Phonetic modification of the syllable /tu/ in two spontaneous American English dialogues

N. M. Veilleux[1] *S. Shattuck-Hufnagel*[2]

[1]Boston University, Boston, MA 02139
[2]Massachusetts Institute of Technology, Cambridge, MA 02139

ABSTRACT

In a pilot study of phonetic modification of function words in 2 spontaneous speech dialogues, 99 utterances of the syllable /tu/ corresponding to *to, two, too, -to* and *to-* included the pronunciation variants [t^hu, t^h ə, ɾə, də, nə, sə, s, t^h, ə, t^hʌ]. Factors influencing phonetic modification included phonetic context, prosody, part of speech, adjacent disfluency and individual speaker. 11% of the acoustic landmarks defining /t/ closure, /t/ release and vowel jaw opening maximum were not detectable in hand labelling. In a separate corpus, 59% of recognition errors involved grammatical or function words like conjunctions, articles, prepositions, pronouns and auxilliary verbs, and for 17 tokens of /tu/, half were misrecognized. Implications of these preliminary results for linguistic theory, cognitive modelling of speech processing and automatic speech recognition are discussed.

1. INTRODUCTION

A striking characteristic of fluent communicative speech is the severe phonetic modification that occurs, making the phonetic form of many words, syllables and phonemic segments massively different from single-word citation forms or even from carefully read continuous speech. These modifications can be extreme, but are believed to be systematic with respect to syntax, prosody, phonemic context and other factors. For example, grammatical words such as conjunctions, articles, prepositions, pronouns and auxilliary verbs are particularly likely to undergo severe modification, as when 'give him' is produced as 'giv'm'. This phenomenon poses a challenge to current models and theories in linguistics, cognitive modelling and automatic speech recognition. Lack of comprehensive information about the acoustic-phonetic details hampers efforts to provide linguistic rules that describe the nature of the changes and the contexts in which they will occur, impedes efforts to build cognitive models of speech planning for production and of speech processing for understanding, and causes problems for automatic speech recognition. The problem is especially acute for grammatical or function words because they are are ubiquitous in spoken utterances. The present study is part of an ongoing effort to survey the range of function word modifications in fluent continuous spoken American English, to identify the factors that govern the observed modification patterns, and to estimate the potential effects of incorporating this knowledge into automatic speech recognition systems. The study focusses on the syllable /tu/, corresponding to a number of morphological shapes (e.g. *to, two, too, -to, to-*), because it occurs frequently and is often modified. These characteristics allow us to examine the influence of adjacent phonemic context, intonational phrase structure, disfluency, part of speech and individual speaker, as some of the factors that govern phonetic modification in running speech. In addition, we analyse the potential effect of these modifications on automatic speech recognition by looking at their influence on phonetic landmarks [4] and their correlation with recognition errors.

2. MATERIALS

The speech was drawn from the Callhome database (dialogues recorded during telephone conversations between pairs of speakers who are close friends or family members), and the Switchboard database (dialogues between pairs of strangers discussing an assigned topic); both are available from the University of Pennsylvania Linguistics Data Consortium. Two Callhome dialogues were selected for close acoustic analysis of /tu/ syllables they contained. One conversation (12 mins) involved 2 female speakers, and the other (4 mins) 2 males. Recognition data were analysed for four Switchboard dialogues.

Tokens of lexically-specified /tu/ syllables in the Callhome database included 99 examples: 82 tokens in one dialogue and 17 in the other. The information recorded for each token included: a segmental phonetic transcription of both the target syllable and its immediate context (two labellers resolved disagreements with discussion), the location of the syllable with respect to intonational phrase boundaries and pitch accents (experienced prosodic labellers used the POSH system developed by Price et al. [2]), part of speech, occurrence of an adjacent disfluency and identity of the speaker.

3. PHONETIC VARIATION

Range of variation Utterances of the syllable /tu/ took the forms [*tʰu*] (24), [*tʰ*ə] (46), [**ɾ**ə] (7), [*d*ə] (5), [*n*ə] or nasalized [**ɾ**ə] (10), [sə] (2), [s] (1), [tʰ] (1), [tʰʌ] (1) and [ə] (1), see Figures 1 - 2. Several factors exhibited a systematic influence on the type of modification that occurred.

