SPEAKER CHANGE DETECTION

A. Satriawan and J.B. Millar

Computer Sciences Laboratory
Research School of Information Sciences and Engineering
The Australian National University

ABSTRACT - An approach to speaker change detection based on speaker discontinuity models is described. It builds on the limited evidence of speaker characteristics in individual phone segments to enable a rapid response when a speaker change is detectable. A baseline system based on a speaker identification task is built and tested on the TIMIT corpus.

INTRODUCTION

Automatic speaker monitoring (ASM), defined as automatic techniques to cope with potentially multi-speaker environments, encompasses a number of related spoken language processing techniques. ASM is concerned with such issues as modelling the number of speakers co-talking, the existence of a dominant speaker, the continuity or discontinuity of speaker identity, and speaker identity itself. There are several important practical applications of ASM : speaker assignment for conversational speech in an interview or telephone conversation, quick identification of speakers in senatorial debate in parliament, assignment of speaker in pilot/co-pilot conversation with tower control (Gish, Siu, and Rohlicek, 1991), detection of an intruding imposter (Schalkwyk *et.al*, 1994), speaker recognition in noisy environment, and better segregation of target speech from background speech.

This paper will present our initial work in the study of ASM. It concentrates on the speaker change detection (SCD) problem which is based on models of speaker discontinuity. The next section introduces a framework for ASM based on a speaker recognition paradigm. The following sections describe our SCD approach and the result of our preliminary experiments. The last section presents our conclusions.

A TAXONOMY OF AUTOMATIC SPEAKER MONITORING

The ASM problem can be regarded as a standard pattern recognition problem. Its solution in this paradigm involves two main phases : the training phase and the testing phase. In the training phase, the characteristics of a speaker are extracted from training utterances and stored as templates for reference purposes. In the testing phase, an input signal which consists of mixture of several speech utterances from a multi-speaker environment will be recognised in terms of its constituent speakers, the number of speakers, and the speaker at certain time. An input signal of this kind will be called a multi-speaker utterance (MSU).

The nature of the speaker population defines four possible task classes of varying complexity : The *speaker population* participating in the training and testing phases may be a *closed* or an *open set* and the *number of speakers* participating in the MSU may be *known* or *unknown.*

In a closed-set task, all possible speakers in the MSU are in the training database. A closed-set task can then be subdivided based on the knowledge of the number of speakers in the MSU. The number of speakers in the utterance may or may not be known beforehand. In the first case, the speakers may be known, and the task is only to detect which speakers are speaking at a certain time. The second case is more difficult since beside the need to recognise the speakers, the number of speakers in the MSU also needs to be detected.

In an open-set task, whether all speakers in the MSU are in the training database is unknown. Some speakers may be in the database but some may not be. The problem becomes more complicated if the number of speakers is unknown. However, even in the case of known number of speakers the problem is still difficult. The open-set task with unknown number of speakers is the most difficult one.

Within the above structures, the problem can also be divided into two cases based on the *degree of speech co-occurrence* within the MSU : *sequential speakers*, when there is no co-occurrence of speech from different speakers, and *concurrent speakers*, when more than one speaker is speaking at the same time.

The SCD problem is an ASM problem with sequential speakers. In this paper however, we adopt a more restricted definition of SCD. We deal only with a closed-set task with the possibility of a known or unknown number of speakers. We also limit the MSU to comprise only speech utterances, that is we do not allow non-speech intrusions. In other words, the speakers in MSU speak one at a time and the shortest possible intrusion is a speech syllable.

One of the main challenges in speaker change detection is to recognise that the speaker has changed as soon as possible. To cope with this, speaker models based on short utterances such as individual phones or phones from broad phonetic classes are needed. This need is also shared with studies aimed at rapid speaker verification. Our approach to SCD that will be described in the following section is to utilise these phonetic event based models to detect the speakers.

SPEAKER CHANGE DETECTION

Our SCD approach works as follows. First, a speaker-independent speech recognition method is used to label the speech utterance into phonetic events. We use broad phonetic class (BPC) categories to label our phonetic events. Second, using these labels the speakers of these phonetic events are recognised, either by identification or by verification, based on the ability of those phonetic events to discriminate the speakers in the speaker space. Third, consecutive phonetic events which are recognised as from the same speaker are combined into a single speech segment. This segment is then assumed to be from that speaker. The speakers have changed if two segments are recognised as from two different speakers. Depending on the task which is performed in recognising a speaker in the second step above, we can define the strategy for SCD as speaker identification based or speaker verification based.

