Speech, Hearing and Language: work in progress

Volume 12

2000

Automatic cue-enhancement of natural speech for improved intelligibility

Marta ORTEGA, Valerie HAZAN and Mark HUCKVALE

Department of Phonetics and Linguistics
UNIVERSITY COLLEGE LONDON

Automatic cue-enhancement of natural speech for improved intelligibility

Marta ORTEGA, Valerie HAZAN and Mark HUCKVALE

Abstract

In previous work, 'cue-enhancement' was found to significantly increase the intelligibility of speech in noise. However, the practical application of the technique was limited by the fact that the regions of the speech signal to be enhanced needed to be manually labelled. The principal aim of this project was therefore to automate the identification and enhancement of 'landmark' regions containing a high density of acoustic cues and to demonstrate improvements in intelligibility at least equal to that obtained for manually-enhanced materials.

We have implemented a technique for automatic cue-enhancement via the automatic identification of potential enhancement regions (PERs), and evaluated intelligibility for automatically-enhanced speech, relative to natural or manually-enhanced speech. Little loss in intelligibility was seen between the manually-tagged and automatically-enhanced materials. However, there was little evidence of statistically-significant improvements as a result of the enhancements. This may have been due in part to the fact that amplification levels across consonantal regions had to be standardised, due to the limitations of the automatic tagging.

1. Introduction

In this work, 'enhancement' refers to processing a clean source speech signal such that its intelligibility is more resistant to subsequent degradation. This kind of enhancement is relevant in at least two application areas: in *telecommunications* where a speech signal is degraded by characteristics of the channel (e.g. noise, band-limits, coding system or reverberation); and in *speech and language therapy* and *second language learning.* In this application, a speech signal can be emphasised in a computer-based training system to help a client develop phonetic discrimination abilities, despite poor phonological awareness or the use of a different phonological system (e.g. that used in another language). Our approach therefore differs from conventional signal enhancement, which is largely concerned with the removal of additive noise through techniques such as spectral subtraction, adaptive filtering, adaptive noise cancellation, and harmonic selection (e.g., Cheng, O'Shaughnessy & Kabal, 1995).

When describing methods that enhance speech prior to degradation, a distinction must be made between two types of techniques. Some apply enhancements automatically to portions of the signal which display certain characteristics (e.g. regions characterised by fast spectral change). Others apply enhancements to specific phonetic segments and, thus, require the speech signal to be annotated in terms of its phonetic components. Automatic enhancement methods such as those involving high-frequency emphasis or removal of the first formant can have a significant effect on intelligibility. However, these improvements have been shown in conditions of substantial distortion such as infinite clipping which have a very substantial effect on signal quality (Niederjohn & Grotelueshen, 1976). More recently, Tallal and her colleagues (Tallal et al., 1996) have applied automatic enhancement techniques which involve amplifying regions of rapid spectral change and manipulating segment durations. Long-term auditory training exercises using such enhanced materials appear to be

beneficial with language- and reading-disordered children. Finally, initial tests evaluating the effect of automatic enhancements based on increases in amplitude of voiceless stops and fricatives and increases in the duration of segments on speech intelligibility for non-native listeners have shown encouraging results (Colotte and Laprie, 2000).

Methods which require signals to be segmented and labelled are more phonetically-motivated and concentrate on enhancing 'landmark' regions of the signal that are known, through perceptual research, to contain a high density of cues to phonetic identity (Stevens, 1985). These 'landmark' regions can be inherently transient and of low amplitude. For example, the perceptually-important formant transitions following plosive release are both brief and of low initial intensity as vocal fold vibration starts. Phonetically-motivated enhancement approaches have been used to increase the salience of these information-bearing regions by increasing their relative intensity or duration. By making it easier for normally-hearing listeners to process acoustic cues contained in these segments, the speech signal could become more resistant to subsequent degradation. Phonetically-motivated enhancement techniques have also been investigated by others with the view to improve speech intelligibility in listeners suffering from different types of hearing disability. For example, Gordon-Salant (1986) explored the effects of increasing consonant duration and consonant-vowel intensity ratio in a set of nonsense syllables presented to normally-hearing and deaf listeners. The manipulation of intensity ratios had the greatest effect on intelligibility. In a number of studies, Revoile, Bunnell and colleagues obtained moderate improvements in consonant intelligibility in hearing-impaired adults as a result of spectral or temporal enhancements (e.g. Revoile et al, 1987; Bunnell, 1990). Jamieson (1986) also used such an approach successfully in auditory training in second-language learners and children with language disorders.

