most from intonation correction, but the effect of the correction has greater impact on the foreign accent for some L1 groups than for others. The Tamil speakers' N2 is more foreign accent reduced by the correction than the Chinese speakers' N2 and the Chinese N2 speech is more foreign accent reduced than the English speakers' N2. Likewise, speakers with French and German as their native languages benefit most from duration correction, but the foreign accent reduction effect is larger for the French L1 group than for the German L1 group.

The native Norwegian listeners that participated in the experiment represented both low-tone and high-tone dialects. No correlation was found between listener dialect and responses in the perception experiment.

References

Boersma, P. & D. Weenink, 2006. *Praat: doing phonetics by computer* (Version 4.4.17) [Computer program]. Retrieved April 19, 2006, from http://www.praat.org/.

Derwing, T. & M.J. Munro, 1997. Accent, intelligibility and comprehensibility: Evidence from four L1s. *Studies in second language acquisition 19*, 1-16.

Flege, J.E., M.J. Munro & I.R.A. MacKay, 1995. Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America 97*, 3125-3134.

Munro, M.J. & T. Derwing, 1995. Processing time, accent and comprehensibility in the perception of native and foreign accented-speech. *Language and Speech 38*, 289-306.

Munro, M.J. & T. Derwing, 2005. Second language accent and pronunciation teaching: A research based approach. *TESOL Quarterly 39*(3), 379-397.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 65–68

The Filler *EH* in Swedish

Merle Horne
Dept. of Linguistics and Phonetics, Centre for Languages and Literature, Lund University
merle.horne@ling.lu.se

Abstract

Findings from a pilot study on the distribution, function and phonetic realization of the filler EH in interviews from SweDia2000 interviews are presented. The results show that EH occurs almost exclusively after function words at the beginning of constituents. The phonetic realization of EH was seen to be of three basic forms: a middle-high vowel (e.g. [e], [e], [ə]), a vowel+nasal (e.g. [em], [əm], [ən]), and a vowel with creaky phonation (e.g.[ḛ], [ə̰]). The vowel+nasal realization occurs as has been shown for English before other delays and is associated with planning of complex utterances. Since creaky phonation is associated with terminality, the creaky vowel realization of EH could be interpreted as signalling the juncture between the filler and an upcoming disfluency.

1 Introduction

The filler, or 'filled pause' *EH* has often been termed a 'disfluency' (e.g. Shriberg, 2001), since it constitutes a delay in the flow of speech associated with referential meaning. However, since it can often be assigned pragmatic functions, such as signalling an upcoming focussed word (Bruce, 1998), or need on the part of the speaker to plan or code his/her speech and thus a desire to hold the floor, *EH* can also be considered to be an integral part of the linguistic system (see e.g. Allwood (1994), and Clark & Fox Tree (2002) who refer to it as a 'word'). In a study on English, Clark & Fox Tree (2002) found that its realization as *Uh* signals a minor delay in speaking, whereas *Um* announces a major delay in speaking.

A number of studies on Swedish have reported some characteristics of *EH* in different speaking styles, but none have focussed on the variation in the phonetic realization of *EH* as far as I know. Hansson (1998), in a study on the relationship between pausing and syntactic structure in a spontaneous narrative, found that the filled pauses (n=22) in her material occurred at clause boundaries after conjunctions and discourse markers and before focussed words. Lundholm (2000) in a study on pause duration in human-human dialogues found that the filler *EH* (n=55) in authentic travel-bureau dialogues occurred in turn non-initial position and had a duration similar to silent planning pauses (mean = 340 ms). Eklund (2004) in a number of studies on simulated human-human and human-machine dialogues found that the most common position of *EH* (n=2601) was utterance-initial before another disfluency and that it was most often followed by *jag* 'I', and *det/den* 'it'. The filled pauses were found to have a mean duration of about 500 ms, thus considerably longer than those found by Lundholm (2000) in authentic task dialogues.