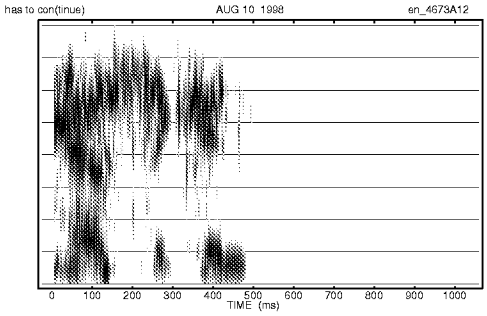

Figure 1: *The /t/ in this token of /tu/ is [s]-like*

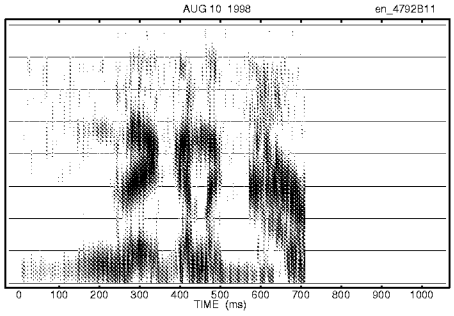

Figure 2: *The /t/ in this token of /tu/ is [n]-like*

Effect of segmental context Initial /t/ could be modified from citation form [tʰ] to [**ɾ**, n] or [s]. All of the [**ɾ**]'s occurred after [r, l, n, m] or a vowel. All of the [n]'s or nasalized [**ɾ**]'s occurred after a nasal, and all of the [s]'s occurred after an [s]. In addition, segmental context influenced the duration of the /tu/. For example, a preceding /t/ or /d/ was associated with a longer /t/ closure, suggesting a gemminate.

Effect of prosodic context Occurrence at the end of an intonational phrase (a break index of 3 or higher in the POSH system) was associated with a longer duration of the vowel /u/ than in phrase medial position. However, tokens of /tu/ associated with a region of disfluency also had longer durations, even when the disfluency did not result in a phrase-final f0 configuration. Although the numbers of tokens are small, this observation suggests that it might be useful to examine the duration lengthening effects of intonation phrase boundaries and disfluencies to see they can be distinguished.

Effect of part of speech Syllables that corresponded to different parts of speech were associated with very different modification rates (although numbers are small.) For example, none of the 9 *two* or *too* tokens were modified from their /tu/ shapes, while 83% of the 90 *to, to-* and *-to* tokens were. The occurrence of disfluency also varied with part of speech: All of the 11 disfluencies occurred in conjunction with *to*, while none occurred with *two* or *too*.

Differences between the speakers For the longer dialogue, the 2 speakers appeared to differ in the rate of flapping (10 vs. 3), and also in the rate of disfluency (7 vs. 2), despite the fact that the two speakers produced roughly proportional numbers of the different morphemes. In addition, the two speakers showed an order of magnitude difference in the range of durations they produced for this syllable. Analysis of a larger database may show whether these speaker differences are inter-related.

Effects on acoustic landmark realization Stevens [4] has proposed that the recognition of words in spoken utterances depends on the ability to locate reliable acoustic landmarks in the speech stream, and use information about the features in these regions of the signal to infer the phonemic segments. These acoustic landmarks can be of several types, including abrupt consonantal landmarks, glide landmarks and vowel landmarks. In citation form, the syllable /tu/ is characterized by an abrupt /t/ closure, an abrupt /t/ release, and a vowel landmark at the maximum amplitude of the /u/. Marking the acoustic landmarks in 99 tokens of /tu/ by hand showed that 34% of /t/ closures were missing, as were 24% of the /t/ releases, but only 2% of the vowel landmarks were missing. When a closure landmark was missing, the preceding segment (e.g. a /d/) was unreleased, or the /t/ apparently assimilated to the preceding segment (as in e.g. *wanna*). The vowel landmark was missing when the /t/ assimilated to a following /yu/ (as in *to you*) or to a preceding /s/ (as in *has to*). Further analysis will show whether, in cases of apparent vowel deletion, duration cues the missing nucleus.

In addition to losing landmarks, /tu/ syllables sometimes were realized with a different set of landmarks. For example, when the /t/ was realized as a [ɾ], the landmark was a non-abrupt minimum in amplitude. An automatic recognition system based on landmark- related cues to lexically contrastive features will need to take account of the modification processes that change or obscure the observable landmarks.

4. RECOGNITION RESULTS

Recognition Error Rates for Function Words Four dialogues from the 1996 Automatic Speech Recognition evaluation Switchboard corpus were used for a preliminary examination of ASR performance on spontaneous speech, using as a measure the relative word error rates for function words. Since there is no consensus on a definition of function words, a conservative list of 267 pronouns, auxilliary verbs, prepositions and adverbs was compiled for use in this study. The sheer number of function words combines witha generally high mis-recognition rate to contribute substantially to the low word recognition rate in automatic systems. For example, in these four Switchboard dialogues, 59% of the mis-recognized words were members of the function word list (37% of the mis-recognized items were content words and 4% were filled pauses or disfluencies.) Although the error rate for function words was similar to that for content words (36% vs. 35%), the greater number of function words contributes to the greater number of errors (787 function words were mis-recognized, compared to 499 content words, from a total of 3572 words). It can be argued that mis-recognitions of content words and function words have different causes, even though the rates are similar. Function words are common[1] and so are expected to be well represented in training data, used to derive parameters for stochastic models. When there are few training tokens of a particular item, the model parameters may not be representative of the variety of possible instantiations, and are therefore not robust in mapping to unseen data. But function words provide abundant training tokens. Thus, errors on function words are errors on extensively-trained models, while errors on content words reflect minimal training. The function word models may represent a compromise among different pronunciations, so that despite their extensive training they do not correspond well to the signal in any particular utterance.

