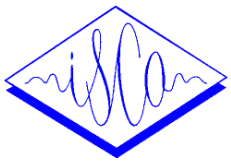

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

STUDIES OF VOWEL AND CONSONANT REDUCTION

Sharon Y. Manuel [1,2], Stefanie Shattuck-Hufnagel[2], Marie Huffman [2,3], Kenneth N. Stevens[2], Rolf Carlson[4], and Sheri Hunnicutt[4]

[1]Dept. of Communication Disorders & Sciences, Wayne State University, Detroit, MI
[2]Massachusetts Institute of Technology, Cambridge MA
[3]Currently at Eloquent Technology, Inc., Ithaca NY
[4]Royal Institute of Technology, Stockholm

ABSTRACT

In normal (casual) speaking modes, speakers often modify, or seemingly delete, segments that are produced in citation forms of the same words. This paper discusses three examples of how attention to the acoustic detail of spoken language can reveal aspects of the articulation which are perhaps not readily apparent in more cursory examinations of the speech signal. A lexical access model which is sensitive to such acoustic detail will find a better match between both normal and citation spoken forms and their shared abstract representation, than one which is not.

I. INTRODUCTION

Recognizing words in natural speech can be difficult, partly because speakers often produce renditions that differ from citation forms. An extreme form of this modification, called *neutralization*, can result in disappearance of the critical difference between two lexical items or strings, e.g. *support* and *sport*, *below* and *blow*, or *in the* and *in a*. In developing a lexical access model [1] that relies on fine-grained acoustic analysis informed by a detailed understanding of the acoustic consequences of articulatory movements, we have examined several types of reduction and apparent neutralization. We find that timing and spectral cues reveal the effects of certain of the articulatory aspects that distinguish between these segments, even when some of their normally more robust cues are absent.

II. SOME MODIFICATIONS OF /ð/

We first consider the case of the English voiced dental (or interdental) fricative /ð/. In normal speech, the phoneme /ð/ is particularly susceptible to contextual modifications such as (at least partial) assimilation to preceding consonants [2,3] or deletion. This susceptibility is presumably due to a combination of factors, any one of which is known to lead to weakening of phonemes. First, /ð/ is a voiced non-strident fricative, and in fact in all but the most emphatic speech, the frication noise created at the dental constriction is very weak in amplitude. Second, /ð/ is a coronal, and in English (as in many other languages), coronals are prone to assimilation and other weakening processes. Finally, /ð/ has a very limited distribution in English. It primarily occurs in word-initial position in function words, e.g. *they*, *them*, *those*, *then*, and, most commonly, *the*. Function words are generally unstressed and subject to weakening processes. It is true that /ð/ also occurs medially in a number of content words (*bother*, *father*, *brother*, *weather*), and word-finally in a few words such as *seethe*, *teethe*, *bathe*. However, in these cases /ð/ is preceded by a vowel, and here we are interested in what happens to /ð/ when it is preceded by a consonant. Consequently, our comments will be restricted to word-initial /ð/ as it appears in function words.

Assimilation of /ð/ to a preceding /n/, for example, could lead to the neutralization of underlying contrasts such as the difference between *one nose* and *won those*. Deletion of /ð/ could neutralize the difference between *in a* and *in the*. If we look at spectrograms of *in the*, the region between the /I/ and the /ə/ appears to be completely nasal, suggesting that the /ð/ has either deleted or assimilated to the preceding /n/. An example is shown in Fig 1. But are the underlying differences totally neutralized? That is, is /ð/ completely deleted (or completely assimilated) in these cases? Our impression as speakers and listeners is that the /ð/ has not lost all of its phonetic properties. What types of objective evidence can we bring to bear on this issue?

