[5] E. Shriberg and R. Lickley, "Intonation of clause-internal filled pauses," in *ICSLP-1992*, 1992, pp. 991–994.

[6] M. Candea, I. Vasilescu, and M. Adda-Decker, "Inter-and intra-language acoustic analysis of autonomous fillers," in *DiSS-2005*, 2005, pp. 47–51.

[7] Y. Zhao and D. Jurafsky, "Tone or toneless: The prosody of mandarin filled pauses," in *Poster presented at the 10th Conference on Laboratory Phonology, Paris*, 2006.

[8] R. Lickley, "Juncture cues to disfluency," in *ICSLP-1996*, 1996, pp. 2478–2481.

[9] J. Cole, M. Hasegawa-Johnson, C. Shih, H. Kim, E.-K. Lee, H.-y. Lu, Y. Mo, and T.-J. Yoon, "Prosodic parallelism as a cue to repetition and error correction disfluency," in *DiSS-2005, Aix-en-Provence, France*, 2005, pp. 53–58.

[10] J. Martin and W. Strange, "The perception of hesitation in spontaneous speech," *Perception and Psychophysics*, vol. 3, no. 4, pp. 427–432, 1968.

[11] R. Lickley and E. Bard, "On not recognizing disfluencies in dialogue," *ICSLP-1996*, pp. 1876–1879, 1996.

[12] R. Lickley, R. Shillcock, and E. Bard, "Processing disfluent speech: How and when are disfluencies found," in *Proceedings of Eurospeech'91*, 1991, pp. 1499–1502.

[13] H. Clark and J. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[14] I. Vasilescu, M. Candea, and M. Adda-Decker, "Perceptual salience of language-specific acoustic differences in autonomous fillers across eight languages," in *Interspeech-2005*, 2005.

[15] C.-T. Chu, Y.-H. Sung, Y. Zhao, and D. Jurafsky, "Detection of word fragments in Mandarin telephone conversation," in *Interspeech-2006*, 2006.

[16] W. Donaldson, "Measuring recognition memory." *J Exp Psychol Gen*, vol. 121, no. 3, pp. 275–7, 1992.

[17] R. Lickley, R. Hartsuiker, M. Corley, M. Russell, and R. Nelson, "Judgment of Disfluency in People who Stutter and People who do not Stutter: Results from Magnitude Estimation," *Language and Speech*, vol. 48, no. 3, p. 299, 2005.