Proceedings of the Institute of Acoustics

Feature Histograms as a Model of Speech Perception

R. Linggard (1), P Linford (1), & J Oglesby (2)

(1) UEA, Norwich
(2) BT Laboratories, Martlesham Heath.

Abstract.

This paper reports the results of an investigation into the use of a discrete set of MFCC features in a speaker-independent, isolated-word recogniser, and discusses the implications for speech perception. In these experiments, a discrete set of feature types is obtained by Vector Quantising frames of speech encoded as Mel-frequency Cepstral Coefficients (MFCC). Since the frames of speech from which the MFCCs are derived are over-lapped to a large extent, sequence information is contained, implicitly, in the frames which occur. Thus an utterance may be represented by a simple histogram of the feature-types it contains. The frame duration and overlap in this scheme, are optimised using a simple recognition paradigm. In training mode, a feature histogram, which acts as a template, is constructed for each word in the vocabulary. In recognition mode, the histogram of the unknown word is compared with the template histograms.

1. Introduction

If speech is considered to be a sequence of phonetic features, then, with an appropriate analysis technique, it should be possible to specify and detect such features. A word can then be represented as an ordered list of the phonetic features which occur in the word. This is the basis of phonetic transcription - order being important because the phonetic features are considered, in theory at least, to be independent and isolated [1]. However, in attempting to recognise speech automatically, it becomes clear that phonetic features seem to overlap and modify each other, and the transitions between some phonetic features can be regarded as phonetic features. Thus, the difficulty with a phonetic representation is that a complete and consistent set of phonetic features is difficult to define.

A more direct approach would be to allow a set of features (phonetic or otherwise) to be specified automatically, so that there is no ambiguity in their definition and no difficulty in their automatic recognition. Perhaps the simplest way to do this is to split the speech signal into frames of a certain duration and quantise them on an statistical basis. That is, if particular frames of approximately the same type occur often enough then they will be given a specific label. A useful mechanism for performing this automatic labelling of frames is Vector Quantisation (VQ). Using a suitable VQ system, significant features in the speech signal can be defined automatically, so that the process of defining features can be performed without any phonetic preconceptions. [2], [3]

Feature Histograms as a Model of Speech Perception

The experiments described here, use a fixed duration frame of spectral parameters (MFCCs) as the basic feature, and a Kohonen Neural Network (K-net) as a Vector Quantiser [4]. The frames are overlapped to a large extent so that the frames representing a particular word will be insensitive to synchronisation with the beginning of the utterance. Given large overlapping of frames, the sequence information is already inherent in the frames which occur, so that short utterances can be defined by a list, or histogram, of the features-types which they contain. Word recognition may then be accomplished by using feature histograms to represent words in a template comparison scheme. In defining a suitable set of features (no longer really phonetic) via this paradigm, two parameters have to be determined, the duration/overlap of the frames, and the number of states in the Vector Quantiser. The accuracy of an isolated-word, speaker dependent, speech recogniser was used to determine optimum values of these parameters.

2. The Framing Algorithm and Database

The data used in these experiments were taken from a database of 10 speakers giving 4 utterances of 100 town names. The experiments were carried out using all the data of one male speaker, two utterances for training and two for testing. The main results were confirmed using data from a female speaker.

The speech signals in the database were collected at 10,000 samples per second and stored with 16 bits per sample. Each word in the database is stored as a separate file and converted to frames by the following process. The time samples are separated into frames of a fixed duration, each frame overlapping those on either side so that a given sample will occur in several consecutive frames. Each of these time frames is then converted to a set of 8 Mel-Frequency Cepstral Coefficients (MFCCs). The particular MFCC conversion process used here is as follows. The frame of time samples is shaped using a Hamming window and then converted to a spectrum using the FFT. The components of this spectrum are then grouped into 20 bands on a subjective frequency (Mel) scale. The amplitude of each band is compressed logarithmically, and the shape of the this compressed Mel-spectrum is encoded as the first 8 coefficients of its cosine transform.

3. The Vector Quantiser.

Kohonen networks are known to perform well as vector quantisers [5]. The training procedure spreads the nodes of the network through the n-space occupied by the data in such a way that they reflect the probability distribution of the data. This is a desirable property for a vector quantiser in these experiments since it guarantees that data is represented with minimum error. The MFCC frames are mainly representative of speech, with some derived from noise and/or silence. It is assumed that similar frames will form clusters in the n-space, most clusters representing speech and some representing noise and silence. The vector quantiser will populate the n-space with nodes in such a way as to reflect the local density of the MFCC frames.

