Speaker Search and Indexing for Multimedia Databases

T. Silva, D. D. Karunaratna, G. N. Wikramanayake
K. P. Hewagamage and G. K. A. Dias.
University of Colombo School of Computing,
35, Reid Avenue, Colombo 7, Sri Lanka.

E-mail: tyro@sltnet.lk, {ddk, gnw, kph, gkd }@ucsc.cmb.ac.lk

Abstract

This paper proposes an approach for indexing a collection of multimedia clips by a speaker in an audio track. A Bayesian Information Criterion (BIC) procedure is used for segmentation and Mel-Frequency Cepstral Coefficients (MFCC) are extracted and sampled as metadata for each segment. Silence detection is also carried out during segmentation. Gaussian Mixture Models (GMM) are trained for each speaker, and an ensemble technique is proposed to reduce errors caused by the probabilistic nature of GMM training. The indexing system utilizes sampled MFCC features as segment metadata and maintains the metadata of the speakers separately, allowing modification or additions to be done independently. The system achieves a True Miss Rate (TMR) of around 20% and a False Alarm Rate (FAR) of around 10% for segments between 15 and 25 seconds in length with performance decreasing with reduction in segment size.

1.0 Introduction

The use of multimedia has become more widespread due to advances in technology and the growing popularity of “rich’ presentation mediums such as the Internet. As a result of this growth we are now faced with the problem of managing ever expanding collections of audio and video content. This problem is particularly acute in the area of e-learning where it is not uncommon to find thousands of archived video and voice clips. Users accessing such multimedia archives need to be able to sift through this wealth of visual and aural data, both spatially and temporally, in order to find the material they really need [19].

The management of such multimedia content has not traditionally been a strong point in the field of databases where numerical and textual data has been the focus for many years. For this purpose such systems utilize descriptive structured metadata, which may not always be available for the multimedia [18].

What is required in this context is a ‘true’ multimedia database with the ability to query the actual content for the purpose of locating a given person, object or topic area.

Currently there are no commercial database products supporting content based querying, although there are several ongoing research projects such as the Combined Image and Word Spotting (CIMWOS) programme focused on this area. Many avenues for research exist such as face recognition, object recognition and keyword spotting. This paper describes the work undertaken in one such avenue, that of speaker recognition, as part of a research programme at the University of Colombo School of Computing, Sri Lanka.

2.0 Background and Architecture

In order to provide the capability to carry out a speaker based search of multimedia clips three main tasks have to be achieved. These tasks are respectively Segmentation, Recognition and Indexing. The architecture of the system incorporating these three tasks is shown in figure 1.

Segmentation is carried out after initial feature extraction on the new multimedia clip. A labeling procedure is used after segmentation to classify segments as voice or silence and the latter are marked and excluded from the later steps involved in querying.

Training speakers within this system is done in a supervised manner by specifying audio files containing speech purely from the given speaker. The mean and variance of the log-likelihood for the specified test data is obtained in order to establish a decision criterion for the later search.

Both activities of training and segmentation result in metadata which is stored in a relational database for the purpose of Indexing. This metadata is combined at query time to give the result of the search.

