Proceedings of The Institute of Acoustics

PROBLEMS WITH DYNAMIC FREQUENCY WARPING AS A TECHNIQUE FOR SPEAKER-INDEPENDENT VOWEL CLASSIFICATION.

W.A. Ainsworth (1), K.K. Paliwal (2), and H.M. Foster (1)

(1) Dept. of Communication and Neuroscience, University of Keele.
(2) Speech and Digital Systems Group. Tata Institute of Fundamental Research, Bombay 5, India

INTRODUCTION

There are a number of sources of variation in the speech produced by different speakers. Some of these, such as dialect, are learned and are under the control of the speaker. Others, such as vocal tract length and glottal source characteristics, have a physical origin and are less easily controlled. Automatic speech recognisers which operate with a number of speakers will need to take account of these factors.

One of the most straightforward of these factors would appear to be vocal tract length. The shorter the vocal tract, the higher the formant frequencies. All that is required is a method of estimating the vocal tract length from the speech of an unknown speaker, then using this to scale the formants to match the speech of a known speaker. This approach was adopted by Wakita [1], but was not particularly successful. This was partly due to the difficulty of obtaining an accurate estimate of vocal tract length, and partly due to the use of linear scaling. Fant [2] has shown that a non-linear transformation is required to change the formant frequencies of a vowel produced by a male speaker with those of a vowel produced by a female speaker.

A different approach to speaker-independent vowel classification is being investigated which requires no prior knowledge of the input speaker. This approach uses the power spectrum of a vowel as the parameters, and employs non-linear warping along the frequency axis using a dynamic programming algorithm. The algorithm is similar to that used by Sakoe and Chiba [3] to deal with variations in the time domain.

METHOD

A set of templates are formed from the vowels of a reference speaker. The log-power spectra of these vowels are uniformly sampled at m frequency values. Let these be:

$$A_1 = (a_{11}, a_{12}, \ldots a_{1i}, \ldots a_{1m})$$

.

.

$$A_2 = (a_{n1}, a_{n2}, \ldots a_{ni}, \ldots a_{nm})$$

and let the log-power spectrum of an unknown vowel by a different speaker be:

$$B = (b_1, b_2, \ldots b_j, \ldots b_m).$$

The problem is to find the minimum distance D(A, B) between the test spectrum and each of the reference spectra. The reference spectrum associated with the smallest difference then indicates the class of the test vowel. The dynamic programming algorithm enables this minimum distance to be calculated.

The distance between the ith point of the reference spectrum and the jth point of the test spectrum can be defined by:

PROBLEMS WITH DYNAMIC FREQUENCY WARPING AS A TECHNIQUE FOR SPEAKER-INDEPENDENT VOWEL CLASSIFICATION.

$d(i, j) = (a_i - b_j)^2.$

This distance is computed for each pair of points within an area called a 'warping window' which constrains the amount of non-linearity. A symmetric warping window is bounded by $j = i + r$ and $j = i - r$ where r is the warping window length.

The cumulative distance, $g(i, j)$, is then computed from the local distance $d(i, j)$ by [3]:

$$g(i, j) = \min \begin{cases} g(i-1, j) + d(i, j) \\ g(i-1, j-1) + 2d(i, j) \\ g(i, j-1) + d(i, j) \end{cases}$$

given the initial condition $g(1, 1) = 2d(1, 1)$. The final cumulative distance is given by:

$D(A, B) = g(m, m) / 2m.$

DATA

The data used in the experiments consisted of eleven English vowels spoken in an /h/-vowel -/d/ context by eight speakers (four male and four female). Ten repetitions of each of the vowels were recorded by each of the eight speakers.

The speech signal was digitised at a 10kHz sampling rate with a 12-bit analogue-to-digital converter. The waveforms of the /h/-vowel-/d/ utterances were displayed and a 25.6 ms segment from the steady-state part was selected with a manually controlled cursor. The vowel waveform was weighted with a 25.6 ms Hamming window and a 10th order linear prediction analysis was performed [4]. The log-power spectrum of each vowel was obtained from the resulting ten linear prediction coefficients using the fast Fourier transform algorithm. Each vowel was thus represented by an 128-point power spectrum.

In order to train the vowel recognition system for a given speaker, the ten repetitions of each vowel were averaged, and the resulting reference spectra for each vowel were stored.

