

Proceedings of Meetings on Acoustics

Volume 2, 2008

<http://asa.aip.org>

154th Meeting
Acoustical Society of America
New Orleans, Louisiana
27 November - 1 December 2007
Session 2aSCb: Speech Communication

2aSCb8. Investigating the consonant-vowel boundary: Perceptual contributions to sentence intelligibility

Daniel Fogerty and Diane Kewley-Port

Speech sounds divide into two broad categories: vowels and consonants. Vowels have a distinct perceptual advantage over consonants in determining sentence intelligibility [Kewley-Port, Burkle, and Lee, J. Acoust. Soc. Am. (2007)]. However, specifying the segmentation of consonants and vowels is problematic because of their extensive acoustic overlap due to coarticulation. The current study used TIMIT sentences to investigate perceptual contributions of consonants and vowels across systematic changes in the consonant-vowel (C-V) boundary. Sentences were presented to listeners with either consonant or vowel information replaced by noise. The C-V boundary was shifted by six specific proportions of the vowel, such that consonant duration increased while vowel duration decreased, yielding 12 different noise replacement conditions. The percent of words listeners repeated correctly was scored. A two-to-one vowel advantage for intelligibility at the traditional C-V boundary was confirmed. Initial results from ten listeners suggest that functions of the shifted C-V boundaries for consonants and vowels differentially contribute to intelligibility. Preliminary comparison of these functions suggests that both consonants and vowels contribute equally (50%) to intelligibility at about a 20% proportional boundary shift. Results will be interpreted in terms of phonological versus acoustic accounts of speech perception. (Supported by NIH)

Published by the Acoustical Society of America through the American Institute of Physics

Investigating the consonant-vowel boundary: Perceptual contributions to sentence intelligibility

Daniel Fogerty and Diane Kewley-Port
Department of Speech and Hearing Sciences, Indiana University
200 S. Jordan Ave, Bloomington, IN 47405
dfogerty@indiana.edu

I. INTRODUCTION

A primary feature common to the sounds of all languages is the use of both consonants and vowels (Ladefoged, 2001). However, the division of speech sounds into these different units of speech based upon the acoustic waveform is somewhat unclear. The assignment of formant transitions to either the consonant or vowel segment is particularly problematic, as these transitions provide information about both consonants and vowels (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Coarticulation creates acoustic cues for consonants and vowels that largely overlap, the perceptual affects of which may be observed over multiple segments (see Kent & Minifie, 1977). However, separation of the meaningful sounds in language into the categories of consonants and vowels is one of the most fundamental principles of how speech is structured (Ladefoged, 2001; Stevens, 2002). While a discrete division between consonants and vowels may be largely a convenience (see Ladefoged, 2001), a view the authors of this paper share, a more detailed analysis of the perceptual consequences for the placement of segmental boundaries will help illuminate the roles that consonants and vowels play in speech perception. Stevens (2002) suggests that prominent acoustic landmarks might be used perceptually to discretely segment the ambiguous speech signal. While some researchers believe that the discrete perceptual segmentation of units is necessary for speech perception, as Blumstein and Stevens (1979) made the case for using spectral templates for stop consonant identification, some have argued for more dynamic accounts (see Strange, 1987). Cooper *et al.* (1952) were among the first to explore in stop consonants the relation between perception and acoustic cues, such as the frequency release of the burst and the direction of the second formant transition. They described the perception of these specific cues as dependent upon their positional relationship to the neighboring vowel. However, the perceptual contribution of consonant and vowel portions remains unclear.

Historically, consonants have been the focus of speech intelligibility research, primarily using methods investigating words or nonsense CVCs. While the speech sound class of consonants represents a heterogeneous category of many different types of sounds, the high frequency, low intensity portion of many consonants (such as stops and nasals) makes them easy targets to mask in natural environmental settings. Indeed, Stevens (2002) cited the low intensity nature of many consonants as one of the defining features that distinguishes them from vowels. Given that individuals with sloping, high-frequency hearing loss typically perform worse on consonant recognition tasks because of these acoustic properties of consonants, it is commonly believed that consonants carry large amounts of the information for speech intelligibility, particularly when only considering CVCs or minimal pairs. Perhaps because of some of these

considerations, Miller (1951) concluded that consonants contribute more to sentence intelligibility than vowels. Research on consonant-vowel ratios, typically with CVCs, have demonstrated that amplifying the intensity of the consonant relative to the vowel enhances recognition of voiced stops (Freyman & Nerbonne, 1989; Freyman, Nerbonne, & Cote, 1991). However, Sammath, Dorman, and Stearns (1999) reached an opposite conclusion using their stimulus materials which controlled for equal audibility of the consonant. In addition, Owren and Cardillo (2006) demonstrated that words with only consonants present (the vowels having been excised and replaced by noise; consonant-only condition), contribute more to word meaning, while words with only the vowels preserved (and the consonants replaced by noise; vowel-only condition) contribute more to indexical properties of the speaker (see Figure 1 for waveforms of a CVC excised from a sentence generated in our study using a similar noise replacement technique). Therefore, evidence does suggest an important role for consonants in speech intelligibility, at least at the word level.

