Proceedings of The Institute of Acoustics

ESTIMATING A VISUAL SPEECH-RECEPTION THRESHOLD

Alison Macleod and Quentin Summerfield

MRC Institute of Hearing Research, University of Nottingham

INTRODUCTION

Much research into speech coding strategies for high technology aids for the profoundly and totally deaf has been aimed at identifying parameters of the speech waveform, which, when extracted and presented via a cochlear implant or tactile aid, will optimally supplement lipreading (1). This approach brings with it the need for accurate tests with which to measure the benefit received from diffferent coding strategies. Ideally, such tests would use stimulus materials representative of everyday fluent speech, such as unrelated sentences. However, large individual differences in the ability to lipread fluent speech exist among both hearing-impaired and normally-hearing subjects. This presents a problem when the benefit from an aid is measured as an increase in %-correct score over lipreading alone. Those subjects whose unaided score falls near 50%-correct appear to gain more from the aid than those who score very poorly or very well, because changes in the scores of the latter two groups may be limited by floor and ceiling effects. The problem arises for two reasons: i) %-correct performance does not constitute an interval scale; ii) unaided lipreading ability and %-correct performance level are confounded in a conventional test.

We have explored a technique for measuring lipreading ability as a visual speech-reception threshold (VSRT). By analogy with the technique for measuring an auditory speech reception threshold (2), we adaptively vary the visual signal-to-noise ratio (SNR) necessary for a criterial level of performance. We do this by varying the amount of visual noise added to the video image of a talker's face. Performance is then measured not as a %-correct value, but as the visual SNR giving 50%-correct performance. The present study examined the viability of such a technique as a tool for investigating individual differences in lipreading ability.

STIMULI

The sentences used for the measurement of a VSRT were generated from the closed set of 20 words shown below.

CATEGORY	1	2	3	4
1	FRANK	FOUND	ONE	PIG
2	JIM	KEPT	TWO	FISH
3	RON	WANTED	THREE	HENS
4	STEVE	SEES	FOUR	RATS
5	PAUL	BOUGHT	MANY	SHEEP

The use of a closed set ensured that all subjects could be trained to lipread the stimulus material with near 100% accuracy in the absence of visual noise. By selecting a word at random from each of the four categories in turn, up to 625 different sentences can be created. Two hundred sentences, generated in this way, were recorded onto videotape. The talker was an adult male speaker of Southern British English.

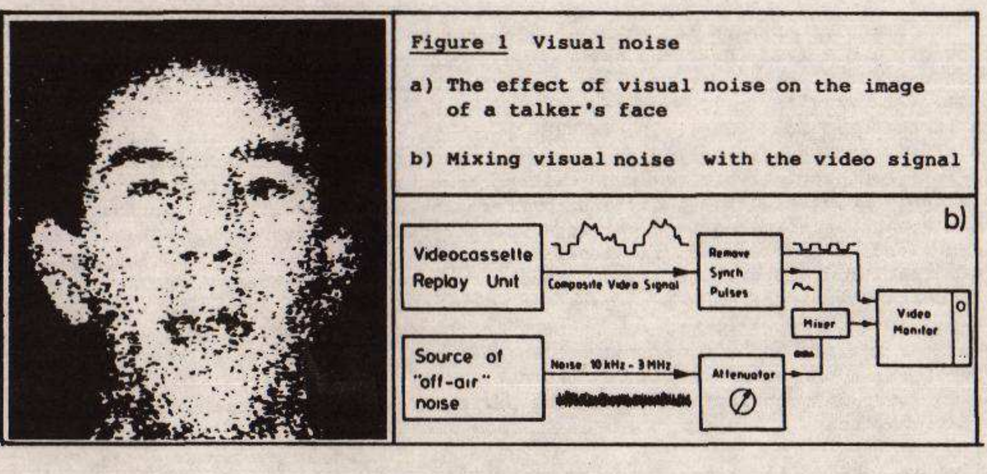


Figure 1 Visual noise

a) The effect of visual noise on the image of a talker's face

b) Mixing visual noise with the video signal

EXPERIMENT 1

We set out to answer two questions:
1) Is VSRT correlated with lipreading ability, measured using a conventional %-correct test? In other words, is it the case that better lipreaders can tolerate a lower visual SNR than poorer lipreaders, when constrained to the same %-correct level of performance.
2) How does the benefit gained from a simulated aid to lipreading relate to lipreading ability measured a) as a %-correct score on a conventional test, and b) as a VSRT?

PROCEDURE

Twenty normally-hearing subjects (mean age = 22 years) took part. All had normal or corrected-to-normal vision, screened with a Snellen acuity chart, and normal contrast sensitivity, screened using the Arden Gratings (3).

Each subject was trained to lipread the stimulus set in the absence of visual noise. Subjects viewed isolated words from each category, then sentences spoken with a pause between each word, and then sentences spoken fluently. At each stage, after a period of practice, the subject received 20 test trials. The level of accuracy necessary for the subject to proceed to the next stage was 19 out of 20 trials correct. All subjects tested achieved this criterion.

In the experiment itself, subjects received a practice block of 30 sentences followed by three blocks of 50 sentences. These were presented with visual noise, consisting of spatially and temporally random fluctuations in luminance, added to the video signal in controlled amounts. The first sentence of the practice block was presented repeatedly, starting at a very disadvantageous SNR, which was gradually improved until the subject could lipread the sentence correctly. This established a starting point for the adaptive procedure. On remaining lipreading-alone trials, the visual SNR necessary for 50%-correct performance was determined using a simple one-up one-down tracking procedure (4) with a fixed step size. On alternate trials, the subject heard a signal known to aid to lipreading, in the form of a series of rectangular pulses synchronised to the closing of the talker's vocal folds (5). The visual noise level was adjusted only on unaided trials.

Lipreading ability was also measured using a conventional test. Subjects lipread 30 sentences, drawn from the BKB audiometric sentence lists (6). The BKB lists have been used frequently to assess lipreading ability (7), and are easily scored as the percentage of designated keywords reported correctly. The 30 sentences are a subset of 60 BKB sentences selected to span a range of lipreading difficulty, from easy to hard (8). They were recorded by the same talker under the same conditions as the sentences used to measure VSRT.

RESULTS

Each subject's VSRT was calculated by averaging the visual noise presentation levels for the unaided trials in the three test blocks. VSRT correlated significantly with lipreading ability measured as %-correct keywords on the BKB sentence test (n=20, r=0.59, p 0.01). This result confirms that the measure of VSRT reflects skills necessary for lipreading unrelated sentences. Benefit from the simulated aid measured as the increase in %-correct above 50% when vision was supplemented, averaged 25% (range = -2 to +36%), and did not correlate significantly with VSRT (n=20, r=-0.08, n.s) or %-correct keywords on the BKB sentence test (n=20, r=0.14, n.s.). The present result suggests that the positive correlation found previously between lipreading ability and subsequent improvement with aiding, when both were assessed using %-correct as the dependent variable (9), may be an artefact of the method used to assess benefit, and that benefit and lipreading ability need not be related.

EXPERIMENT 2

In Experiment 1, the method of adding visual noise to the video signal caused brightness and contrast to increase undesirably as a function of visual noise level. Accordingly, a new method was adopted, shown in Figure 1, in which the video signal was attenuated with increasing visual noise level, such that overall brightness remained constant. Experiment 2 was run to ensure that this change did not affect the correlation between VSRT and lipreading ability.