Feature Histograms as a Model of Speech Perception

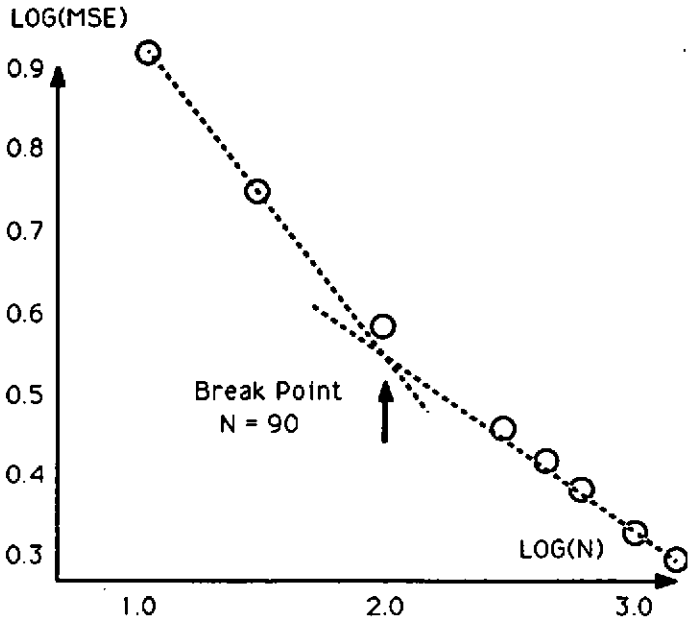

Fig 1. Mean Square Quantisation Error against increasing K-net size (number of VQ nodes).

It is necessary, initially, to determine how many nodes to use in the VQ system. If the VQ has too few nodes then the quantisation error in representing each frame by its nearest VQ node will be, on average, quite large. As the number of nodes in the VQ is increased, this error will fall until the number of nodes approaches an optimum for representing the data. After this point the VQ will have redundant nodes which it may then use to split existing clusters. As this occurs it would be expected that the rate of decrease in quantisation error will fall off.

An experiment to investigate the change in quantisation error with increase in VQ nodes was carried out with un-endpointed speech data from one male speaker, using two utterances of the 100 words in the vocabulary. The MFCC frames from this data were processed using the Kohonen algorithm for several K-nets each having a different number of nodes. The results are summarised in the graph of Fig 1, which shows means square quantisation error against the number of nodes in the K-net. As expected, the error falls off as the number of nodes is increased, with a break in the actual slope at approximately 90 nodes. This is taken to indicate that the data can be represented efficiently using about 100 nodes. In order to allow a safety margin, 256 nodes were used in all the following experiments.

