FEATURES FOR A COMPUTER WORD RECOGNITION SYSTEM

Tracy M Clark, W K Kennedy, and R H T Bates.
Department of Electrical and Electronic Engineering,
University of Canterbury, New Zealand

Abstract

Any review of the extensive literature on word recognition reveals that a large variety of speech features is used in computer based word recognition. However, most of the work focusses on a limited set of features adapted to a selected recognition method.

We report on a series of experiments designed to isolate the relative merits of a range of features. The results for each feature or set of features are standardised by testing them with female and male speakers having a New Zealand accent using a vocabulary zero to nine. A dynamic time warping algorithm is used. The features tested include root mean squared, zero crossing rate, linear predictive coefficients, cepstral coefficients, and transitional data in the form of dynamic cepstral coefficients. It is found that the best performance is achieved with the cepstral information. Addition of other features to this set gives only marginal improvement.

INTRODUCTION

To objectively evaluate the utility of computer word recognition features with respect to each other, they must be tested using the same algorithm while maintaining database parameters such as speaker, recording techniques and vocabulary constant. In this way recognition results for each feature can be compared for a particular speaker.

To evaluate the effect on recognition accuracy with respect to the speaker a number of different speakers have been tested, both male and female. These speakers all have New Zealand accents. A standard database containing the ten digit words zero to nine is used. These words have been spoken on average 20 times by each speaker.

The recognition algorithm is an unconstrained endpoint dynamic time warping(DTW) scheme from which the calculated distances are used for deciding recognition outcome.

This paper begins by discussing the recognition algorithm used for the tests. The following section presents the features used for recognition which included root mean squared(RMS), zero crossing rate(ZX), linear predictive coefficients(LPC), cepstral coefficients(CEP), and dynamic cepstral data. Results for optimum performance of the features for each speaker are given in the final section.

RECOGNITION ALGORITHM

The recognition algorithm used for testing all the features discussed, is a speaker dependent, isolated word scheme. The input speech is blocked into frames of samples from which the acoustic features to be tested are calculated. These acoustic features are stored to be used for either recognition, as a test word, or for training, as a reference word. For recognition, the reference and test features are time aligned via a dynamic time warping procedure described by Itakura(1975) , and a Euclidean distance is calculated between the test word and each reference word in the dictionary. After the test word has been matched against all the reference words the set of distances is scanned to select the minimum which identifies the recognized word. A block diagram of the recognition algorithm is shown in Figure 1.

Each speaker recorded the ten words up to twenty times in a quiet room. The words were digitized at 10 kHz after being passed through a 4.5 kHz low pass anti-aliasing filter. The digitized words

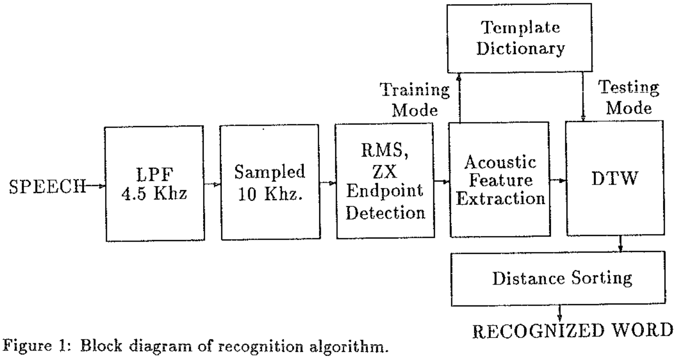

Figure 1: Block diagram of recognition algorithm.

were stored on disk for future algorithm testing. Ten repetitions were used for training while the other repetitions for each speaker were kept for testing.

The features tested using the recognition algorithm were root mean square value, zero crossing rate, linear prediction coefficients, cepstral coefficients, and dynamic cepstral coefficients. These features were calculated with a range of pre-processing variables to study the effect of window size, window type and pre-emphasis. Window lengths were varied over a range of 150 to 400 samples using either a rectangular or Hamming window. If pre-emphasis was used a scale factor of 0.95 was chosen.