Twenty-two new subects took part (mean age = 21 years). Their vision was screened in the same way as before. Training was reduced to 10 trials at each stage of practice, with a criterion of 9 out of 10 trials correct before progressing to the next stage. To measure VSRT, subjects received 25 sentences for practice, followed by 2 test blocks of 25 sentences each. All but two subjects were given feedback throughout the test. Lipreading ability in terms of a %-correct score was measured using all 60 of the selected BKB sentences.

VSRT correlated significantly with %-correct keywords on the BKB sentence test (n=20, r=0.78, p 0.01), confirming that the new configuration of the visual noise did not affect the relationship between VSRT and lipreading ability.

EXPERIMENT 3

Adding visual noise to the image of a talker's face lowers performance by degrading the visual cues normally available to the lipreader, in either or both of two ways. Visual noise might increase the likelihood of all phonemic confusions uniformly. Alternatively, noise might radically alter the pattern of confusions by selectively degrading cues necessary for distinguishing particular consonants or vowels. The first objective of Experiment 3 was to distinguish these alternatives.

The second objective was to examine the relationships between scores on sentence-based tests of lipreading, and scores on lipreading tests involving vowels or consonants presented in nonsense syllables. It has been found that sentence lipreading ability correlates more highly with vowel lipreading scores than with consonant lipreading scores (10, 11). One interpretation of this pattern is that vowel lipreading ability is a more important component of the ability to lipread sentences than is consonant lipreading ability. However, it has been pointed out (12, 13) that consonants fall into well-defined visual categories between which all lipreaders can perceive differences (e.g. of place of articulation), but within which not even the best lipreaders can make distinctions (e.g. of voicing and nasality). This restricts individual differences. Vowels, in comparison, provide a continuously graded range of potentially discriminable lip shapes, allowing individual differences to appear, and providing sufficient variability to permit correlations with other measures of performance to emerge. By excluding visually identical contrasts, we generated a consonant set involving fine distinctions in lip shape. If the low correlations observed between consonant- and sentence-based lipreading tests are caused by insufficient variability in consonant scores, the use of a set of consonants selected to encourage the emergence of individual differences in performance could produce correlations with sentence lipreading that equal or exceed those found between sentence and vowel lipreading scores.

STIMULI AND PROCEDURE

The 15 vowels chosen were /iː, I, e, ʌ, æ , aː, ɒ , əʊ, ʊ, uː, aʊ, eI, aI, oI, /. These occur in most dialects of British English. The vowels were spoken in b-vowel-b syllables.

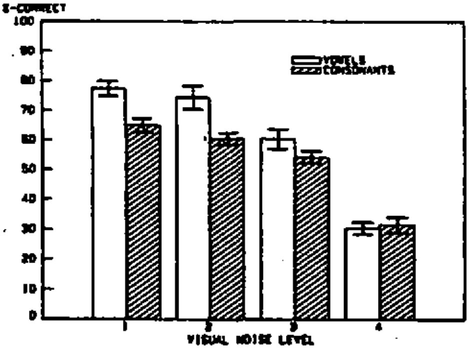
The consonants /b, w, ð , v, r, l, z, d, ʒ , dʒ, g, j/ were chosen to represent each of the 12 consonant categories distinguishable under optimum conditions (15). Consonants were spoken in a-consonant-a syllables.

Separate recordings were made for the vowels and consonants. Ten occurrences of each stimulus, in random order, were recorded onto videotape. The talker and recording conditions were the same as those used in experiments 1 and 2.

Stimuli were presented for identification in four visual noise conditions. These were: 1) no noise; 2) a low noise level, below most subjects' VSRTs; 3) a moderate noise level, close to the VSRT of many subjects; and 4) a high noise level, above the VSRT of any subject.

Figure 2

Vowel and consonant identification scores as a function of visual noise level



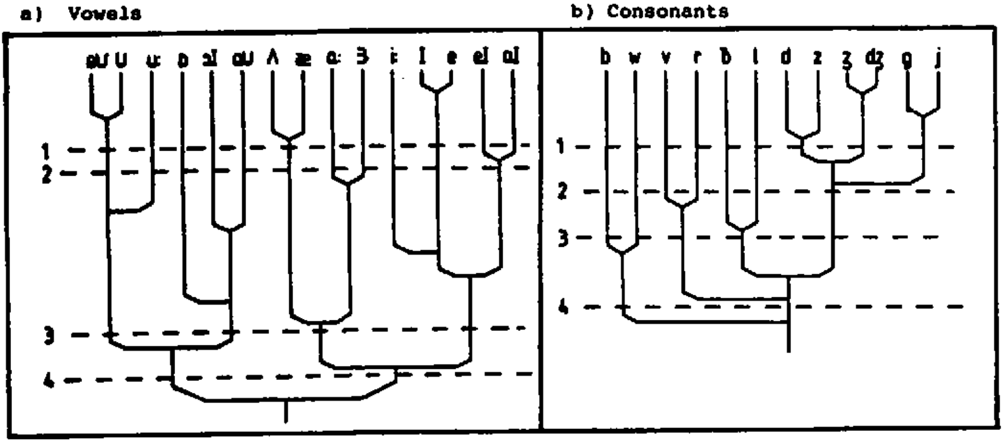
Twenty-two subjects took part. All had previously participated in Experiment 2. They responded to each trial by pressing a button labelled with the orthographic approximation of the syllable they thought most like the one they had seen. At the start of the consonant and vowel sessions, subjects were familiarised with the approximations used and given practice, with feedback, in identifying the syllables without visual noise. Before receiving each condition, subjects received a short practice block, without feedback, containing syllables presented at that visual noise level.

RESULTS

Figure 2 shows that group mean vowel and consonant identification scores fell as the level of visual noise increased. Separate analyses of variance on the root-arcsin transformed scores (16) revealed a main effect of noise level for both the vowel condition (F3,84)=43.8, p 0.001 and the consonant condition (F(3,84)=59.6, p 0.001). Post-hoc Scheffe tests revealed significant differences between noise conditions 3 and 4 (p 0.01), and between noise condition 3 and noise conditions 1 and 2 taken together (p 0.05), for both the vowel and the consonant sets. No other pairs of values differed significantly.

Hierarchical Clustering Analysis Group confusion matrices for the vowel and consonant sets, pooled over visual noise level, were analysed using the hierarchical clustering program HICLUS (17). Hierarchical clustering is a technique which simplifies a matrix of confusions among a set of objects, by converting it into a a set of clusters reflecting the similarities between the objects. The analysis starts by combining the two most similar objects into one cluster. It then re-computes the similarity matrix, including the new cluster as a single object. This process is repeated until all objects are contained within a single cluster. The scheme is hierarchical in that once an object has been incorporated into a cluster, it may not leave that group at a later stage. This enables hierarchical clustering schemes to be represented very simply, in a tree diagram which illustrates the history of cluster formation, and thus the pattern of similarities among the original objects.

