Enhancement techniques to improve the intelligibility of consonants in noise : Speaker and listener effects

Valerie Hazan, Andrew Simpson and Mark Huckvale

Department of Phonetics and Linguistics, UCL, 4, Stephenson Way, London NW1 2HE.
http://www.phon.ucl.ac.uk

ABSTRACT

The aim of our work is to increase the intelligibility of speech in noise by modifying regions of the signal that contain acoustic cues to consonant identity in order to make it more resistant to subsequent degradation. Two instances of each of 36 vowel-consonant-vowel (VCV) stimuli comprising the consonants /b,d,g,p,t,k,f,v,s,z,m,n/ in the context of the vowels /ɑ,i,u/ were recorded by two male and two female speakers without any phonetic training. These tokens were manually annotated; the vowel onset/offset and consonantal constriction/occlusion regions were then selectively amplified, combined with speech-shaped noise at 0 dB SNR and presented to a group of 14 native-English listeners. Significant increases in intelligibility between the natural and enhanced conditions were obtained for all speakers but the extent of the improvement was greater for the initially least intelligible speakers.

In a second experiment, speech material for two of the four speakers was presented to three new groups of native English, native-Japanese and native-Spanish L2-learners of English. For all groups, consonant intelligibility was significantly higher in the enhanced condition. The extent and patterns of errors were related to the ‘distance’ between the phonological systems of the listeners’ L1 and L2 for the set of consonants under investigation. Results of these two experiments demonstrate the robustness of our enhancement techniques across speaker and listener types.

1. INTRODUCTION

The aim of our work is to increase the intelligibility of speech in noise by enhancing key regions of the speech signal before it is contaminated by noise. The regions that are amplified are those that contain acoustic cues to consonant identity: the consonant constriction/occlusion regions, i.e. the burst transient and aspiration, friction or nasality regions, and the vowel onset and offset regions which contain formant transitions. Our previous work, reported at ICSLP96 [1] showed the perceptual benefits of these phonetically-motivated enhancement techniques as significant increases in intelligibility were shown for nonsense word (VCV) and sentence materials produced by a single male speaker, a trained phonetician.

It is well known that speakers differ significantly in their intelligibility, which may be related to certain acoustic-phonetic characteristics of their speech [e.g., 2]. It is therefore imperative to demonstrate the effectiveness of an enhancement technique with a range of speakers. To this end, perceptual tests were carried out using natural and enhanced tokens produced by four speakers with no phonetic or voice training (Experiment 1). Robustness of an enhancement technique needs also to be determined by its effect on a wide range of listeners. First, it should be effective for a large proportion of listeners within a given subject population. Secondly, as a potential application of cue-enhanced speech is in improving speech intelligibility for non-native listeners, it is also important to evaluate whether such enhancements would be effective with these subjects who may not be using the same acoustic cues to phonemic contrasts as native listeners. Speech enhancement techniques have been used in auditory training with non-native listeners [e.g., 3] but not, to our knowledge, to improve speech intelligibility in noise for such listeners. It is known that even highly-fluent non-native speakers have particular difficulties in understanding speech in noise [e.g. 4]. In Experiment 2, natural and enhanced stimuli were therefore presented to two groups who did not have English as their first language: native Japanese and native Spanish listeners.

2. EXPERIMENT 1: SPEAKER EFFECTS

2.1. Speech material

Two instances of each of 36 vowel-consonant-vowel (VCV) stimuli comprising the consonants /b,d,g,p,t,k,f,v,s,z,m,n/ in the context of the vowels /ɑ,i,u/ were recorded by 4 speakers. The speakers were aged between 25 and 30 years old, 2 were male (MH, MS), 2 were female (AO, DJ) and none had received any phonetic training. Speakers AO, DJ, and MS had south-eastern British English accents; speaker MH’s accent was north-eastern but slight; all speakers had British English as their first and dominant language. Stimuli were recorded in an anechoic room and were digitised at a 16 kHz sampling rate with 16-bit amplitude quantisation. Digitised stimuli were then annotated using a waveform-editing tool to mark the regions for amplification.

For the vowel onset/offset regions, the reduced amplitude as the consonant constriction/occlusion was formed or released was counteracted by progressively amplifying the final five cycles of the first vowel, or the initial five cycles of the second vowel, by between 2 and 4 dB. The burst, friction or nasality regions were amplified by 6 dB and the aspiration regions in plosives was amplified by 12 dB.