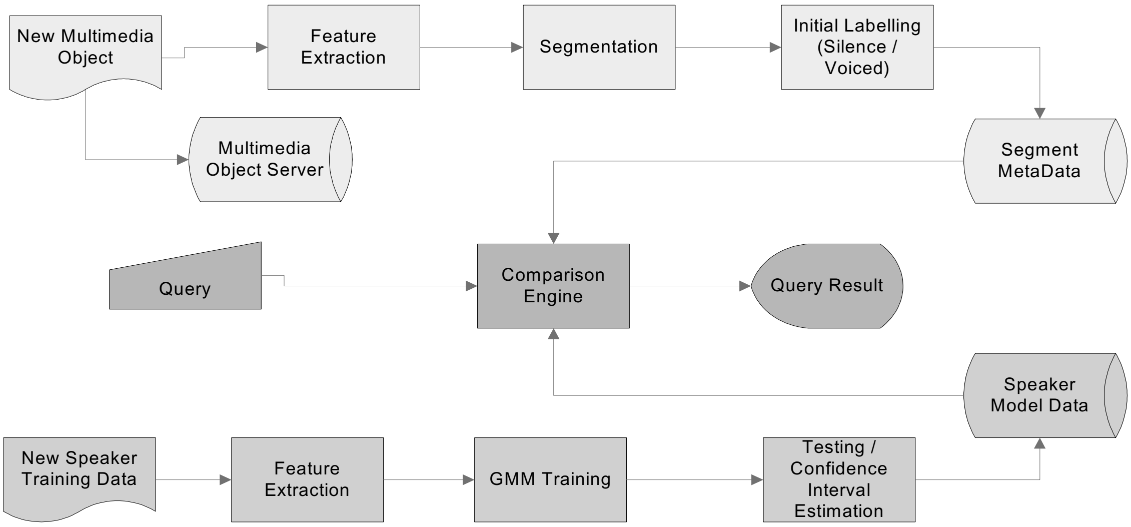

Figure 1 : Architecture of the Speaker Search and Indexing System

The components of this system are described in more detail in the following sections.

2.1 Feature Extraction

The success of all three tasks identified previously is dependent on the extraction of a suitable feature set from the audio. Features such as intensity, pitch, formants, harmonic features and spectral correlation have been used in the past [1],[17] but in recent speech-related research such as [10],[12],[5] the two dominant features used are Linear Predictive Coding Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC).

Both methods utilize the concept of the “cepstrum” based on successive fourier transforms and a log transform applied to the original signal. LPCC uses linear predictive analysis whereas MFCC utilizes a triangular filter bank to derive the actual coefficients. Of these LPCC is the older method and has been cited as being superior for certain types of recognition topologies such as simplified Hidden Markov Models [2]. MFCC features are however more noise resistant and also incorporate the mel-frequency scale, which simulates the non-uniform sensitivity the human ear has towards various sound frequencies [16].

In this system the audio is downsampled to 11025 Hz and cepstral coefficients are drawn from each 128 byte frames. The zeroth (energy) coefficient is dropped from this feature vector. Non-overlapping hamming windows are used and a feature vector size of 15 and 23 is used for recognition and segmentation respectively. At the initial stages of this research project various combinations of MFCC and LPCC coefficients were experimented with and it was discovered empirically that a pure MFCC feature vector gave the best results.

2.2 Segmentation

For the task of segmentation three broad methods have been identified [3]. These are namely Energy Based Segmentation, Model Based Segmentation and Metric Based Segmentation.. The first approach is more traditional and based on identifying changes in amplitude and power which occur at speaker change points. The second method requires prior speaker models to be trained and segmentation carried out as a classification procedure on each section of audio as it is being processed.

The final method, metric based segmentation uses estimated “differences” in adjacent audio frames in order to discover speaker change points. Typical metrics used include the Euclidian, Kullback-Leiblar [4], Mahalanobis and Gish distance measures [5],[6] as well the Bayesian Information Criterion (BIC). The BIC is a form of Minimum Descriptor Length criterion that has been popular in recent times. Various extensions such as automatically adjusting window sizes [7] and multiple passes at decreasing granularity levels [8] have also been proposed.

BIC is employed to test the hypothesis that there is a significant different in speakers in two adjacent sections of audio. The equation used is shown below:

$$BIC = N*\log|\Sigma| - (i*\log|\Sigma_1| + (N-i)*\log|\Sigma_2|) \\ - \tfrac{1}{2}\lambda(d+\tfrac{1}{2}d(d+1))\log N$$

where **N(μ, ∑)** denotes a normal/Gaussian distribution with covariance ∑ and mean μ, **N** is the total number of frames in both segments, **i** is the number of frames in the first segment and **d** is the dimensionality of each sample (feature vector) drawn from the audio. $\boldsymbol{\lambda}$ is a penalty factor which is set to unity in statistical theory.