The identification method has several advantageous. It does not depend on our knowledge of the speakers in the MSU. It can be applied to the known and unknown number of speaker cases without any major modification. The disadvantage of this method is that speaker identification requires more computation as tests against all speakers are needed.

In the speaker verification based strategy, instead of an identification task, a speaker verification task is performed. In this task a speaker of a phonetic event is verified based on a claim made about its speaker. Therefore, in the second step above, a claim about the speaker of the current phonetic event is needed and the verification task is to verify that claim. A simple approach is to claim that the speaker of the current phonetic event is that of the previous phonetic event (i.e. the speaker of the phonetic event prior to the current phonetic event). If the claim results in a positive answer the current phonetic event has the same speaker as that of the previous event. Otherwise, if the current phonetic event fails to be verified as from the same speaker as that of the previous event, that means the current speaker is not the same as the previous speaker and therefore the speaker changed prior to this current phonetic event.

If the speaker is judged to be different from the claimed speaker, the next problem is to decide the identity of the speaker of the current phonetic event. At this point an identification must be performed.

EXPERIMENTAL DESIGN

The experiments described in the next section are based on the speaker identification method. There are several design issues we face in using this method. The first is how to represent a BPC segment. One approach is to use a single feature vector. For example Fakotakis, Tsopanoglou, and Kokkinakis (1993) used cepstral coefficients extracted in the middle of vowels (correspond to the highest energy content) as the elements of the feature vector. Our approach is to represent a BPC segment by a set of feature vectors. Our feature vectors correspond to cepstral or delta-cepstral analysis on the frames of the BPC segment. A problem of this method is that if the identification is performed based on a single frame the frames of a BPC segment could be identified as from several different speakers. These feature vectors, however, should be used optimally and be recognised only as from a single speaker. There are several ways to do this which depend on the classification method used. In this work we used Hidden Markov Model (HMM) based method to perform this frame integration.

Another issue is how to model a speaker in a BPC. The simplest one is to use a Normal distribution for each speaker. Another one is to use a mixture of Normal distributions to cater for the possibility of clusters caused by the individual phones within the BPC. Both of these methods operate on a frame by frame basis. A more complex one is to use an HMM for each speaker. This allows us to model the temporal variations of the speech signal through the transition probabilities. Through these transition probabilities also, the frame integration is performed in this model. We investigated these three possibilities in our experiments.

The next issue is the BPC segment integration described in the third step of this approach. Our experiments are still based on the individual BPC segment recognition. No segment integration is performed in this initial work.

The last design issue concerns the speaker-independent speech recognition method used in the first step of this approach. All of our experiments were done without using an automatic speech recogniser except for the HMM based modelling where automatic recognition was also performed independently. We use the BPC labels of the manually labelled test utterances (provided in the TIMIT corpus) to choose the phonetic event based models for our experiments instead of from automatically labelled utterances.

EXPERIMENTAL RESULTS

A baseline system using the speaker identification task was constructed to test the strategies outlined above. The identification was done framewise and no frame integration was performed. Therefore, the results described below would be the worst case. The baseline was tested on 26 male speakers of almost the same ages from dialect region 1 and 5 of the TIMIT database. Five broad phonetic classes were used and extracted from the manually labelled data. They are vowel, nasal, fricative, plosive, and lateral. The training set consists of the SA2 sentence and six other sentences. The testing set is the SA1 sentence and two other sentences not chosen in the training set.

Acoustic features used in this experiment are LPC-based 12-th order cepstral and delta-cepstral coefficients and normalised energy. They are extracted from Hamming windowed frames of 32 ms, each separated by 8 ms and pre-emphasised. The LPC analysis was performed using the correlation method. The delta-cepstral is calculated from two frames prior to current frame to two frames after the current frame. The normalised energy is always included in a feature vector.

Discriminant Analysis Method

In the baseline system the identification task is performed using the discriminant analysis method. Two experiments were performed which correspond to whether a parametric or non-parametric discriminant analysis method was used. In the parametric case, feature vectors of a speaker in a BPC are assumed to follow a normal distribution. Therefore the problem is to estimate the parameters of the distribution. When the covariance matrices of all these distributions are assumed to be the same (covariances are pooled across speakers within a BPC), the problem reduces to a linear discriminant method (LDA). When different covariance matrices are used for different speakers, the problem reduces to a quadratic discriminant method (QDA).