The amplification was applied digitally by scaling the regions' sample values. In order to avoid waveform discontinuities, 5 ms raised-cosine ramps were used to blend adjoining sections together. After manipulation, stimuli were combined with noise that had the same spectral envelope as the long-term average spectrum of speech (conforming to CCITT Rec. G227) A signal-to-noise ratio of 0 dB was 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 200 ms before the onset of the first vowel and lasted 1.5 s, to ensure that all stimuli had the same duration after the noise had been added.

2.2 Listeners

14 listeners took part in the experiment. All were aged between 20 and 30 years, had British English as their first and dominant language, and had hearing thresholds <= 20 dB HL in the range 125 Hz - 8 kHz. Listeners took part in two sessions, each lasting an hour, and were paid for their participation.

2.3 Test procedure

Stimuli were presented binaurally at a comfortable listening level in a sound-proof room through Sennheiser HD414 headphones using a computer-controlled procedure. After the presentation of each nonsense word, the listener had to identify the consonant heard by selecting with a mouse-controlled cursor one of twelve consonant symbols displayed on a computer monitor.

Listeners heard 4 repetitions of a natural and corresponding enhanced version of each of 2 different tokens of each of the 36 VCVs spoken by each of the 4 speakers. This gave a total of 2304 stimuli. Stimulus presentation order was completely randomised. Listeners received 10 minutes of familiarisation with the task before starting the experiment.

2.4 Results

Overall intelligibility scores . Mean intelligibility over the four speakers improved from 73.8% in the natural condition to 82.9% in the enhanced condition. Analyses of Variance carried out on the complete data set revealed that there was a significant effect of condition (natural vs enhanced) [F(1, 13)=315.39; p<0.0001], speaker [F(3, 39)=78.61; p<0.0001), vowel [F(2, 26)=104.69; p<0.0001) and an interaction between speaker and condition [F(3, 39)=74.77; p<0.0001**).** Duncan's post-hoc multiple range test showed that mean scores for each speaker differed significantly from all others.

Analyses of variance were then carried out separately on the data obtained for each speaker to evaluate the main effects of condition, vowel context and token (two different tokens presented for each VCV). For all four speakers, the effect of condition was significant at the 0.0005 level or more in the expected direction. For all, the effect of vowel context was significant at the same level. The effect of token was non-significant for all speakers.

Effect of speaker. Next, the effect of speaker was examined in more detail. Consonant intelligibility per speaker is presented in Figure 1. The mean improvement in intelligibility scores as a result of enhancement ranged from 5% for Speaker DJ to 19% for Speaker MH. The difference in consonant intelligibility between the least and most intelligible speakers was 23 % for the natural stimuli but only 8% for the enhanced stimuli as a result of a much greater effect of enhancement for the originally less intelligible speaker. The highest scores were obtained for the two female speakers.

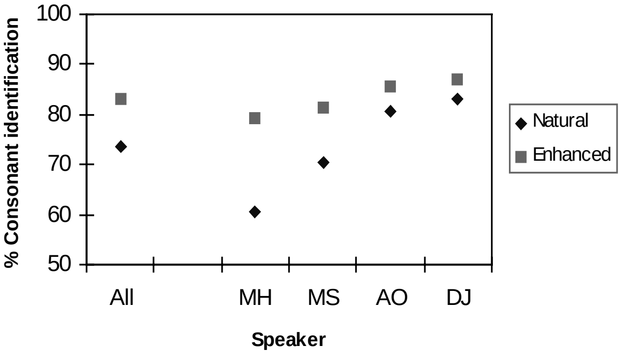

Figure 1: Mean intelligibility scores for natural and enhanced stimuli averaged across all listeners.

Individual consonant identification was also examined to see whether particular consonants contributed to the difference in overall intelligibility per speaker. The analysis centered on a comparison between the most (DJ) and least (MH) intelligible speaker (see Figure 2). The greatest differences across these two speakers was in the perception of the plosives and non-sibilant fricatives.

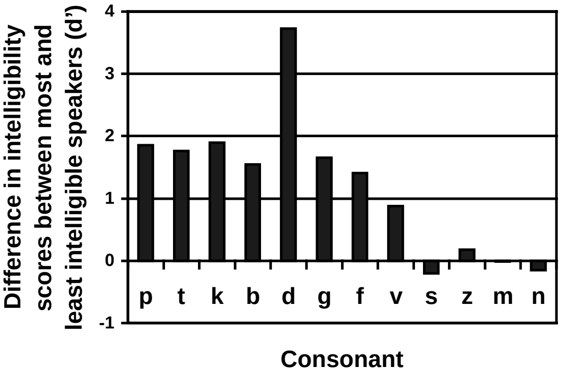

Figure 2: Difference between intelligibility scores obtained for natural stimuli for the most and least intelligible speakers. Scores are transformed to d' to reduce the effect of response bias.