Recognition Results for /tu/ Four minutes of the shorter Callhome dialogue were part of the DARPA 1996/1997 recognition test and development sets, so that Nbest speech recognition data from the BBN/BU recognizer was available. The top hypothesis was used as the recognizer output string.

About half (53%) of the tokens were mis-recognized. The recognizer performed equally well across the limited set of phonetic and syntactic contexts examined. However, as shown in Figure 3, as the duration of the [ə] in 14 tokens increases, the performance of the recognizer appears to improve. Of the 7 tokens with a vowel

[1] Together, the words I, AND, THE, YOU, TO, A, THAT, OF, IT, KNOW, and IN make up 19.9% of all word tokens in a 1995 switchboard training set. YEAH, UH and KNOW are the only non-function words with over 1% of the data.

duration less than 40 ms, only one was correctly recognized. Of the remaining 7 tokens with longer [ə] durations, all but two were correctly recognized; these 2 were lengthened preceding a disfluency.

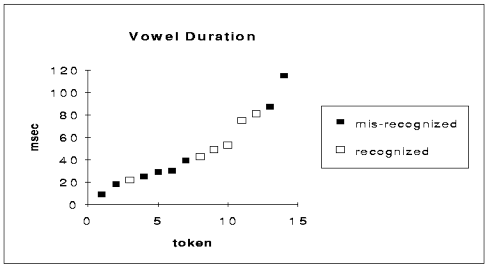

Figure 3: *Comparison of [ə] duration in tokens of /tu/. The 2 rightmost tokens are disfluent.*

Although none of the fluent /tu/ tokens in this dialogue were phrase-final, the differences in duration may be linked to a more fine-grained phrasing structure [7], represented in the POSH labeling system as the difference between word junctures labeled with 0-2 break indices [2]. How these break indices relate to other linguistic structures such as clitic groups, prosodic words and phonological phrases is an open research question.

The three [t^h u] tokens had longer vowel durations (64 - 229 ms), as expected from the intrinsically longer [u]. The 229 ms vowel was the only pre-disfluency [t^h u] that was correctly recognized, possibly because of a strong language model (*looking forward* might predict *to*), or other phonetic factors that will be explored in future work.

5. DISCUSSION

Implications for automatic recognition Although word error rates have improved steadily on the Switchboard corpus to approximately 40%, they still compare unfavorably with the 3-5% rate for read speech. One of the often-cited reasons for the degradation of ASR performance is the prevalence of severely modified, even deleted, phonemes [1] [3] or pronunciation variability, which may reflect syntactic and prosodic context dependencies. A joint acoustic / syntactic / prosodic model of speech has been proposed [5] and successfully used to leverage prosodic information in automatic speech recognition tasks [6]. This framework can be extended to include phonetic modification features in the acoustic model.

Implications of these results for landmark-and-feature-based models of lexical access are also significant. It may be necessary to 'parse' a string of landmarks into a set of segments even when some expected landmarks are missing. For example, the long closure durations for some [dt] and [tt] sequences, where only one closure and one release landmark were observed, might suggest two segments even if not all 4 landmarks are found. There are additional implications for speech synthesis and cognitive processing models.

References

1. Cohen, J. 1992. The Summer of our Discontent, *International Conference on Spoken Language Processing* , Philadelphia.

2. Price, P., Ostendorf, M., Shattuck-Hufnagel,S., Fong, C., 1991. The Use of Prosody in Syntactic Disambiguation, *Journal of the Acoustical Society of America* 90, 2956-2970.

3. Jelineck, F., 1996. Summary of LVCSR Workshop, *CLSP/JHU Workshop on Innovative Techniques for Large Vocabulary Continuous Speech Recognition,* John Hopkins University, Baltimore.

4. Stevens, K., Manuel, S.Y., Shattuck-Hufnagel, S., Liu, S., 1992. Implementation of a Model for Lexical Access Based on Features, *International Conference on Spoken Language Processing* , Banff, I, 499-502.

5. Veilleux, N. 1994. Computational Models of the Prosody/Syntax Mapping for Spoken Langua ge Systems Ph. D. Thesis, Department of Electrical, Computer and Systems Engineering, Boston University.

6. Veilleux, N. 1996. Stochastic Models of Prosody for Automatic Spoken Language Systems *Proceedings of the Acoustical Society of America* Honolulu, Hawaii, 2849.

7. Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M. & P. Price. 1992 “Segmental Durations in the Vicinity of Prosodic Phrase Boundaries.” *Journal of the Acoustical Society of America 91,* 3, pp. 1707—1717.

This work is supported by NIH RO1-DC02125/ RO1-DC02978 and NSF Sub. No. C-Q0198