However, research at the sentence level has revealed a very different picture. In a preliminary report, Cole, Yan, Mak, Fany, and Bailey (1996) used a segment replacement paradigm (see Figure 2) to investigate the contribution of consonants and vowels to sentence intelligibility under three replacement conditions. Either the consonants or vowels in the sentence were replaced with speech-shaped noise, a harmonic complex, or silence. Their results suggested that vowel-only sentences maintained a two-to-one advantage over consonant-only sentences when measuring the percent of words repeated correctly. This advantage remained the same regardless of the replacement condition. Furthermore, a substantial vowel advantage was observed, even when 10ms was deleted from the onset and offset of the vowels.

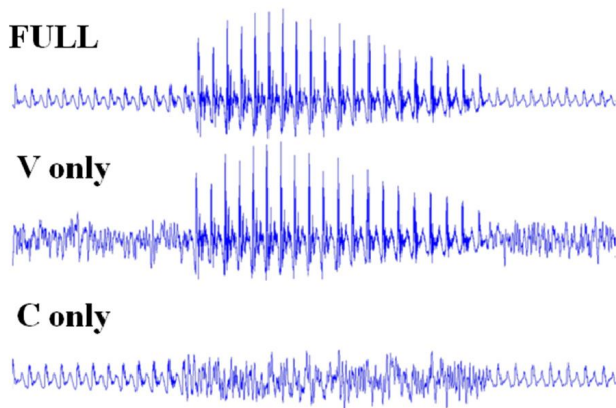


Figure 1. Waveforms demonstrating two noise replacement conditions and the full waveform of the word “mean”. V only = vowel-only, C only = consonant-only.

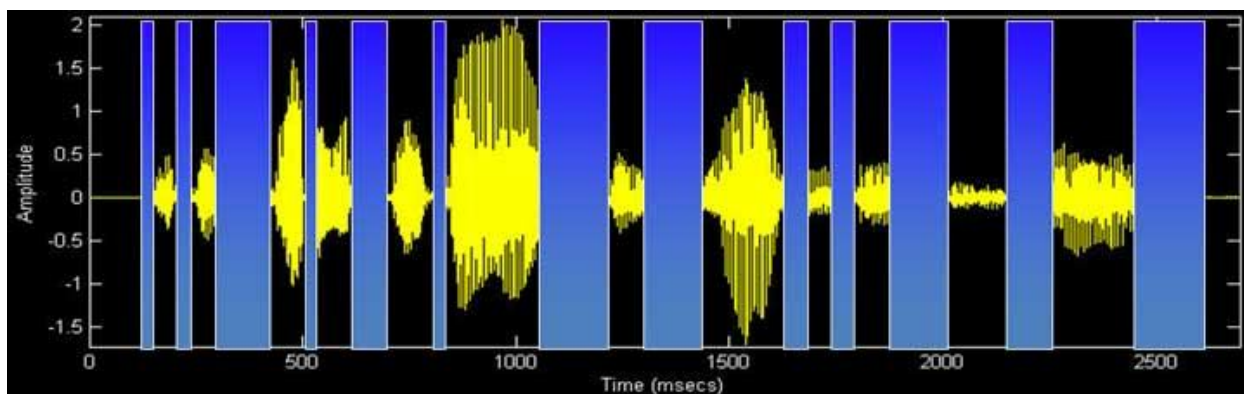


Figure 2. Speech waveform of an experimental sentence schematizing the segments where noise replaced all consonants (blue bars). Only the speech information between these noise replacement bars is available for identifying the words in the sentence.

Previous work in our lab confirmed these preliminary results from Cole *et al.* (1996). When only vowels are presented in sentences, vowels have a two-to-one advantage over when only consonants are present (Kewley-Port *et al.*, 2007). Thus, it appears that vowels (as traditionally defined) contribute more information to speech intelligibility than consonants. This work has led to several questions, some of which were addressed in the present study, and are listed below.

- 1) How does shifting the consonant-vowel (C-V) boundary change the perceptual intelligibility of sentences when only consonant or vowel information is available?
- 2) How might analysis of different linguistic levels describe these perceptual changes?
- 3) What placement of the C-V boundary yields equal perceptual contributions of vowels and consonants?

The present study investigates these questions in sentences using a noise replacement paradigm, similar to that used by Kewley-Port *et al.* (2007) and Lee and Kewley-Port (2006). It was designed to explore the perceptual contributions of consonant and vowel information to sentence intelligibility, as well as to examine how these perceptual contributions are distributed across the C-V boundary. This study was not designed to draw discrete C-V boundaries, but rather explain the perceptual contributions of these two fundamental categories of speech sounds.

II. METHODS

A. Participants

Twenty young (age 19-30, mean 24) normal hearing listeners (6 male, 14 female) were paid to participate in the study. All participants were native speakers of American-English and passed a pure-tone audiometric screening.