RECOGNITION FEATURES

Root Mean Square (RMS)

The root mean square value of the signal characterises the loudness of the speech sound and is calculated as

$$(RMS)_k = \sqrt{\frac{\sum(S_i^2)}{\Upsilon}}$$

where S_i is the ith sample in the kth frame and Υ is the number of samples in the kth frame. To obtain results that are independent of average signal level this measure must be normalized across the word such that each word has the same maximum value. The set of RMS values of a word gives a simple, rapidly calculated feature that is capable of distinguishing words with unique envelopes, such as the words 'six', 'seven' and 'eight'. Unfortunately, however, many other words tend to look very similar using this measure, examples of these are the words 'one' and 'nine', or the words 'three' and 'four'. Relatively low levels of microphone noise, spikes and breath noises from the speaker can cause large errors.

Energy measurement has been used for some time in word recognition schemes, often as an addition to some frequency representation, such as LPCs (Rabiner *et al.*, 1984a) (Rabiner, 1984) (Rabiner *et al.*, 1984b), or zero-crossings (Lau and Chan, 1985). However when used as the only parameter for recognition it has been reported to give a very low accuracy. Rabiner(1984) claimed only 30% accuracy with energy alone. This is verified by our study in which the RMS parameter gave recognition accuracies ranging from 16% to 60%. Optimum conditions were very speaker dependent but recognition tended to improve with a window size between two to three pitch periods and with no pre-emphasis of the speech. Overlapping the windows gave a slight increase in accuracy.

Although the performance of RMS as a recognition parameter is poor it does have an important role in the endpointing of words (Lamel *et al.*, 1981).

Zero crossing rate (ZX)

It was established by Licklider *et al.*(1948) that infinitely clipped speech, that is speech that retains only its zero-crossing positions, remains highly intelligible to human listeners. Since then zero-crossing measures have been used as a recognition parameter. Although the time intervals between successive crossings of the zero line are related in a complicated way to the frequencies present in the sound, it is possible to obtain a frequency related measure by recording the rate of these crossings within a frame of the sound. The principle application of this measure is in separating the relatively low frequency voiced sounds from the high frequency unvoiced components.

Zero-crossing rate in some time interval is the most widely used measure in zero-crossing analysis(Niederjohn, 1975). Using the same notation as Niederjohn the zero-crossing rate can be written as

$$Z_k = M_k/\Upsilon$$

where M_k is the number of zero crossings in the frame k, and Υ is the number of samples in the frame.

The ability of zero-crossings to distinguish unvoiced sounds from background silence makes this measure useful for endpointing words. However breath noise along with other higher frequency sounds can cause erroneous identification of the endpoints.

We found that the optimum conditions when zero-crossing rate was used alone for word recognition occurred with a window of 200 samples (approximately two to three pitch periods), with pre-emphasis and with 30 percent overlap of frames. The increase in recognition accuracy obtained with pre-emphasis is due to the increased amplitude of the unvoiced component.

Linear Prediction coefficients (LPC)

LPC coding is one of the most common techniques used in the processing of speech, and in particular in the field of speech recognition. Representation of speech using LPCs is equivalent to approximating its short term power spectrum by an all-pole speech production model (Makhoul, 1975). Thus LPCs (a_k) can be written as the filter coefficients of the linear system representing the vocal tract and given by

$$H(z) = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$

Modelling by all-pole methods has severe drawbacks for particular sounds, especially those which contain spectral nulls, such as occur in nasalised sounds. LPC analysis on this type of sound gives results which vary significantly for nominally similar signals, particularly around the region of the spectral zeros (Juang *et al.*, 1987). Another problem is that the order of the model must be chosen a priori. The fixed number of poles thus enforced can lead to the addition of spurious resonant peaks.

Recognition accuracies reported using LPCs have been in the high 80 or low 90% region. We were unable to reproduce that level of accuracy. Our recognition results ranged from 20 to 80% depending on the speech pre-processing used and the speaker. The optimum conditions occurred when frames of 200 samples of the speech were pre-emphasised and Hamming windowed. Overlapping the analysis frames did not give significant improvements, and in some cases reduced recognition accuracy. Recognition accuracy also increased when silence and low amplitude fricative sounds were removed from within the word.