EXPERIMENT 1

In the first experiment the warping window was made symmetrical bounded by $j = i + r$ and $j = i - r$. The window size, r, was varied from zero (which allows no frequency warping) to 320 Hz. Recognition scores were obtained for the system trained with vowels from male speaker 1 and tested with vowels from male speaker 2 and female speaker 1. The results are shown in Table 1. It can be seen that with the male test vowels the recognition score reduced as the warping window size increased. With the female test vowels the recognition score increased slightly as the warping window size increased.

Observations of the original spectra and the warped spectra showed that the dynammic programming algorithm performed the frequency warping appropriately [5]. The problem was that although the spectrum of a test vowel could be warped to match the spectrum of a reference vowel, it could also be warped to match different reference vowels. Sometimes the cumulative distance was less between vowels from different classes than from the same class, so a misclassification resulted.

PROBLEMS WITH DYNAMIC FREQUENCY WARPING AS A TECHNIQUE FOR SPEAKER-INDEPENDENT VOWEL CLASSIFICATION.

Table 1. Vowel recognition scores for symmetric warping windows bounded by $j = i+r$ and $j = i - r$. The recognition system was trained for male speaker 1.

Warping window, r (Hz)	Recognition score (%)	
	Male speaker 2	Female speaker 1
0	82.7	23.6
40	80.0	25.5
80	77.3	24.5
120	72.7	24.5
160	68.2	26.4
200	66.4	24.5
240	62.7	25.5
280	58.2	26.4
320	51.8	24.5

As the size of the warping window appeared to have only a small effect on the recognition scores, it was fixed at 160 Hz and a complete set of experiments was performed in which the vowels of each speaker served in turn as the reference patterns and all of the vowels acted as the test patterns [5]. The results showed that in speaker-dependent mode, where the reference and test vowels came from the same speaker, the recognition score averaged 96%. In speaker-independent mode where the reference vowels came from a different speaker from the test vowels the score was about 40%. When the reference vowels were produced by a male speaker and the test vowels by a female speaker, or vice versa, the average recognition score was only 33%.

EXPERIMENT 2

In Experiment 1 the warping window was symmetric about the axis. This is appropriate if nothing is known about the speakers. However, although the test speaker is unknown, the identity of the speaker who provides the reference vowels can be known. If he is a male speaker it might be expected that vowel spectra from another male speaker would need to be expanded or compressed with equal probability. With test spectra from a female speaker, however, it would be expected that more compression than expansion would be required. To test this hypothesis an experiment was performed with asymmetric warping windows, with male reference spectra and female test spectra.

It has been found that limiting the frequency range of the spectra from 250 Hz to 3200 Hz, and with a symmetric warping window of 160 Hz, a recognition score of 39.3% was obtained [6].

With the warping window bounded by $j = i + 320$ and $j = i$, the score increased to 47.0%, and with the warping window bounded by $j = i$ and $j = i - 320$, the score fell to 24.4%. Thus for a male reference spectra and female test spectra an appropriate asymmetric warping window can increase the recognition score.

A parametric study of the window size was performed [6]. The warping window was bounded by $j = i + q$ and $j = i$, and q ranged from 0 to 480 Hz. The results are shown in Table 2. A maximum recognition score of 48.6% was achieved with a warping window size of 160 Hz.

Table 2. Vowel recognition scores for asymmetric warping windows bounded by $j = i + q$ and $j = i$. The system was trained for each of the male speakers and tested for each of the female speakers.

Warping window, q (Hz)	Recognition score (%)
0	38.4
80	44.4
160	48.6
320	47.0
480	43.8

EXPERIMENT 3

Just as the identity of the speaker of the reference vowels is known, and this knowledge can be used to constrain the warping window, so is the identity of each of the reference vowels known. There is no reason to suppose that each vowel is best served by the same warping window. Suppose each vowel has a warping window bounded by $j = i + p_n + r_n$ and $j = i + p_n - r_n$ where p_n is the asymmetry factor and r_n is the window size. The problem is to determine the optimum values of p_n and r_n. (In general the warping windows might have arbitrary shapes, so many more parameters might be involved.)