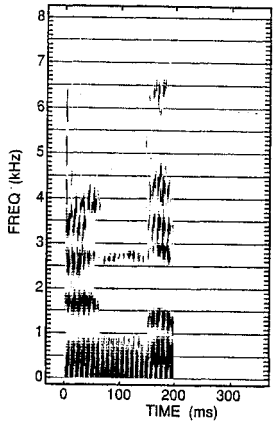

*Fig. 1 Spectrogram of the utterance **in the**. The consonant region appears nasalized throughout.*

We first examine the contrast between the utterances *in a* and *in the*. Two facts suggest that the /ð/ has not been deleted. First, deletion of the /ð/ would make the /ə/ of *the* word-initial. Word-initial vowels are sometimes produced with a glottal attack, and we never see this for the vowel of *the*. Second, if the /ð/ had been deleted, we would expect the /n/ to be produced as a nasal flap, as /n/ usually flaps when followed by an unstressed vowel. On the other hand, *in the* has a long nasal consonant duration. We used sentences from the TIMIT database to compare the duration of intervocalic consonant portions for *in a* and *in the*. This database contains a variety of sentences read at self-selected rates by a number of speakers. The results are shown in Fig. 2 in the form of a histogram. The intervocalic consonant region was generally quite short for *in a*, but long for sequences that had an underlying ***the*** following the /n/. There is very little overlap in the two distributions of the duration of the consonant region. For *in a*, the consonant region was almost always less then 55 msec, and with one exception, for *in the* it was always 55 msec or longer. It seems likely that listeners (or a machine) could use the duration of the nasal murmur to fairly reliably distinguish between *in a* and *in the* (for an explicit test of this possibility, see [2]).

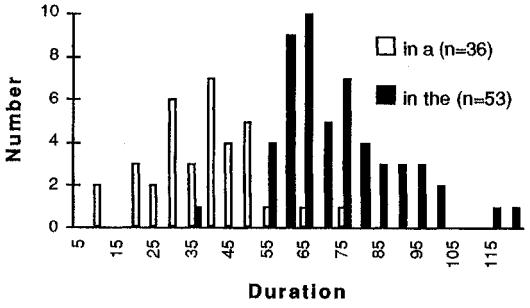

*Fig. 2 Histogram showing duration of consonant intervals for **in a** and **in the** in read sentences.*

We played these sentences to seven naive listeners, providing them with a written script for each sentence. For the relevant portion of each sentence, the script included a choice between *a* and *the*. We asked the listeners to indicate which determiner the speaker had used. Out of 644 responses, there were only 6 errors. Of course, listeners might have been relying on information other than, or in addition to, duration.

We now consider whether the /ð/ completely assimilates to the preceding /n/, or only partially assimilates. In order to avoid the issue of flapped /n/, it is helpful here to compare /Vn#nV/ and /Vn#ðV/ sequences. We have analyzed a number of acoustic recordings, made in our laboratory, of sentences with such sequences. Examination of spectrograms and waveforms of minimal pairs suggests the /ð/ in /Vn#ðV/ is, for the most part, nasalized. Yet, when we listen to these sentences we have a strong impression that the speaker produced a /ð/. This leads us to suspect that assimilation of /ð/ to the preceding /n/ is incomplete.

In thinking about what would constitute acoustic evidence of incomplete assimilation, we consider the expected articulatory configurations associated with both /n/ and /ð/ in intervocalic position, and with the expected acoustic consequences of these articulatory configurations. While /n/ and /ð/ are both made with the tongue blade, /n/ is made at the alveolus, whereas /ð/ is made more forward, at or between the teeth. One expected acoustic correlate of a dental versus alveolar place of articulation is that F2 should be lower in frequency going into and out of a /ð/ constriction than it is going in and out of an /n/ constriction [4]. So if the /ð/ in a /Vn#ðV/ sequence has not assimilated in place to the preceding /n/, and is still somewhat dental, we would expect a lower F2 at the consonant-vowel release than we would find in a /Vn#nV/.

We have begun to test this expectation with minimal pairs such as *win those* and *win no's*. We measured the F2 frequency at the release of the consonant into the second vowel. Some preliminary results are shown in Fig. 3. These data are averaged over 5 to 8 tokens, and data are shown separately for different speakers and different minimal pairs, because F2 will be dependent not only on place of articulation, but also on individual speaker and vowel context. The important point is that for each set of data, at the consonant release F2 is always considerably lower for the utterance with an underlying /ð/.

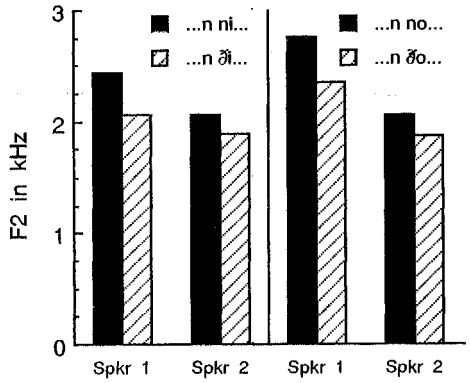

Fig. 3 F2 at release of second consonant. F2 is always lower for underlying /ð/ than for underlying /n/, indicating that /ð/ has maintained its dental place.

