ENHANCING INFORMATION-RICH REGIONS OF NATURAL VCV AND SENTENCE MATERIALS PRESENTED IN NOISE

Valerie Hazan and Andrew Simpson

Department of Phonetics and Linguistics, University College London,
4, Stephenson Way, London NW1 2HE, UK

ABSTRACT

Two sets of experiments to test the perceptual benefits of enhancing information-rich regions of consonants in natural speech were performed. In the first set, hand-annotated consonantal regions of natural VCV stimuli were amplified to increase their salience, and filtered to stylize the cues they contained. In the second set, natural semantically unpredictable sentence (SUS) material was annotated and enhanced in the same way. Both sets of stimuli were combined with speech-shaped noise and presented to normally-hearing listeners. Both sets of experiments showed statistically significant improvements in intelligibility as a result of enhancement, although the increase was greater for VCV than for SUS. These results demonstrate the benefits gained from enhancement techniques which use knowledge about acoustic cues to phonetic contrasts to improve the resistance of speech to noise.

1. INTRODUCTION

This paper reports some work carried out as part of a program investigating the effects of acoustic cue enhancement on the intelligibility of natural and synthetic speech. Certain portions of the speech signal, notably the boundary regions between phones, are information-rich as they carry cues about the identity of phones on either side of the boundary [1]. These regions, believed to be important in cueing the identity of intervocalic consonants, can be inherently transient and of low amplitude. For example, the formant transitions following plosive release have the dual drawback of being both brief and of low initial intensity as vocal fold vibration starts. This work attempts to counteract some of the limitations in the encoding of consonant identity by increasing the salience of these information-rich regions. We hypothesise that this will make it easier for listeners to process acoustic cues contained in these segments and therefore make speech stimuli more resistant to subsequent interference from background noise. This approach is therefore different from classical speech enhancement techniques which aim to process the signal after it has been degraded by noise, filtering or reverberation.

Previous work has demonstrated benefits from different types of cue-enhancement for hearing-impaired listeners [2,3] and for the perception of synthetic speech [4]. This study looks at the enhancement of natural speech for normal hearing listeners with the view to improving the intelligibility of speech output systems in poor listening conditions.

In the first experiment, the effect of cue-enhancement was examined using controlled nonsense Vowel-Consonant-Vowel (VCV) material, in which all contextual information was removed. In this way, segmental intelligibility based on the perception of acoustic information only can be evaluated. In the following experiments, cue-enhancements were implemented in sentence-length material which is characterized by a much higher degree of variability in vocalic context and degree of coarticulation.

2. VCV MATERIAL ENHANCEMENT

2.1. Method

36 vowel-consonant-vowel stimuli comprising the consonants /b,d,g,p,t,k,f,v,s,z,m,z/ in the context of the vowels /a,i,u/ spoken by a male speaker were recorded and digitized at 48 kHz sampling rate with 16-bit amplitude quantization. Annotations were made manually using a waveform editing tool to segment the stimuli into different sections. The relative levels of sections of the stimuli were then manipulated before the stimuli were reassembled by abutting adjoining segments and then down-sampling the resultant stimuli to 16 kHz to smooth any waveform discontinuities at segment boundaries. Amplitude manipulations were made by calculating the mean RMS level of each segment of the stimulus; with reference to this level sample values within a segment were then scaled to either produce a relative amplitude increase, or to set the mean RMS level of a number of segments to the same value.

After manipulation, stimuli were combined with noise whose spectral envelope was the same as the long-term average spectrum of speech. Signal-to-noise ratios (SNRs) of 0 and -5 dB were calculated on a stimulus by stimulus basis and took into account any change in the amplitude of the stimulus produced as a result of enhancement. The noise started 100 ms before the onset of the first vowel and finished 100 ms after the end of the second.

For all stimuli a distinction was made between (a) the transition regions between vowel and consonant, and (b) the consonantal constriction/occlusion regions, i.e. the burst transient, burst and aspiration, frication or nasality portions. For the transition portions, the problem of reduced amplitude as the consonant constriction/occlusion was formed or released was counteracted by amplifying the final cycles of the first vowel, or the initial cycles of the second vowel by 3 dB. The amplitude of the consonant occlusion/constriction region was amplified by either 6 or 12 dB according to consonant category (see Table 1).