An attempt has been made to estimate these values by setting r_n = 80 Hz and varying the asymmetry factor, p_n. The asymmetry factor p_n was first set at 80 Hz (which corresponds to the maximum score obtained in Experiment 2). Then p_1 (the asymmetry factor for the first vowel) was increased by 80 Hz and the new recognition score obtained. If this improved the overall score p_1 was increased again and the new score determined, but if the score decreased p_1 was reduced by 80 Hz and the process repeated. Once the optimum value for p_1 had been determined, the process was repeated for p_2, etc. This cycle was repeated twice for all eleven vowels. The score at the end of the second cycle was not much different from that at the end of the first.

Table 3. Recognition scores for vowel-dependent warping windows. The average score for test vowels from all female speakers is shown for each male reference speaker.

Male reference speaker	Recognition score (%)
1	56.6
2	62.3
3	52.0
4	65.9
Mean	59.2

PROBLEMS WITH DYNAMIC FREQUENCY WARPING AS A TECHNIQUE FOR SPEAKER-INDEPENDENT VOWEL CLASSIFICATION.

The results are shown in Table 3 for each of the male speakers. In each case the vowels from the female speakers were employed as the test patterns. The mean score was 59.2% and the best score was 65.9% with male speaker 4.

Table 4. Mean values of the asymmetry factor for each of the vowels in Experiment 3.

Vowel	Asymmetry factor (Hz)
/i/	120
/ɪ/	240
/ɛ/	180
/æ/	220
/ɑ/	180
/ɒ/	80
/ɔ/	100
/ʊ/	160
/u/	40
/ʌ/	140
/ɜ/	80

The mean values of the asymmetry factor for each vowel are shown in Table 4. These tend to be larger for open and front vowels and smaller for close and back vowels, but the correlation is not very great. They also tend to be larger for short vowels than for long vowels, but the reason for this is not known.

CONCLUSIONS

A dynamic programming algorithm is being investigated as a technique for introducing non-linear frequency warping in a speaker-independent vowel classifier. It has been found that although the algorithm is capable of warping the spectrum of a vowel produced by one speaker to match the spectrum of the same vowel produced by another speaker, it is also capable of warping it to match a different vowel. In its unconstrained form, it functions poorly as a vowel classifier. With male reference spectra and female test spectra the recognition score was only 33%. Using an asymmetric warping window increased this score to 48.6%, and using vowel-dependent warping windows increased the score to 59.2%. If the vowels from the best male speaker are chosen as the reference patterns, the average score becomes 65.9%. With a speaker-dependent system, however, the score obtained was 96.1%.

Acknowledgements

The work was supported by a grant from the Science and Engineering Research Council.

REFERENCES

[1] H. Wakita, 'Normalization of vowels by vocal tract length and its application to vowel identification', I.E.E.E. Trans., ASSP-25, no.2, 183-192, (1977).

[2] G. Fant, 'Non-uniform vowel normalization', STL QPSR, no. 2-3, 1-19, (1977).

[3] H. Sakoe and S. Chiba, 'Dynamic programming algorithm for spoken word recognition', I.E.E.E. Trans., ASSP-28, no.6, 623-635, (1978).

[4] J. Makhoul, 'Linear prediction: a tutorial review', Proc. I.E.E.E., vol. 63, no.4, 561-580, (1975).

[5] K.K. Paliwal and W.A. Ainsworth, 'Dynamic frequency warping for speaker adaptation in automatic speech recognition', to be published.

[6] W.A. Ainsworth and H.M. Foster, 'The use of dynamic frequency warping in a speaker-independent vowel classifier', Proc. Nato ASI on New Systems and Architectures for Automatic Speech Recognition and Synthesis, Bonas, France, (1984).

Proceedings of The Institute of Acoustics

A NOISE COMPENSATING SPECTRUM DISTANCE MEASURE APPLIED TO AUTOMATIC SPEECH RECOGNITION

J.S. Bridle, K.M. Ponting, M.D. Brown and A.W. Borrett

Joint Speech Research Unit, Princess Elizabeth Way, Cheltenham, GL52 5AJ

INTRODUCTION

One important practical problem in automatic speech recognition (ASR) is to deal with interfering noise. The types of noise of interest include environmental noise, such as air conditioning or generator whine, non-speech sounds made by the speaker, such as coughs or breath, and intrusive sounds of short duration, such as the banging of a door. This paper is concerned only with additive noise which can be characterised by its amplitude and spectrum shape, both of which are assumed to be slowly-varying functions of time. It is hoped that the techniques based on these assumptions will be useful in dealing with background noise in many real applications of ASR, such as its use in aircraft cockpits and in some noisy industrial environments.