A result less than zero for BIC indicates that the two segments are similar.

The actual technique used in this system is based on the variable window scheme by Tritschler [8]. A speaker change point is sought within an initial window, and if such a point could not be found the window is allowed to grow in size. When a speaker change point is detected then that location is selected as the starting point for the next window. To avoid the calculations becoming too costly it was decided to introduce an upper limit to the size of the window. Beyond this limit the window starts "sliding" rather than growing.

An intentional bias towards over-segmenting was introduced into the BIC equation to ensure than all possible speaker change points were accounted for. This was done by setting λ to a value less than 1. Since this caused a lot of spurious boundaries to be detected a second pass was carried out in which adjacent segments were compared once again using the BIC and merged if they were found to be similar.

After segments are obtained the Zero Crossing Frequency (ZCF) is used to detect silence segments. Voice segments have a very much higher variance for the ZCF and a simple threshold was used to differentiate them from the segments containing silence.

2.3 Speaker Model Training

Several methods have been utilized in the literature for the purpose of speaker recognition. Text-dependent recognition, such as that required for voice password systems, has been carried out using Hidden Markov Models (HMM) and Dynamic Time Warping (DTW). In the text-independent approach, which is more relevant to the research question addressed in this project, the Vector Quantization method has been used in the past [9], and more recently combined with neural networks in the form of Learning Vector Quantization (LVQ) [10].

However Gaussian Mixture Models (GMM) are at present the most popular method used. This method models the features for each speaker by using multiple weighted Gaussian distributions [11]. The speaker dependent parameters for each Gaussian (mean, variance and the relative weight) can be estimated through the Expectation Maximization (EM) algorithm. Various forms of this model have been tried, including variations in the type of covariance matrix used [12],[13] and a hierarchical form [14].

In this system GMM's with full covariance matrices are used with an initial 20 mixtures per GMM. Low-weight mixtures are dropped from the model after training. Rather than utilizing a single GMM per speaker it was decided to incorporate three such models within an ensemble classifier, as the stochastic nature of the EM algorithm resulted in diverse models at each run. By averaging three models it was hoped that any bias that occurred during training due to the random initialization of the EM algorithm could be reduced.

Following the training phase the model is tested with data for the same speaker. The mean and standard deviation of the log-likelihoods obtained for the test data is calculated and used for the recognition criteria as described in the next section.

2.4 Indexing

One method used for the Indexing task involves the use of pre-trained anchor or reference models. Distance or Characterization vectors are calculated for each utterance based on its location relative to the anchor models. Similar vectors are used to specify the relative location for speakers within the feature space. Searching for a specific speaker is simply a matter of finding the utterances that have a characterization vector close (in terms of Euclidean or any other distance measure) to that of the target speaker [15]. While anchor models are efficient in terms of search time, they have a much higher reported error rate than direct utilization of GMM recognition.

In the design of the indexing scheme for this system, the following assumptions and decisions were made

1. A query was assumed to always be for a known speaker. Hence a trained GMM classifier would have to exist beforehand for any speaker for whom a query was formulated. Queries using example audio tracks could also be incorporated into this system, as it would simply require that a speaker model be trained using the given sample audio track before the query process was carried out.

2. No prior labeling on the basis of speaker would be carried out on the audio. The reason for this was that such a labeling would be inflexible as it would not allow new speakers to be added to the system without extensive recalculation and annotation of the existing audio. Therefore it was decided to defer the work involved in labeling to query time itself.