It appears that these speakers had a more dental constriction at the end of the consonant sequence in /Vn#ðV/ than they did in /Vn#nV/. That is, the /ð/ did *not* assimilate in place to the preceding /n/. Furthermore, if we look more fully at F2 at the moment of consonant implosion, there is evidence that the tongue is more fronted even at the *beginning* of the consonant sequence for /Vn#ðV/ than it is for /Vn#nV/. It would appear that the /n/ has at least partially assimilated in place to the following /ð/, rather than the other way around (see also [2, 5]).

Another difference between /ð/ and /n/ has to do with the abruptness of the release of the constriction. Nasal and oral stops, which are made with complete oral constrictions, generally have more abrupt releases than do fricatives, which are made with incomplete oral constrictions. Even for full stops, dentals are expected to have slower releases than alveolars [4]. If /ð/ maintains its fricative-like degree of constriction and dental place, in /Vn#ðV/ sequences, we would expect to see a less-abrupt release into the second V than we see in /Vn#nV/ sequences. In spectrograms of a number of such utterances produced by several speakers, the F1 rise at release generally does appear to be less abrupt for /Vn#ðV/. As an example, in Figure 4, we show F1 frequencies for /n/ and /ð/ releases into the vowel /o/. The measurements were made from spectra calculated from a 5 msec window centered on consecutive pitch pulses. F1 rises more gradually for the consonant release in the utterance *win those* than in *win nose*.

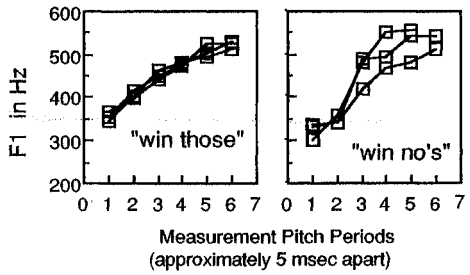

Fig. 4 F1 at release of second consonant. F1 rises more gradually for the utterance with an underlying /ð/ .

Normally, of course, the most robust articulatory difference between /n/ and /ð/ is the position of the velum. For intervocalic /n/ the velum lowers in the preceding vowel, begins to rise during the /n/, but is still somewhat low as the oral constriction for the /n/ is released into the following vowel. In contrast, the velum is normally relatively high throughout a /Vðv/ sequence. The acoustic consequences of nasalization of vowels and consonants are complicated, but in part involve the appearance of extra poles and zeros, and also the dampening of formant bandwidths. As we have noted, the entire /nə/ interval in a /Vnðv/ appears to nasalized. But, if the velum were to raise somewhat earlier in /Vnðv/ sequence than in a /VnnV/ sequence, we might expect to see a diminution of nasalization effects toward the end of the consonant interval or relatively early in the vowel. We have just begun to explore this possibility, and hope to report on results in the oral version of this paper.

Both the perceptual and acoustic results support our intuitions about our own articulation, and indicate the following scenario when speakers produce /Vn#ðV/ sequences. The tongue initially makes a tight dental constriction (rather than the alveolar constriction made in /Vn#nV/). This constriction is then slightly weakened and is slowly released. The initial portion of the consonant region retains aspects of the normal constriction degree for the nasal /n/, while the final portion seems to reflect the normally expected constriction degree for the fricative /ð/. The velum is low during the /n/ and continues to have a low position throughout most of time in which the oral constriction is maintained. In a feature-based model [6, 7] we would say that it is primarily the feature [+nasal] which has been assimilated by the /ð/; its other features remain as they would in a non-nasal context. Using a coproduction model [8], we would say that the oral gestures for the /n/ and /ð/ have incompletely overlapped, and that the velar lowering gesture for the /n/ more completely overlaps with the /ð/.

III. VOWEL REDUCTION

We turn now turn to examples of vowel reduction. As is well known, in English, unstressed vowels like the first vowel in the words *support* and *below* are often reduced in duration, and may sometimes appear to be deleted. If these vowels are completely deleted, leaving no traces, then *support* might be expected to become homophonous with *sport* and *below* homophonous with *blow*.