The requirement is for a method which will allow a speech recognition algorithm to continue functioning even when the background noise obscures the differences between some speech sounds and masks some sounds altogether. We are concerned with "template matching" ASR methods (e.g. [1]), where the problem to be solved is that of comparing two spectra ("template" and "input") taken from speech in different background noise conditions. From the two spectra and the separately computed estimates of the corresponding background noise spectra we want to compute a "distance" measure indicating the dissimilarity between the two spectra. (The method of obtaining the estimates of the background noise spectra is not directly relevant to the algorithm presented in this paper.) The algorithms presented do not depend on the precise representation used for the spectra, but for the purposes of illustration, we assume spectra to be composed of logarithmically coded spectrum amplitudes from a filter bank analyser.

Two main approaches have been suggested for comparing speech spectra in the presence of background noise. Firstly, the method of noise subtraction has been suggested [2]. The simplest version of this technique involves subtracting from each speech spectrum the amount of power in each channel that is thought to be due to the background noise. The resulting spectra are then compared using a standard distance calculation as if no noise had been present. The most serious objection to this procedure is that when the noise level in some channel is close to the level of the speech, then the level resulting from the subtraction contains essentially no information about the original speech level.

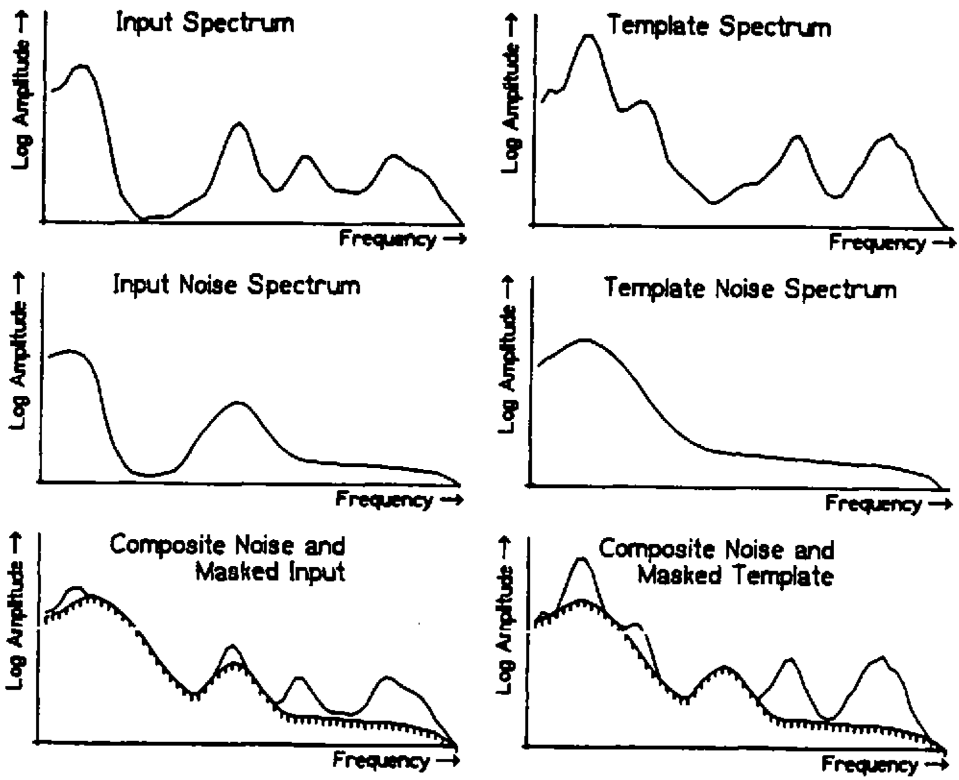
The second approach is that of "noise masking", suggested by Klatt [3]. This approach removes any spurious differences due to the different noise backgrounds by setting any channel amplitude in either speech spectrum which is lower than the larger of the corresponding channel noise estimates to that noise level. However, this method is not ideal and suffers from several drawbacks described below.