Using the linear discriminant method on normalised energy and cepstral coefficients alone the error rates based on the broad phonetic classes were very high. The two lowest posterior error estimates (Fukunaga and Kessel, 1973) were 45.64% and 58.96% for nasals and vowels respectively. Using QDA, the error rates were reduced. Posterior error estimate is used here since our test set is very small due to the size limitation of TIMIT corpus for speaker recognition. Error count estimate can have a large variance in a small test set. The posterior error estimates (PEE) obtained using LDA and QDA methods with cepstral coefficients are shown in Table 1. These two results are inline with previous results of Cohen and Froind (1989) which show that the use of speaker dependent covariance will improve recognition rate.

BPC	LDA	QDA
Nasal	45.64	25.30
Vowel	58.96	42.09
Lateral	61.16	37.97
Fricative	78.09	55.77
Plosive	84.42	65.80

Table 1. Posterior error percentages using parametric method with cepstral coefficients.

In the parametric method a normal distribution is used to model feature vectors of a speaker within a BPC. This assumption may not be good enough since the feature vectors of the same phones within a BPC may form clusters which can cause several bumps in the density. In this case the normality assumption will be violated. A non-parametric method can be used to alleviate this problem by estimating the form of the speaker probability distribution from the data (Hand, 1982). In this case the parameters of the distribution may be unknown.

There exist several non-parametric methods (Hand, 1982). In this experiment we used a kernel based method with a normal kernel for our non-parametric method. With this kernel, the distribution of a speaker is modelled by a mixture of normal distributions. The covariance matrix used for the normal kernel can be the same for all speakers (pooled covariance) or can be different for different speakers. The results of using non-parametric method with cepstral coefficients are given in Table 2.

Comparing the results of the parametric and non-parametric methods, we can see that the posterior error can be reduced by more than 50%. This improvement suggests that our assumption that a speaker can be modelled by a Normal distribution is not good enough. The normality assumption cannot take into consideration the possibility of several bumps caused by the different fine phonetic classes within each BPC. This is also inline with the better results of speaker identification performance using a single state HMM with larger number of mixtures or using vector quantisation with larger codebook size in Matsui and Furui (1992). The different results of using pooled and different covariance matrices also confirm the need to use speaker dependent covariance as already mentioned above.

BPC	Speaker independent covariance matrix	Speaker dependent covariance matrix
Nasal	13.15	8.16
Vowel	22.73	16.24
Lateral	18.87	13.21
Fricative	25.51	19.71
Plosive	35.05	30.16

Table 2. Posterior error percentages using non-parametric method with cepstral coefficients.

Use of Dynamic Information

Cepstral coefficients contain only the instantaneous information of the speech signal. Transitional information of these cepstral coefficients has been found to be important for speaker recognition especially the speed of these variations (Soong and Rosenberg, 1988). Therefore, before improving our baseline system, we perform another experiment with transitional features but still using the simple model above (parametric and non-parametric). We add delta-cepstral coefficients in our feature vectors. This additional feature represents the dynamic information within the BPC. This results in better performances with about 20 - 50% decrease in posterior error estimate in the parametric method and more than 50% in the non-parametric method.

Based on these results, our HMM-based model, described in the next subsection, will utilise both the cepstral and delta-cepstral features.

HMM Based Method

The better results of models based on speaker dependent covariances compared to those of speaker independent suggest that models which can account for speaker dependent variations are preferable. The improvement obtained in using delta-cepstral also suggest that models which can cater for time variations are also needed. Therefore in our attempt to further reduce the error of the baseline system these kinds of models will be used. Specifically, HMM will be used to model each speaker within each BPC.

In this work the HMM can be seen as a way of integrating the frames of a BPC and a way of modeling a BPC of a speaker as a mixture of Gaussian distributions. The former is obtained through the use of the state transition probabilities of the HMM. The latter is by the use of states and of mixtures of Normal distributions to model the output probability of states. The states and transition probabilities between states allow us to account for the time variability of the speech signal.

Three state left-to-right with no skip HMMs were used to model the BPCs of a speaker. A mixture of three Gaussian distributions with diagonal covariance matrices was used for modelling the output probability of the states. Each HMM which models a BPC of a speaker was trained on the corresponding BPC segments extracted from the training sentences of that speaker using Baum-Welch re-estimation.