2 Current study

The present study has been carried out to pursue the investigation of *EH* in spontaneous data to get some better idea as to its distribution, function, and phonetic realization in authentic interviews where the speech is basically of a monologue style. Spontaneous speech from the

SweDia 2000 interview material was used for the study (<http://www.swedia.nu/>). The speech of two female speakers from Götaland (a young woman from Orust and an older woman from Torsö) was transcribed and all *EH* fillers were labeled.

3 Results

3.1 Distribution of EH

The spontaneous *SweDia* data showed that *EH* occurs principally in non-utterance-initial position. There were only two cases of *EH* in utterance-initial position in the data studied and their mean duration was 899 ms.

EH occurs almost exclusively as a clitic to a preceding function word: 127 of the 137 instances of *EH* were cliticized to a preceding function word. The most frequent function words preceding *EH* were the coordinate conjunctions *och* 'and' and *men* 'but' which often have discourse functions, e.g. introducing topic continuations, new topics, etc. 52 cases of *EH* occurred after these two function words. *Och EH* 'and *UH*' was the most common function word+filler construction and was often (in 30 of 38 cases) preceded or followed by an inhalation break, a clear indicator of a speech chunk boundary (see Horne et al., 2006). The second most frequent function word category preceding *EH* was the subordinate conjunction *att* 'that' which also sometimes functions as a discourse marker introducing a non-subordinate clause. 24 instances *EH* occurred after *att*. The other instances of *EH* were found after the following function words: preposition (n=13), articles (n=9), pronouns (Subject) (n=9), basic verbs or auxiliary verbs (n=9), demonstrative article (n=5), indefinite adjective (n=3), subordinate conjunction (other than *att* 'that') (n=2), negation (n=1). Content words preceded *EH* in only 7 cases. Finally, there was 1 case where *EH* was a repetition.

3.2 Phonetic realization of EH in non-initial position

Three basic realizations of the filler EH have been observed: (1) a middle-high front or central vowel: e.g. [ɛ], [e], [ə] (see Fig. 1), (2) a nasalized vowel or vowel+nasal consonant: e.g. [ɛn], [əm], [ə^{n}] (see Fig. 2), (3) a glottalized or creaky vowel: e.g. [ə̰], [ɛ̰] (see Fig. 3).

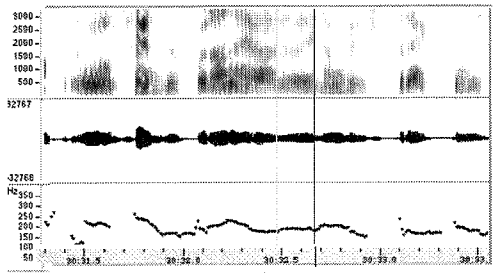

Figure 1. Example of the realization of *EH* as the middle high vowel [ə].

The vowel realizations of *EH* were the most frequent (n=61) and had a mean duration of 268 ms and a SD of 136 ms. The nasalized or vowel+nasal realizations were second in frequency (n=43), and had a mean duration of 436 ms and a SD of 185 ms. These showed a distribution like the vowel+nasal fillers in English that Clark & Fox Tree (2002) analysed, i.e. they were

always followed by other kinds of 'delays', sometimes several in sequence as in Figure 2 with SWALLOW, SMACK, INHALE following *EH*.

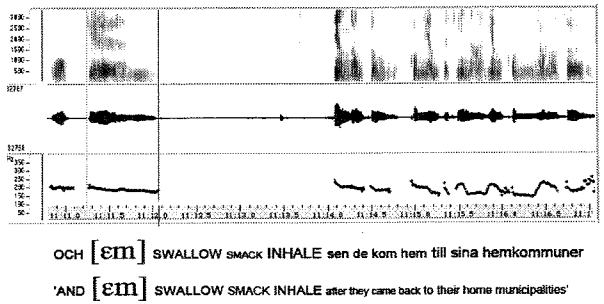

Figure 2. Example of the realization of *EH* as a vowel+nasal [ɛm]. Notice the other delays (SWALLOW, SMACK, INHALE) following [ɛm].

