AUTOMATIC SPEAKER RECOGNITION FOR USE OVER COMMUNICATION CHANNELS

Melvyn J. Hunt, John W. Yates and John S. Bridle

Joint Speech Research Unit
Eastcote Road
Ruislip, Middlesex, U.K.

SUMMARY

A system is described for the automatic comparison of speakers given short samples of their speech. The method does not depend on knowing what is being said, and is to a large extent independent of the degradations likely to be suffered by the speech during transmission. A small computer has been used to generate statistics on fundamental frequency and spectral shape information produced by a real-time cepstrum processor. Fundamental frequency is intrinsically resistant to most transmission degradations, and the spectral statistics taken are independent of linear spectral shaping. A useful level of recognition performance has been obtained using a total of 154 20s samples of read speech from thirteen typical speakers of British English.

1. INTRODUCTION

The work described in this paper is part of an attempt to develop a completely automatic means of speaker identification. Such a facility might be useful, for example, in the early stages of a police investigation where a telephoned message is to be compared with the recorded voices of a large number of suspects. The requirements, then, are that the system should be text independent and that as far as possible it should work on speech that has been degraded by transmission over communications channels such as the telephone. Satisfying these requirements entails some disadvantage in terms of accuracy of identification on high-quality speech and in terms of the length of speech sample needed. We are aiming for a system that would take as input a twenty second sample of telephone quality speech from an unknown speaker. The output would be a list of the five or ten percent of the speakers known to the system whose voices are most similar to that of the unknown speaker. Atal (1) has provided a clear and comprehensive account of the general principles and considerations involved in this sort of investigation.

Speaker recognition research requires the use of large quantities of speech, so for reasons of speed and storage limitations it is desirable to interpose a hardware speech processor between the analogue speech recording and the general purpose computer. At the outset of the work described in this paper we were aware that the fundamental frequency of the voice was likely to be useful for speaker recognition (2), and that the periodicity of the speech signal was not affected by the usual channel distortions. We considered that the 'cepstrum' method described by Noll (3) would be well suited to estimating fundamental frequency in telephone speech, so we processed the speech with a hardware implementation of Noll's method. With a 5 kHz sampling rate this 'cepstrum processor' produces an estimate of the fundamental period every 20 ms. It also provides the intermediate results of short-term log power spectrum and its cosine transform, the cepstrum.

2. THE SYSTEM

2.1 Overview

Because of the difficulty of selecting specific phonetic features automatically in telephone speech, we have confined ourselves to considering the overall statistical properties of the speech. We have written programs to reduce a twenty second speech sample to a set of about thirty numbers describing its statistical behaviour. We refer to these numbers as 'input parameters', and we hope that a suitably transformed set of them will characterise the speaker.

After the input parameters have been produced the rest of the system provides facilities for testing, displaying and transforming them, adding and deleting parameters, dividing data into sets containing selected speakers or selected groups of recordings and finally attempting recognition. By using small, modular programs all accessing files of a standard format, we have a flexible vehicle for carrying out pattern recognition experiments. Beyond the point where the input parameters are produced the system is quite independent of what features in the speech produced them, or indeed of the fact that they are concerned with recognising people from their voices at all.

In a speaker recognition experiment a group of speech samples is taken from each of a number of speakers. The speech samples are subsequently divided into two groups - a group to be used to train the system and a group representing unknown samples to be used to test recognition. The samples are reduced to parameter sets and recognition performance will depend on how tightly the sets from each speaker - or more generally, recognition class - cluster in the parameter space relative to the separation between speakers or classes. The training set is used to select useful input parameters and to find a transformation of them that will lead to

tight clustering.

2.2 Parameter Selection and Transformation

An estimate of the usefulness of a parameter taken in isolation is provided by the ratio of its between-class to within-class variance, usually known as the F ratio. Low values of this ratio can be used to weed out unpromising input parameters.

Frequently, though, several apparently promising parameters will be highly correlated with one another, so that their contribution to overall recognition performance cannot be assessed individually. Insofar as the correlations are linear, the technique known as linear discriminant analysis (LDA) can be used. In the quantity known as 'divergence' (1) LDA provides an estimate of the likely recognition performance of a subset of the input parameters. At the same time it specifies the linear transformation of the input parameters which makes the within-class covariance for the training set isotropic and which condenses the discriminating power of the input parameters into as few transformed parameters as possible.