In our previous work, we aimed to improve the intelligibility of speech presented in a noisy background by normally-hearing listeners. The cue-enhancement technique targeted less robust cues involved in the decoding of consonant identity. The first phase of the project evaluated the intelligibility of spoken nonsense Vowel-Consonant-Vowel stimuli presented in noise at signal-to-noise (SNR) ratios of 0 dB and -5 dB. Enhancements led to a significant increase in intelligibility of the order of 10%, equivalent to an increase in signal-to-noise ratio of 5 dB. Significant increases in intelligibility were also obtained for sentence material (Hazan and Simpson, 1998). In a further phase of the work, test materials included nonsense words recorded by two male and two female speakers without any phonetic training. 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 less intelligible speakers. In a second experiment, speech material for two of the four speakers was presented to 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 (Hazan & Simpson, 2000). This phonetically-motivated enhancement approach has therefore clearly been successful but the need for pre-annotated material is a serious limitation of the use of these techniques.

The principal aim of this work was therefore to automate the identification and enhancement of 'landmark' regions. There were three main objectives: (i) to find good methods for the automatic identification of potential enhancement regions (PERs), (ii) to investigate the effect of errors in automatic PER identification on the intelligibility of enhanced speech (iii) to compare the intelligibility of natural, manually-enhanced and automatically-enhanced speech using more natural speech materials.

2. Signal processing

2.1 Estimation of potential enhancement regions (PERs)

The process for the automatic identification and location of regions for enhancement was based on a broad-class hidden-Markov model classifier described in Huckvale (1997). Briefly, this classifier uses a mel-scale cepstral coefficient acoustic vector and six context-free HMMs with three states and five gaussian mixtures. The six models represented silence (SIL), vocalic regions (VOC), fricative regions (FRC), nasal regions (NAS), stop-gaps (GAP) and stop-aspiration (ASP). A bigram phone language model was used with the hard constraint that ASP events could only occur after GAP events. Potential enhancement regions were recovered by rule from the recognised transcription.

2.2 Recognition data for PERs on typical material

The models were trained on half of the phonetically annotated portion of the SCRIBE corpus. This material consists of read passages and sentences from a small number of different speakers of British English recorded in an anechoic room using a high quality microphone. Evaluation was performed on two sets of materials: (1) on the other half of the SCRIBE corpus, and (2) on a subset of the experimental stimuli used in Experiment 3. Rates of misses and false alarms were calculated by comparing manual enhancement labels with automatic ones as far as was practicable. Performance figures below are based on whether the reference and test enhancement regions overlapped in time by at least 50% of the duration of the shortest region (See Table I).

Enhancement Region	**SCRIBE corpus [portion not used for training]**		**Materials from Experiment 3**	
	Misses %	**False-alarms %**	**Misses %**	**False-alarms %**
Bursts	20	25	10	35
Fricatives	25	9	15	40
Nasals	10	24	22	28
Vowel Onsets	22	16	20	19
Vowel Offsets	24	17	53	53

Table 1: *Rates of misses and false alarms the automatic detection of PERs for two types of manually-annotated materials: (1) the portion of the SCRIBE corpus not used for training the models and (2) a sub-set of the materials used in Experiment 3.*

Class labelling accuracy on the SCRIBE corpus material was 71%, with most errors involving vocalic regions being identified as nasals, or fricative regions being

identified as aspiration. High false-alarm rates on the materials from Experiment 3 were partly due to the fact that not all possible events were manually labelled. The high miss rate and false alarm rate for offsets on the experimental stimuli were due to the fact that recognised offsets had a large temporal variability so that while offsets were recognised, they did not always overlap in time with the manual labels. In the speech perception experiments, aspiration and frication regions were enhanced equally (see section 2.3), so this latter classification error had no enhancement significance.