Figure 3 Hierarchical clustering schemes



Results of the HICLUS analysis of the vowel set are shown in Figure 3a). A viseme group was defined as a cluster for which intra-cluster identification scores were greater than 75%-correct (18). This analysis revealed that, in condition 4, the highest noise level, subjects could discriminate rounded from unrounded vowels, but could not make finer distinctions. At the next noise level, subjects could discriminate four vowel categories: two groups of rounded vowels, the open, unrounded vowels /u:, æ, ɛ:, a:/ and the unrounded close front vowels /i, I, e, aI, eI/. There was little difference between the categories discriminable in conditions 1 and 2. The short vowels /e, I/ and /u:, a/ remain difficult to identify even in the absence of visual noise.

The pattern of clusters for the consonant set is shown in Figure 3b). The major distinction is between the relatively discriminable front consonants /b, w, v, r,ð , l/ and the less easy-to-distinguish back consonants /d, z, ʒ, dʒ, g, j/. In condition 4, subjects could distinguish only the labials /b, w/ from the rest of the set. In condition 3, subjects could discriminate five groups, /b/, /w/, /v, r/, /ð, l/ and /d, z, ʒ, dʒ, g, j/. In condition 2, all the "front" consonants were identified with greater than 75% accuracy. With no noise present, eight groups achieved this criterion, but there were only 66% intracluster responses to the cluster /g, j/.

Analysis of the clustering schemes produced for each noise level, pooled over subjects, revealed little difference in the order or pattern of cluster formation across the four noise levels for either the consonants or the vowels. Thus, visual noise level, while determining overall number of confusions, does not determine the internal patterns of confusions.

Table 2. Product-moment correlations between consonant and vowel scores, and VSRT and BKB scores. (** = p 0.05, * = p 0.01).

			Vowels				Consonants			
		BKB	1	2	3	4	1	2	3	4
VOWELS										
	1	0.53**								
	2	0.55**	0.42*							
	3	0.78**	0.47*	0.58**						
	4	0.43*	0.45*	0.53**	0.38*					
CONSONANTS										
	1	0.60**	0.54**	0.26	0.57**	0.43*				
	2	0.52**	0.38*	0.25	0.52**	0.57**	0.62**			
	3	0.64**	0.60**	0.22	0.56**	0.45*	0.66**	0.70**		
	4	0.24	0.43*	0.37*	0.27	0.56**	0.41*	0.34	0.22	
VSRT		0.78**	0.46*	0.30	0.55**	0.28	0.42*	0.50**	0.51**	-0.03

Comparison with VSRT and BKB score The results of correlating subjects' consonant and vowel identification scores at each visual noise level with their VSRT and BKB scores are shown in Table 2. VSRT correlated significantly with consonant identification and vowel identification in the absence of visual noise. VSRT correlated most highly with syllable recognition in noise condition 3, which was closest to the mean level of the subjects' VSRTs.

Consonant and vowel recognition scores also correlated significantly with lipreading scores for the unrelated BKB sentences. In condition 1, the correlation between vowel score and BKB score is comparable with that reported by others (9, 10), but the correlation between consonant score and BKB score is higher than any previously reported. Again, the highest correlation present was between BKB score and syllable identification in condition 3.

The pattern of results obtained confirms our hypothesis that the low correlations previously found between consonant lipreading and sentence-based measures were likely to have been the result of insufficient variability in consonant scores. By selecting a consonant set which excludes visually identical contrasts, the ability to lipread sentences is shown to be reflected in both vowel and consonant lipreading scores.

CONCLUSION

VSRT provides a measure of lipreading ability that is free from floor and ceiling effects that plague the interpretation of conventional tests. We have verified that this measure correlates with scores on conventional sentence tests of lipreading, and that adding visual noise to the image of a talker's face increases the overall level of phonemic confusions, but does not distort their internal pattern.

REFERENCES

(1) R.A. Schindler and M.M. Merzenich (eds), Cochlear Implants, Raven Press, New York, (1985).

(2) R. Plomp and A. Mimpen, "Improving the reliability of testing the speech-reception threshold for sentences", Audiology, Vol. 18, 43-52, (1982).

(3) G. Arden and A.G. Gukukoglu, "Grating test of contrast sensitivity in patients with retrobulbar neuritis", Arch. of Opthalm., Vol 62, no. 7, (1979).

(4) H. Levitt, "Transformed up-down methods in psychoacoustics", J.A.S.A., Vol. 49, 467-477, (1971).

(5) A.J. Fourcin, "Laryngographic assessment of phonatory function", ASHA Reports, Vol. 11, (1981).

(6) J. Bench and J. Bamford, "Speech-hearing tests and the language of hearing-impaired children", Academic Press, (1979).

(7) S. Rosen and T. Corcoran, "A video-recorded test of lipreading for British English", Br. J. Audiol., Vol. 16, 245-254, (1982).

(8) A. Macleod and Q. Summerfield, "Quantifying the benefits of vision to speech perception in noise", Br. J. Audiol, (in press).

(9) M. McGrath and Q. Summerfield,"Intermodal timing relations and speech recognition by normal-hearing adults", J.A.S.A., Vol. 77, 678-683, (1985).

(10) F. Heider and G. Heider, "An experimental investigation of lipreading", Psychol. Monographs, Vol. 52, 124-153, (1940).

(11) M. Breeuwer and R. Plomp, "Speechreading supplemented with auditorily presented speech parameters", J.A.S.A., Vol. 79, 481-499, (1986).

(12) G. Plant, "Visual identification of Australian vowels and diphthongs", Austral. J. Audiol., Vol. 2, 83-91 (1980).

(13) M. McGrath, "An examination of cues for visual and audio-visual speech perception using natural and computer-generated faces". Unpublished Ph. D. thesis, University of Nottingham, (1986).

(14) E. Owens and B. Blazek, "Visemes observed ny hearing-impaired and normally-hearing adult viewers", J.S.H.R., Vol. 28, 381-393, (1985).

(15) K. Berger, Speechreading: principles and methods, National Educational press Inc, Baltimore, (1972).

(16) B. Winer, "Statistical principles in experimental design", McGraw-Hill, New York, (1971).

(17) S.C. Johnson, "Hierarchical clustering schemes", Psychometrika, Vol. 32, 241-254, (1967).

(18) B.E. Walden et al., "Some effects of training on speech recognition by hearing-impaired adults", J.S.H.R., Vol. 24, 207-216, (1981).

VOISCOPE ANALYSIS OF ABNORMAL LARYNGEAL EXCITATION

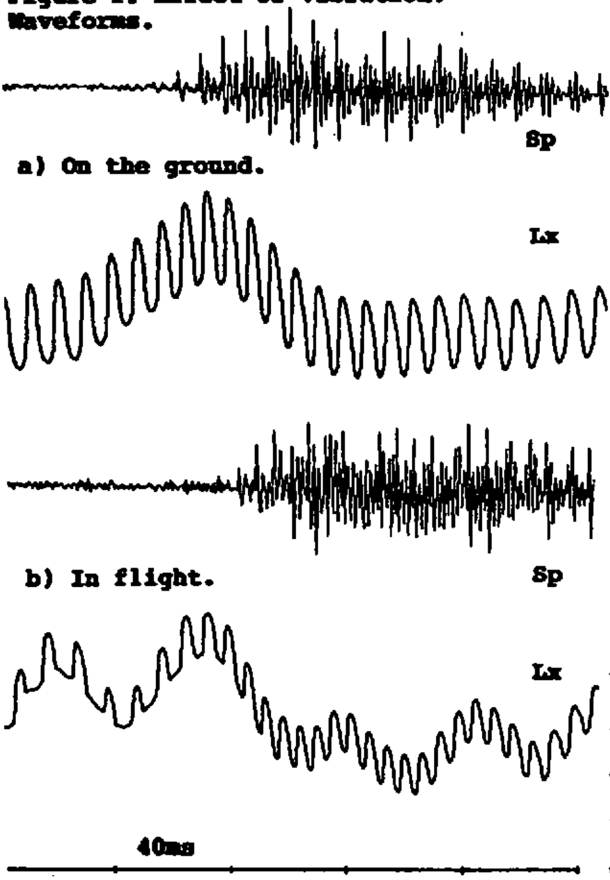
D. J. Miller (1), M. R. Taylor (2) and A. J. Fourcin (3).