To overcome the drawbacks of Klatt's approach, we have developed a noise compensation algorithm [4] which allows for the presence of noise when calculating the distance between spectra. To set the scene, we first briefly describe Klatt's noise masking algorithm with its inherent disadvantages. We then present the principles of the JSRU noise compensation algorithm and report some preliminary experiments using a hardware implementation of this algorithm.

KLATT'S NOISE MASKING ALGORITHM

The intention of Klatt's noise masking approach is to modify the two speech spectra so that they appear to have the same background noise. The modified spectra can then be compared directly using a standard distance calculation.

Figure 1: Klatt's noise masking algorithm



From the two noise estimates, m and n, corresponding to the two speech spectra to be compared, a composite noise spectrum N or "mask" is calculated from

$$N = \mathrm{Max}\,(m, n) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

The two speech spectra, x and y, can then be masked by this composite noise spectrum to produce the two modified spectra:

$$X = \mathrm{Max}\ (\ x\ ,\ N\ ) \qquad (2)$$

$$Y = \mathrm{Max}\ (\ y\ ,\ N\ ) \qquad (3)$$

It can easily be shown that, for similar speech spectra in different background noise conditions, the method will indeed produce two similar spectra for comparison.

However, there are problems with Klatt's method, one of which is illustrated by Figure 1. This shows how meaningful and perhaps important differences (such as, in this case, the amplitude and frequency of the second formant) may be reduced or even completely obscured when genuine information about the input speech spectrum is masked by the template spectrum noise estimate or vice versa.

For time-varying background noise, the noise estimates would have to be stored with the templates during template acquisition, and the distance calculation would be made rather complicated by the need to calculate equations (1-3) for each channel whenever an input frame is compared with a template frame.

JSRU NOISE COMPENSATION ALGORITHM

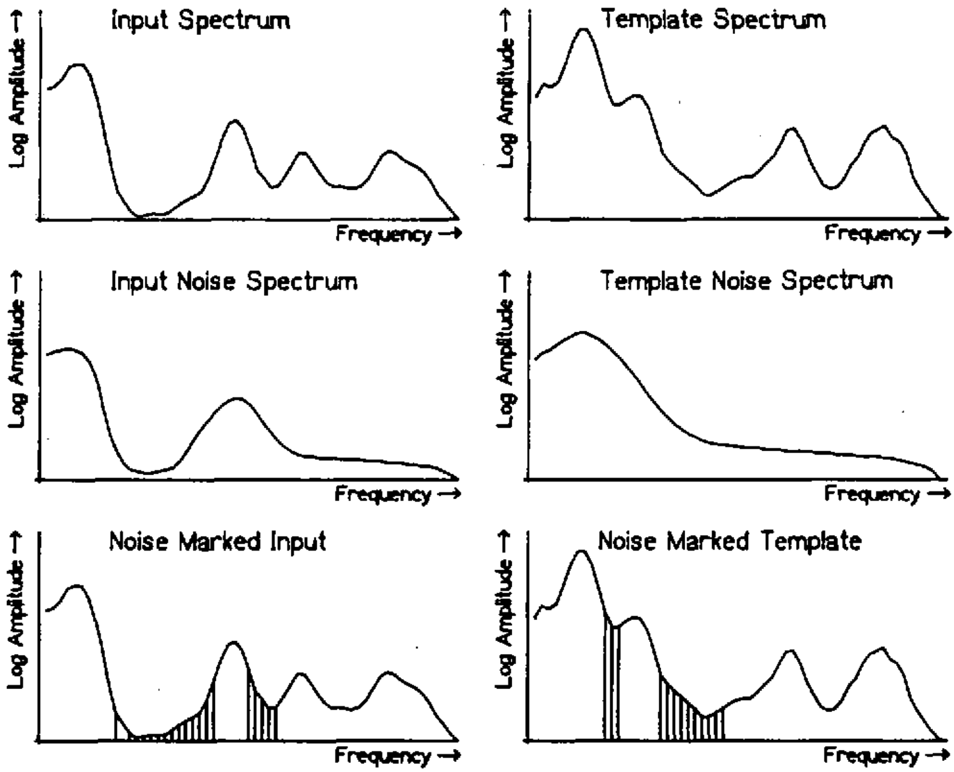
The ideal spectrum distance measure would preserve the advantage of Klatt's method over the noise subtraction approach and solve the anomaly of Klatt's method, but be simpler to implement. Our solution to the problem is a noise compensation method which consists of two separate parts: the noise estimation and noise marking performed in the acoustic analysis; and the modified distance calculation, which takes account of the information about noise levels. This noise compensation algorithm can be implemented simply in hardware.