Cepstral Coefficients(CEP)

Cepstral coefficients are now the most widely used feature for word recognition. The real cepstral coefficients are defined as the Fourier transform of the log spectrum and are calculated recursively from the LPCs as (Furui, 1986)

$$-kc_k = ka_k + \sum_{n=1}^{k-1}(k-n)c_{k-n}a_n \qquad k > 0$$

Cepstral processing produces a smoothed version of the LPC spectrum with smoothness dependent on the number of coefficients chosen - ten are usually used.

Previous studies have given recognition results, using cepstral coefficients, well into the 90% region and usually around 98 to 99%(Juang *et al.*, 1987) (Rabiner *et al.*, 1989) (Furui, 1986). In this study the performance of cepstral recognition was heavily dependent on the speaker, with results ranging from 16% to 90% but mostly clustered around 80%. Optimum conditions occurred when two or three pitch periods of speech were used, Hamming windowed and pre-emphasised. Overlapping data frames did not give any significant improvement.

Dynamic Cepstral Coefficients (DCEP)

Spectral transitions as well as instantaneous spectral features are believed to be important for sound recognition. Dynamic spectral feature analysis is still in its infancy and only a few researchers are using this information for word recognition, although it has been used more widely as a speaker recognition tool (Soong and Rosenberg, 1988). Simple first order finite differences are far to noisy to be used as dynamic information so the method used in this study followed that of Soong*et al.*(1988). Using the LPC-based cepstral coefficients ($c_m(t)$), orthogonal polynomials which characterise the time trajectories of the cepstral coefficients over a finite number (2k+1) of fixed length frames are calculated. For a first order orthogonal polynomial two coefficients are calculated. The zeroth order or constant term is given by

$$c_m^{-}(t) = \frac{\sum_{k=-K}^{K} h_k c_m(t+k)}{\sum_{k=-K}^{K} h_k}$$

where h_k is a symmetric window function of length (2k+1) frames. The first order orthogonal polynomial coefficient, or spectral slope is

$$\Delta c_m(t) = \frac{\sum_{k=-K}^{K} k h_k c_m(t+k)}{\sum_{k=-K}^{K} h_k k^2}$$

When Furui(1986) used these measures he showed the dynamic coefficients extracted from nine frame intervals to give only slightly better recognition performance than the instantaneous cepstrum coefficients. In contrast, we obtained recognition accuracies higher than the cepstral coefficients alone by up to 10%, with results ranging from 60% to 96%. Optimum conditions occurred, as with the cepstral coefficients previously, with frames of data of two to three pitch periods, pre-emphasised and Hamming windowed. The dynamic coefficients gave best results when calculated over seven to nine frames, with this window being moved forward one frame at a time. Window function h_k was tested as constant ($h_k = 1$), linear ($h_k = k$), and squared ($h_k = k^2$) with the highest accuracy occurring with a linear window function.

RESULTS

Testing has so far been completed for five male and two female speakers. The results in Table 1 are given for the most useful subset of the variables tested.

The best results, across all speakers, were obtained using dynamic cepstral coefficients. We are therefore led to believe the representation of the the transitional information of the speech is an important recognition feature.

The second greatest increase in recognition resulted from applying a Hamming window to the data. This pre-processing step gave recognition increases for all speakers.

Pre-emphasised data was found to work better for most speakers and for all features except RMS, while overlapping data frames did not seem to give any significant improvement.

Feature (number used)	samples/ frame	sample overlap	window type	pre-emphasis	Male Speakers					Females Speakers	
					AE	AM	CC	CP	MC	DR	TC
RMS(1)	200	0	rect	No	30	44	40	54	42	31	57
	200	0	rect	Yes	26	36	26	48	36	32	46
	200	0	Hamm	No	30	48	33	59	38	29	44
	200	0	Hamm	Yes	23	36	34	50	45	37	-
	200	60	rect	No	33	42	-	51	43	30	55
ZX(1)	200	0	rect	No	33	31	22	61	33	39	25
	200	0	rect	Yes	45	35	32	70	58	54	55
	200	60	rect	Yes	49	32	39	83	60	59	61
LPCs(10)	200	0	rect	No	21	44	33	54	35	33	38
	200	0	rect	Yes	20	48	47	61	38	51	-
	200	0	Hamm	No	33	50	43	69	32	40	74
	200	0	Hamm	Yes	40	70	54	74	40	52	73
	200	60	Hamm	Yes	45	50	-	56	-	26	78
CEP(10)	200	0	rect	No	45	60	47	71	52	48	-
	200	0	rect	Yes	47	58	49	77	43	69	-
	200	0	Hamm	No	65	70	56	79	58	60	81
	200	0	Hamm	Yes	79	76	67	87	53	76	89
DCEP	window $h_k = 7 frames, linear$										
Zeroth order(10)	200	0	Hamm	Yes	84	90	86	96	75	86	95
First order(10)	200	0	Hamm	Yes	75	82	86	93	72	86	90

Table1. Percentage recognition accuracies for each speaker.