(1) LARYNGOGRAPH Ltd. 1 Foundry Mews, London, NW1 2PE.
(2) SMITHS INDUSTRIES AEROSPACE & DEFENCE SYSTEMS,CHELTENHAM.
(3) UNIVERSITY COLLEGE LONDON

In a wide range of normal and pathological voices, four main physical factors associated with vocal fold vibration are of special importance. First, the sharpness of closure of the vocal folds determines the spectral spread of the excitation which is provided for the illumination of the resonances of the vocal tract. Second, the duration of vocal fold closure and the nature of closure itself are important factors in respect of the definition of the events which immediately follow the initial excitation and in the preservation of a degree of clarifying isolation between the vocal tract proper and the subglottal cavities. Third, the duration of the open phase is of especial interest since, when this is large, the effects even of a good initial excitation impulse can be vitiated. Finally, but not of least importance, regularity of vocal fold closure from cycle to cycle is crucial to the clarity of the pitch percept which defines intonational contrasts.

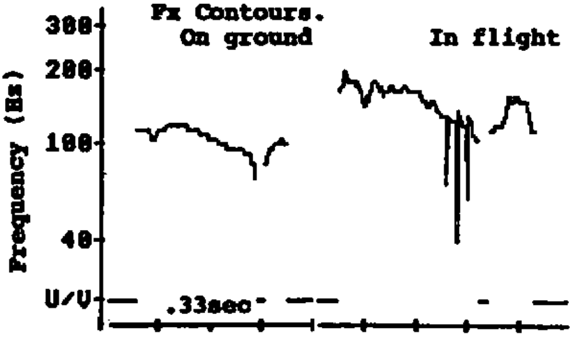
In the present discussion data will be presented which has been obtained from the direct monitoring of excitation activity - for normal voice; for voice production in which the speaker is subjected to appreciable mechanical vibration; for two examples of pathological voice, organic and "functional"; and finally for synthetic speech. In this last case the excitation information has been derived directly from the excitation generator by electrical connection. In the case of the human speech sources, an electrolaryngograph has been employed. The laryngograph is a simple electrical device which responds to the change in electrical conductance due to vocal fold contact and measured between two electrodes placed superficially on the wings of the speaker's thyroid cartilage. Its output waveform has been directly correlated with synchronous stroboscopic examination of vocal fold vibration, related to the inverse filtered acoustic signal, and used in the triggering of X-flash pulses of radiation in studies of phonation. All these investigations confirm the basic electrical interpretation of its output and make it possible in consequence to examine voice quality almost directly in terms of the four factors above. Sharpness of closure corresponds to the rapidity of increase of conductance; the duration of the peak of conductance is related to the duration of vocal fold closure; the trough in the conductance waveform has some correspondence with the open phase and finally, waveform regularity relates directly to excitation periodicity and so to pitch. In what follows, Lx refers to the conductance waveform, and Fx to the reciprocal of its period. Lx and Fx together with the analyses of period range, Dx, and scatter, Cx, which are discussed have been obtained from the use of the Voiscope® operating in conjunction with a BBC computer.

Figure 1. Effect of vibration. Waveforms.



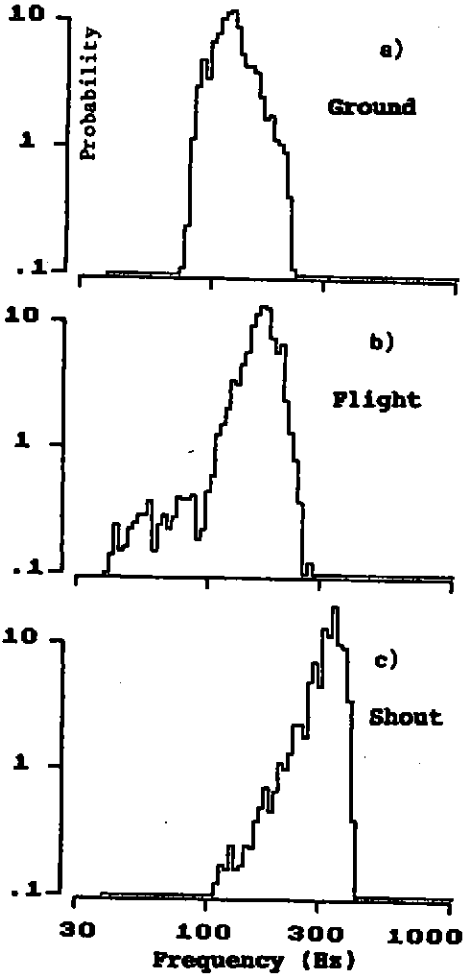
The first example of abnormal laryngeal excitation effects is drawn from a situation in which a normal speaker is subjected to unusual bodily vibration. The male adult speaker in Figure 1 is speaking initially in quiet conditions on the ground. A standard text has been recorded and the same utterance has been examined in Figure 1b when the speaker was in flight. Helicopter vibration components were measured with three accelerometers so that the main bodily vibratory disturbance could be spectrally assessed, a peak at 22.5 Hz was found. Simultaneously to the acoustic recording from the speaker, the laryngograph electrodes were used to monitor vocal fold vibratory activity. The sequence "ZED" in 1a shows the gross laryngeal adjustment for an initial voiced fricative going into a vowel accompanied with normal laryngeal vibration. In figure 1b the gross laryngeal adjustments are now associated with a 22.5 Hz vibratory component and the detailed form of the individual vocal fold vibrations is quite noticeably changed both in respect of periodicity and shape. The higher fundamental frequency clearly evident in Lx is characteristic

Figure 2. Effect of Vibration. Fx Contours.



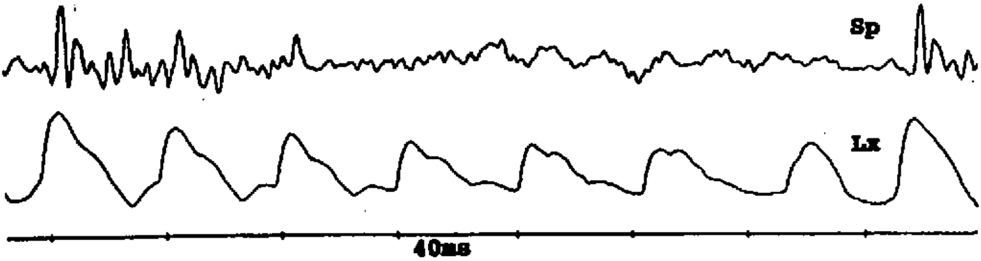
of the speaker's response to the higher ambient noise conditions. Figure 2 indicates the influence of vibration on another aspect of analysis based on Lx which gives the fundamental frequency contour, Fx. The sequence is centred around the normal modal value of the speaker and quite smooth with characteristic consonantal Fx jumps. In flight, Fx is far more irregular and occasionally broken by octave jumps produced by the vibratory interference with larynx activity.

