The intelligibility of consonants in noisy vowel-consonant-vowel sequences when the vowels are selectively enhanced

Georg F. Meyer and Robert P. Morse
Centre for Cognitive Neuroscience
Liverpool University, L69 7ZA
E-mail: g.meyer@liv.ac.uk, r.morse@liv.ac.uk

ABSTRACT

The performance of speech enhancement algorithms deteriorates rapidly with decreasing signal-to-noise ratio (SNR). At a low SNR, high intensity phonemes such as vowels are therefore more likely to be enhanced than low intensity speech segments such as many consonants. Although the selective enhancement of vowels enhances transitional cues for consonant recognition, it simultaneously degrades relative-amplitude cues. Experiments with normal-hearing subjects were performed to determine the overall effect of selective enhancement of vowels on the intelligibility of consonants in consonant-vowel-consonant utterances.
In quiet, a 12-dB enhancement of the vowels did not significantly reduce consonant intelligibility compared with an unenhanced control condition at 6 dB (A). When unenhanced utterances were presented in background noise with an average SNR of –6 dB at the vowel segments, 50.1 % of the consonants were correctly identified while 69.8 % of consonants were recognised in a condition where the consonant SNR remained unchanged but where the vowels were selectively amplified by 12 dB. Equal enhancement of the vowels and consonants by 12 dB, however, led to91.5 % consonant recognition. We conclude that speech-enhancement algorithms should enhance all speech segments to the greatest possible extent, even if this leads to selective enhancement of some phoneme categories over others.

1. INTRODUCTION

Many speech enhancement algorithms have been proposed that exploit knowledge of the target speech signal to separate it from the background noise. Some algorithms exploit the regularity of voiced speech [1], some use the speaker position [2], while others rely on the stationary nature of noise compared with the rapid variation of speech [3]. The maximum improvement in signal to noise ratio (SNR) that these algorithms can achieve has been reported to lie around 10 - 12 dB [4,5] but the performance of all speech enhancement strategies declines dramatically at low SNRs. All speech enhancement algorithms will enhance the high-intensity speech segments, but low-intensity speech segments will not be enhanced when the noise estimation fails or the signal energy is mistaken for background noise and removed. Speech enhancement at low SNR will consequently lead to a change in the amplitude of a consonant relative to the amplitude of an adjacent vowel. Previous studies have also shown that changes in these relative amplitude features can, in isolation, be perceived as a change in the place of articulation [6,7]. Similarly, changes to the burst amplitude of voiceless stop-consonants in the F4 and F5 regions relative to a following vowel affects the /p/ to /t/ contrast [7,8]. Given that relative amplitude is a cue for consonant discrimination, disproportionate enhancement of vowels compared with consonants at low SNR may decrease speech intelligibility.
Consonants in a vowel context may also be identified by the transitions of the formants from the initial and final vowels [9-11]. The direction and degree of F2 movement is the strongest transitional cue for the place of articulation because it reflects the locus position of the consonant [12]. Sussman [13] presents data that shows a robust linear correlation between the F2 frequency measured at the onset consonant and the vowel F2 frequency across speakers and languages. Similarly, frequency transitions are amongst the cues that have been implicated in the perception of nasals [14] and fricatives [15,16].
Current speech enhancement algorithms, which selectively enhance high intensity, voiced segments, will improve the representation of the transitions between the vowels and consonants by decreasing the noise in the vowel segments. Therefore, even if a consonant is masked by noise, the transition information in the enhanced vowels might be expected to contain sufficient information to allow the consonant to be identified.
The selective enhancement ofthe high intensity elements of speech therefore potentially has two opposing effects on consonant intelligibility: degradation of relative amplitude cues and enhancement of transitional cues. If amplitude cues are the predominant cues then it would be appropriate to choose a speech enhancement strategy that ensures their preservation, even if this means that high-intensity segments are enhanced only to the maximum extent possible for low-intensity segments. If transition cues are the dominant cue, then enhancing all segments to the greatest extent that is locally possible would be the appropriate choice although this would disrupt amplitude cues. The experiments reported in this paper were designed to investigate the tradeoff between these two strategies.

ISBN 1-876346-48-5

2. METHODS

A set of experiments were conducted to test the hypothesis that *selective* enhancement of some speech segments can lead to a gain in intelligibility for the remaining segments. To test this hypothesis and not confound the results by the implementation details of specific speech enhancement algorithms the experiments were based on a hypothetical speech enhancement algorithm that enhanced the signal-to-noise ratio of the vowels but not the consonants in vowel-consonant-vowel (VCV) syllables.