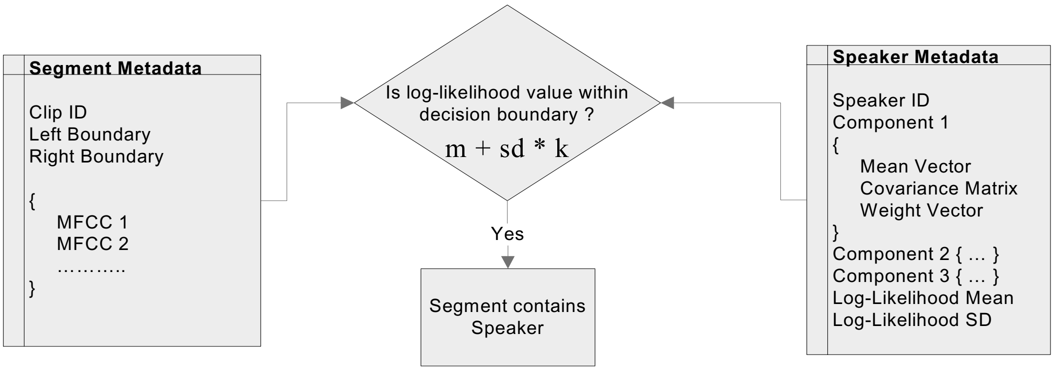

Figure 2 : The Metadata used for Indexing

Following the segmentation phase, information on segment boundaries discovered through the BIC procedure is held as metadata, along with a label specifying if the segment is silence. The MFCC features of that segment are also extracted and saved. This constitutes the metadata for the audio clips.

In order to avoid the excessive storage requirements inherent in storing all the MFCC features of a segment, a sampling procedure is carried out for segments larger than a certain size. Feature vectors are obtained from the segments at linear intervals. Experiments show that the effect on log-likelihood caused by using the sampled feature vectors is relatively low.

Metadata for each speaker is derived from the speaker training procedure explained in section 2.3 , This consists of the parameters derived for the GMM's as well as the mean and the standard deviation of the log-likelihood test data for that speaker

Carrying out a search is simply a case of retrieving the GMM data for the queried speaker and subsequently calculating the probability of the MFCC vectors for each segment against this GMM. The overall structure of the indexing scheme is shown in figure 2.

The constant K shown in the decision criteria in figure 2 is actually a user specified value which represents how much "relaxation" the system is allowed when evaluating the query. Higher values for K result in a higher probability that existing segments for that speaker are found, along with an increased probability that false speakers are also discovered.

One feature of this system is that metadata for speakers and audio are maintained independently of each other. This allows either one to be changed (i.e. when adding or removing audio clips and speakers) without requiring changes to be done to the other.

2.5 Architecture

While recent research into speaker recognition has often utilized MFCC feature vectors and the BIC metric, in the majority of these cases these systems depend on having pre-trained models of all possible speakers in the domain. Hence the segmentation phase is followed by an annotation of the segments with the id of known speakers, and then carrying out text based searches of this field.

In this respect our system provides the same capability as the anchor modeling done by Sturim [15], although there is no indication whether his system utilized automated segmentation. In addition, the use of direct evaluation of speaker GMM against sampled MFCC features of the segments has not been encountered as an indexing technique in the literature.

Similarly no studies have reported utilizing an ensemble GMM classifier to reduce the possible bias introduced during the stochastic training process.

3.0 Evaluation

The algorithms for Speaker Recognition and Segmentation were implemented in Matlab, translated into C code and were compiled as COM objects. The indexing and the user interface were created using C#.NET and integrated with the COM objects.

Experiments were performed using a small sized speaker database consisting of 60 second audio clips and 9 speakers. Each speaker had 3 clips of lone speech. In addition 8 clips of mixed speech were obtained, ensuring that each speaker was represented in at least 3 mixed clips. The recording environment contained background conversations and noise due to activities of students in the laboratory and air conditioning hum. Both segmentation and recognition were tested on this data.

The criteria used for evaluation were the False Alarm Rates (FAR) and True Miss Rates (TMR). They are defined as shown below:

$$\text{FAR\%} = \frac{\text{Erroneously Detected Segments} \times 100}{\text{Total Detected Segments}}$$

$$\text{TMR\%} = \frac{\text{Missed True Segments} \times 100}{\text{True Segments}}$$

The results obtained for segmentation were quite good with an FAR of 3.2% and a TMR of 6.7%.