Feature Histograms as a Model of Speech Perception

4. Comparison of Histograms

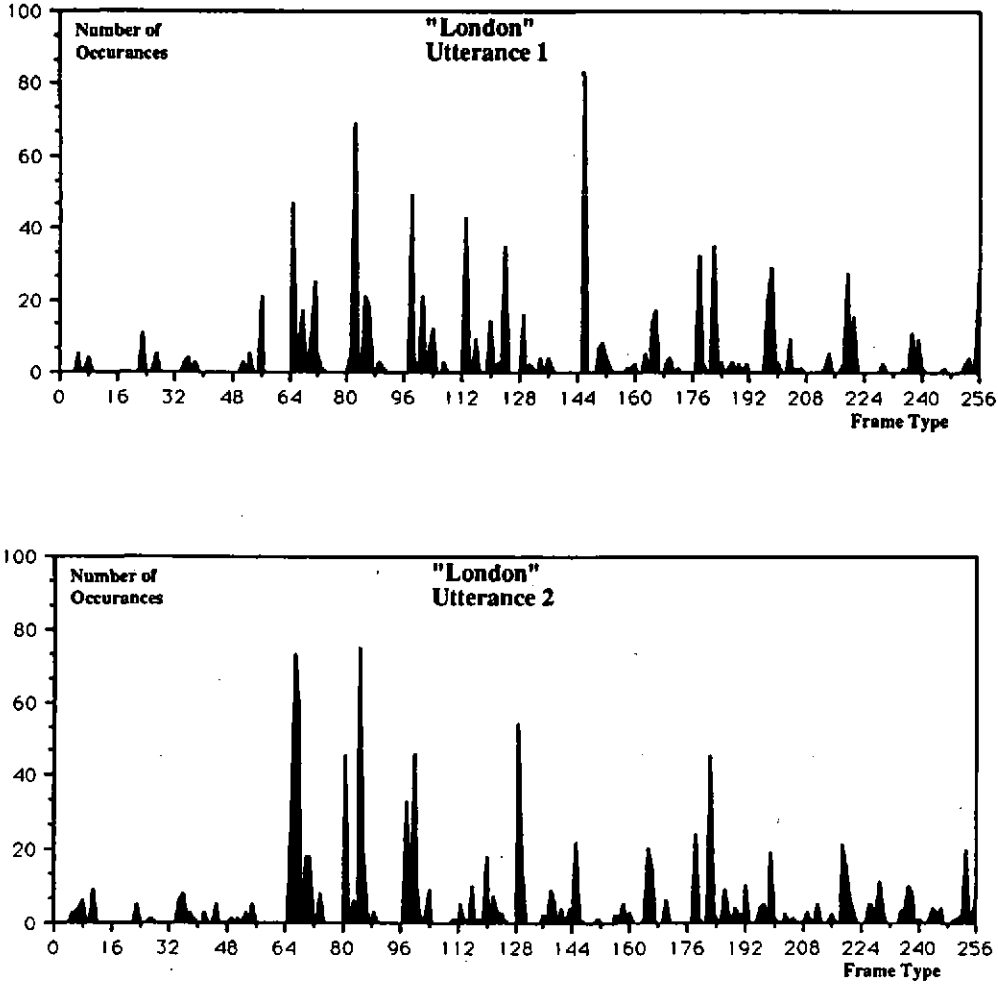

Fig 2 Frame-Type Histograms of the word "London"

Feature Histograms as a Model of Speech Perception

Histograms of particular words can be constructed by using the K-net to classify each MFCC frame in the word, and recording the number of occurrences of each frame type. Two histograms of the word "London", spoken by the same speaker, are shown in Fig 2. Each of the 256 divisions of the horizontal axis represents a frame-type, and the vertical height represents the number of occurrences of that particular frame-type in the word. In this scheme, all frames have been classified and are represented in the histogram, thus the area of the histogram is proportional to the number of frames in the word, that is, to its duration.

It can be seen that a single word does not contain an even distribution of frame-types. Some types do not occur at all and others occur many times causing large peaks in the histogram. The apparent periodicy is due to the fact that the K-net is two dimensional and the 256 nodes are arranged as 16 x 16. Thus the peaks occuring at intervals of 16 are part of the same group. The presence of peaks in the histogram is due to the occurrence of segments in the word where the same frame-type repeats several times, this is mainly the situation with vocalics, particularly vowels.

The variation in height of some spikes in these two utterances of the same word is due, partly, to the natural variation in vowel length, and partly to the slight difference in vowel quality which shifts the peak to an adjacent nodes. In devising schemes to measure the distance between histograms, it is important that these "vowel spikes" should not dominate the distance measurement. Methods to compensate for this effect are equivalent to a normalising process, similar to that of dynamic time warping.

5. Frame Duration and Overlap.

In order to find reasonably optimum values for the frame duration and overlap, three values of each were used in the following experiments. Frame duration was either 25.6 mS, 12.8 mS, or 6.4 mS, and the overlap was either 50%, 75%, or 90%, giving nine complete sets of frame data.

A 256 node K-net was constructed using the training data from the male speaker which consists of two utterances each of 100 town names. Two template histograms were prepared for each word in the vocabulary using the training utterances. The test utterances were converted to histograms and compared with the templates using a simple Euclidean distance, nearest-neighbour classifier. This was carried out for each of the nine sets of duration/overlap parameters. The results of these experiments are shown in Table 1.

Frame Duration	Window Overlap (%)		
	50	75	90
6.4mS	74.5%	80.5%	82%
12.8mS	68%	71%	60.5%
25.6mS	58%	67.5%	70%

Table 1. % Recognition Accuracy

In terms of overlap, the highest (90%) gives the best result. This was to be expected since the larger the overlap the greater the correlation between successive frames and the more they will embody the sequence information. It was also expected that longer frame durations would give a more coherent representation of the speech signal since these would correspond more nearly to the size of conventional phonetic events. However, as the table shows, the shortest frame length produced the best results despite the fact that a window of 6.4mS gives poor frequency resolution. This is in contradiction to the strategy of most MFCC front end processing schemes where frame durations of 20 to 30 mS are typical. The optimum framing condition of 6.4 mS with 90% overlap corresponds to a frame of 64 samples sliding 7 samples. These values were adopted for all subsequent experiments.