2.3 Adaptation of previous enhancement levels for automatic-annotation

In the manually-tagged material used in previous studies (e.g. Hazan and Simpson, 2000), the aspiration, friction and nasal regions were amplified by 6 dB and the burst transients in plosives by 12 dB. The five vocalic cycles preceding and following the consonant were also amplified. Given the lower level of accuracy of PERs for the differentiation of plosive bursts, aspiration and fricatives, a single FRC label was used to cover the three types of regions. As a consequence, the amplification level for the three types of regions had to be standardised. The techniques for enhancing vowel onset and offset also had to be adjusted due to the difficulty in identifying individual vocalic cycles: a 40 ms region was automatically detected at vowel onset and offset and the amount of amplification decreased from 4 to 1 dB in 10 ms slices.

3. Perceptual evaluations of enhanced speech materials

3.1 Experiment 1

The aim of this experiment was to compare intelligibility rates obtained for manually and automatically enhanced materials for speech materials ranging from single words to spontaneous speech.

3.1.1 Speech Materials

Three types of test material were used.

a) *FAAF* test (Foster and Haggard, 1979), which consists of a list of 80 monosyllabic words. The test involves a closed-set forced-choice response, the four alternatives differing by a single phoneme. An analysis of errors made can give information about confusions in terms of the phonetic features of voicing, place and manner of articulation.

b) *SUS test* (Benoit, Grice and Hazan, 1996), which consists of 50 Semantically Unpredictable Sentences (SUS). These 4-5 word sentences were syntactically correct and semantically anomalous.

c) *Telephone information task – real words (info – N) and nonsense words (Info – NS)* These tests were designed at UCL and used spontaneously-produced speech material by 4 speakers (2 male, 2 female). Recordings were made during a simulation of a telephone-based tourist information service. In the listening tasks, listeners were presented with short extracts of these recordings and were asked a question to elicit a real word (‘Info-N’) or nonsense street name (Info-NS) contained in the extract.

3.1.2 Task Methodology

Three versions of the materials were prepared: natural, manually-enhanced (ME) and automatically-tagged enhanced (AU) stimuli using PER evaluations. In order to create the ME materials, speech files were manually annotated using a waveform-editing tool to mark the (1) burst, (2) aspiration, (3) frication, and (4) nasal consonantal regions, (5) the first five cycles of the vowel following the consonant, and (6) the last five cycles of the vowel preceding the fricative or nasal consonant. For the vowel onset region, the reduced amplitude around the consonant constriction was counteracted by progressively amplifying the five cycles of the vowel. Degree of amplification decreased from 4 to 1 dB over the five initial cycles. For stop consonants, the burst transient was amplified by 12 dB and aspiration regions by 9 dB. For fricatives, the friction region was amplified by 9 dB and for nasals, the nasal regions by 9 dB. The amplification was applied digitally by scaling the regions' sample values. To avoid waveform discontinuities at region boundaries, 5 ms raised half-cosine ramps were used to blend adjoining sections together.

The AU enhanced stimuli differed from the ME stimuli both in respect of the regions marked for enhancement and the amount of enhancement applied to them. While nasal regions were the same in both enhancements, frication regions in the automatic-enhanced copies (FRC) included stop bursts, and aspiration, as well as the friction noise of fricative sounds. These FRC regions were enhanced by 9 dB, thus eliminating the difference in amplitude between burst and the remaining consonantal regions that was present in the ME enhancement method. As for vowels, the tagging of five individual cycles at vowel onset and offset in the ME stimuli was replaced by the identification of a 40 ms segment. The 10 ms closer to the enhanced consonant were amplified by 4 dB, the next 10 ms by 3 dB, followed by 2 dB, and 1 dB in the last 10 ms. Since the difference between fricative and stop sounds was lost in the new labels, the 40 ms preceding a stop sound were also enhanced. Glides were labelled as vowels.

3.1.3 Listeners

52 listeners were tested. They were divided into four groups of 13 subjects. Two of the groups heard the natural and ME stimuli whilst the other two heard the natural and AU stimuli. Order of presentation of natural vs enhanced, and the half of the word/sentence list to be enhanced (A or B) was counter-balanced across groups. Each group heard half of the sets of test materials in the reference natural condition and half in one of the enhanced conditions.