Apparent deletions of this sort can be explained in gestural overlap models [8] as being the result of the relative timing of

the release of the first consonant and the implosion of the second consonant. If the constriction for the second consonant is achieved prior to the time at which the constriction of the first consonant is released, then there will be no period in which the oral tract is unconstricted. Therefore, the primary aspects of the underlying vowel will be missing.

Vowel reduction between two voiceless phonemes.

The case of words like *support* is particularly interesting to us because both of the relevant consonants are voiceless and made with an open glottis, whereas the intervening vowel is made with a narrowed glottis suitable for voicing. Furthermore, as the /p/ is syllable-initial, it is aspirated. In terms of oral-glottal coordination, this aspiration results from the fact that the glottal opening gesture is timed so that the glottis is maximally open at the release of the labial constriction for the /p/ [9,10]. This type of oral-glottal coordination contrasts with the type seen in syllable-initial /sp/ clusters (e.g. *sport*). In the latter case, the glottis has a single opening and closing movement, and the glottis is fairly well closed by the time the /p/ is released. A question arises as to what happens to the glottal gestures for *support* when the speaker times the oral gestures so that the /s/ and /p/ are contiguous. Elsewhere [11,12] we have reported on the results of relevant acoustic and perceptual studies we have done on the *support - sport* pair. Here we will highlight the major results.

The acoustic analysis was based on the words *support* and *sport* produced in careful, reading, conversational and very fast speaking modes by four speakers. For *sport*, the labial closure for the /p/ was produced while the /s/ constriction was still in place. The acoustic evidence for this was a strong labial tail on the /s/ frication, and/or the lack of any voiced or voiceless formants between the /s/ frication and the /p/. The /p/ release was always characterized by a very brief period of aspiration. An example is shown in Fig 5a.

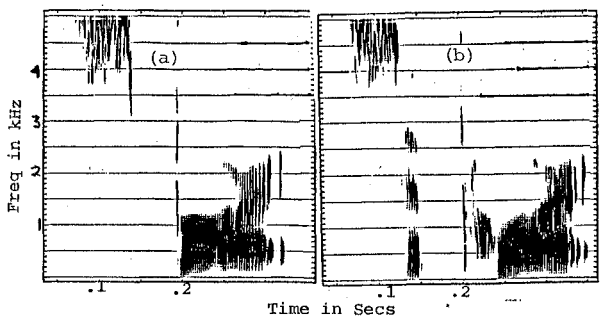

Fig 5 Spectrograms of ***sport*** *(a) and* ***support*** *(b).*

In most of the tokens of the word *support*, there was evidence of a period of oral tract opening between the /s/ and the /p/, as shown in Fig. 5b. However, this interval was often quite brief, and sometimes nonexistent, especially in fast or conversational speech. We show some examples in Figs. 6-8.

In the token shown in Fig. 6, there is a brief period of aspiration and then a short voiced schwa between the /s/ release and the /p/ closure. The aspiration indicates that the alveolar constriction has been released, and the pitch periods following the aspiration show that the glottis is fairly well adducted. The /p/ release is followed by a long interval of aspiration.

For some tokens, we saw no voicing after the /s/. In Fig. 7a we show a token which has only a brief period of aspiration after the /s/. Fig. 7b shows spectra made during the /s/ frication (dotted line) and during the aspiration noise (solid line) which follows the /s/ frication. There is a strong peak at 2200 Hz in the post-frication noise, and this is an indication that the sound source is post-alveolar, rather than alveolar. So while it may first appear that no vowel had been produced, in fact the /s/ released into an open vocal tract. Again, in this token, the /p/ release is heavily aspirated.