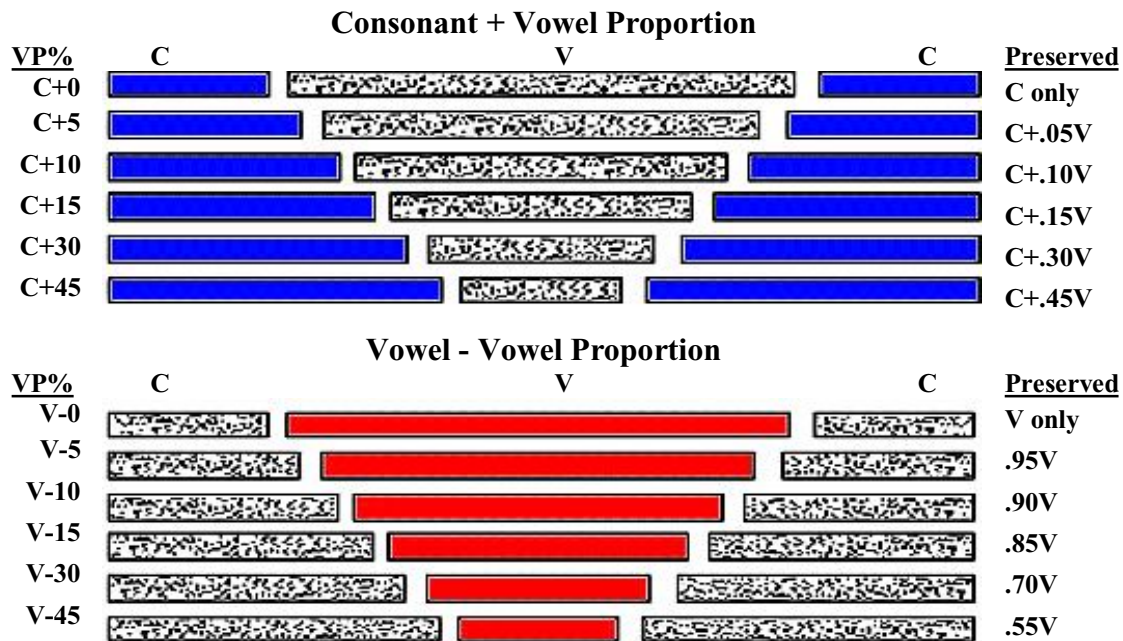
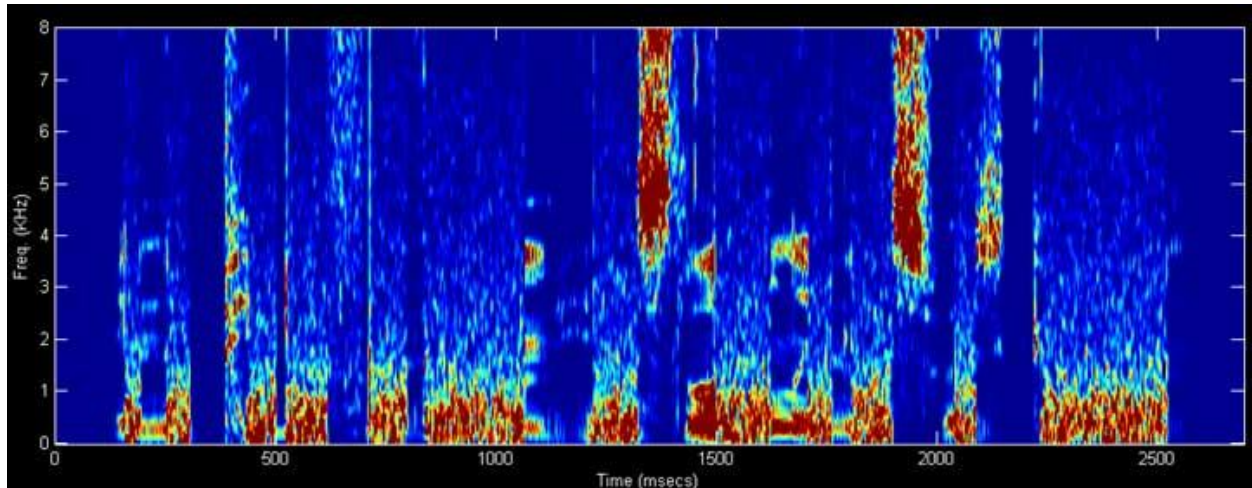