3.1.4 Results

	FAAF	*SUS*	*Info-N*	*Info-NS*
Natural[1]	**79.93**	**70.00**	**36.54**	**14.34**
s.d.	*5.75*	*5.97*	*15.42*	*10.31*
ME	**77.13**	**70.23**	**37.36**	**12.09**
s.d.	*6.49*	*7.76*	*11.3*	*7.51*
AU	**74.15**	**68.07**	**42.6**	**11.99**
s.d.	*7.90*	*9.17*	*9.61*	*9.25*

Table 2: *Mean intelligibility rates and standard deviation measures for the four types of speech material (FAAF words, SUS sentences and telephone information task with real and nonsense words). Test conditions included Natural (unenhanced) speech, manually-tagged enhanced speech (ME) and automatically-tagged enhanced speech (AU).*

The effects of enhancement were generally very small and sometimes negative (See Table II). The variability associated with each condition, as shown by standard deviation values, was greater than the difference between conditions and a statistically-significant improvement relative to the natural condition was only obtained for the spontaneous 'Info-N' for one order of presentation. Listeners' performance differed across list orders, and these differences were significant in some conditions. The two halves in each list were not phonetically balanced and the effect is likely to be due to a strong learning and list effect.

3.1.5 Discussion of experiment 1

There was no significant difference between intelligibility rates obtained for the manually-tagged and automatically-tagged materials; this suggests that errors in alignment did not have too great an effect on intelligibility. However, there was no significant difference in scores between the natural and enhanced conditions. In this experiment, listeners were tested on different materials in different test conditions in order to avoid word learning effects; the variability associated with materials could have contributed to this lack of significant difference, given the generally small effect of enhancement. Another potential cause for the reduction in the effect of enhancement was that the levels of amplifications had been altered relative to the previous studies. In the Hazan and Simpson (2000) study, the frication, aspiration and nasal regions had been amplified by 6 dB only rather than by 9 dB. It could have been the case that the higher level of amplification led to an increase in certain consonant confusions. This would have been likely to affect intelligibility scores in the FAAF materials especially, as these word-level materials included response alternatives differing in a single consonant. The aim of Experiment 2 was therefore twofold: (1) to see whether previous significant results with manually-tagged materials could be replicated using the same speech materials but without the differentiation in the level of amplification of burst and aspiration portions and (2) to establish whether the

[1] The means reported for the Natural condition were those obtained for the group that also heard the manual enhancement condition. The means obtained for both listener groups for the Natural condition were typically within 2% of each other.

difference in levels of amplification could have been the cause of a reduction in the enhancement effect.

3.2 Experiment 2: Comparison of manually enhanced and natural VCVs at two different enhancement levels (6 dB and 9 dB).

A perception test was carried out using the same VCV tokens as Experiment 2 in Hazan and Simpson (2000).

3.2.1 Speech materials

VCV materials included the 12 consonants / b d g p t k m n f v s z/ in the context of the vowels /ɑ,u/ produced by Speakers AO and MS (i.e., neither the most or least intelligible speakers identified in Experiment 1 of that study). Two different levels of amplification were compared: 6 dB and 9 dB. The two experiments differed slightly in that a differentiation between level of amplification of burst and aspiration was made in Hazan and Simpson (2000) but not in the current study. Stimuli were mixed with speech-shaped noise at a signal-to-noise level of 0 dB.

3.2.2 Test Methodology

Test materials consisted of 576 stimuli (4 repetitions * 3 conditions * 12 consonants * 2 vowels * 2 speakers) presented in four blocks of 144 stimuli. Stimuli were randomised within each block. Before the test, there was a 144-item training block to familiarise listeners with the task. The training and testing sessions were carried out in a sound-attenuated room, using computer-based speech perception testing software. Stimuli were presented at a comfortable level over headphones.

3.2.3 Results

	VCV
Natural	**76.33**
s.d.	4.09
Enhanced (6dB)	**80.79**
s.d.	5.01
Enhanced (9 dB)	**79.98**
s.d.	4.51

Table 3: *Mean intelligibility rates and standard deviations for VCV stimuli presented in three conditions.*

Mean intelligibility rates for the three test conditions are presented in Table III. Analyses of variance were carried out to evaluate the main effect of test condition. There was a significant improvement between both enhancement conditions and the natural condition. However, the increase in intelligibility was smaller than that obtained in the Hazan and Simpson (2000) study (from 8.7% to 4.4% for the 6 dB condition). This could be due to the lack of differentiation in amplification levels between burst and aspiration. Indeed, in previous studies, increases in intelligibility had been shown to be primarily due to increased identification of plosive consonants. It is therefore not unlikely that reducing the prominence of the burst would have led to a reduction in the effect of enhancement both in this and the previous experiment. No

difference was obtained between the 6 dB and 9 dB levels of amplification. The 6dB level was therefore retained for the next experiment.