6. Biased Correlation.

Small differences between large histogram components can have the same effect as large differences between small components, which will bias the Euclidean distance metric in favour of vowel differences. The actual calculation of the Euclidean distance is

$$E^2 = \Sigma_i E_i^2 = \Sigma_i (A_i^2 - B_i^2)$$

where A_i and B_i are corresponding components in two histograms.

If the histogram components are large then their difference will tend to dominate the distance measure, which is not desirable if the large components are due to vowels. A more meaningful way of comparing two histograms would be to normalise the difference between two components by their magnitude. This can be achieved using a simple correlation metric of the form

$$C_{T_i} = 2A_iB_i/(A_i^2 + B_i^2)$$

Since A_i and B_i are always positive, the correlation is unity if A_i and B_i are equal, and less than unity if they are unequal. The total correlation between two histograms is then the sum of the simple correlation between components.

This expression for correlation has two weaknesses. If A_i and B_i are both zero then C_{T_i} will be indeterminate, whereas it should equal one, as an indicator of maximum similarity. If A_i = 0 and B_i = a small value, then C_{T_i} will equal zero, whereas it should have a value which reflects the fact that a small number and zero may correlate quite well. These difficulties may be overcome if a small element β is added to each component in the histograms.

$$C_{B_i} = \frac{2(A_i+\beta)(B_i+\beta)}{((A_i+\beta)^2+(B_i+\beta)^2)}$$

With this modification, it is obvious that if $A_i = B_i$ then C_{B_i} is still unity, so long as $\beta>0$. The bias term, β, acts as a non-linear compression parameter, and its value can be used to maximise the similarity between histograms of the same word.

Feature Histograms as a Model of Speech Perception

The result of using this "biased correlation" measure on the data with 6.4mS frames and 90% overlap, processed by a K-net with 256 nodes, is given in Table 2.

Bias Value β	% Correct
10^{-10}	85.5
1.0	92.0
5.0	94.5
10.0	96.5
20.0	96.5

Table 2. Recognition Accuracy using Biased Correlation

7. Discussion

The original objective of this investigation was to find a more objective definition of phonetic features in the speech signal by clustering MFCC frames of a standard duration and overlap. The expectation was that frames of durations in the region 20 to 30 mS would cluster strongly on centres which would correspond, approximately, to conventional phonetic categories. Though frames of duration 25.6 mS did cluster with reasonable efficiency, the duration of frames which showed the most useful clustering properties was 6.4 mS, much too short to correspond to a complete phonetic segment. This surprising result suggests that these features are acoustic, rather than phonetic, and correspond to units which would be detected at the level of the auditory periphery rather than in the cortex.

The model of speech perception which is implied by this recognition scheme is radically different to that on which conventional speech recognisers are based. Obviously, the sequence of sound units is critically important in human recognition of speech utterances. However, the success of the scheme presented here suggests that sequence over short periods of time is perceived implicitly rather than explicitly. That is, the perception of a single time frame persists long enough for it to be integrated with the perception of several other frames, and this multiple perception allows the nervous system to form higher level percepts based on the integrated perception of several frames. This figure of 6 mS is roughly equal to the time required for an impulse to travel the length of the basilar membrane (BM), and it is interesting to speculate that these basic frames correspond to instantaneous patterns of vibration on the BM, which are detected by the neurons of the spiral ganglion acting as an associative memory.

8. Acknowledgements

The work described here is supported by the Speech Applications Division at BT Laboratories, Martlesham Heath, and forms part of a joint project between BT Labs and UEA on basic research into speech and speaker recognition.

Feature Histograms as a Model of Speech Perception

9. References

[1] J D O'CONNOR "PHONETICS" PENGUIN 1991, ISBN 0-14-013638-X.

[2] J S MASON, J OGLESBY & XU "Codebooks to optimise Speaker Recognition" European Conference on Speech Technology, pp 207-270, Sept 1989.

[3] LIN-CHENG LIU, DENNIS CHOU & WSIAO-CHUAN WNG "A Speech Recognition method based on Feature Distributions", Pattern Recognition Vol 24, Number 8, pp 717-722, 1991.

[4] T KOHONEN, "Clustering, Taxonomy, and Topological Maps of Patterns." Proc. Int. Conf on Pattern Recognition, Oct. 1982.

[5] G D TATTERSALL, P W LINFORD & R LINGGARD "Neural Arrays for speech recognition". BT Technical Journal Vol 6 Number 2 April 1988.