In two further conditions filtering was used to change the spectral content of perceptually-important regions in order to make them more discriminable. For plosives, the burst spectrum was examined to locate the greatest concentration of energy; the precise location varied depending on the vowel context but was around 300 Hz for labials, between 1.2 and 3 kHz for velars, and between 2.5 and 4 kHz for alveolars. The burst was then band-pass filtered to retain energy at and around this frequency with the width of the pass-band set to four times the ERB [5] at the center frequency. For the fricative stimuli, the frication region was filtered to enhance the contrast in its lower-cut-off frequency, a cue to place of articulation in fricatives. The fricatives /f,v/ were high-pass and band-stop filtered respectively so that frication only appeared above 1 kHz; /s,z/ were filtered so that aperiodic energy only appeared above 4 kHz. No filtering was performed on nasal consonants.

In summary, the following test conditions were used: in condition B, only the occlusion/constriction region was amplified; in condition BT, both the occlusion/constriction and format transition regions were amplified; in condition BF, the occlusion/constriction region (for plosives and fricatives) was filtered before being amplified; in condition BFT, all types of manipulations were applied.

	B	BF	BT	BFT
P	Burst:+12dB	Burst:filtered, +12dB	Burst:+12dB Transitions:+	Burst:filtered, +12dB, Transitions+
F	Friction:+6dB	Friction:filtered +6dB	Friction:+6dB Transitions:+	Friction:filtered +6dB Transitions+
N	Nasality:+6dB	Nasality:+6dB	Nasality:+6dB Transitions+	Nasality:+6dB Transitions+

Table 1: Manipulations applied in VCV Experiment (P=plosives, F=fricatives, N=nasals)

Examples: Natural/enhanced (BFT) pairs of VCVs in quiet [SOUND A622S01.WAV] and in noise at 0 dB SNR [SOUND A622S02.WAV].

Subjects. 13 listeners aged between 20 and 35 with pure tone thresholds below 20 dB HL were tested.

Test procedure. Listeners were tested individually in a sound-attenuating room, using a computer-based testing procedure. Stimuli were presented binaurally via AKG240DF headphones, and listeners responded by pointing at a consonant only on the screen using a mouse. Listeners heard three blocks of each condition, with each block containing five repetitions of the 36 stimuli; the presentation order was randomized across listeners. All listeners heard stimuli at 0 dB and -5 dB SNRs.

Results. Figure 1 shows the intelligibility scores for all conditions. ANOVAs revealed that the effect of test condition was significant at -5 dB SNR [$F(4,48)=41.54$; $p<0.0001$] and at 0 dB SNR [$F(4,48)=16.04$, $p<0.0001$]. At both SNRs all enhanced conditions gave significantly higher intelligibility scores than the unenhanced condition. Filtering combined with amplitude manipulations did produce a significant additional improvement at the worse SNR. The highest mean increase was of the order of 12% for -5 dB SNR and 6% for 0 dB SNR. The scores obtained for the BFT condition at -5 dB were nearly identical to those obtained for the unenhanced stimuli at 0 dB SNR. The effect of the enhancement therefore corresponds to an increase in signal-to-noise ratio of approximately 5 dB.

The main effects of subject and vocalic context were also significant at both SNRs. Duncan's Multiple Range tests revealed that consonant perception in the /u/ context was significantly poorer than in the /i/ and /a/ contexts. ANOVAs were carried out on the scores obtained for the perception of the features of place and manner of articulation and voicing. Enhancements led to a significant increase in the correct perception of the place and manner of articulation but had little effect on voicing, although this could well be due to a ceiling effect.

In summary, all types of enhancement were successful in producing improvements mainly in the encoding of the place of articulation feature but also in the manner feature.

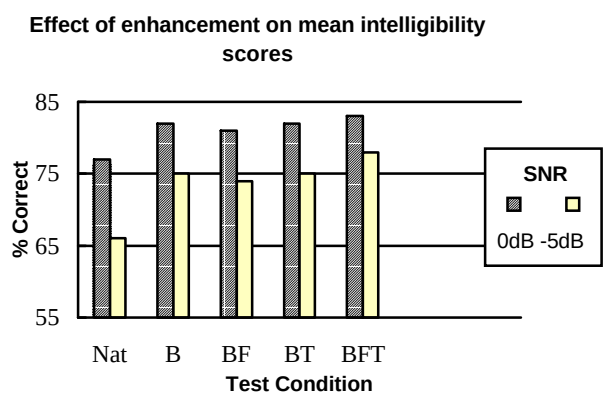

Figure 1 - Intelligibility scores for the VCV experiment.

3. SENTENCE MATERIAL ENHANCEMENT

3.1. General Method

The second set of experiments applied the same enhancement techniques to natural sentence materials. 50 semantically unpredictable sentences (SUSs) [6] read by the same male speaker used in the VCV experiment were recorded and digitized at 16 kHz with 16 bit amplitude quantization. Sentences were annotated to identify the information-rich regions described below. SUS material was used in order to limit the amount of contextual information present; sentences were syntactically correct but had words with no semantic relationship. Sentences were constructed using five different grammatical structures, and each sentence contained four key words. A greater range of

consonants including affricates and approximants were manipulated than in the VCV experiment; consonants annotated were /b,d,g,p,t,k,f,s,T,h,v,z,n,N,tS,dZ,l,r,w/ (SAMPA notation).