3.3 Experiment 3: Perceptual evaluation of enhanced speech using materials graded in terms of their complexity

In Experiment 3, different enhancement conditions were used, including one that replicated enhancement levels and techniques previously shown to be successful in increasing intelligibility. In order to reduce variability linked to the use of different sub-lists of materials per condition, the same materials were used for each test condition, but each condition was heard by different groups of listeners.

3.3.1 Test material

Three tests were used in the experiment: two containing real words (FAAF and LDW) and the third containing nonsense words (no lexical or other contextual information).

a) VCV test: this consisted of 24 VCV nonsense word stimuli which combined 12 English consonants, i.e. /p, t, k, b, d, g, m, n, f, v, s, z/ within two vocalic contexts, i.e. ɑC'ɑ, and uC'u.

b) FAAF test (Foster and Haggard, 1979). As in Experiment 1.

c) Lexical Density Word (LDW) test (e.g. Bradlow and Pisoni, 1999). This test consists of 148 words. Half of these words belong to high density lexical neighbourhoods and so are expected to be harder to recognise as they are confusable with phonologically-close neighbours, whilst the other half belong to sparse lexical density neighbourhoods and so are expected to be easier to identify.

A 48 year-old male Londoner with an East London accent served as a speaker. He read the test materials at a normal speech rate in an anechoic room. The stimuli were digitised at 16 bit resolution using a 44.1 kHz sampling rate.

3.3.2 Stimuli

Three enhancement conditions were prepared for each set of materials: (1) manually-tagged enhanced materials based on previously-used enhancement methods and levels (ME), (2) manually-tagged enhanced materials (NME) using PER labels (FRC, NAS, VOC), and (3) automatically-enhanced materials using PER labels (AU). A fourth condition consisted of natural (i.e. unenhanced) stimuli,

The labelling procedures applied to the ME and AU conditions in Experiment 3 were identical to those described in Experiment 1, but all labelled regions that were amplified by 9 dB in Experiment 1 were amplified by 6 dB in Experiment 3. In the NME condition, labelling and enhancement procedures were identical to the AU condition with the exception that the labelling was applied manually. After being enhanced, all tokens were mixed with speech-shaped noise at –5 dB SNR. Afterwards, the rms amplitude of each copy was equated.

3.3.3 Experimental design

So that the same speech material could be used in each condition, four groups of 12 listeners heard all materials in one condition only. In order to control for possible

learning effects, the percent of correct transcriptions of the first and the last quartile in each test were compared. Because each listener within a group received a different randomisation of the words of each test, differences due to particular items were controlled over each group of listeners. The randomisation of the LDW test included the same number of hard and easy words in each first and fourth quartile.

3.3.4 Experimental Tasks

In a sound attenuated room, listeners performed the four tests by following the instructions displayed on a PC screen while listening to the stimuli over headphones. First, the goal of the experiment was explained. Then, a 40-second long spontaneous speech sample of the speaker was played to them. All the subjects performed the tests in the same order, i.e. VCV, FAAF and LWD. The first ten items of the VCV task were used as familiarisation material for speech under noise conditions.

In the VCV test, listeners were asked to identify the consonant they heard in a closed set of 12 consonants. In the FAAF test, subjects listened to 80 different words within the carrying sentence 'Can you hear ____ clearly?'. They were asked to identify the word they heard in a four-alternative forced-choice response mode. Finally, they listened to the LWD words in the carrying sentence 'Say___again' and transcribed the word they heard. Transcriptions were written, not typed, in order to avoid errors due to typos.

3.3.5 Subjects

Forty-eight listeners with no phonetic training participated in the experiment. They had normal hearing thresholds as shown by pure-tone hearing thresholds of 20 dB HL or better at octave intervals between 125 and 8000 Hz.

3.3.6 Results

In the VCV test, data was scored in terms of the percentage of consonants correctly identified. In the LDW, FAAF and SUS tests, the percentage of correctly identified words was calculated (See Table IV).