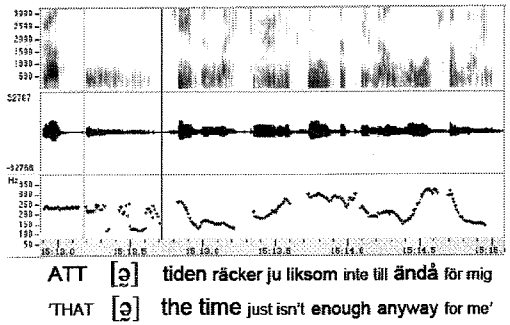

Figure 3. Example of the realization of *EH* as the creaky vowel [ə̰].

The creaky vowel realizations of *EH* were the fewest (n=31) and had a mean duration of 310 ms and a SD of 150 ms. Their duration thus overlaps with the durations of the vowel and vowel+nasal realizations. Unlike the vowel+nasal realizations, the only other delay that was observed to follow the creaky filler was a silent pause. Creaky fillers are in some sense unexpected, since *EH* is often assumed to be a signal that the speaker wants to hold the floor, whereas creak, on the other hand, is assumed to be a signal of finality (Ladefoged, 1982). Nakatani & Hirschberg (1994), however, have observed that glottalization is not uncommon before a speech repair, and thus the creaky *EH* could, therefore, be interpreted as a juncture signal for an upcoming disfluency. Observation of the *SweDia* data shows in fact that the creaky realizations have a tendency to occur before disfluencies, as in the following examples: a) *men den såg ju inte ut [ə̰] det var någon* 'but it did not look [ə̰] it was somebody', b) *å då var det en [ə̰] en kar som heter Hans Nilsson som blev ordförande* 'and then there was a [ə̰] a guy named Hans Nilsson who became chairman'. Creaky fillers also occur in environments where the speaker seems to be uncertain or have problems in formulating an utterance, e.g. *för*

att då blev det ju så att PAUSE [ə] PAUSE Johannesberg det skulle ju läggas ner 'since it happened that PAUSE [ə] PAUSE Johannesberg it was going to be shut down'.

4 Summary and conclusion

This study on the distribution, function and phonetic realization of the filler *EH* has shown that the occurrence of *EH* in the *SweDia* spontaneous speech studied here is mostly restricted to a position following a function word at the beginning of an utterance. This supports and generalizes the findings of Hansson (1998) and Lundholm (2000) who found the filler *EH* most often in utterance internal position after conjunctions/discourse markers in spontaneous speech, both of a monologue and dialogue type. This differs from the findings for simulated task-related dialogues in Eklund (2004), where the filler *EH* occurred almost exclusively in utterance-initial position. This difference is most likely due to the simulated nature of the speech situation where the planning and coding of speech is more cognitively demanding.

As regards the phonetic realization of the filler *EH*, the patterning in Swedish is seen to be partially similar to the findings of Clark & Fox Tree (2002) for English: A vocalic realization of *EH* occurs before shorter delays in speech whereas a vowel+nasal realization correlated with relatively longer delays in speech. The duration of the vocalic realizations in the present data (mean = 268 ms) corresponds rather well with the median duration for *EH* found by Lundholm (240 ms) in spontaneous dialogues; thus, we would expect that the fillers in her data were realized mainly as a vowel such as ([ɛ], [e], [ə]). A third realization, with creaky phonation, whose distribution overlaps with the other two realizations would appear to be associated with relatively more difficulty in speech coding; the creaky phonation, associated with termination, perhaps signals that the speaker has problems in formulating or coding his/her speech, and was observed to sometimes occur before repairs and repetitions. More data is needed, however, in order to draw more conclusive results.

Acknowledgements

This research was supported by a grant from the Swedish Research Council *(VR)*.