CONCLUSIONS

We tested a range of feature types using different forms of pre-processing on the speech of a set of New Zealand speakers. The best results were obtained using dynamic cepstral coefficients which represents the transitional information of the speech. The results also showed the single most useful pre-processing step was to Hamming window the data.

Using New Zealand speakers allowed us to compare results with those from American speakers discussed in the literature. It also allows us to tentatively compare accent effects on recognition performance. Firstly, an examination of the confusion matrices shows that the results for New Zealand speakers are slightly different to those of their American counterparts. For New Zealand speakers the major confusion is between the words 'one' and 'nine', while the results from American speakers give major confusion between 'five' and 'nine'. Secondly, those speakers with a stronger 'New Zealand' accent, that is with very nasalised vowels, gave lower accuracies, such speakers were AE, AM, CC and MC.

It might be expected that higher accuracy would be achieved when sets of features are combined for recognition. Our initial testing has revealed that overall recognition accuracy is improved by 2 to 3%, equivalent to a 75% reduction in error, when the 'best' features (cepstral and dynamic cepstral) are combined.

ACKNOWLEDGEMENTS

We thankfully acknowledge the financial assistance for our research from Unisys-Linc Development Centre (Christchurch).

REFERENCES

Furui, S. (1986), 'Speaker-independent isolated word recognition using dynamic features of speech spectrum', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-34, No. 1, Feb, 52–59.

Itakura, F. (1975), 'Minimum prediction residual principle applied to speech recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-23, No. 1, Feb, 67–72.

Juang, B., Rabiner, L.R. and Wilpon, J.G. (1987), 'On the use of bandpass liftering in speech recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-35, No. 7, July, 947–954.

Lamel, L.F., Rabiner, L.R., Rosenberg, A.E. and Wilpon, J.G. (1981), 'An improved endpoint detector for isolated word recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-29, No. 4, August, 777–785.

Lau, Y. and Chan, C. (1985), 'Speech recognition based on zero crossing rate and energy', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-33, No. 1, February, 320–323.

Licklider, J. and Pollack, I. (1948), 'Effects of differentiation, integration and infinite peak clipping upon the intelligibility of speech', *Journal of the Acoustical Society of America*, Vol. 20, No. 1, January, 42–51.

Makhoul, J. (1975), 'Linear prediction: A tutorial review', *Proceedings of the IEEE*, Vol. 63, No. 4, April, 561–580.

Niederjohn, R.J. (1975), 'A mathematical formulation and comparison of zero-crossing analysis techniques which have been applied to automatic speech recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-23, No. 4, August, 373–379.

Rabiner, L. (1984), 'On the applications of energy contours to the recognition of connected word sequences', *Bell Systems Technical Journal*, Vol. 63, No. 9, November, 1981–1995.

Rabiner, L.R., Pan, K.C. and Soong, F.K. (1984a), 'On the performance of isolated word speech recognizers using vector quantization and temporal energy contours', *AT & T Technical Journal*, Vol. 63, No. 7, Sept, 1245–1260.

Rabiner, L., Sondhi, M. and Levinson, S. (1984b), 'A vector quantizer combining energy and LPC parameters and its application to isolated word recognition', *AT & T Technical Journal*, Vol. 63, No. 5, May, 721–736.

Rabiner, L., Wilpon, J. and Soong, F. (1989), 'High performance connected digit recognition using hidden Markov models', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-37, No. 8, August, 1214–1225.

Soong, F.K. and Rosenberg, A.E. (1988), 'On the use of instantaneous and transisional spectral information in speaker recognition', *IEEE Transactions in Acoustics, Speech and Signal Processing*, Vol. ASSP-36, No. 6, June, 871–879.