	Natural		ME		NME		AU	
	Mean	*S.D.*	Mean	*S.D.*	Mean	*S.D.*	Mean	*S.D.*
%FAAF	81.25	*6.37*	84.79	*3.87*	84.17	*4.53*	84.58	*2.68*
%LWD	48.87	*11.87*	50.84	*6.30*	51.91	*5.62*	52.25	*4.65*
%VCV	73.78	*5.75*	76.39	*6.56*	72.74	*6.69*	72.40	*6.93*

Table 4: *Intelligibility rates and standard deviations for three types of materials (FAAF words, LDW words, and VCV nonsense words) in a control 'natural' condition and three enhanced conditions: manually-tagged using previously-used methods, manually-tagged using PERs and Automatically-tagged.*

There was a general trend for higher intelligibility rates to be obtained for the enhanced conditions and there was no significant decrease in intelligibility across manually-tagged and automatically-enhancement conditions (See Figure 1). However, the difference between natural and enhanced conditions did not reach statistical significance for any of the speech materials. For the Lexical Density Word test, the expected difference in intelligibility rates was obtained between 'Easy' words from

sparse phonological neighbourhoods and 'Hard' words from dense phonological neighbourhoods (See Figure 2). However, the enhancement conditions had little effect on higher word category.

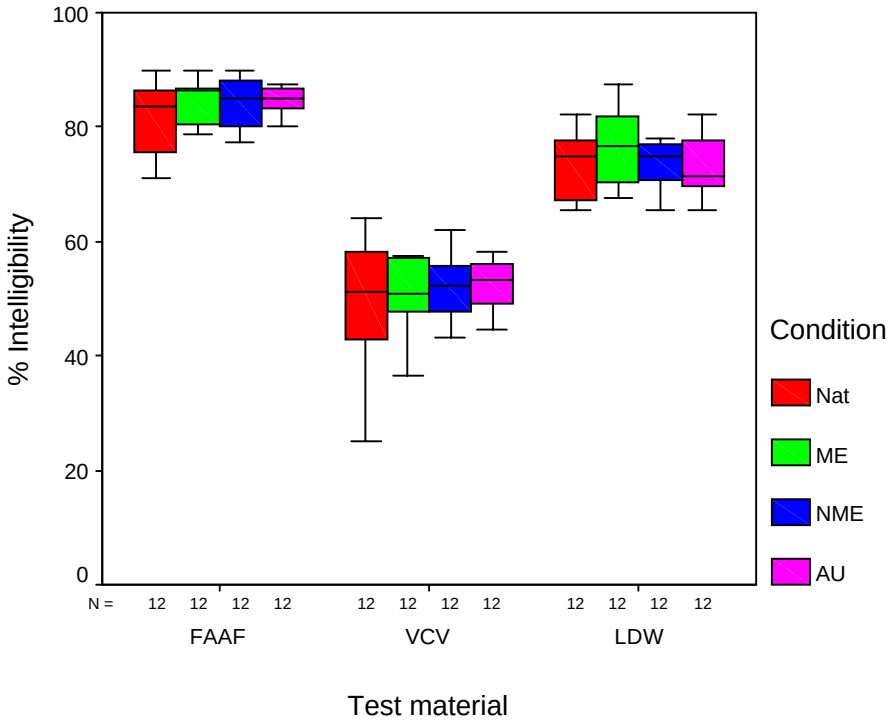

Figure 1: *Box-plots of intelligibility rates for FAAF, VCV and LDW materials in the control (Natural) and three enhancement conditions.*