Figure 3. Effect of vibration. Distributions 1st order.



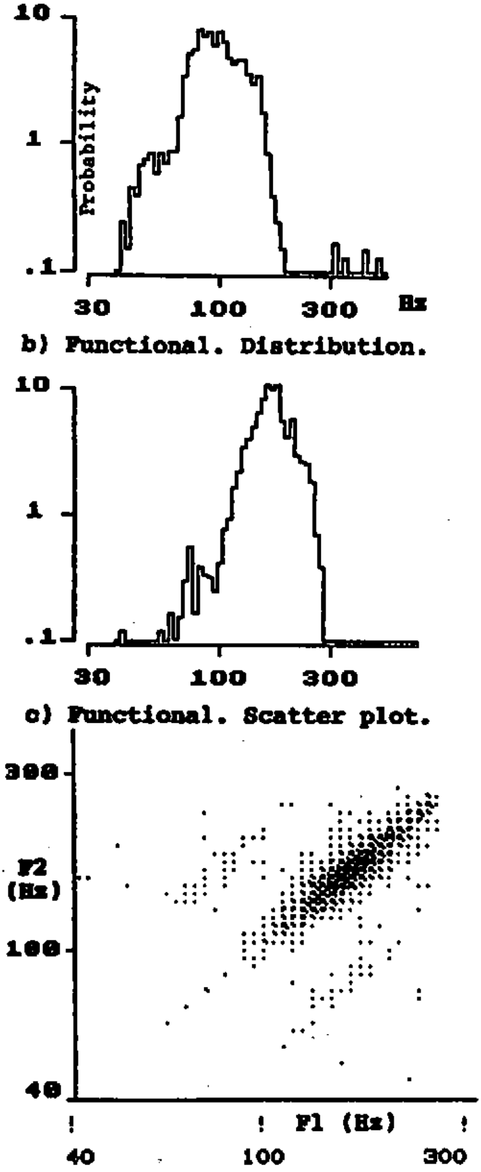
Different but related effects of vibration are evident at every level of analysis. Figure 3 shows the results of analyses directed towards the determination of the long term frequency distribution based on recordings of a text approximately 2 minutes' duration (5K to 8K periods). In each case the analysis makes use of the determination of laryngeal period and it is the number of periods corresponding to any frequency which is associated with the probability estimates. The ground analysis is quite typical of a healthy adult male and shows a prominent mode at 123Hz. There is here no low frequency irregularity of the sort which is typically found with creaky voice. In flight, however, there is a very marked area of low frequency irregularity in the histogram and a marked upward shift of the main mode to 169 Hz. These two effects are largely separable and in fig. 3(c) shouting alone gives a complete upward displacement of the mode to 338 Hz together with a reshaping of the distribution. Another technique of analysis is discussed in relation to figure 6 in respect of these samples. The main features of the in-flight distribution are to a degree predictable in terms of two separate effects - the low frequency skirt corresponding to octave jumps in laryngeal frequency and the upward shift of the main mode corresponding to the change in Fx with ambient noise intensity.

Figure 4. Pathology - waveforms.



However, the frequency modulation of laryngeal vibration which will also be associated with the vibration has almost certainly an influence on the shape of the distribution in the in-flight condition. In the second example of abnormal laryngeal excitation, a pathological condition has been examined. In figure 4 the speech and laryngograph waveforms coming from a brief sample recorded by a patient with Reinke's Oedema are shown. In this condition the vocal folds are grossly swollen so that although the glottis is not obstructed the normal vibration of the vocal folds is impaired. The impairments arise not only from the asymmetries of effective mass and stiffness but also and more particularly from the irregular nature of the contacting surfaces. At the extreme right-hand end of the pair of waveforms, the last laryngeal closure is associated with a prominent peak of acoustic response. Immediately prior to this peak of acoustic response, however, there is a laryngeal closure which is not associated with a correspondingly evident pressure peak response from the vocal tract. Speech output from vocal fold input is dependent not merely on the magnitude of closure of the vocal folds but upon the acoustic spectrum of the associated excitation, this can only have a broad frequency range if there is a relatively rapid change in the volume velocity and this in turn can only occur if there is a rapid effective closure of the vocal folds. In this example the variability of vocal fold closure is shown fairly clearly by the Lx waveform in between the beginning and end Lx closure peaks of the sample. The perturbations in shape of the Lx waveform similarly result primarily from the oedematous nature of the vocal folds, these variations are typically associated with a corresponding variation in the periodicity of the speech which has a quality random component apparently superimposed on it. In much of the clinical work supported by Voiscope/BBC microcomputer analyses of this type, it is not possible to make use of the pcm recording techniques used for the helicopter studies - partly as a matter of cost and partly of convenience. Here, as in most of the other associated clinical studies, a simple two channel cassette recorder has been used and the waveforms of figure 4 are derived from its outputs after digital phase correction on the same BBC as is used for both the other analyses and interactive voice therapy.

Figure 5.



a) Pathological. Distribution.

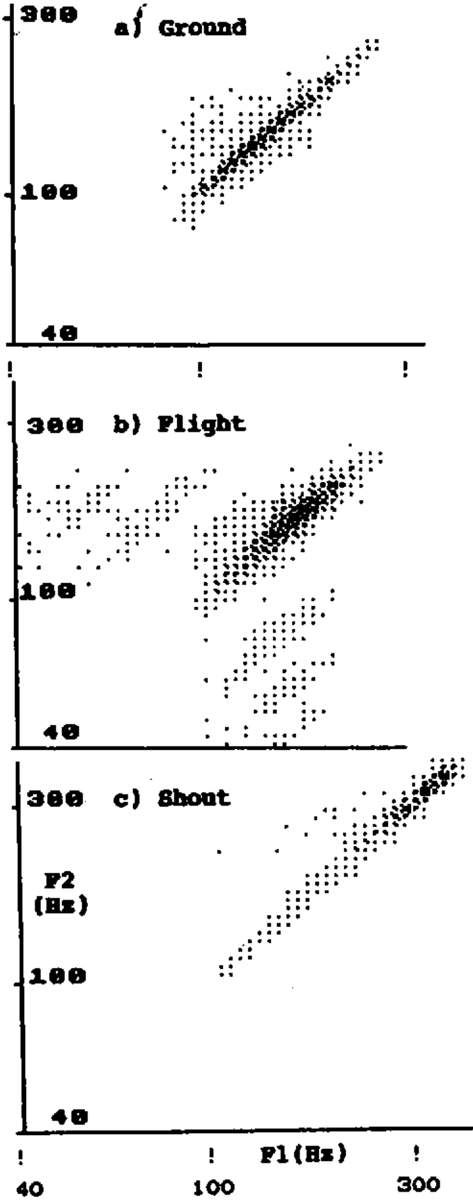
b) Functional. Distribution.

c) Functional. Scatter plot.