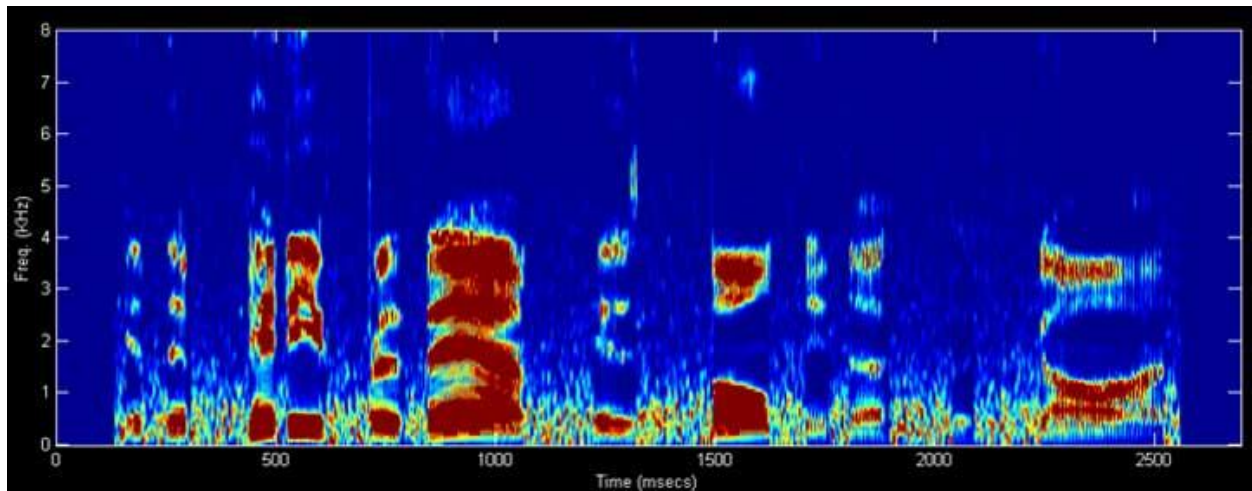
Figure 3. Schematic of noise replacement conditions depicted for a single CVC within the sentence. The C+0 and V-0 (for C only and V only conditions) uses the original TIMIT C-V boundaries.

B. Stimuli

Forty-two sentences were selected from the TIMIT database (Garfalo *et al.*, 1993, www.ldc.upenn.edu) to be used in the study. These sentences were spoken by 21 male and 21 female talkers from the North Midland dialect region (DR3), which corresponds with the region that participants were recruited from. Sentences contained an average of 8.2 words (range 6-10 words) and 33 phonemes (11 vowels, 22 consonants). The consonant-vowel boundaries are specified by the TIMIT database. These boundaries are typical of traditional segmental boundaries, although TIMIT employed CASPAR automatic segmentation that was confirmed by