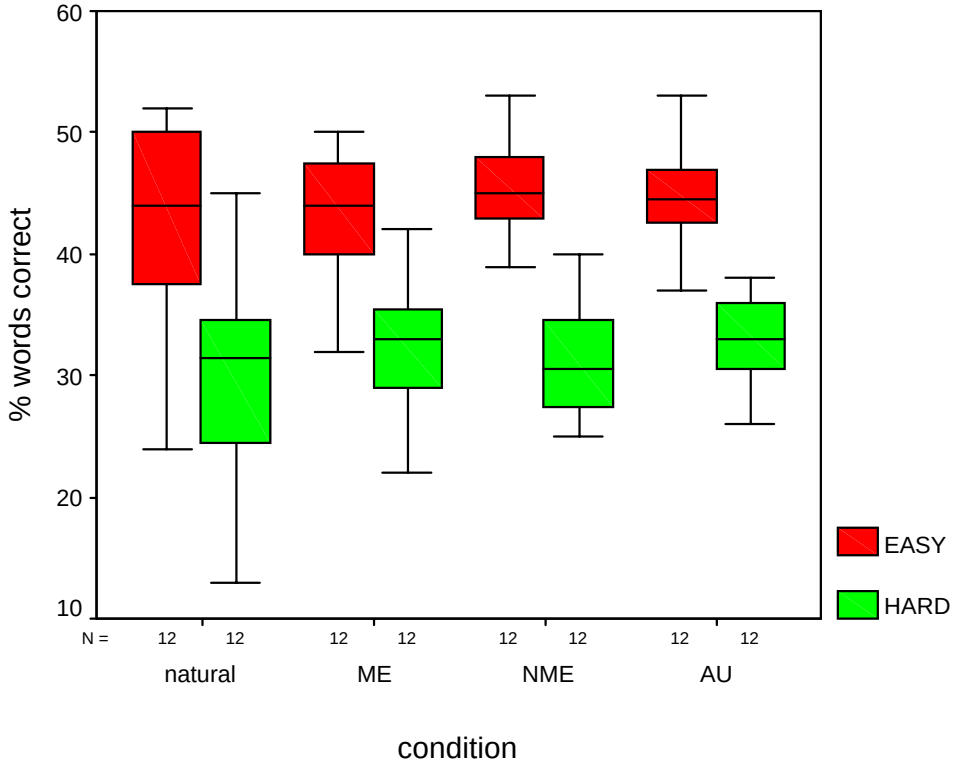

Figure 2: *Box plots showing the word intelligibility rates for the Lexical Density Word test. In all test conditions, the 'Easy' words, which belong to sparse neighbourhoods were more easily recognised than the 'Hand' words, which belong to dense neighbourhoods.*

3.3.7 Discussion of Experiment 3

As in Experiment 1, there was no loss in intelligibility between the manually-tagged and automatically-enhanced conditions, but there was no statistically significant improvement between the natural and enhanced stimuli for any of the speech materials. These results were disappointing given the positive effects obtained for cue-enhanced stimuli for VCV and sentence materials in previous studies (Hazan and Simpson, 1998, 2000).

For VCV stimuli, the increase in intelligibility was marginally lower than that obtained in experiment 2, and was in this case non-significant. The difference in the effect of enhancement may have been due to a speaker effect or to differences in the methodology used in both studies. In Experiment 2, the same listeners heard both the natural and enhanced condition whereas in Experiment 3, different groups of listeners heard one condition only. All our previous studies have shown clear individual variability in the effect of enhancement, with some listeners showing little to no improvement in score in the enhanced condition whilst others showed significant increases. For example, in one study using VCV materials, the increase in scores for individual listeners between natural and enhanced conditions varied between 2 and 14% (median: 8%). The variability linked to the use of different listener groups per condition could therefore have led to a reduction in the effect of enhancement. Also, in the previous study, listeners had a fairly extensive familiarisation session with

enhanced stimuli before the beginning of the test whereas in this study, the familiarisation was limited to 10 tokens.

In tests other than the VCV lists, the use of meaningful word and sentence materials further increases variance as lexical and listener effects on intelligibility rates are magnified. In Experiment 3, we aimed to reduce the variability linked to the presentation of different list items per condition. However, this entailed using different listener groups per condition. Thus, one source of variability was reduced but another was increased, again making it more difficult to show significant differences across conditions.

4. General discussion

We have implemented a technique for automatic cue-enhancement via the automatic identification of potential enhancement regions (PERs) and evaluated PER detection accuracy for a range of materials. A Windows-based implementation of this technique is freely available from our departmental website (www.phon.ucl.ac.uk/enhance). This software allows a user to automatically enhance speech materials using our phonetically-motivated cue-enhancement approach or standard techniques (amplitude compression, spectral substraction).