3.2. Sentence Experiment 1

Method. Some small adjustments were made to the enhancement techniques used in the VCV experiments. Plosive and affricate bursts were filtered, but it was necessary to use wider pass-bands given the greater variation of center burst frequency in this continuous material. The degree of amplification of the burst was reduced to 9 dB. No filtering was applied to fricatives due to the increased variability in cut-off frequency in these phones in the sentences. In the formant transition regions, the five final and initial voicing cycles before and after the consonant occlusion/constriction region were boosted by 3 dB. After being manipulated, stimuli were combined with speech-shaped noise at 0 dB and 5 dB SNR.

In addition, in order to check that these adjustments in the enhancement techniques did not significantly affect the results obtained in the VCV experiment, the same manipulations that were used in the SUS material were also applied to the VCV material described above.

Class	Manipulations
Plosives	burst: +9dB,filtered; aspiration: +9dB; transitions+
Fricatives	friction: +6dB; transitions+
Affricates	burst: +9dB, filtered; friction: +6dB; transitions+
Approximants	constriction: +3dB; transitions+
Nasals	nasality: +6dB; transitions+

Table 2: Manipulations applied in Sentence Experiment 1.

Examples: Natural/enhanced pairs of sentences in quiet [SOUND A622S03.WAV] and in noise at 0 dB SNR [SOUND A622S04.WAV].

Subjects: Separate groups of listeners were used for each SNR condition. All were aged between 20 and 35 with pure tone thresholds below 20 dB HL. 12 listeners were tested in the 0 dB SNR condition and 13 in the 5 dB SNR condition.

Test procedure. Listeners were tested individually in a sound-attenuating room, using computer-controlled sentence presentation. Sentences were presented binaurally via AKG240 DF headphones, and listeners responded by writing down the sentence heard on a response sheet. Each listener heard 25 SUS sentences in the natural condition and 25 in the enhanced condition. Sentence order within a block was randomized, and which half of the sentence list a block was drawn from, and whether a subject heard the enhanced or natural condition first were counterbalanced across subjects.

Results. Figure 2 shows the intelligibility scores for all conditions. Sentences were scored in terms of the number of key-words correctly transcribed. Intelligibility scores were then obtained by calculating the percentage of key-words correctly transcribed in each 25-sentence block (total of 100 key-words).

At 5 dB SNR, the effect of enhancement was significant [$F(1,8)=6.08$, $p=0.039$]. The order in which conditions were presented, and the sentence blocks used did not significantly affect test scores. At 0 dB SNR, the enhanced condition did not produce significantly higher scores than the natural condition. Results obtained for the VCV tests replicated those obtained above. At 0 dB SNR, mean intelligibility scores showed a significant increase from 76% to 83% (paired-difference t-test $p<0.001$).

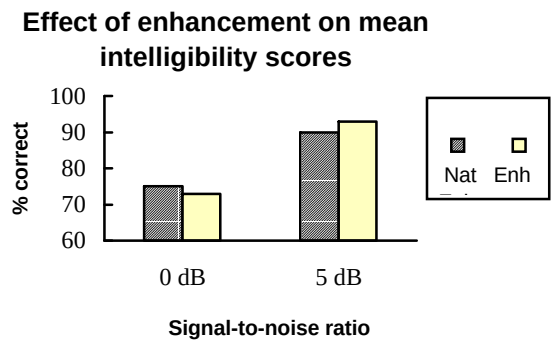

Figure 2 - Intelligibility scores for Sentence Experiment 1.

No consistent benefit of cue-enhancement on intelligibility was obtained for this sentence material. This experiment varied from previous ones in three important respects: First, the type of material itself was radically different: the sentence-length material imposed a greater cognitive load on the listeners, especially as the sentences used were semantically-unpredictable. Second, a wider range of consonant classes with a greater variety of vocalic contexts were manipulated compared to previous experiments. Third, a different set of subjects was tested.

The replication of the VCV results make it unlikely that listener effects might be the cause for this difference. A detailed examination of sentence results did suggest that some of the enhancements made to affricates and approximants had led to an increased number of errors for words containing those sounds. In order to test whether this was the cause of the poorer results compared with those obtained for the VCV material, a further experiment was set up using the same SUS material, with manipulations only made to plosives, fricatives and nasals, as in the VCV experiment.