a) Consonant-only

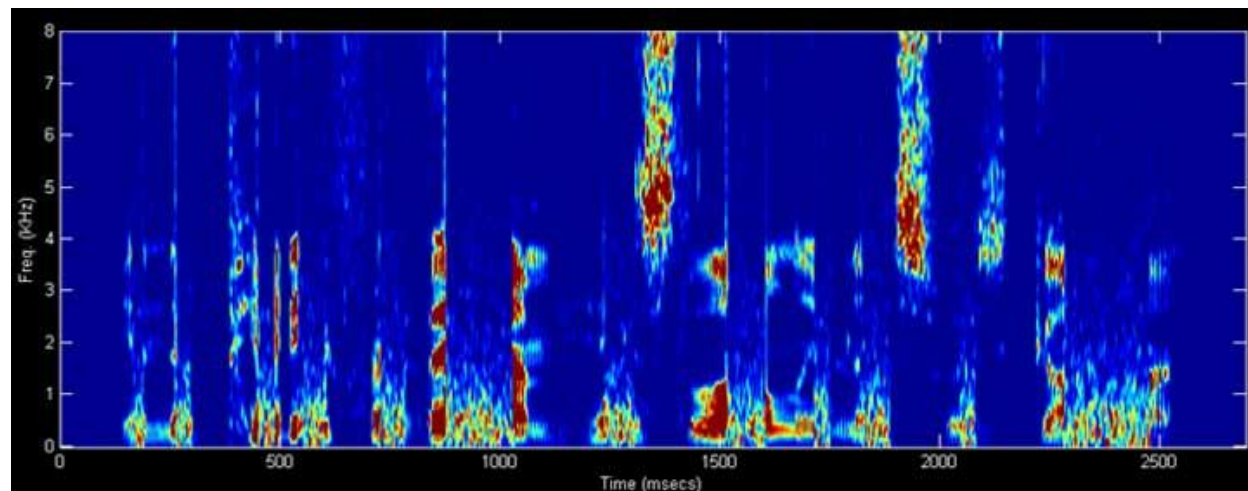


b) Vowel-only

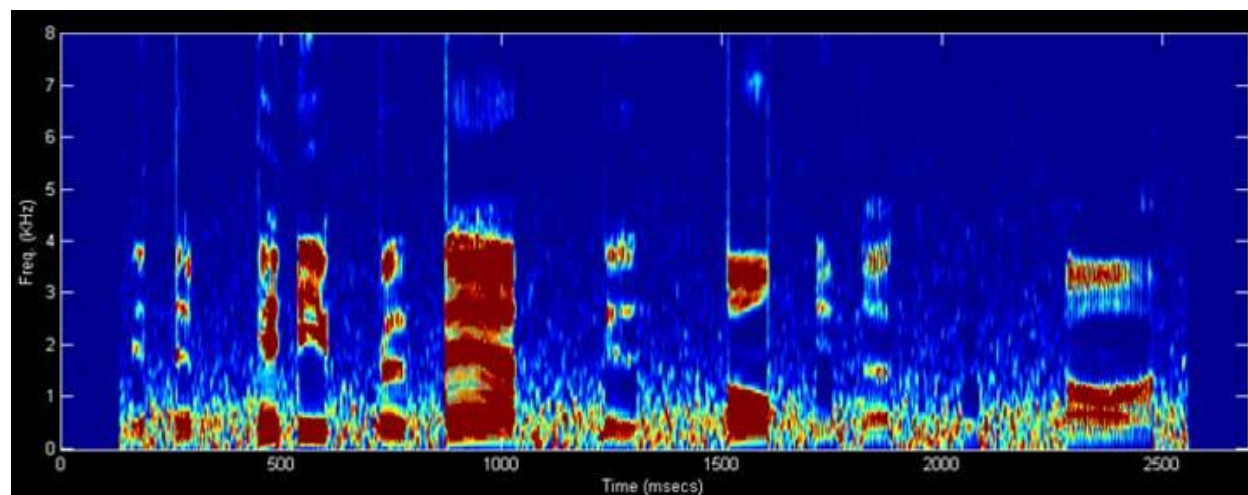
Figure 4. Spectrograms of the noise replacement using the original TIMIT C-V boundary. The experimental sentence is: “In the pity for them his loneliness was gone.” a) preserves only the consonants and replaces the vowels with noise, b) preserves only the vowels and replaces the consonants with noise.

experienced phoneticians. The TIMIT segmentation methods included rules such as using highly salient and abrupt acoustic changes to mark phoneme boundaries (such as can be seen in Figure 1) and dividing formant transitions in half during slow periods of change (as these transitions provide information regarding both phonemes, see Liberman *et al.*, 1967). Each sentence was digitally scaled to a constant RMS value and presented at 70 dB SPL.

The consonant-vowel acoustic boundary was manipulated by a proportion of the vowel duration (VP) to examine the perceptual consequences of systematically shifting the boundary into the vowel. For the purposes of the noise replacement paradigm used here and elsewhere



a) Consonant + 30%VP (C+.30V)



b) Vowel - 30%VP (.70V)

Figure 5. Spectrograms of the sentence “In the pity for them his loneliness was gone” are shown under two different noise replacement conditions. These two conditions are the inverse of each other: a) when the consonant plus 30% of the vowel are preserved, 70% of the center portion of each vowel is replaced by noise, b) when the middle 70% of the vowel is preserved, and the consonant plus 30% of the vowel at the boundary is replaced with noise.

(see Kewley-Port *et al.*, 2007), consonant strings were treated as a single unit for replacement, as were vowel strings. Two processing strategies each created a different stimulus type; one type preserved only the consonant information and replaced the vowels with noise (C+VP), and one type preserved only the vowel information, replacing the consonants with noise (V-VP). These processing strategies were manipulated by shifting the consonant-vowel boundary by 6 vowel proportion amounts (2 X 6 design), yielding a total of 12 conditions. Figure 3 shows a schematic of these 12 conditions. The schematic shows the noise replacement for only a single CVC; however, this method was applied to all segments across the entire sentence. The conditions are labeled in two ways, first according to how the C-V boundary was shifted (e.g., C+5, meaning the original consonant segment plus 5% of the vowel), and second, the proportion of the segments preserved (e.g., C+.05V). The rest of this paper will use the proportion of segments preserved in discussing these conditions.

A speech-shaped noise based on a standard long-term-average speech spectrum (ANSI, 1969) was used for the noise replacement. It was designed to be flat from 0-500 Hz and had a -9 dB/octave roll-off above 500 Hz. The noise level was presented at 16 dB below the average sentence level (21dB below the vowel with loudest RMS). This level was used in a previous study for consonant noise replacement (Kewley-Port *et al.*, 2007). Example sentences are shown in Figures 4 and 5. During the experiment, sentences were presented in the presence of a continuous, low level background noise (approximately 50 dB SNR) to mask any residual transients from noise replacement.

C. Design & Procedure

Participants were seated in a sound attenuated booth and listened to sentences via ER-3A insert earphones. Each participant was presented with six of the twelve conditions. Each participant heard fourteen sentences (114-115 words; ~462 phonemes) presented for each of the three vowel proportions they were presented with. These vowel proportions were presented under both processing strategies, where either the consonant or vowel segments were preserved. Thus, each participant heard a total of forty-two sentences per segmental condition. Each sentence was presented twice, once each under a different condition. Past research demonstrated little effect of hearing the sentence presented twice (Kewley-Port *et al.*, 2007), and the same sentence was never presented in the same block to participants. Therefore, each participant listened to a total of 84 sentences.

The task for each subject was to repeat each sentence aloud. The experimenter scored responses online and audio recorded them for offline reanalysis by an independent scorer (Inter-rater agreement on 10% of responses: 98.8%). Only words repeated exactly correct were scored (i.e., including endings such as plural *s*).