In figure 5a the distribution of Fx for this patient is shown (based on the probability of larynx period occurrence, and obtained from a 3 minute fluent speech sample). It can be clearly seen that the low frequency skirt on the histogram is quite different from that of the normal distribution in figure 3. In these cases successful treatment of the oedema leaves the main body of the distribution effectively unchanged and removes the low frequency irregularity.

Quantitative analysis makes it possible to gain some insight into the condition of voices which are not associated with any evident pathology, and in figure 5b "functional" disorder is associated with another adult male sample. The relatively restricted range and the presence of low frequency components are quite evident, and the physical nature of the disorder is glimpsed in the associated plot in figure 5c, where the scattering produced by laryngeal vibrational irregularity is displayed by plotting Fx1 against Fx2 where 1 and 2 relate to immediately adjacent fundamental periods of vibration. A random vibration would simply produce a diffuse distribution, but in figure 5c the central diagonal is associated with two minor parallel diagonals, produced by period doubling in the vibration. The clinical diagnosis of a functional, non-organic, disorder is not confirmed in that there are quite clear physical abnormalities shown in this analysis. A patient may present with no visually obvious symptoms of disorder on laryngoscopic examination and yet have precursive indications of pathology in the quantitative assessments discussed here.

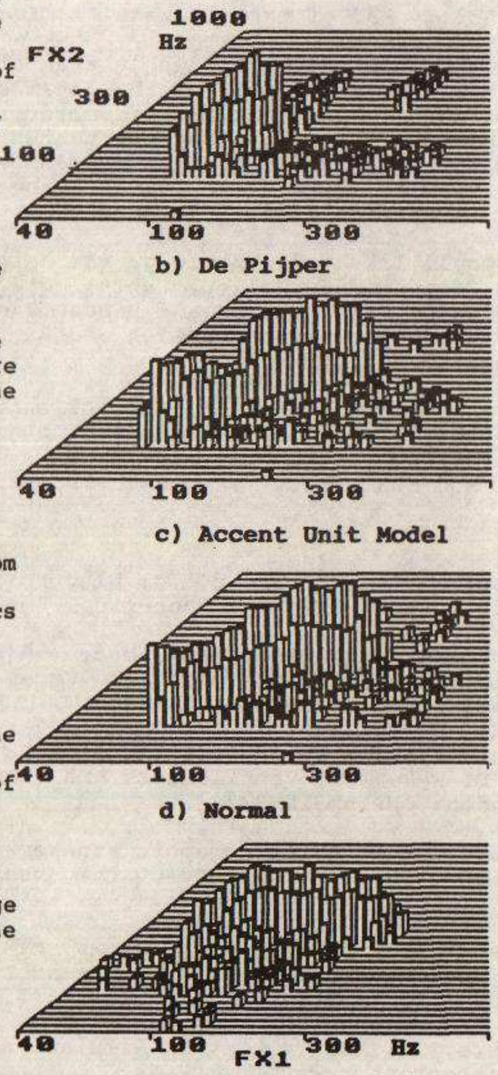
Figure 6. Vibration.
Scatter plots



In figure 6a the normal voice of the distribution of figure 3, "ground", is shown using the same method of analysis for the same sample. The main well-formed diagonal indicates a uniformly produced voice and the slight tendency to an asymmetrical scatter is a function of the prosodic charactistics of the speech. The occurrence, mainly below the modal value, of a preponderance of rising intonation contours gives a slight bias for the second period in any doublet to be smaller. In figure 6b the same prosodic characteristics are now superimposed on a restricted and heightened modal frequency due to the increased acoustic intensity of the utterance - the Lombard effect is really perceived in practice perhaps somewhat paradoxically not by an increase in loudnesss but rather by an increase in perceived pitch - but the main feature of the analysis is associated with the parallel clusterings which in figure 6b are much more pronounced than in the pathological condition of 5c. The lobes, however, still result from integral period spacings and correspond to octave relationships. Vibration has interfered with larynx movement so that very occasionally triplets and slightly more frequently doublets and singlets are missed out in the sequence of vocal fold closures. In figure 6c, it can be seen that shouting whilst it has substantially moved the main mode, still nevertheless leaves the essential range of larynx frequency and with a fairly small degree of scatter. Shouting has not here involved a ventricular component and it is evident that the irregularity of fig. 6(b) is only due to externally applied vibration.

In figure 7 the speech is synthetic and produced by rule, and the analyses are based on the Voiscope processing simply of the synthesiser excitation waveform. As before, in the scatter plots of figures 5 & 6, the first period frequency is plotted horizontally and Fx2, the second frequency, corresponds to the other side of the base of the presentation. Vertically the log of the probability of the occurrence in any Fx1, Fx2 cell is shown by the height of the histogram. Whilst this method of printed presentation has the disadvantage that parts of the distribution are of necessity concealed, it has the advantage of greater probability range than in the simple density scatter plot. In figure 7a, the signal is from a version of the original JSRU synthesis by rule system; scatter features are relatively evident which come from the nature of the implementation and mode and range characteristics from the prosodic rules. The DePijper prosodic algorithm (devised for English by Michael Johnson & Jill House at UCL) has been used for the synthesis of the same passage in figure 7(b) and here the notably bicuspate form of the distribution is clear. This results from the 'top-hat' intonation contours, originally used for Dutch by t'Hart & Cohen. In figure 7(c) an expressive range of intonation has been used in the new model (MJ & JH) but the excessive dichotomy has been reduced - although the distribution is still bimodal. Finally for natural speech, in figure 7 (d) a single mode dominates and the presence of low frequency creaky voice contributes to a marked scattering of low frequency Fx activity.

Figure 7. 3-D Scatter plots.



a) JSRU original

b) De Pijper

c) Accent Unit Model

d) Normal

The association of the analyses presented here with a tabulation of simple numerical statistical measures of dispersion, skewness, kurtosis, makes it possible to provide a useful basis for the comparison both of the effects of treatment in the case of pathology and of the the results of development in the case of synthesis. More importantly, however in the longer term is the contribution which this whole area of excitation analysis will make to the definition of the detailed features of voice in a rule governed fashion, capable of embracing, to a degree, all of the situations which have been presented here.
The following references have been selected simply to give an overview of related work directly bearing on the application and interpretation of Lx information.

We would like to acknowledge the help of the H.Q. Director Army Air Corp. Middle Wallop, Hants, Sgt. Bruce Stewart and W.O. Jack Robson of D&T Squadron and Duncan Tincello of Smiths Industries in the production of the speech and Lx recordings for the helicopter vibration study.

We are most grateful to many colleagues in University College for all the help and programming which has contributed to the usefulness of the analyses which have been discussed, and to the MRC and SERC/Alvey for basic support.

REFERENCES

Abberton,E and Fourcin,A.J: Electrolaryngography. In C.Code and M.Ball (eds) Clinical Phonetics. London: Croom Helm 62-78.

Childers, D.G, Smith, A.M, Moore, G.P:Relationships between Electroglottograph, Speech and Vocal Cord Contact. Folia Phoniatrica 36:105-118 (1984).

Fourcin,A.J.: Laryngographic examination of vocal fold vibration, in B. Wyke (ed) Ventillatory and Phonatory Control systems, London: OUP pp315-333.

Fourcin A.J: Laryngographic assessment of phonatory function, in C L Ludlow (ed) Conference on the Assessment of Vocal Pathology, Maryland: ASHA Reports 11.