Noise Marking

In our approach, an important first step is taken as the spectra arrive. Using the current estimate of the background noise, we compare each "speech" spectrum amplitude with the corresponding noise estimate and decide whether the "speech" amplitude is more likely to be due to the noise than to the speech. These "noise marking" decisions, along with the spectrum amplitudes, contain all the information about the noise estimate that we shall need in the distance calculation. (In general we should compute a likelihood of the amplitude being due to noise rather than speech, but in all our implementations we have used a hard decision and a single bit to record it.) Figure 2 illustrates this procedure. Those sections of the spectra marked as noise are shown by vertical shading.

A NOISE COMPENSATING SPECTRUM DISTANCE MEASURE APPLIED TO ASR

FIGURE 2: JSRU noise compensation algorithm - marked channel amplitudes



Noise Distance

The standard spectrum distance calculation is modified to include the information about the noise. Let Cx and Cy denote corresponding spectrum channel amplitudes, and f(Cx,Cy) the usual distance function for a single channel, typically a squared Euclidean distance measure. If the larger of Cx and Cy is marked as noise, as in Regions 2, 4 and 6 of Figure 3, then nothing can be deduced about the difference between the two underlying speech spectra in this channel. Instead of assigning a zero value (which denotes a perfect match) to the distance, D, a suitably-chosen non-zero "noise distance", d*, is used:

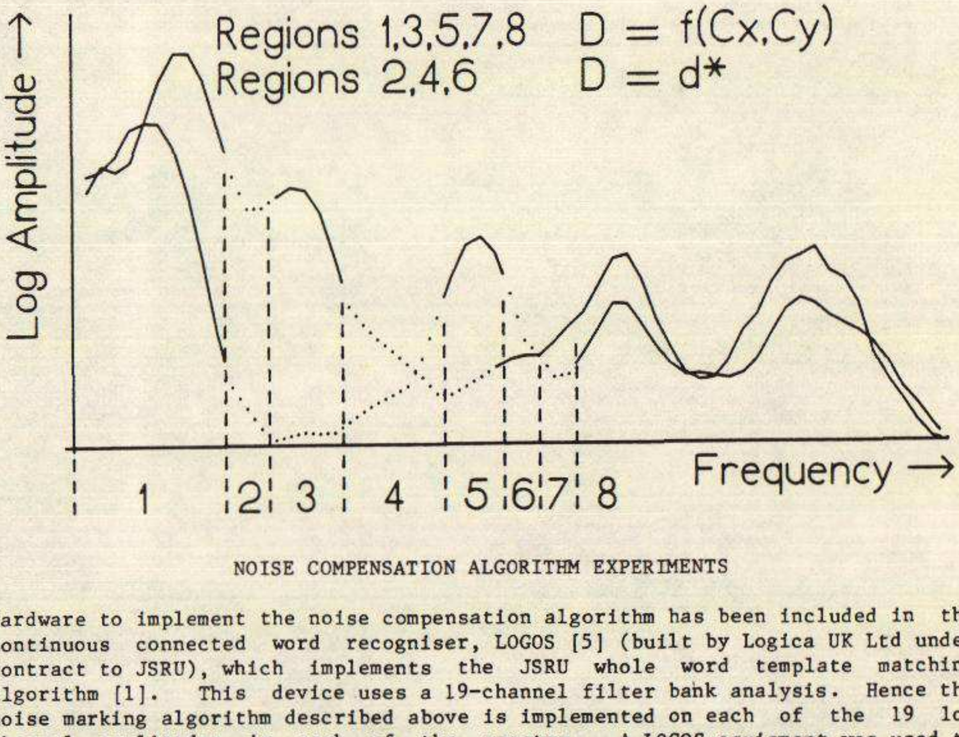
$$D = d^* \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

In this way a perfect match can only be found between speech spectra which are identical and have all their values marked as speech and not from a comparison of spectra which have one or more values marked as noise. If the larger of Cx and Cy is marked as speech, as in the other regions of Figure 3, then the distance is calculated from the channel amplitudes in the usual way:

$$D = f(Cx, Cy) \quad \ldots (5)$$

This equation uses all the available information because, even if the lower level is marked as noise, the distance between the two underlying signal levels must be at least that in (5). The distance between the two spectra can now be found by adding together over all channels the values of d* or f(Cx,Cy) as appropriate.