The five stimulus conditions given below were used:

Baseline: Natural VCV syllables presented at a baseline level of 65 dB (A) and without background noise.

Enhanced-Q: VCV syllables with the vowels at the baseline level and without background noise, but with the consonants attenuated by 12 dB. The label enhanced-Q stands for 'enhanced vowels in quiet'.

Enhanced-N: VCV syllables presented with the vowels at the baseline level, the consonants attenuated by 12 dB and with a background noise level of 6 dB SNR. The label enhanced-N stands for 'enhanced vowels in noise'

-6 dB SNR: VCV syllables presented at the baseline level but in background noise with a SNR of -6 dB.

6 dB SNR: VCV syllables presented at the baseline level but in background noise with a SNR of 6 dB.

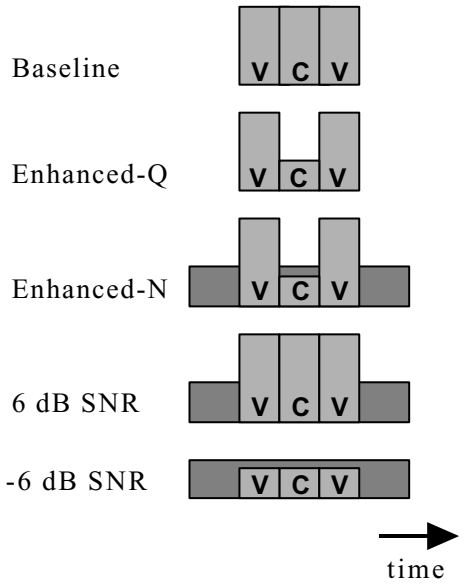

Fig 1: Schematic diagram of the stimulus conditions used.

The purpose of the baseline condition was to determine whether the consonant recognition performance of native English listeners is comparable to that of American listeners when listening to pre-existing test material of American VCVs [17]. Consonant intelligibility for the baseline and the enhanced-Q conditions were measured to determine the effect of manipulating the relative-amplitude in quiet. The key comparison, however, is between the enhanced-N and the –6 dB SNR condition (Fig. 1). For these conditions, the SNR of the consonants was the same, but the vowel energy in the enhanced-N condition was 12 dB higher than in the –6 dB SNR condition. We hypothesized that the enhanced condition would lead to greater consonant intelligibility compared with the –6 dB SNR condition, even though for both conditions the SNR of the consonants was identical. The 12-dB enhancement of the vowels was chosen because it is about the maximum enhancement that can currently be expected from a speech enhancement algorithm (e.g. [1-5]).

To determine the contribution of transitional cues to the intelligibility of the consonants, we also investigated consonant intelligibility when the VCV syllables were presented at the baseline level but in background noise with a SNR of 6 dB. In this condition, the SNR seen at the vowels was the same as in the enhanced-N condition (6 dB SNR), but the consonants in the 6 dB SNR condition were 12 dB more intense than in the enhanced-N condition. Greater speech comprehension for the 6 dB SNR condition compared with the enhanced-N condition would indicate that transition information in the vowels is not the only cue for identifying consonants and other cues such as stop-consonant burst or relative amplitude also contribute.

Stimuli

The VCV syllables were taken from a publicly available database [17]. The stimuli consisted of 20 consonants (/p/, /b/, /t/, /d/, /k/, /g/, /f/, /v/, /s/, /z/, /ʃ/, /ð/, /tʃ/, /dʒ/ , /m/, /n/, /l/, /r/, /j/, /w/) presented in a VCV context with the same vowel, /a/, /i/, or /u/, at the start and end of the syllable. The database contains 200 VCV syllables for each vowel context; each stimulus was spoken once by five male and five female speakers. The speech data was sampled at 44.1 kHz with 16 bit resolution.

To enable manipulation of consonant amplitude, the consonants in the VCV syllables were manually segmented. The segment boundaries were selected by visual inspection of the spectrogram and time domain waveform of the signal as well as using auditory feedback.

The initial and final boundaries of the fricatives /s/,/z/ and /ʃ/ were determined by the onset and offset of strong frication. The segment boundaries for the fricatives /f/,/v/ and /ð/ as well as for the nasals /n/ and /m/ were set to coincide with the intensity decrease relative to the surrounding vowels.