Two experiments were performed for testing this HMM method. The first experiment is similar to the previous experiment where the labels are assumed to be known. In this case, the known BPC labels are used to form a network of HMMs with zero inter-model transition probabilities. This network is then used to perform a Viterbi search on the input utterances. The size of this network, therefore, depends on the number of BPC segments in an MSU. The performance of this experiment is given in terms of error count estimates (ECE) which are obtained by counting the number of frames of BPC segments which are falsely recognised. This ECE has two drawbacks. The first one, as mentioned previously, it can have a large variance. The other one is that the ECE of a BPC may be affected by other BPCs in its left or right contexts due to segment boundary errors. This results in a biased estimate of the error which can only be avoided if there is no boundary error. Manual examination of the recognition results shows, however, that this boundary error effect is minimal. The posterior error estimate of this experiment is not easily be obtained.

BPC	QDA	HMM
Nasal	58.16	36.47
Vowel	52.94	37.63
Lateral	68.46	54.94
Fricative	74.19	59.53
Plosive	80.53	73.03

Table 3. Error count percentages using QDA and HMM with known label.

The results based on cepstral and delta cepstral coefficients are shown in Table 3. For comparison with the previous results, we also include the ECE obtained using QDA method with cepstral and delta cepstral coefficients.

Comparing the results of Table 3, we observe that the integration of frames and refinement of speaker model within a BPC reduce the error rates. However, it is not clear from this experiment which causes the improvement. It can be seen also that speakers are still difficult to recognise based on plosives.

The second experiment is different from the first one in that we do not assume the knowledge of the BPC labels of the MSU. The broad phonetic event and its speaker are obtained by performing a Viterbi search using all possible HMM models to find the optimum average log probability per frame. The ECE of this experiment based on cepstral and delta-cepstral are shown in Table 4.

Comparing the results in Table 4 and the second column of Table 3 shows that knowledge of the labels reduces the error rates except in the lateral class. These results, however, have larger bias compared to the known label case. This suggests that a good BPC recognition strategy is needed in order that this method can be used reliably. How good the BPC recognition strategy is, in terms of the segment boundary errors, is still unclear in this experiment. Correct recognition of segment boundary is also important for quick detection of speakers.

BPC	HMM unknown label
Nasal	38.95
Vowel	40.05
Lateral	53.27
Fricative	61.08
Plosive	76.34

Table 4. Error count percentages using HMM with unknown label.

DISCUSSION

The improvement of the results of the baseline system with the introduction of better speaker model and frame integration is encouraging. However, comparing the results obtain with the current state of speaker recognition, these results still need improvement.

There are several ways to do this, some of them are described below. The first one is to use a better database which is specifically design for speaker recognition. The TIMIT corpus has several limitations in the richness of within speaker variations in a BPC which shows up in this experiment. There are several phones which are not well represented in the training or testing sets of a speaker. A good database to test this approach is the one which contains a balance of phone segments in the training and testing sets with rich speaker variations.

The second one is to use a better HMM topology. The results reported here are not based on an optimum choice of this topology. A better approach may be to use different topologies for different BPCs and optimised over all speakers.

The third is to implement the segment integration and find a better frame integration method. The use of verification based method should also be investigated.

CONCLUSIONS

A speaker change detection approach has been presented along with the results of experiments based on the approach. It has been shown that a change of speaker can be detected more reliably in the nasal and vowel parts of the speech especially when speaker dependent covariance matrices are used. The integration of frames of a BPC also improves the recognition results as well as the reliable recognition of the BPC.

REFERENCES

Cohen, A. and Froind, I. (1989) *On text independent speaker identification using a quadratic classifier with optimal features*, Speech Communication 8, 35--44.

Fakotakis, N., Tsopanoglou, A., and Kokkinakis, G. (1993) *A text-independent speaker recognition system based on vowel spotting*, Speech Communication 12, 57--68.

Fukunaga, K. and Kessel, D.L. (1973) *Nonparametric Bayes error estimation using unclassified samples*, IEEE Trans. on Information Theory, 19, 434--440.

Gish, H., Siu, M.H, and Rohlicek, R. (1991) *Segregation of speakers for speech recognition and speaker identification*, Proc. ICASSP 1991, 873--876.

Hand, D.J. (1982) *Kernel discriminant analysis*, (Research Studies Press: New York).

Matsui, T. and Furui, S. (1992) *Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs*, Proc. ICASSP 1992, 157--160.

Schalkwyk, J., Barnard, E., Cole, R.A., and Sachs, J.R. (1994) *Detecting an imposter in telephone speech*, ESCA workshop on speaker identification, verification, and recognition, (Martigny, Switzerland), 119--122.

Soong, F.K. and Rosenberg, A.E. (1988) *On the use of instantaneous and transitional spectral information in speaker recognition*, IEEE Trans. on Acoustics, Speech, and Signal Processing, 36, 871--879.