The transformation specified by LDA will be optimal for recognition if the training set is representative of the properties of the test material, and if the distributions within classes are multivariate normal and differ between classes only in their means. LDA cannot, for example, deal with non-linear correlations, and the transformation it specifies can be distorted by single uncharacteristic examples within a class. To check for these possibilities we can inspect scatter plots of selected pairs of input or transformed parameters on a display screen. We also carry out crude tests (skew and kurtosis) of normality for each parameter individually. Non-linear transformations of the parameters can be effected and tested using a BASIC program in which the transformation required is specified at run time.

2.3 Recognition Phase

The transformations determined from the training set are applied to both the test and training sets. The test set can then be used to measure recognition performance. The sets of parameters in the training set for each speaker are averaged to produce a centroid intended to characterise that speaker. A sample from the test set is then ascribed to the speaker whose centroid is closest to its parameter set. Since LDA makes the estimated within-speaker covariance isotropic, there is no need to weight the parameters in the recognition phase and a simple Euclidean distance metric can be used.

Instead of using the distances from the test sample to the centroids, one could look at the distances to all the individual samples in the training set and use a modified k-nearest-neighbour rule as the recognition criterion. We have a fairly large number of classes with a small number of approximately normally distributed samples within each class. In this situation the advantage of considering the training samples separately would probably be small and would not justify the increased computational effort involved.

3. CHOICE OF FEATURES

3.1 Fundamental Frequency

The repetition rate of glottal excitation in voiced speech, appearing as harmonic structure in the spectrum, is a feature which is resistant to moderate amounts of almost all the distortions encountered in communications channels. Statistics of the values of fundamental frequency estimated by the cepstrum processor during a speech sample should therefore be reasonably channel independent. This independence can be helped by a sensible choice of statistics. For example, high-order moments should be avoided since these will be heavily affected by outlying, probably spurious, points. In work carried out so far we have also avoided statistics concerned with longer-term features in the intonation pattern. This is because reliable estimation of such statistics requires long samples of speech; they are probably more dependent on the situation the talker is in than short-term statistics and it is difficult to devise measures of long-term features which are not liable to be disrupted by sudden bursts of noise.

3.2 Spectrum Envelope Measures

The spectrum envelope characteristics of speech are a function of the glottal pulse shape, the intrinsic dimensions of the vocal tract and the state of the articulators. Such characteristics are therefore promising sources of speaker identification information. The cepstrum processor provides a compact description of the spectrum envelope in the low-order cepstrum coefficients.

Spectrum shape measures are in general very susceptible to transmission distortions. It is, however, possible to devise measures derived from low-order cepstrum coefficients which are unaffected by moderate amounts of linear filtering. The principle on which these measures are based has already been described by Atal (1). Briefly, any spectrum shaping effect will appear as a multiplicative factor in the power spectrum, and so will be additive in the log power spectrum and in any linear transformation of this spectrum such as the cepstrum. If the spectrum shaping effects of the channel are constant over the speech sample, any measure of the variability in the cepstrum coefficients about their means will be unaffected by spectrum shaping. In choosing statistics of low-order cepstrum coefficients, then, one should only use measures which are unaffected by the addition of a constant value to all samples, for example, moments about the mean and differences between consecutive values of a coefficient.

In practice, even with high signal-to-noise ratios, the finite word lengths used in computing the cepstrum coefficients result in the invariance property holding only for moderate amounts of spectrum shaping. Moreover, such measures have no special resistance to the effects of non-linear distortions and additive noise, though some degree of immunity to short-term high-level noise bursts can be obtained by taking statistics only on sections of recordings judged to be voiced speech from their harmonic structure.