Recognition tests were carried out on these same segments and the results are shown in figures 3 and 4. The audio segments were divided into three classes, namely large (60 seconds), medium (15-25 seconds) and small (under 15 seconds). These segments were those obtained as the result of the previous segmentation algorithm applied to the speech data. Each segment therefore corresponded to the continuous speech of one speaker. The results for large and medium segments are promising with the former showing error rates close to 0% and the latter having a TMR of around 20-30% along with an FAR of under 10% for values of the relaxation constant around 1.5. Small segments however showed a much higher error rate with the TMR being as high as 50% and FAR between 20-30%.

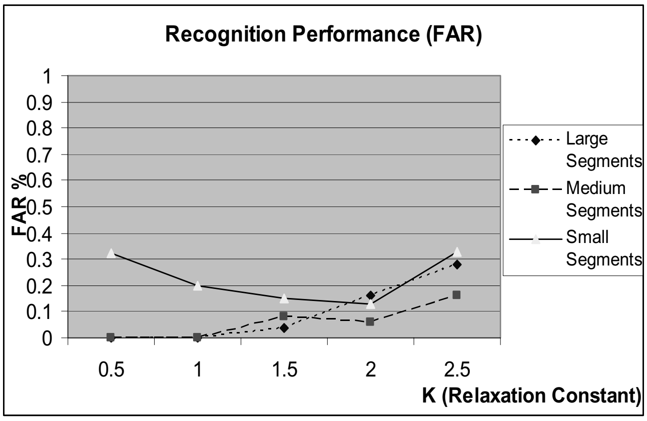

Figure 3 : The FAR achieved for recognition according to segment size and the value of the relaxation constant

4.0 Conclusions

This paper has addressed the problem of speaker recognition and indexing using a segmentation procedure based on BIC, a GMM recognition system and an indexing scheme utilizing metadata based on sampled MFCC features. The techniques outlined here have given good segmentation results and a satisfactory recognition rate when used with medium to large sized segments recorded under the same conditions.

Due to the independence of the segmentation and speaker training components the actual work of speaker recognition may be delayed till a query is carried out. This means that knowledge of speakers is not required when a new audio clip is segmented. We only require that a speaker model exist at the time of the actual search. This creates the possibility of extending the system so that a search can be carried out on the criteria that the speaker in the returned segment matches that of an “example” audio clip provided by the user at search time.

However some further evaluation has shown that the performance of this system gets worse when applied to segments extracted from audio recorded under different conditions. This shows that this method is sensitive to channel and noise effects in the audio. There remains much scope to explore various forms of filtering as well as channel normalization in order to improve the recognition rates of this system.

In addition the indexing scheme outlined is not as fast as the use of anchor models. A method by which the speed advantage of anchor models and the accuracy of GMM based indexing could be to utilize the anchor model based indexing scheme as the first stage of the search procedure in order to cut down on the number of segments considered.

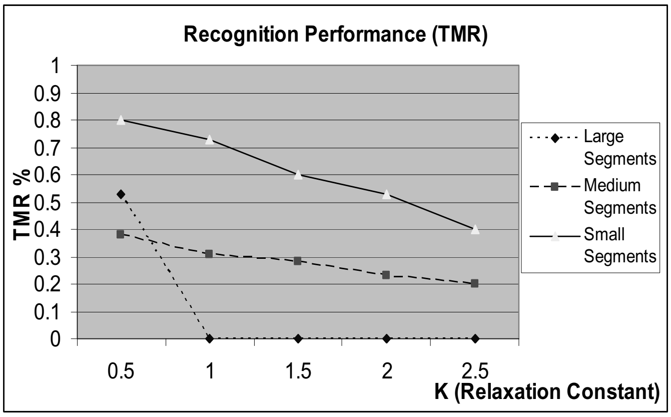

Figure 4 : The TMR achieved for recognition according to segment size and the value of the relaxation constant

References

[1] Pool, J. (2002) "Investigation on the Impact of High Frequency Transmitted Speech on Speaker Recognition", Msc. Thesis, University of Stellenbosch