FIGURE 3: JSRU noise compensation algorithm - distances



NOISE COMPENSATION ALGORITHM EXPERIMENTS

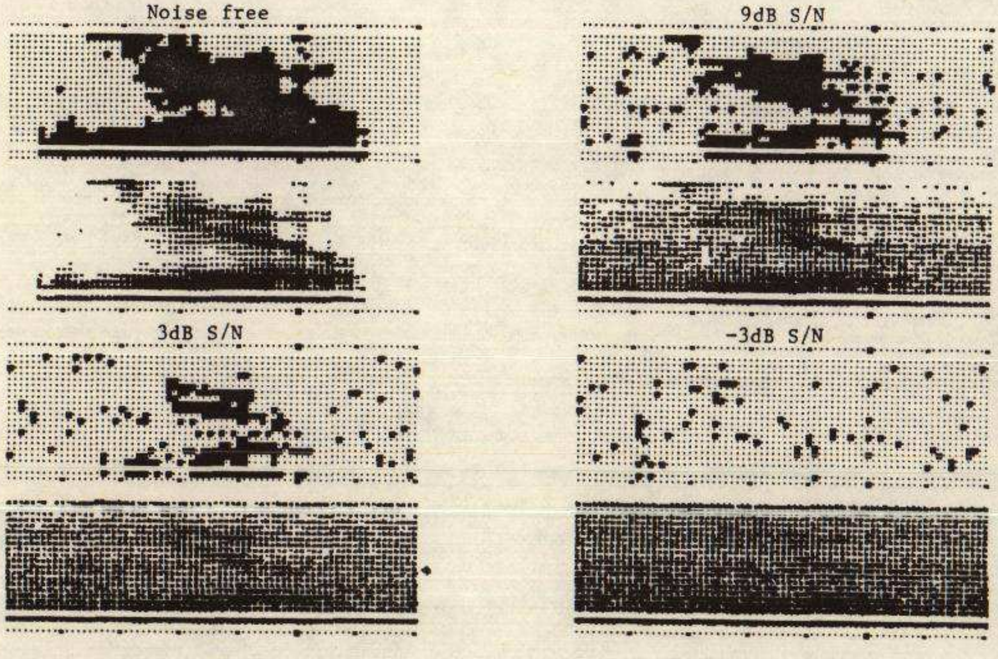
Hardware to implement the noise compensation algorithm has been included in the continuous connected word recogniser, LOGOS [5] (built by Logica UK Ltd under contract to JSRU), which implements the JSRU whole word template matching algorithm [1]. This device uses a 19-channel filter bank analysis. Hence the noise marking algorithm described above is implemented on each of the 19 log channel amplitudes in each of the spectra. A LOGOS equipment was used to investigate the effectiveness of the noise compensation algorithm on the test material described below.

Test material

For this experiment we used data which, though very unrealistic, is well defined and compatible with the assumptions behind our method. We used recordings prepared at the Institute for Perception, TNO, The Netherlands as part of a cooperative research project by the NATO Research Study Group concerned with automatic speech recognition. It is planned that these same noisy recordings will be used for other experiments in our own and other laboratories. The test material consists of a series of isolated and connected digits produced by one speaker, including two "training tables" (SA and SB) of 100 isolated digits in non-random order and three "test tables" (1A, 1B and 1C) of 100 isolated digits in random order.

A long-term average power spectrum of the original speech was measured, and stationary noise with roughly the same spectrum shape was added to the original speech, to produce conditions referred to as 9dB, 3dB and -3dB signal-to-noise ratios. Examples of the same utterance of "zero" in the three noise conditions and without noise are shown in Figure 4. For each condition, noise marking decisions are shown for the 19 channels and separately for the peak channel above the corresponding spectrum amplitude picture.