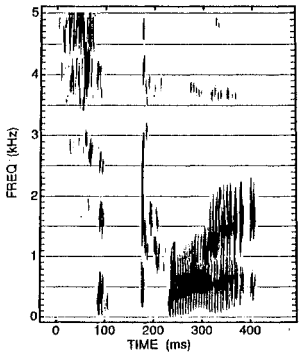

Fig. 6 The word ***support*** *produced with a short schwa.*

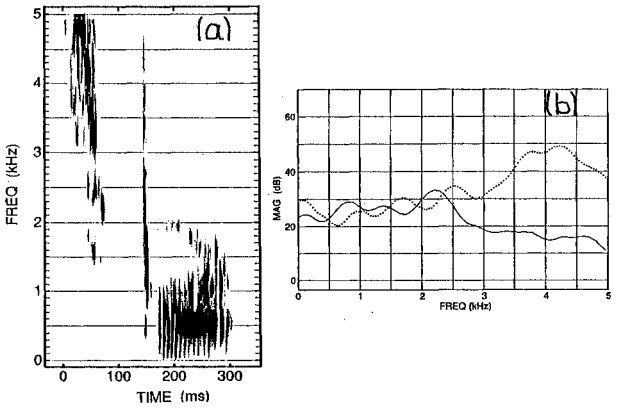

Fig. 7 The word ***support*** *produced with aspiration between the /s/ and the /p/. The left panel (a) shows the spectrogram. The right panel (b) shows spectra calculated during the /s/ frication (dotted line) and the post-frication aspiration (solid line).*

In some tokens there was a brief period of voicing after the /s/ release. Initially we had assumed this voicing was associated with an open oral tract - a vowel. However, as shown in Figure 8a and 8b, the spectrogram and power spectra of some tokens show very little energy in the higher frequencies for this interval. We conclude that these are examples of voice bars and that the speaker in fact moved directly from the /s/ oral constriction into the /p/ oral constriction. The glottal narrowing gesture, however, was retained for the underlying vowel, as indicated by the voice bar. The /p/ in this case is aspirated, showing that the same type of oral and glottal relationship exists here as exists in careful speech.

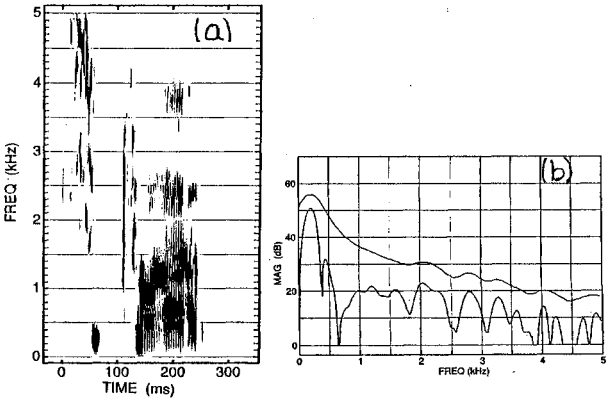

Fig. 8 The word ***support*** *produced with a voice bar between the /s/ and the /p/. The left panel (a) shows the spectrogram. The right panel (b) shows spectra calculated during the voiced interval between the /s/ and the /p/.*

These data suggest that speakers may allow quite a bit of variability in the timing of the /s/ release and the closure for the /p/ in *support*, even going to the extreme of making the /s/ release directly into the /p/ closure. However, the glottal gestures for the underlying schwa are maintained. Furthermore, the speakers have the same kind of glottal-oral timing for the /p/ (/p/ is always aspirated) in casual as in careful speech.

The acoustic consequences of these production patterns are such that *support* and *sport* remain acoustically distinct. Several experiments show that aspiration of the /p/ has cue value in signaling *support* [13]. Our own work shows that listeners are sensitive not only to the aspiration after the /p/ release, but also to the presence of aspiration or a voice bar after the /s/.

Vowel reduction between two voiced phonemes.

Our final example of reduction involves the behavior of the initial schwa vowel in words like *below* and *derive*. Again, apparent deletion can come about if the speaker happens to achieve the second consonantal constriction before the initial consonant is released. If this happens, these words will potentially be homophonous with *blow* and *drive*, respectively.

As we have seen, when the oral gestures of the schwa are absent, as in some productions of words like *support*, glottal gestures and their coordination to oral gestures can remain as cues to the underlying disyllabicity of the word. However, in words like *below* and *derive*, the underlying schwa vowel is surrounded by two voiced phonemes. In fact, the glottis should be in a narrowed, voicing state throughout *below* and *blow*, and *derive* and *drive*. If the oral gestures for the schwa are essentially omitted, we might expect the difference between *below* and *blow*, and *derive* and *drive*, to be completely neutralized.