III. RESULTS & DISCUSSION

A. Descriptive Analysis

Results indicated that the vowel-only (V only) condition (64.7%) had a two-to-one advantage over the consonant-only (C only) condition (30.5%), replicating Cole *et al.* (1996) and Kewley-Port *et al.* (2007). T-tests between C+VP and V-VP conditions were significantly

different at all vowel proportions, except when $VP=15\%$ ($p=0.35$), indicating where consonant and vowel conditions approach similar intelligibility.

Figure 6 shows trends of the consonant and vowel conditions drawn with equal fit ($R^2=0.98$). A linear function approximates the consonant conditions (C+VP), while a cubic function is required to provide the same fit for the vowel conditions (V-VP). The consonant conditions (C+VP) appear to increase linearly with the addition of proportional vowel information. In contrast, vowel conditions (V-VP) appear to remain robust against proportional decreases in information until 30% is removed. Based upon these results, an

empirically determined perceptual boundary between consonants and vowels can be proposed. For vowels, sentence intelligibility is constant until a 30% shift in vowel proportion. This occurs when 70% of the vowel center is preserved, corresponding to when intelligibility is 56% correct for only vowels and 69% for only consonants. Thus, a perceptual boundary can be observed where the summed contributions from consonants and vowels represent a maximum combined intelligibility. This perceptually based C-V boundary maximizes the independent contribution of consonants and vowels to sentence intelligibility.

B. Sentence Level Analysis

Figure 7 shows on average, how much of the total sentence is preserved across the 12 conditions. The total sentence duration on average was 2.48 seconds. Sentence-level measurements for the original TIMIT boundaries demonstrated that consonants comprised 55% of the total sentence duration, while vowels only comprised 45%. While consonants have a shorter duration, more consonants (22.1) occurred per sentence than vowels (11.5). On average, 74% of consonants occurred in consonant strings, while only 12% of vowels occurred in vowel strings. When 30% of the vowel is removed in the .70V condition performance is at 56% correct, despite only 1/3 of the total duration of the sentence being presented. Performance in the consonant conditions does not surpass this performance level until a 30% shift in the boundary for the C+.30V condition, where performance reaches 69% correct. This is despite the fact that 2/3 of the sentence is presented in this consonant condition, compared to the 1/3 presented in the vowel condition just mentioned.

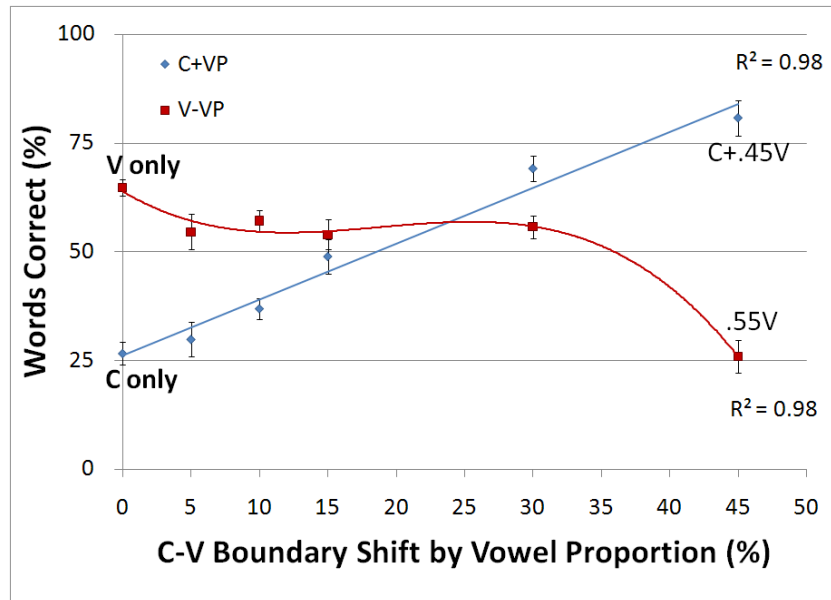


Figure 6. Results for the 12 experimental conditions plotted according to the vowel proportion at which the C-V boundary was shifted. The original TIMIT C-V boundary is at a 0%VP shift.