4. SPEECH MATERIAL

The material used in this study consisted of radio weather forecasts read by professional meteorologists and recorded from an FM broadcast receiver. Weather forecasts form a particularly convenient body of speech material: they consist of two or three minutes of read speech produced by typical speakers of British English who are not professional announcers or actors; they occur at regular times of day and the speaker's name is always announced; finally, individuals normally remain part of the team of weather forecasters for several years, so that long-term changes in voices can be studied. Although no special care was taken in making the recordings, they are certainly of higher quality than might be expected on communications channels such as telephones.

We collected recordings from thirteen speakers, eleven men and two women. For one of the male speakers it was unfortunately impossible to collect more than one recording. Consequently, this speaker had to be excluded from many of the experiments.

Two separate data sets were made from the recordings. The first consisted of just the fundamental frequency information for 330 thirty-second extracts from the forecasts. This corresponds to about 25 extracts per speaker taken from an average of six weather forecasts. The second contained both fundamental frequency information and the first eight cepstrum coefficients. In this set there were a total of 149 twenty-second extracts taken from exactly two forecasts for twelve speakers, plus five twenty-second extracts from the single forecast of the thirteenth speaker.

5. EXPERIMENTS

5.1 General Considerations

There can be two quite separate products of speaker identification experiments: one can use them to discover ways of obtaining improved recognition performance, and one can use them to test the performance of an existing system. Although they are superficially similar, the two kinds of experiments often have conflicting requirements. Performance testing demands strict separation of test and training sets. Not only must no speech extract occur in both sets, but also all extracts from the same recording session must be kept together in one set or the other. Moreover, the test set can only be used once if it is not to provide feedback which would help to train the system and thus give unrealistically optimistic results. By contrast, performance improvement often requires repeated use of the same test set; to do otherwise would be very expensive on data and would be an unnecessary source of variability in the results. A less rigorous separation of test and training sets may also provide a more convenient experimental procedure and a more effective use of a limited data set.

Our own experimental work so far has been primarily directed towards performance improvement and the data base was set up with this in mind. Although we have attempted some rigorous performance testing the unsuitable data base leads to results which probably understate the capabilities of the present system.

Visual inspection of scatter plots of values of pairs of the parameters we have tried has revealed no obvious non-linear correlations or bad values and our simple tests have not indicated peculiar distributions for any of our parameters. Consequently, we have so far confined ourselves to linear transformations of the input parameters.

5.2 Performance Improvement

The procedure adopted in the performance improvement work was to use the whole of the data set to determine the linear transformation to be applied. Each sample is then taken in turn and its effect removed from its own centroid before recognition of that sample is attempted.

The data set containing only fundamental frequency information was used exclusively for performance improvement. The first point we investigated was the form in which the fundamental frequency, or 'F_0', information should be used, ie whether one should take statistics on the distribution of linear frequency, linear period or log frequency. To do this we chose some simple input parameters and compared for the three cases the recognition performance and the values of divergence produced by the LDA. Log frequency turned out to be better than linear period, and, rather to our surprise, linear frequency was best of all. In seeking to explain this result it may be relevant that the mel scale of subjective frequency approximates to linear objective frequency in the range of human F_0.

We next investigated the most suitable F_0 input parameters to use. Statistics were taken of the individual values of F_0 and of first and second-order differences of adjacent values of F_0. First we tried some simple statistics: means, mean deviations and some low-order moments about the mean. At a slightly more complex level we investigated parameters based on possible correlations between F_0 and its rate of change, and parameters indicating the proportions of time that F_0 is rising or falling. Finally, we tried taking simple statistics of F_0 with F_0 first partitioned into periods when it was rising or falling or when its curvature was in a particular range.

In general, the simplest statistics were found to be the most useful: statistics of F_0 were more useful than those involving differences in F_0, means were more useful than measures of variation about the mean, and partitioning the F_0 track before taking statistics of it did not prove worthwhile.

The group of six input parameters which appeared to give the best recognition performance was determined. This consisted of the mean, the mean deviation, the second and third moments about the mean, the proportion of time F_0 was falling, and the second moment about the mean with each value multiplied by the sign of the deviation from the mean. These parameters were then used in the validation experiments with the data set

which included cepstrum coefficients. Obviously, the long-term mean could only be used for F_0 and not for the cepstrum data, but the other five parameters were computed for each of the eight cepstrum coefficients. There were therefore 46 input parameters in all.