2.2. Discussion of Experiment 1

Significant improvements linked to our enhancement technique which had been obtained with speech material produced by a single phonetically-trained male speaker have now been replicated with four untrained speakers and a different group of listeners. Although the extent of the enhancement effect varied across speakers, the difference between natural and enhanced scores was significant for all of them. The lower the intelligibility score for natural stimuli, the greater the effect of enhancement. This had the result of levelling out the intelligibility scores obtained for the enhanced stimuli across speakers. The enhancements were most effective in increasing the intelligibility of plosive and non-sibilant fricative consonants.

3. EXPERIMENT 2: LISTENER EFFECTS

3.1 Test material

Listeners were tested on a subset of the stimuli used for Experiment 1. Here the material included the same 12 consonants in the context of the vowels /ɑ,u/ produced by Speakers AO and MS (i.e. neither the most or least intelligible speakers) and again presented in noise at 0 dB SNR. Two blocks of 192 stimuli containing randomly ordered natural and enhanced stimuli for both speakers (4 repetitions per token) were recorded onto a DAT tape with a fixed inter-stimulus interval.

3.2. Listeners

The experimental group comprised native Japanese and native Spanish listeners who were attending a Summer School in the UK. Information about their first and second language background and self-assessment of fluency and comprehension was gathered via a questionnaire. All listeners reported normal hearing.

The Japanese group included 22 listeners with a median age of 19 years; the median age at which they started learning English was 13 years. On a range of 1 (poor) to 7 (excellent), their mean self-assessment of comprehension of English was 2.45 and of English fluency was 2.14. The Spanish group included 16 listeners with a median age of 22 years; the median at which they started learning English was 11 years. On the same scale, their mean self-assessment of comprehension of English was 4.87 and of English fluency was 4.07. The control data was obtained from a group of 18 native English listeners, all students in the first year of a Speech Sciences degree at UCL. The median age for this group was 19 years.

2.2.1 Test procedure

Listeners were tested in a quiet classroom in groups with stimuli presented through headphones at a comfortable listening level. After the explanations on test procedure had been given, in the native language if necessary, listeners heard 20 examples of the VCV stimuli before testing began. Listeners heard the stimuli in two blocks separated by a five-minute interval. They responded to each presentation by writing the consonant on the grid provided. The twelve possible consonant responses were printed at the top of each sheet.

3.3 Results

Overall scores

	Natural		Enhanced		Difference Nat/Enh	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Controls	73.2	11.7	81.9	9.1	8.7	3.2
Japanese	59.8	12.9	65.9	12.2	6.1	3.2
Spanish	52.9	15.0	61.4	16.4	8.6	3.4

Table 1. Mean intelligibility scores per listener group for the natural and enhanced test conditions

Analyses of variance for unbalanced groups (general linear models procedure) were carried out on the intelligibility data to test for the effects of test condition (natural vs. enhanced), language background (Spanish, Japanese or English) and speaker. The effect of test condition was significant [F(1, 53)=317.80; p<0.0001] and in the expected direction. The interaction between test condition and L1-background was not significant which suggest that the three language-background groups did not differ significantly in the way in which they were affected by test condition.

The main effect of language background was significant [F(2,224)=90.32 p<0.0001] and Duncan's multiple range test showed that the three listener groups differed significantly from each other (in the following order: native listeners, Japanese-L1 listeners, Spanish-L1 listeners).

Effect of speaker. Mean intelligibility scores are presented below for the three listener groups for female speaker AO and male speaker MS (See Figure 3). The effect of speaker was significant [F(1,53)=148.09; p<0.0001] with female speaker AO being more intelligible than male speaker MS. As in Experiment 1, the difference in intelligibility between speakers was much reduced in the enhanced relative to the natural condition.[i]

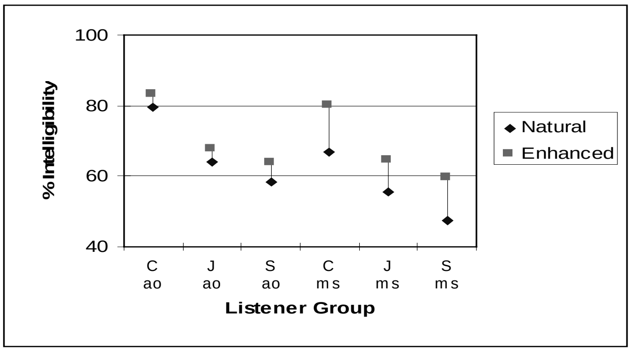

Figure 3: Mean intelligibility scores for speakers AO and MS for control listeners C, Japanese-L1 listeners (J) and Spanish-L1 listeners (S)

Effect of individual listener. The effect of enhancement was consistent for a large majority of listeners: in the non-native groups, only two listeners showed less than 2% improvement and none obtained lower scores for the enhanced condition. Increases in intelligibility for individual listeners ranged from 0.5 to 12.2% in the Japanese-L1 group, 3.2 to 16.2% in the Spanish-L1 group, 2.3 to 13.7 % in the native group.