The consonant onset boundary for all plosives was set at the start point of the occlusion for voiced and unvoiced stops. The offset boundary for voiced plosives was placed immediately after the plosive burst while for unvoiced plosives the segment boundary was placed at the voice onset.

The segment onset for affricates was defined as the onset of the occlusion while the segment end coincides with the frication offset.

Liquids and approximants are characterised by gradual formant transitions and intensity changes so that the segment boundaries were set at approximately the mid-point of the formant transitions between the consonant and surrounding vowels.

The segmentation was verified by a phonetically trained listener for a subset of the signals.

The amplitude of each stimulus was scaled so that the average RMS in the two most energetic 50 ms segments, one from each vowel in the VCV, was scaled to 65dB (A) measured at the listeners' position (Brüel & Kjær 4133 microphone, 2619 pre-amplifier and 2608 measuring amplifier). This procedure was similar to that used by Miller

ISBN 1-876346-48-5

and Nicely [18] and Dorman *et al.* [19]. Because the vowel amplitudes were normalized, the consonants were presented at different RMS levels depending on the vowel context. The average consonant-level relative to the vowel-level for the vowels /a/, /i/, /u/ was -10.3 dB, -4.7 dB, and -3.9 dB respectively.

In both enhanced conditions (noise and quiet), the consonant was attenuated by 12 dB. The transitions between attenuated and unattenuated signal components was smoothed using 25 ms half-Hanning windows that were centered on the consonant segment-boundaries.

The stimuli were presented using a Tucker-Davis Technologies (TDT) RP2 signal processor. The normalized signals were resampled and rescaled to allow playback at the default 24 bit dynamic range and 50 kHz sampling rate of the processor. The sound quality was not affected by this process.

For experiments in which the speech signals were presented in background noise, the noise was software generated using the TDT RP2 system. Noise with a Gaussian amplitude distribution was filtered by a first-order low-pass filter such that the long-term spectrum had a –3 dB point at 500 Hz (corresponding to the long-term spectrum of speech, e.g. [20]).

The noise was presented continuously during all experiments. Subjects had at least 20 seconds to adapt to the noise before the experiment started.

Procedure

Signals were generated under computer control and output to four loudspeakers (Cambridge Soundworks four point surround) using a TDT HB7 headphone buffer. The loudspeakers were arranged in a rectangle 1.6 m high, 1.30 m wide, and 2 m in front of the subject in an IAC 1204A soundproof room.

The order of vowel contexts used for each subject was randomized. For each vowel context, the subject was first tested with the baseline condition, i.e. the VCV stimuli for the vowel context were presented at 65 dB (A) in quiet. Presentation of this condition first ensured some familiarity with the test material and therefore reduced the probability that consonants were mislabeled. The other conditions with the same vowel context were then presented in a pseudo-random sequence. For each stimulus condition, the 200 stimuli for the particular vowel context were presented once. After each stimulus presentation, the subject was asked to indicate the consonant heard by pointing with a pen on a graphics tablet that contained labels, such as 'ooSHoo' for /uʃu/ or 'eewee' for /iwi/. The experiments were self-timed; as soon as a choice was made the next stimulus was played. Each session took approximately seven minutes to complete.

Six native English speakers, three female and three male, aged between 20 and 40, participated in the experiments. As shown later, the decision to use a relatively small number of subjects was justified by the significance of the results across conditions and no significant differences across subjects. All subjects were judged to have normal hearing, based on audiograms in which hearing loss was always less than 20 dB from 100 Hz to 8 kHz.

3. RESULTS

The consonant intelligibility results for the five conditions, three vowel contexts and six subjects are shown in Fig. 2. An analysis of variance showed significant main effects for the five experimental conditions ($F(4,60) = 814.0$, $p < 0.001$) and for the three vowel contexts ($F(2,60) = 29.11$, $p < 0.001$); vowel context and experimental condition were treated as within subject effects. The analysis also showed that there was a significant interaction between the two factors ($F(8,60) = 19.11$, $p < 0.001$). We found no significant differences between subjects ($F(5,60) = 2.17$, $p = 0.067$) and we have therefore pooled data across subjects in the results given. Further to the analysis of variance, significant factors were analysed post-hoc using the simultaneous Tukey test with adjusted p-levels. Only significantly different comparisons are reported here.