However, this does not seem to be the case. We examined acoustic recordings of a number of minimal pairs. In each pair, the initial consonant was a voiced stop which was either followed immediately by a sonorant (/l/ or /r/), or by a stressless schwa and then /l/ or /r/. These recordings were from 8 speakers reading a series of short paragraphs. In many cases, the initial schwa vowel in words like *below* appeared to be omitted. We compared the durations of the sonorant constrictions for these words with the duration of the sonorant constrictions of words without an underlying schwa. In almost all cases the sonorant duration was longer for words with an underlying schwa. Work by others [14,15] has shown that duration of such sonorants can have important cue value for the /CVl/ versus /Cl/ and /CVr/ versus /Cr/ distinctions.

What of the cases in which there does not appear to be an added sonorant duration? Are they in any way distinct from the way they would have been had they not contained an underlying schwa? One possibility is that in careful speech, the actual articulation of syllable-initial /l/ and /r/ is somewhat different than the articulation of those phonemes in syllable-initial clusters. For example, the /l/ in *below* may have a different tongue body position, and a somewhat different release, than the /l/ in *blow*. If so, and if these differences are maintained despite the fact that in casual speech the /l/ of *below* may be achieved before the /b/ is released, *below* and *blow* would remain articulatorily distinct. We are currently examining the acoustics of these words to try to determine if such differences can be found. For some speakers, there do appear to be spectral distinctions.

IV. DISCUSSION

Our results demonstrate that while normal speaking modes may produce forms which are rather different than citation forms, speakers often do not alter their speech as radically as it might at first appear. A detailed "microanalysis" of the acoustic signal, informed by theories of expected articulatory-to-acoustic relationships, reveals that a number of robust durational and (sometimes subtle) spectral cues remain constant across speaking modes. Such cues will presumably be useful in lexical access models, particularly those which are based on phonetic and acoustic features [1].

Acknowledgments

This work was supported by grant IRI-8910561 from the National Science Foundation. Assistance was provided by Anders Lofqvist and Haskins Laboratories through NIH grant DC-00865. We would like to thank Peggy Li for help with data analysis and Lorin Wilde for editorial support.

References

[1] K. N. Stevens, S. Shattuck-Hufnagel, S. Y. Manuel, and S. Liu. Implementation of a model for lexical access based on features, *This Volume*.

[2] L. Shockey. Perceptual test of a phonological rule. *Haskins Status Report on Speech Research* SR 50, pp 147- 150, 1977.

[3] A. C. Gimson. *An Introduction to the Pronunciation of English*. London: Edward Arnold. 1989.

[4] K. N. Stevens, S. J. Keyser, and H. Kawasaki. Toward a phonetic and phonological theory of redundant features. In J. S. Perkell and D. H. Klatt (eds.) *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Lawrence Erlbaum, pp. 426-449, 1986.

[5] V. W. Zue and S. Shattuck-Hufnagel. The palatalization of alveolar fricatives in American English. *Proceedings of the IXth International Congress of Phonetic Sciences*, Vol. 1, pp. 215, 1979.

[6] N. Chomsky and M. Halle. *The Sound Pattern of English*. New York: Harper and Row, 1968.

[7] G. N. Clements. The geometry of phonological features, *Phonology Yearbook*, 2, pp. 223-52, 1985.

[8] C. P. Browman and L. Goldstein. Gestural specification using dynamically defined articulatory structures. *J. of Phonetics*, 18, pp. 299-320. 1990.

[9] H. Yoshioka, A. Lofqvist, and H. Hirose. Laryngeal adjustments in the production of consonant clusters and geminates in American English. *J. Acoustic. Soc. Am.*, 70, 1615-1623.

[10] K. Munhall and A. Lofqvist. Gestural aggregation in speech: laryngeal gestures. *J. of Phonetics*, 20, pp. 111-126, 1992.

[11] S. Y. Manuel. Recovery of "deleted" schwa. *Perilus*, XIV, pp. 115 - 118, 1991.

[12] S. Y. Manuel. Vowel reduction and perceptual recovery in casual speech. *J. Acoust. Soc. Am.*, 91: 4, Pt 2, p. 2388, 1992.

[13] J. Fokes and Z. Bond. Perception of syncope in native and non-native American English. *Proceedings of XII International Congress of Phonetic Sciences*, V, pp. 58-61, 1991.

[14] P. J. Price. Sonority and syllabicity: Acoustic correlates of perception. *Phonetica*, 37, 1980.

[15] S. F. Taub. The effect of rate of speech on the perception of syllabicity. *J. Acoust.. Soc. Am.*, 88 S1, p. 128, 1990.