Noscoe,N.J, Fourcin A.J., Brown, M.A. and Berry, R.J: Examination of vocal fold movement by ultra-short pulse X-radiography. British Journal of Radiography, 56, 641-645.

Titze,I.R, and Talkin, D: Simulation and interpretation of glottographic waveforms. In C.L. Ludlow (ed) Conference on the Assessment of vocal pathology. ASHA report 11.

These arguments lead naturally to an approach to speech pattern processing which is founded on *information theory* and on speech pattern *modelling*; information about speech and speech patterns is encoded in a suitable model, and appropriate algorithms are used to compute the output of the model for a specified input (for recognition or synthesis).

In general, the problem of modelling physical systems can be decomposed into four sub-problems [9]: (a) *representation*; i.e. the type of model (e.g. static or dynamic, linear or nonlinear, deterministic or stochastic, discrete or continuous etc.), (b) *measurement*; i.e. which physical properties should be measured and how, (c) *estimation*; i.e. the determination of those physical quantities that cannot be measured from those that can, and (d) *validation*; i.e. the demonstration of confidence in the model.

The research issues in speech pattern modelling are therefore concerned with the modelling paradigms (knowledge representation), the representation of acoustic data, the definition of a 'good' interpretation or output, the search strategies for finding a good interpretation (or output) of a model given some input data (optimality), methods for model construction and parameter estimation, and performance assessment procedures.

At the present time the most computationally useful modelling paradigm is based on stochastic generative models and particularly *hidden Markov models* (HMMs) [10]. In this case a-priori speech knowledge is expressed in the structure and parameters of a finite-state machine. The definition of the goodness of any particular interpretation is in terms of the likelihood of the model generating the observed data, and the most likely interpretation (or output) may be found using an optimal search procedure such as dynamic programming. Another key property of such a model is that there exists a parameter re-estimation algorithm (the Baum-Welch algorithm for HMMs) which is guaranteed to increase the likelihood of generating a particular set of observation data by suitable adjustments to the probabilities embedded in the model.

The main advantages of this particular modelling paradigm are that (a) it generalises the non-linear time alignment approach [11,12], (b) it adheres to Marr's *principle of least commitment* [13], (c) by being statistically based it is possible to account for 'unseen' data gracefully (so called 'ignorance based' modelling [14]), and (d) it may be applied at the level of sound segments, words and grammar simultaneously.

The disadvantage of this approach is that, although it is relatively easy to tune the details of a model, it is difficult to derive the its overall topology (structure). However, new modelling paradigms are already appearing which are capable of learning higher order *hidden* structural properties of patterns based on *adaptive parallel distributed processing networks* [15]. This means that it is likely that stochastic generative models will themselves become a special case in a much more general combined structural and stochastic speech pattern modelling paradigm [16].

Key Work Areas

In order to reflect the foregoing research methodology, the work of the Unit is partitioned into five main research areas:-

- speech signal processing
- acoustic-phonetic modelling
- linguistic constraints
- pattern processing principles
- speech systems

The first three are directly related to speech pattern modelling, the fourth provides the theoretical underpinning for the whole programme and the fifth is concerned with practical and implementation issues associated with performing and exploiting the research. The work in all five areas is described in the following section.

RESEARCH PROGRAMME

Speech Signal Processing

The main objective of the Unit's work in speech signal processing is to develop techniques for 'high-resolution' signal analysis in order to provide a more informative representation of speech and speech-related signals. Both acoustic and non-acoustic signals are of interest. These 'rich' representations are needed in order to facilitate higher accuracy speech pattern modelling and to aid the separation of speech from competing signals.

Current work in this area is concerned with two contrasting approaches to analysing the fine temporal and spectral structure of speech signals; the behaviour of a computational model of the human peripheral auditory system [17] is being compared with that from a more mathematical approach based on the Wigner distribution [18]. Work is also in progress on the development of imaging techniques for sensing articulator position (e.g. to provide information about lip movement as an additional cue for automatic speech recognition), and on optimal approaches to low level speech pattern analysis (e.g. formant tracking using dynamic programming [19]).

In the future it is expected that work in the speech signal processing area will move towards the development of mechanisms for making the low-level patterning more explicit, the integration of acoustic and visual speech data, and the active separation of speech from interfering signals and noise.

Acoustic-Phonetic Modelling

This area is the focus for the work of the entire Unit. The main aim is to develop the principles of speech pattern modelling in order to derive *algorithms* for high accuracy recognition and high quality synthesis. This is achieved by means of research into alternative paradigms for modelling both speech and non-speech signals, different algorithms for searching the models and techniques for parameter re-estimation.

A considerable amount of experimental work has been undertaken in this area over recent years, most of it directed towards the development of techniques for overcoming the inherent variability of speech patterns. In particular, effort has been concentrated on algorithms for discriminating accurately between similar sounding words where the phonetic distinction is based on temporal [20] or spectral [21] cues. Also, algorithms for *continuous connected word recognition* have been developed to the stage of commercial exploitation [22].

Current work is concerned with the accurate modelling of the temporal structure of isolated whole-word patterns using hidden *semi-Markov* models (HSMMs) [23,24]. Work is also in progress on models with higher-order Markovian properties [8], the automatic derivation of sub-word structures [21], and new approaches such as error back-propagation networks [25,26].

In the future the modelling and algorithm work will be extended to accomodate higher order properties of a speech signal such as speaking rate, and then applied to more fluent speech. Techniques for adapting models to new speakers will also be investigated.

Linguistic Processing

This area is concerned with the application of the principles of speech pattern processing to the design and construction of structured models which reflect the constraints imposed by relevant a-priori information about the phonological and linguistic structure of speech.

Current work includes a study of stochastic grammars [27] and their integration with existing acoustic-phonetic modelling.

Pattern Processing Principles

This area of work provides the theoretical underpinning to the whole speech research programme and also connections with related work in other disciplines (such as image processing, natural language understanding and self learning machines). Particular aims include the understanding of the principles and inherent limitations of diverse established and newer methods of pattern processing, to develop formal relationships between them, and where possible to put them in the context of of general theories.

Current work supports each of the three main speech pattern processing areas: specifically the study of the Wigner distribution for speech signal processing, parallel distributed processing networks for acoustic-phonetic modelling, and the relationships between hidden Markov modelling and stochastic context-free grammars. Work is also in progress on alternative formalisms for modelling *dialogue* [28].

Speech Systems

The objective of the final area is to provide a working hardware and software environment for speech technology research. This involves the management of general purpose VAX/VMS-based computer facilities consisting of a central VAX 11/8600 cluster for off-line batch processing, two VAX 11/750s primarily for real-time work, and several workstations based on MICRO-VAX II for real-time work and programme development. All the machines are networked using DECNET (via ETHERNET) and three have FPS array processors attached. All of the machines (apart from the 8600s) have AED high resolution colour graphics displays. Special purpose facilities such as the ICL-DAP are also available to the Unit [29].

The speech systems area is also concerned with the management of databases of speech material (e.g. the NATO RSG10 spoken digit database [30] and the RSRE speech database [31]) including advanced methods for storage and retrieval such as optical discs. The work also involves the definition and implementation of internationally agreed assessment procedures.

Finally, this area provides the crucial interface between the fundamental research conducted by the Unit and the systems oriented applications research conducted at RSRE and elsewhere.