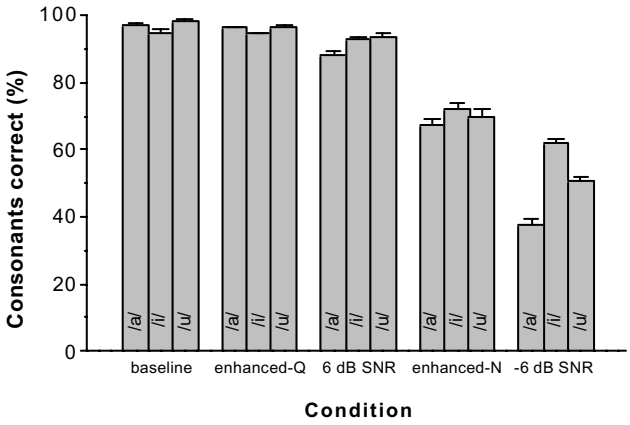

Figure 2: Consonant recognition rates by condition and vowel context. The labels for the conditions are defined in the main text. The error bars are standard errors of the means for the six subjects.

For the baseline condition, the consonant recognition performance of our English subjects was very similar to that reported by Shannon *et al.* (1999) for American listeners with the same stimuli. The mean recognition performance over the three vowel contexts was 96.6% for the English listeners compared with 97.3% for American listeners. Both English and American listeners, who all had normal hearing, confused similar consonants. Both groups of listeners found the consonants /ð/ and /dʒ/ most difficult to recognise (around 85 % correct); The American listeners also had trouble with the consonant /ɹ/ (91 % correct), which was recognised in 97 % of presentations to the English listeners. The vowel context had no effect.

The mean recognition rate for consonants in the enhanced-Q condition was 95.7 % and was not significantly affected by vowel context. The mean intelligibility of the consonants was marginally less than the mean intelligibility for the baseline condition (96.6 %), but the difference was not significant ($t=0.94$, $df=5$, $p=0.88$). In quiet, therefore, the intelligibility of consonants is not substantially affected by changes to the relative-amplitude cue when transitional formant-cues are still available.

Consonant recognition performance for the 6 dB SNR noise condition, the -6 dB SNR condition and the enhanced-N

ISBN 1-876346-48-5

condition (enhanced relative to the –6 dB condition) are shown in Fig. 2. The mean scores for these conditions were 91.5%, 50.1% and 69.8%, respectively. The performance differences between these conditions and for each vowel context were all significant with adjusted p-levels less than 0.001. The vowel context had a significant effect on consonant recognition performance only for the –6 dB SNR condition with significant performance differences between the /a/ and /i/ context ($t = 13.69$, $df = 5$, $p < 0.001$), the /a/ and /u/ context ($t = 7.08$, $df = 5$, $p < 0.001$) as well as between the /i/ and /u/ vowel context ($t = 6.61$, $df = 5$, $p < 0.001$).

4. CONCLUSIONS

The results support the hypothesis that selective enhancement of the vowels alone significantly improves the intelligibility of the consonants when VCV stimuli are presented in noise (69.8 % recognition compared with 50.1 %). In the enhanced-N condition, intelligibility was high for many consonants, e.g. /p/ (87 %), /t/ (94 %), /s/ (84 %) and /ʃ/ (93 %) or /tʃ/ (88 %); other consonants were still difficult to identify, e.g. /k/ (55 %), /v/ (42 %), or /l/ (31 %).

Although selective enhancement of the vowels did enhance consonant intelligibility, even higher intelligibility was obtained when the vowels and consonants were both enhanced to the same degree, as occurred in the 6 dB SNR condition. When both the consonants and the vowels were enhanced, the consonant intelligibility increased from 69.8 % (enhanced-N) to 91.5 % (6 dB SNR).

This finding has implications for the design of speech-enhancement algorithms for automatic speech recognition and telecommunications: speech-enhancement algorithms that selectively enhance speech segments with high SNRs but fail for low energy or unvoiced speech segments are still likely to produce significant gains in speech intelligibility. The recognition performance, however, for VCV utterances in 6 dB SNR background noise was significantly better than the performance for the enhanced-N condition; for both these conditions the SNR of the vowels was identical (6 dB SNR), but the consonants in the enhanced-N condition were 12 dB less intense than for the 6 dB SNR condition. This means that information coded in the vowels surrounding the consonant alone is not the only cue for consonant recognition in noise. The implication, however, for speech-enhancement algorithms is that vowels and consonants should be enhanced to the greatest possible extent, even if this leads to the SNR of the vowels being enhanced preferentially compared with the consonants.