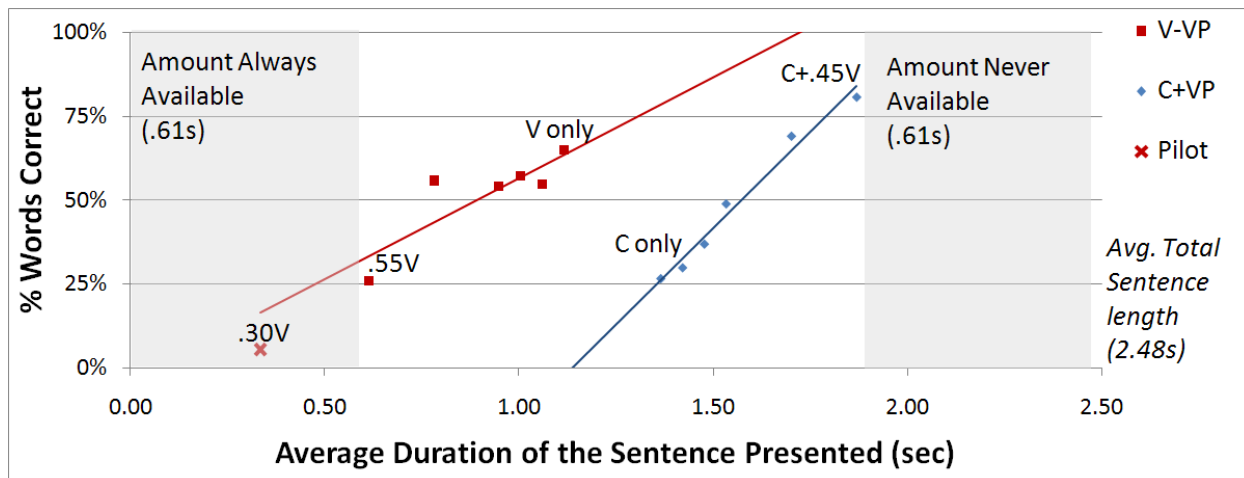


Figure 7. The average amount of the sentence preserved across the 12 conditions (plus one pilot data point at .30V), the rest of the sentence was replaced by noise. The amount of the sentence preserved was distributed over the entire sentence according to the specific noise replacement condition.

C. Segmental Level Analysis

Figure 8 displays the duration of signal information present in the average speech segment calculated for each condition. For the V only and C only endpoint conditions, this corresponds to the average vowel or consonant duration based on the original TIMIT boundaries. On average, the consonant-only condition provided a shorter segment duration (62ms) than the vowel-only condition (97ms). However, for the rest of the conditions, a segment is defined according to the noise replacement paradigm used in this study. For example, a segment in the C+.45V condition is the original consonant plus 45% of the vowel duration, with an average duration per segment of 105 ms. Each individual segment was used in this analysis, regardless of whether it occurred in a segment string. Thus, this corresponds to the average segmental duration rather than average duration of each glimpse of the speech waveform.

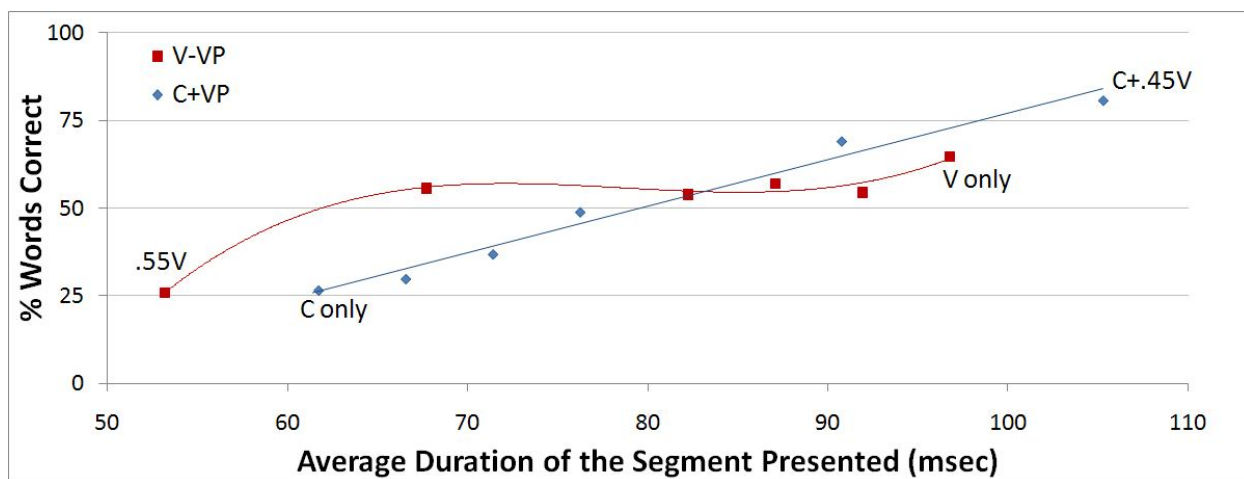


Figure 8. The average duration of an individual segment in each condition (see text for details).

When the amount of signal information was varied by a proportion of the vowel, intelligibility as a function of segment duration for the consonant and vowel conditions crossed. Specifically, at the segmental level, equal perceptual contributions to sentence intelligibility occurred when the signal duration for consonants and vowels was equal to 83ms. The perceptual equivalence of the two conditions occurred when ~15% of the vowel was deleted, or when ~22% of vowel information was added to the consonant. This corresponded to about 55% accuracy in word identification. For a 15% shift in VP, there was no statistical difference in the C+.15V and .85V conditions, and this shift was similar to the perceptual equivalent cross-over in the segmental analysis. This 15% shift in VP suggests a C-V boundary where there is a perceptually equivalent contribution of consonants and vowels at 55% sentence intelligibility.