Before carrying out the performance-testing experiments we undertook some preliminary investigations using the 'performance improvement' procedure on the new data set. To retain the validity of the performance tests, nothing discovered here was used to modify those tests.

The six F_0 parameters taken as a group were found to be much more effective for recognition than the five parameters of any one cepstrum coefficient. This result can be attributed to the fact that for both F_0 and cepstrum parameters the overall mean is the most useful single parameter, but in the cepstrum case it cannot be used if resistance to spectrum shaping effects is required. When, however, twenty or more of the better cepstrum parameters are taken together they are much more effective than the six F_0 parameters, and experience with the other data set showed that recognition based on F_0 statistics would not be much improved by adding more parameters. The cepstrum parameters appeared to be quite independent of the F_0 parameters in that the error rates for the two groups taken separately could be multiplied together to predict the error rate for the combined set.

By using the five parameters for each cepstrum coefficient in eight separate recognition experiments we compared the usefulness of each of the eight coefficients. While the usefulness did vary between coefficients there was certainly no tendency for the higher quefrency coefficients to be less useful. It appears, therefore, that performance might be substantially improved by including more coefficients.

In order to investigate correlations between the instantaneous values of the different cepstrum coefficients we computed covariance matrices for each speaker. Significant correlations were found, and the correlations were similar for all speakers. The correlations were therefore removed by carrying out a principal components analysis of the covariance matrix averaged over all speakers. The parameters of the principal components corresponding to low variability were neither better nor worse than those corresponding to high variability. Moreover, a subset consisting of the best parameters of the principal components gave no better recognition performance than a subset made up of the best parameters of the untransformed cepstrum coefficients.

5.3 Performance Testing

Using techniques developed on the F_0 - only set, peformance testing experiments were carried out on the data set containing cepstrum information. Since we had speech data from at most two weather forecasts per speaker the training set could contain information from only one forecast for the speaker being tested. The input parameters to be used were chosen by taking the 31 out of 46 whose F-ratios in the training set were highest. We tried two ways of organising the performance testing experiments.

The first way was to split the data into two sets each containing data from one weather forecast for each speaker, the thirteenth speaker with only one forecast being omitted. The two sets were then alternately used as test and training set. 133 samples out of 149 (89%) were correctly identified. If the system had produced a list of two possible names for the unknown speaker, it would have included the correct name 147 times out of 149.

The second way of organising the experiment was to make the test set consist of just one forecast with the rest of the data forming the training set. This has the advantage that an unknown sample can be identified as one of 13 speakers rather than one of 12, and it provides more information for the LDA, though in an unbalanced way because the speaker to be recognised contributes roughly half as much information to the LDA as the other speakers do. In fact, the results were very similar, with 132 (89%) correct identifications and 143 correct identifications when a list of two names was allowed.

There is good evidence that extracts taken from the same forecast are much more similar than extracts taken from different forecasts: if extracts from both forecasts were included in the training set (ie performance improvement mode) performance rose from 89% to 100%; and if the system was asked to identify the particular forecast that an extract was taken from, recognition performance was 75%. It seems likely, then, that the testing experiments would have given better and more realistic recognition scores if it had been possible to include data from several forecasts in the training set.

6. CONCLUSIONS

The set of statistical features used here gives a useful level of speaker recognition performance with some resistance to transmission distortions, and this performance can probably be improved by including more cepstrum coefficients. It remains to be seen whether the performance can be maintained when the system is tested on telephone quality conversational material.

ACKNOWLEDGEMENTS

We wish to thank the British Meteorological Office for letting us use the weather forecast material, and Miss P. S. Shillabeer for helping with the experiments.

REFERENCES

1. B. S. Atal, "Automatic Recognition of Speakers from their Voices", IEEE Proceedings, Vol 64, pp 460-475, April 1976.
2. K. O. Mead, "Identification of Speakers from Fundamental Frequency Contours in Conversational Speech", JSRU Research Report No 1002, March 1974.
3. A. M. Noll, "Cepstrum Pitch Determination", J. Acoust. Soc. Amer., Vol 41, pp 293-309, February 1967.