References

Allwood, J., 1994. Om dialogreglering. In N. Jörgenson, C. Platzack & J. Svensson (eds.), *Språkbruk, grammatik och språkförändring*. Dept. of Nordic Lang., Lund U., 3-13.

Bruce, G., 1998. *Allmän och svensk prosodi*. Dept. of Linguistics & Phonetics, Lund U.

Clark, H. & J. Fox Tree, 2002. Using *uh* and *um* in spontaneous speech. *Cognition 84*, 73-111.

Eklund, R., 2004. *Disfluency in Swedish: Human-human and human-machine travel booking dialogues*. Ph.D. Dissertation, Linköping University.

Hansson, P., 1998. *Pausering i spontantal*. B.A. essay, Dept. of Ling. & Phonetics, Lund U.

Horne, M., J. Frid & M. Roll, 2006. Timing restrictions on prosodic phrasing. *Proceedings Nordic Prosody IX*, Frankfurt am Main: P. Lang, 117-126.

Ladefoged, P., 1982. The linguistic use of different phonation types. *University of California Working Papers in Phonetics 54*, 28-39.

Lundholm, K., 2000. *Pausering i spontana dialoger: En undersökning av olika paustypers längd*. B.A. essay, Dept. of Ling. & Phonetics, Lund U.

Nakatani, C. & J. Hirschberg, 1994. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America 95*, 1603-1616.

Shriberg, E., 2001. To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association 31*, 153-169.

SweDia 2000 Database: http://www.swedia.nu/.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics 69
Working Papers 52 (2006), 69–72

Modelling Pronunciation in Discourse Context

Per-Anders Jande
Dept. of Speech, Music and Hearing/CTT, KTH
jande@speech.kth.se

Abstract

This paper describes a method for modelling phone-level pronunciation in discourse context. Spoken language is annotated with linguistic and related information in several layers. The annotation serves as a description of the discourse context and is used as training data for decision tree model induction. In a cross validation experiment, the decision tree pronunciation models are shown to produce a phone error rate of 8.1% when trained on all available data. This is an improvement by 60.2% compared to using a phoneme string compiled from lexicon transcriptions for estimating phone-level pronunciation and an improvement by 42.6% compared to using decision tree models trained on phoneme layer attributes only.

1 Introduction and background

The pronunciation of a word is dependent on the discourse context in which the word is uttered. The dimension of pronunciation variation under study in this paper is the phone dimension and only variation such as the presence or absence of phones and differences in phone identity are considered. The focus is on variation that can be seen as a property of the language variety rather than as individual variation or variation due to chance.

Creating models of phone-level pronunciation in discourse context requires a detailed description of the context of a phoneme. Since the discourse context is the entire linguistic and pragmatic context in which the word occurs, the description must include everything from high-level variables such as speaking style and over-all speech rate to low-level variables such as articulatory feature context.

Work on pronunciation variation in Swedish has been reported by several authors, e.g. Gårding (1974), Bruce (1986), Bannert & Czigler (1999), Jande (2003; 2005). There is an extensive corpus of research on the influence of various context variables on the pronunciation of words. Variables that have been found to influence the segmental realisation of words in context are foremost speech rate, word predictability (or word frequency) and speaking style, cf. e.g. Fosler-Lussier & Morgan (1999), Finke & Waibel (1997), Jurafsky et al. (2001) and Van Bael et al. (2004).

2 Method

In addition to the variables mentioned above, the influence of various other variables on the pronunciation of words has been studied, but these have mostly been studied in isolation or together with a small number of other variables. A general discourse context description for recorded speech data, including a large variety of linguistic and related variables, will enable data-driven studies of the interplay between various information sources on e.g. phone-level pronunciation. Machine learning methods can be used for such studies. A model of pronunciation variation created through machine learning can be useful in speech technology applications, e.g. for creating more dynamic and natural-sounding speech synthesis. It is possible to