D. Future Directions

Ongoing research in the Speech Psychophysics Laboratory at Indiana University is continuing to examine this topic using similar methods. One study proposes to examine contributions of the glimpse window that is presented, independently of segmental information. A second study will examine how periodic and aperiodic noise replacement (using average segmental durations measured from this study) affects perceptual intelligibility of these experimental sentences.

IV. SUMMARY AND CONCLUSIONS

Phoneticians have traditionally defined the C-V boundary at clear acoustic landmarks, usually points of rapid change where consonant and vowel perceptual cues are distributed across the boundary, such as the TIMIT C-V boundaries used here. These boundaries were altered in a noise-replacement paradigm by moving the boundary into the vowel. Results confirmed a two-to-one vowel advantage to sentence intelligibility at the original C-V boundary.

The individual perceptual contribution of consonant and vowel segments to sentence intelligibility was analyzed in two ways. First, sentence intelligibility increased linearly as the consonant duration increased. For the same proportional decrease in vowel duration, sentence intelligibility was relatively constant before a sharp decrease at 30% VP. This suggests that vowel information remains robust against shortening the vowel signal. Furthermore, this vowel condition presented only 1/3 of the total sentence duration; whereas, the consonant condition which had the closest perceptual performance presented 2/3 of the sentence duration. Second, segmental level analysis suggests that performance in vowel and consonant conditions cross when the average signal duration per segment is 83ms. This corresponds to 55% of words repeated correctly and is near the 15% shift into the vowel, where consonant and vowel conditions were statistically equivalent.

Two perceptual definitions of C-V boundaries were empirically determined according to the relative contributions of vowels and consonants to sentence intelligibility. First, shifting the boundary 30% into the vowel, largely removing vowel formant transitions, maximizes the independent contributions of vowel and consonant segments to sentence intelligibility. Such a shift maintains a high level of vowel contributions, while maximizing consonant contributions. Second, shifting the boundary 15% into the vowel where visible changes in the vowel formant transitions are small yields a perceptual equivalence of the contributions of vowels and

consonants to sentence intelligibility when measured at the individual segmental level.

Based upon these results, it appears clear that consonants and vowels contribute differently to the perceptual intelligibility of sentences. This line of research is not intended to define new segmental boundaries, but rather to investigate the perceptual roles of consonants and vowels in speech perception.

ACKNOWLEDGEMENTS

This work was supported in part by NIH-NIDCD Training Grant No. T32-DC00012.

REFERENCES

- Blumstein, S. & Stevens, K. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *J Acoust. Soc. Am.*, 66, 1001-1017.
- Cole, R., Yan, Y., Mak, B., Fandy, M., & Bailey, T. (1996). The contribution of consonants versus vowels to word recognition in fluent speech, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, Atlanta, GA, May, 1996, pp. 853-856.
- Cooper, F., Delattre, P., Liberman, A., Borst, J. & Gerstman, L. (1952). Some experiments on the perception of synthetic speech sounds. *J. Acoust. Soc. Am.*, 24, 597-606.
- Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., & Dahlgren, N. (1990). DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM, *National Institute of Standards and Technology*, NTIS Order No. PB91-505065.
- Kent, R. & Minifie, F. (1977). Coarticulation in recent speech production models. *Journal of Phonetics* 5, 115-133.
- Kewley-Port, D., Burkle, T.Z. & Lee, J.H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *J. Acoust. Soc. Am.*, 122, 2365-2375.
- Ladefoged, P. (2001). *Vowels and Consonants: An introduction to the sounds of languages*. Oxford: Blackwell Publishers.
- Lee, J.H., & Kewley-Port, D. (2006). Sentence comprehension when formant transitions are present or absent by normal-hearing and hearing-impaired listeners. Presented in June at the 151 the meeting of the Acoustical Society of America, Providence. *J. Acoust. Soc. Am.*, 119, 3340.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code, *Psychol. Rev.* 74, 431-461.
- Miller, G. A. (1951). *Language and Communication* (McGraw-Hill, New York).
- Freyman, R. L. & Nerbonne, G. P. (1989). The Importance of Consonant-Vowel Intensity Ratio in the Intelligibility of Voiceless Consonants. *J. Speech Lang. Hear. Res.*, 32, 524-535.
- Freyman, R. L., Nerbonne, G. P. & Cote, H. A. (1991). Effect of Consonant-Vowel Ratio Modification on Amplitude Envelope Cues for Consonant Recognition. *J. Speech Lang. Hear. Res.*, 34, 415-426.

- Owren, M. J., & Cardillo, G. C. (2006). The relative roles of vowels and consonants in discriminating talker identity versus word meaning, *J. Acoust. Soc. Am.* 119, 1727–1739.
- Sammeth, C. A., Dorman, M. F., & Stearns, C. J. (1999). The role of consonant-vowel intensity ratio in the recognition of voiceless stop consonants by listeners with hearing impairment. *J. Speech Lang. Hear. Res.*, 42, 42–55.
- Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am.*, 111, 1872-1891.
- Strange, W. (1987). Information for vowels in formant transitions. *Journal of Memory and Language*, 26, 550-557.