Effect of L1 background. Consonant perception in a second language is determined to a certain degree by the perceptual 'distance' between the phonological systems of the listener's first and second language [e.g.,5]. It was predicted that both the Japanese and Spanish speakers would show confusions between /b/ and /v/ as these sounds are not contrastive in Spanish or Japanese; this would lead to an increase in errors in the perception of the feature of manner of articulation. However, Spanish listeners would additionally show increased errors in the perception of the voicing feature due to the phonetic similarity between English voiced and Spanish voiceless plosives. A predicted confusion between /s/ and /z/ which are not contrastive in Spanish would also lead to an increase in voicing errors. Results indeed show the difference between the two L2 groups as being linked to greater confusion in the perception of the voicing feature by Spanish listeners. In all listener groups, enhancement increased the intelligibility of plosives and non-sibilant fricatives most.

3.4 Discussion of Experiment 2

A first point to note is that the non-native listeners did indeed obtain significantly lower scores than native listeners for this simple consonant intelligibility task that did not involve any lexical or other contextual knowledge. Nevertheless, the enhancements applied did lead to a significant improvement in performance for a great majority of L2 listeners. For all listener groups, the difference in intelligibility between the two speakers was considerably narrowed in the enhanced condition.

The Japanese-L1 group generally obtained higher scores despite having started learning English later than the Spanish group and having lower self-assessment scores of fluency and comprehension. This appears to be due to the fact that the Spanish-L1 group was more greatly disadvantaged in terms of greater 'distance' between L1-L2 phonological categories than the Japanese-L1 listeners.

4. CONCLUSION

In conclusion, these results confirm the success of our enhancement techniques in increasing speech intelligibility in noise for the natural speech of differing clarity. Even though the extent of the enhancement effect was speaker-dependent, the fact that the effect was statistically significant for all speakers tested so far is encouraging.

Speaker effects are of serious concern for speech technology applications. Here, the fact that differences in intelligibility across speakers were reduced in the enhanced condition is encouraging as regards the future practical application of this enhancement technique.

It is also noteworthy that our enhancement techniques lead to improved intelligibility by non-native listeners for consonants degraded by noise even though the listeners received no training nor prior exposure to these stimuli. This was achieved even though the enhancements themselves were based on our knowledge of acoustic cues used by native-listeners, which may differ from acoustic cues used by L2 listeners. It is likely that enhancements more carefully targeted to L2 listeners and based on cue-weighting perceptual experiments with these listeners may be even more successful in improving intelligibility.

5. ACKNOWLEDGMENTS

This work was funded by an EPSRC project grant (GR/L25639). We wish to thank Dr Masaki Taniguchi and Dr Mercedes Cabrera-Abreu for their help in recruiting Japanese and Spanish participants in this experiment.

6. REFERENCES

1. Hazan, V. And Simpson, A. "Enhancing information-rich regions of natural VCV and sentence materials presented in noise", *Proceedings of International Conference of Speech and Language Processing*, Philadelphia, October 1996, vol. 1, 161-164.

2. Bradlow, A.R., Torretta G.M. and Pisoni, D.B. "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics", *Speech Communication*, vol. 20, 1996, p. 255-272.

3. Pruitt, J.S., Kawahara, H., Akahane-Yamada, R. & Kubo, R "Methods of enhancing speech stimuli for perceptual training: Exaggerated articulation, context truncation, and "STRAIGHT" re-synthesis", *Proceedings of ESCA STiLL workshop*, Marholmen, May 1998, p. 105-108, 1998.

4. Mayo, L.H., Florentine, M. and Buus, S. "Age of second-language acquisition and perception of speech in noise", *Journal of Speech, Language and Hearing Research*, vol. 40, 1997, p. 686-693.

5. Flege, J.E. "Second language speech learning: Theory, findings, and problems", In W. Strange (Ed.) *Speech perception and linguistic experience.* Baltimore: York Press: Baltimore, 1995.

[i] The mean scores for the control listeners were 73.2% for the natural and 81.9% for the enhanced condition (8.65% difference). Mean scores for the same subset of speech material by the set of 14 listeners in Experiment 1 were 74.7% for the natural and 83.7 for the enhanced condition (9.1% difference). This shows further evidence of the robustness of the effect despite differences in listener group, range of material and test procedure.