The automatic tagging of regions to be enhanced led to a reduction of the enhancement effect, relative to previous studies. There may be a number of reasons for this. First, it was necessary to standardise the levels of amplification across all consonantal regions, as the models used did not differentiate between plosive burst and aspiration, for example. Data from Experiment 2 showed that the effect of enhancement was reduced as a result. Also, a wider range of regions were amplified relative to our previous studies, as there was no differentiation made in the models between vowels and approximants, and between fricatives and affricates, for example. Our experience suggests that amplifying approximants and affricates may lead to an increase in errors (e.g. Hazan and Simpson, 1998). Finally, the relatively high rates of misses and false alarms for certain consonant categories would have meant that certain key regions were not adequately amplified. It is therefore suggested that the future success of automatic cue-enhancement depends to a great extent on further improvements in the sensitivity and reliability of models to be used for the detection of PERs. For applications where finite amounts of material are needed, manual or semi-automatic tagging of PERs would be preferable at the present time.

We have shown in related experiments (e.g. Hazan and Simpson, 2000; Ortega and Hazan, 1999) that cue-enhancement can be successful in improving consonant intelligibility for second language (L2) learners. The primary aim of enhancement in this application is to target specific sound contrasts, which are likely to be difficult for second-language learners because they do not occur or have a different phonological status in the listener's first language. Enhancement can help attract the learner's attention to the region containing important acoustic cues to the phonemic contrast to be made. In recent work, the need to tailor the enhancement approach to the acoustic cues relevant to particular contrasts was shown. Indeed, in work on the discrimination and identification of a plosive voicing contrast (/d/-/t/) by Spanish-L1 speakers of English, greatest improvements were obtained when the amplitude of the /d/ burst was decreased rather than increased. Manual tagging which allows greater freedom for the setting of consonant-specific levels of amplification is therefore far preferable for such

applications. The use of cue-enhancement in improving the efficacy of auditory-visual training for L2 learners is now being investigated.

5. References

Benoit, C., Grice, M. & Hazan, V. (1996) The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, 18, 381-392.

Bunnell, H.T. (1990) On enhancement of spectral contrast in speech for hearing-impaired listeners. *Journal of the Acoustical Society of America*,88, 2546-2556.

Bradlow A.R., Pisoni D.B. (1999) Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *Journal of the Acoustical Society of America*, 106, 2074-2085.

Colotte, V. and Laprie, Y. (2000) Automatic enhancement of speech intelligibility. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP'2000*, Istanbul

Foster, J.R. and Haggard, M.P. (1979) An efficient analytical test of speech perception. Proc. IoA, IA3, 9-12.

Hazan, V. and Simpson, A. (1998) The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication*, 24, 211-226.

Hazan, V. and Simpson, A. (2000). The effect of cue-enhancement on consonant intelligibility in noise: speaker and listener effects. *Language and Speech*, 43 (3), 273-295.

Hazan, V., Simpson, A. and Huckvale, M. (1998) Enhancement techniques to improve the intelligibility of consonants in noise: Speaker and listener effects. *Proc. ICSLP, Sydney, Australia*, December 1998, 5, 2163-2167.

Huckvale, M. (1997) A syntactic pattern recognition method for the automatic location of potential enhancement regions in running speech. Speech, Hearing and language: UCL Work in Progress. http://www.phon.ucl.ac.uk/home/mark/papers/shl97.pdf

Liu, S. 1994. Landmark detection of distinctive feature-based speech recognition.JASA, 96, 5, Part 2, 3227.

Merzenich, M.M., Jenkins, W.M., Johnston, P., Schreiner, C., Miller, S., Tallal, P. (1996) Language Comprehension in Language-Learning Impaired Children Improved with Acoustically Modified Speech. *Science*, 271.

Ortega, M. and Hazan, V. (1999) Enhancing acoustic cues to aid L2 speech perception. *Proc.ICPhS*, San Francisco, 1-7 August 1999, 1, 117-120.

Revoile, S.G., Holden-Pitt, L., Pickett, J.M., Brandt, F. (1986) Speech cue enhancement for the hearing impaired: I. Altered vowel durations for perception of final fricative voicing. *Journal of Speech and Hearing Research*, 29, 240-255.

Robinson, T., Hochberg, M. & and Renals, S. (1996) The use of recurrent neural networks in continuous speech recognition in *Automatic Speech and Speaker Recognition - Advanced Topics* (Lee, Paliwal and Soong, editors), Kluwer Academic Publishers.