3.3. Sentence Experiment 2

Method. Further adjustments were made to the enhancement techniques used in SUS experiment 1. First, bursts were no longer filtered as it was found that the filter-bandwidths could not be reliably set due to the greater variability in burst center frequency in continuous speech. The degree of amplification of the burst and aspiration was also changed relative to SUS

experiment 1. Second, a change was made in the way in which the initial and final vocalic cycles were amplified to avoid discontinuities in the speech signal; vocalic cycles were amplified by between 4 and 2 dB; amplification was gradually altered and cycles nearest the occlusion were amplified most. After being manipulated, stimuli were combined with speech-shaped noise at 0 dB SNR.

Class	Enhancement
Plosives	burst: +12dB; aspiration: +6dB; transitions+3dB
Fricatives	friction: +6dB; transitions+3dB
Affricates	not manipulated
Approximants	not manipulated
Nasals	nasality: +6dB; transitions+3dB

Table 3: Manipulations applied in Sentence Experiment 2.

Examples: Natural/enhanced pairs of sentences in quiet [SOUND A622S05.WAV] and in noise at 0 dB SNR [SOUND A622S06.WAV].

Subjects. 12 listeners were tested. All were aged between 20 and 30 with pure tone thresholds below 20 dB HL.

Test procedure. As used in Sentence Experiment 1.

Results. The effect of enhancement was significant [F(1,8)=19.66, p=0.002]. Order of presentation of test conditions and sentence block used did not significantly affect results but there was significant interaction between order of presentation and enhancement [F(1,1)=21.45; p=0.002]; listeners who heard the enhanced sentences second showed a greater increase in intelligibility scores.

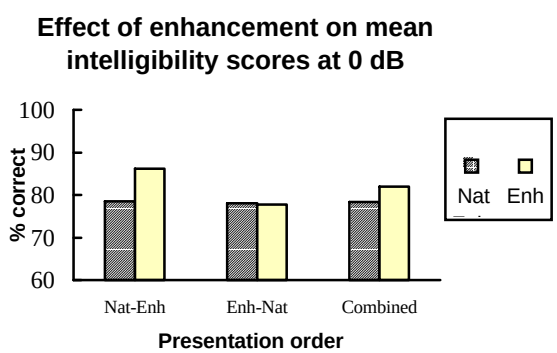

Figure 3 - Intelligibility scores for Sentence Experiment 2.

3.4. Sentence material discussion

The extension of enhancement techniques from highly-controlled VCV material to sentence-level material did lead to a need for refinements of the enhancement techniques. This was due to the fact that consonants appeared in a much wider variety of vocalic contexts and were also inherently more variable in their spectral and temporal characteristics; as a result, the enhancements which were appropriate in the VCV experiments were not always so in the continuous speech material. Manipulations which worked well for VCV material sometimes led to abrupt changes in amplitude and other discontinuities in the sentence material which had a deleterious effect on intelligibility. Results obtained in Experiment 2, however, showed that more careful adjustments made to the degree of amplification of certain constriction/occlusion and transition portions did lead to significant increases in sentence intelligibility as a result of cue-enhancement.

4. OVERALL DISCUSSION

Speech pattern extraction and enhancement has long been shown to be of benefit to hearing-impaired listeners [1]. The work reported here has shown the benefit of speech pattern enhancements in improving perception by normally-hearing listeners in poor listening conditions. Despite the relatively gross manipulations made to the stimuli in this study, a significant improvement in intelligibility was achieved.

5. ACKNOWLEDGMENTS

This work was funded by an EPSRC project grant (GR/J10426).

6. REFERENCES

1. Stevens, K.N. “Evidence for the role of acoustic boundaries in the perception of speech sounds”, in Fromkin, V. (Ed.) *Phonetic Linguistics: Essays in honor of Peter Ladefoged*. Academic Press, Orlando, 1985.

2. Gordon-Salant, S. “Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing.” J. Acoust. Soc. Am., 80 (6), 1599-1607, 1987.

3. Revoile, S., Holden-Pitt, L., Pickett, J., Brandt, F. “Speech cue enhancement for the hearing-impaired: I altered vowel durations for perception of final fricative voicing.” *J. Speech Hearing Res*., 29, 245-255.1986.

4. Yoshizumi, Y., Mekata, T., Yamada, Y., Suzuki, R. *Speech enhancement algorithm for compensating temporal masking effect.* Proceedings of the 14th ICA, Beijing, September 1992, vol. 3, pp. G5-10.

5. Glasberg, B.R. and Moore, B.C.J. “Derivation of auditory filter shapes from notched-noise data.” Hearing Research, 47, 103-138, 1990.

6. Benoit, C., Grice, M. and Hazan, V. “The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences”, *Speech Communication* 18: in press.