REFERENCES

[1] T.W. Parsons, "Enhancing noisy speech" in *Voice and speech processing*. McGraw-Hill New York, pp. 345-364, 1987.

[2] S. Nordholm, B. Claesson, and B. Bengtsson "Adaptive array noise suppression of handsfree speaker input in cars" *IEEE Trans. Veh. Technol*. vol. 42, pp. 514-518, 1993.

[3] S.F. Boll, "Suppression of noise in speech using spectral subtraction" *IEEE Trans. Acoust. Speech Signal Process*. vol. 27, pp. 113-120, 1979

[4] B. Widrow and F.L. Luo, "Microphone arrays for hearing aids: An overview" *Speech Commun.,* vol. 39, pp. 139-146, 2003.

[5] G. Meyer, D.Yang and W.A. Ainsworth, "Applying a Model of Concurrent Vowel Segregation to Real Speech" in: *Computational Models of Auditory Function*, Eds. Greenberg, S. and Slaney, M. IOS Press, pp. 297-310, 2001.

[6] M.S. Hedrick and R.N. Ohde, "Effect of relative amplitude of frication on the perception of place of articulation". *J. Acoust. Soc. Am.*, vol. 94, pp. 2005-2026, 1993.

[7] M.S. Hedrick and W. Jesteadt, "Effect of relative amplitude, presentation level, and vowel duration on perception of voiceless stop consonants by normal and hearing impaired listeners" *J. Acoust. Soc. Am.,* vol. 100, pp. 3398-3407, 1996.

[8] R.N. Ohde and K.N. Stevens, "The effect of burst amplitude on the perception of consonant place of articulation" *J. Acoust. Soc. Am.,* vol. 74, pp. 706-714, 1983.

[9] M. Joos, "Acoustic Phonetics" *Language,* vol. 24, pp. 1-136, 1948

[10] A.M. Liberman, P.C. Delattre, F.S. Cooper and L.J. Gerstman, "The role of consonant-vowel transitions in the perception of the stop and nasal consonants". *Psychol. Monogr.*, vol. 68, pp. 1-13, 1954.

[11] W. Strange, J.J. Jenkins and T.L. Johnson "Dynamic specification of coarticulated vowels" *J. Acoust. Soc. Am.*, vol. 74, pp. 695-705, 1993.

[12] P.C. Delattre, A.M. Liberman and F.S. Cooper, "Acoustic loci and transitional cues for consonants" *J. Acoust. Soc. Am.* vol. 27, pp. pp. 769-773, 1955.

[13] H.M. Sussman, D. Fruchter, J. Hilbert and J. Sirosh. "Linear correlates in the speech signal: the orderly output constraint" *Behav. Brain Sci.,* vol. 21, pp. 241-259, 1998

[14] R.N. Ohde, "The development of the perception of cue to the /m/ - /n/ distinction in CV syllables", *J. Acoust. Soc. Am.,* vol. 96, pp. 675-686, 1994.

[15] S. Soli, "Second formants in fricatives: Acoustic consequences of fricative-vowel articulation", *J. Acoust. Soc. Am.,* vol. 70, pp. 976-984, 1981.

[16] D.H. Whalen, "Perception of the English /s/ - /integral/ distinction relies on fricative noises and transitions, not on brief spectral slices", *J. Acoust. Soc. Am.*, vol. 90, pp. 1776-1785, 1991.

[17] R.V. Shannon, A. Jensvold, M. Padilla, M.E. Robert and X. Wang, "Consonant recordings for speech testing",*J. Acoust. Soc. Am.,* vol. 106, pp. L71-L74, 1999.

[18] G.A. Miller and P.E. Nicely, "The analysis of perceptual confusions among some English consonants", *J. Acoust. Soc. Am.*, vol. 27, pp. 338-362, 1955.

[19] M.F. Dorman, S. Dori, K. Dankowski, L.M. Smith, G. McCandless and J. Parkin, "Acoustic cues for consonant identification by patients who use the Ineraid cochlear implant", *J. Acoust. Soc. Am.,* vol. 88, pp. 2074-2079, 1990.

[20] N.R. French and J.C. Steinberg. "Factors governing the intelligibility of speech sounds", *J. Acoust. Soc. Am.,* vol. 19, pp. 90-119, 1947.