COLLABORATION

In order to augment its internal research programme, or to exploit the results of its research, the Unit is able to enter into *collaborative agreements* with UK firms and universities. In this context the Unit can sponsor work, participate in consortia, enter into special bi-partite arrangements, host summer studentships and accomodate industrial attachments.

At present the Unit sponsors a number of contracts and research agreements with various firms and universities (e.g. the work at UCL on 'modelling acoustic and phonetic variability' [32] and on 'the modelling of nuclear tone for speech synthesis' [33], and the work at Keele on 'psychoacoustic constraints for speech recognition' [34]). The Unit is also involved in two Alvey Speech Technology projects (one of which involves an industrial secondment to the Unit) and in a proposed ESPRIT project on 'speech technology assessment'. The Unit is also a key laboratory in the DTI's National Electronics Research Initiative on Pattern Recognition (RIPR). This is based at RSRE and again involves industrial attachments to the Unit.

CONCLUSION

This paper has presented an overview of the RSRE Speech Research Unit. The main objectives of the integrated JSRU and RSRE speech research programmes have been described and it has been shown how the problems of both speech recognition and speech synthesis are being tackled using a combined structural and stochastic approach to *speech pattern processing*.

REFERENCES

[1] R.A. Cole, R.M. Stern and M.J. Lasry, 'Performing fine phonetic distinctions: templates vs. features', *Invariance and Variability in Speech Processes*, J. Perkell and D.H. Klatt (eds.), Erlbaum, (1984).

[2] D.H. Klatt, 'Review of the ARPA speech understanding project', JASA, Vol.62, 1345-1366, (1977).

[3] T. Kohonen, H. Riittinen, E. Renhkala and S. Haltsonen, 'On-line recognition of spoken words from a large vocabulary', Info. Sciences, Vol.33, 3-30, (1984).

[4] L.R. Bahl, F. Jelinek and R.L. Mercer, 'A maximum likelihood approach to continuous speech recognition', IEEE Trans. PAMI, Vol.5, 179-190, (1983).

[5] R.K. Moore, 'Systems for isolated and connected word recognition', NATO ASI Series, Vol.F16, *New Systems and Architectures for Automatic Speech Recognition and Synthesis*, R. DeMori and C.Y. Suen (eds.), Springer-Verlag, (1985).

[6] G. Bristow (ed.), *Electronic Speech Synthesis*, Granada, (1983).

[7] J.N. Holmes, I.G. Mattingly and J.N. Shearme, 'Speech synthesis by rule', Language and Speech, Vol.7, 127-143, (1964).

[8] J.S. Bridle and M.P. Ralls, 'An approach to speech recognition using synthesis by rule', *Computer Speech Processing*, F. Fallside and W. Woods (eds.), Prentice Hall, (1985).

[9] J.M. Mendel, *Discrete Techniques of Parameter Estimation*, Marcel Dekker, (1973).

[10] R.K. Moore, 'Computational techniques', *Electronic Speech Recognition*, G. Bristow (ed.), Collins, 130-157, (1986).

[11] J.S. Bridle, 'Stochastic models and template matching: some important relationships between two apparently different techniques for automatic speech recognition', Proc. IoA Autumn Conf., (1984).

[12] M.J. Russell, R.K. Moore and M.J. Tomlinson, 'Dynamic programming and statistical modelling in automatic speech recognition', J. Operational Research Society, Vol.37, No.1, 21-30, (1986).

[13] D. Marr, 'Early processing of visual information', Phil. Trans. R. Soc. Lond B, Vol.275, 483-519, (1976).

[14] J. Makhoul and R. Schwartz, 'Ignorance modelling', *Invariance and Variability in Speech Processes*, J. Perkell and D.H. Klatt (eds.), Erlbaum, (1984).

[15] J.S. Bridle and R.K. Moore, 'Boltzmann machines for speech pattern processing', Proc. IoA, Vol.6, Part 4, 315-322, (1984).

[16] J.S. Bridle, 'Adaptive networks for speech pattern processing', Proc. NATO ASI on Pattern Recognition Theory and Applications, Spa, Belgium, (1986).

[17] S.W. Beet, R.K. Moore and M.J. Tomlinson, 'Auditory modelling for automatic speech recognition', Proc. IoA Autumn Conf., (1986).
[18] D. Lowe, R.K. Moore and M.J. Tomlinson, 'The Wigner distribution as a speech signal processing tool', Proc. IoA Autumn Conf., (1986).
[19] B.C. Dupree, 'Formant coding of speech using dynamic programming', Electronic Letters, Vol.20, No.7, 279-280, (1980).
[20] M.J. Russell and R.K. Moore, 'Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition', Proc. IEEE Int. Conf. on Acoustics, Speech and Signal processing, 5-8, (1985).
[21] R.K. Moore, M.J. Russell and M.J. Tomlinson, 'The discriminative network; a mechanism for focusing recognition in whole-word pattern matching', Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 1041-1044, (1983).
[22] J.S. Bridle and R.M. Chamberlain, 'Continuous connected word recognition using whole word templates', Radio and Electronic Engineer, Vol.53, No.4, 167-175, (1983).
[23] M.J. Russell and A.E. Cook, 'Experiments in isolated digit recognition using hidden Markov models', Proc. IoA Autumn Conf., (1986).
[24] A.E. Cook and M.J. Russell, 'Improved duration modelling in hidden Markov models using series-parallel configurations of states', Proc. IoA Autumn Conf., (1986).
[25] S. Peeling, R.K. Moore and M.J. Tomlinson, 'The multi-layer Perceptron as a tool for speech pattern processing research', Proc. IoA Autumn Conf., (1986).
[26] S. Peeling and J.S. Bridle, 'Experiments with a learning network for a simple phonetic recognition task', Proc. IoA Autumn Conf., (1986).
[27] J.K. Baker, 'Trainable grammars for speech recognition', 97th meeting of Acoust. Soc. America, D.H. Klatt and J.J. Wolf (eds.), 547-550, (1979).
[28] P.J. Goillau, 'Pattern processing and machine intelligence techniques for representing dialogues', Proc. NATO Workshop on 'Structure of Multimodal Dialogue Including Voice', Venaco, France, Sept., (1986).
[29] P. Simpson and J.B.G. Roberts, 'Speech recognition on a distributed array processor', Electronic Letters, Vol.19, No.24, 1018-1020, (1983).
[30] R.S. Vonusa, J.T. Nelson, S.E. Smith and J.G. Parker, 'NATO AC/243 (Panel III RSG10) language data base', Proc. US National Bureau of Standards Workshop on Standardisation for Speech I/O Technology, 223-228, (1982).
[31] M.J. Russell, R.K. Moore, M.J. Tomlinson and J.C.A. Deacon, 'RSRE speech database recordings 1983: part II', RSRE Report No.84008, (1984).
[32] M.A. Huckvale, 'Modelling acoustic and phonetic variability of speech', Proc. IEE Int. Conf. on Speech Input/Output; Techniques and Applications, 54-58, (1986).
[33] J. House and M. Johnson, 'Natural nuclear tone modelling for speech synthesis by rule', Proc. IEE Int. Conf. on Speech Input/Output; Techniques and Applications, 210-215, (1986).
[34] P. Cosgrove and J.P. Wilson, 'A study of frequency transition detection using synthetic vowels', Proc. IoA Autumn Conf., (1986).