FIGURE 4: Example utterance of "zero": noise free and in the 3 noise conditions



Template extraction

This experiment was primarily concerned with the performance of the noise compensation algorithm and was not concerned with template preparation or optimum speech recognition performance. One occurrence of each digit from table SB was used as the template for that digit. These were manually extracted from digitised spectrum data captured using LOGOS. A "silence" template is required to explain the between-digit silences. This was obtained for each condition under test using the standard LOGOS "silence training" procedure, which, when invoked, calculates an average spectrum from half a second of input.

Experiments on three noise conditions

Some pilot runs on the training tables (SA and SB) of the test material were undertaken to "tune" LOGOS prior to the experiments on the "unseen" test tables (1A, 1B and 1C). Once suitable parameters of the LOGOS recognition algorithm had been established, they were held constant throughout the experiments.

The experimental runs reported here were performed on all the isolated digit test material available. For each experimental run, Logos was used with the 10 digit templates and the appropriate silence template to recognise the speech material for all 5 isolated digit tables.

To test the effectiveness of the noise compensation algorithm, recognition performance was measured for the three noise conditions with and without the noise compensation algorithm in operation.

RESULTS

The results obtained from the two experimental runs on each of the three different signal-to-noise conditions are summarised in Table 1. Each of the RSG10 tables contains 100 spoken digits (10 of each).

Table 1: Numbers of digits correctly recognised

	With noise compensation						Without noise compensation					
	RSG10 Table						RSG10 Table					
S/N	Training		Test			Total	Training		Test			Total
ratio	SA	SB	1A	1B	1C		SA	SB	1A	1B	1C	
9dB	98	95	92	96	100	481	98	97	92	96	98	481
3dB	85	84	65	79	86	399	44	45	38	46	48	221
-3dB	67	67	64	69	74	341	11	10	10	10	10	51

We make the following observations on the results:

1. The results for the -3dB condition without noise compensation were entirely spurious. We found that several hundred 9's were the recognition outputs for each 100-digit table. This is shown as 10% correct, which is the most charitable interpretation possible.

2. There were almost no extra digits recognised (insertions), except for the -3dB condition without noise compensation.

3. Since the results were obtained from analogue signals there would be some variation between successive runs even if all else were constant.

4. We should discount the results for the 2 training tables as the recogniser had been "tuned" to produce "optimum" results on this data, and the templates were derived from table SB.

5. Any detailed analysis of these results is beyond the scope of this paper.

It is clear from the above limited experiments that speech recognition performance using the JSRU noise compensation algorithm as implemented in LOGOS is considerably better than the performance without it, particularly at the high background noise levels.

It is intended that a JSRU Research Report [6] will contain full details of the results presented here and also the results of related experiments planned for the future.

CONCLUSIONS

A noise compensation method is described for comparing two noisy spectrum cross sections. This method ensures that differences which are probably due to the presence of known amounts of noise in the spectra contribute very little to the distance measure. The method combines both simplicity and power to provide an algorithm suitable for hardware implementation. In our noise compensation experiments reported here, the noise compensation makes little difference for high signal-to-noise ratios, but for low signal-to-noise ratios there is considerable improvement in performance.

ACKNOWLEDGMENT

R.M. Chamberlain contributed substantially to the early work on the development of the principles described in this paper, and was responsible for producing the patent application embodying this algorithm.

REFERENCES

[1] J.S.Bridle, M.D.Brown and R.M.Chamberlain, 'Continuous connected word recognition using whole word templates', The Radio and Electronic Engineer, Vol.53, No.4, 167-175, (1983).
[2] B.Beek and R.S.Vonusa, 'General review of military applications of voice processing', NATO Advisory Group for Aerospace Research and Development (AGARD), Lecture Series No.129 on Speech Processing, (1983).
[3] D.H.Klatt, 'A digital filter bank for spectral matching', Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-76, 573-576, (1976).
[4] J.S.Bridle and R.M.Chamberlain, 'Noise compensating spectral distance processor', U.K. Patent No.8233119, (1983).
[5] J.B.Peckham, J.R.D.Green, J.V.Canning and P.Stephens, 'A real-time hardware continuous speech recognition system', Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-82, 863-866, (1982).
[6] J.S.Bridle, K.M.Ponting, M.D.Brown and A.W.Borrett, 'A noise compensating spectrum distance measure applied to automatic speech recognition', JSRU Research Report No.1012, (1985).