[2] Bimbot, F., Hutter, H. & Jaboulet, C. (1998) "An Overview of the Cave Project Research Activities in Speaker Verification", *Speech Communication*, Vol 31, Number 2-3, pp. 155-180

[3] Kemp, T., Schmidt, M., Westphal, M. & Waibel, A, (2000), "Strategies for Automatic Segmentation of Audio Data", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP*, Volume 3, pp. 1423-1426

[4] Lu, L. & Zhang, H. (2002), "Real-Time Unsupervised Speaker Change Detection", *Proceedings of the 16th International Conference on Pattern Recognition (ICPR)*, Vol. II, pp. 358-361

[5] Pietquin, O, Couvreur, L & Covreur, P (2002), "Applied Clustering for Automatic Speaker-Based Segmentation of Audio Material", *Belgian Journal of Operations Research, Statistics and Computer Science (JORBEL) - Special Issue on OR and Statistics in the Universities of Mons*, vol. 41, no. 1-2, pp. 69-81

[6] Siegler, M.A., Jain, U., Raj, B. & Stern, R.M. (1997), "Automatic Segmenting, Classification and Clustering of Broadcast News Audio", *Proceedings of Darpa Speech Recognition Workshop*, February 1997, Chantilly, Virginia

[7] Tritschler, A. & Gopinath, R (1999), "Improved Speaker Segmentation and Segment Clustering Using the Bayesian Information Criterion", *Proceedings of Eurospeech 1999*, September 1999, Budapest

[8] Delacourt, P, Kryze, D & Wellekens, C.J (1999), "Speaker Based Segmentation for Audio Data Indexing", ESCA W*orkshop on Accessing Information in Audio Data*, 1999, Cambridge, UK

[9] Orman, O.D. (1996), "Frequency Analysis of Speaker Identification Performance", MSc Thesis, Bogazici University

[10] Cordella, L.P., Foggia, P., Sanson, C. & Vento, M. (2003) "A Real-Time Text-Independent Speaker Identification System", *Proceedings of the 12th International Conference on Image Analysis and Processing*, IEEE Computer Society Press, Mantova, pp. 632-637

[11] Reynolds, D.A. (1995), "Robust Text-Independent Speaker Identification Using Gaussian Mixture Models", *IEEE Transactions on Speech & Audio Processing*, Volume 3, pp.72-83

[12] Sarma, S.V. & Sridevi V. (1997), "A Segment Based Speaker Verification System Using SUMMIT", *Proceedings of Eurospeech 1997*, September 1997, Rhodes, Greece

[13] Bonastre, J., Delacourt, P., Fredouille, C., Merlin, T. & Wellekens, C. (2000), "A Speaker Tracking System Based on Speaker Turn Detection for NIST Evaluation", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, June 2000, Istanbul, Turkey, pp. 1177-1180

[14] Liu, M., Chang, E. & Dai, B. (2002), "Hierarchical Gaussian Mixture Model for Speaker Verification", *Microsoft Research Asia publication*

[15] Sturim, D.E., Reynolds, D.A., Singer, E. & Campbell, J.P. (2001), "Speaker Indexing in Large Audio Databases Using Anchor Models", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 429-432

[16] Skowronsk, M.D. & Harris, J.G. (2002), "Human Factor Cepstral Coefficients", *Proceedings of the First Pan American/Iberian Meeting on Acoustics*

[17] Matsui, T. & Furai, S. (1990) "Text Independent Speaker Recognition Using Vocal Tract and Pitch Information", *Proceedings of the International Conference on Spoken Language Processing*, Vol. 4, pp. 137-140

[18] Wen, J., Li, Q., Ma, W. & Zhang, H. (2002) "A Multi-Paradigm Querying Approach for a Generic Multimedia Database System", *SIGMOD Record*, Vol. 32, No. 1

[19] Klas, W. & Aberer, K. (1995), "Multimedia Applications and Their Implications on Database Architectures",