SPEECH, SILENCE, MUSIC AND NOISE CLASSIFICATION OF TV BROADCAST MATERIAL

A. Samouelian, J. Robert-Ribes† and M. Plumpe≠*
* Department of Electrical, Computer and Telecommunications Engineering,
University of Wollongong, Wollongong, NSW 2522, Australia.
† Digital Media Information Systems, CSIRO Mathematical and Information Sciences
North Ryde NSW 1670, Australia
≠ Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA

ABSTRACT

Speech processing can be of great help for indexing and archiving TV broadcast material. Broadcasting station standards will be soon digital. There will be a huge increase in the use of speech processing techniques for maintaining the archives as well as accessing them.

This paper starts with a review of several techniques used for classification of speech, music and noise. Generally, approaches that use Neural Networks (NN) or Hidden Markov Modelling (HMM) do not allow to "look inside" the network or models to determine which aspect of the sounds are similar to each other. This makes it difficult for the researcher to determine the features of the audio that are important and which ones can be ignored [1]. Furthermore, for archiving TV broadcast material, the segment time accuracy does not need to be as precise as when labelling speech corpora to be used for speech recognition research. Here, it is more important to have the correct label than to have the precise start and finish times of each segment.

We present an application of information theory to the classification and automatic labelling of TV broadcast material into speech, music and noise. We use information theory to construct a decision tree from several different TV programs. This is known as the training data. We then apply this decision tree to a different set of TV programs, known as test data. We present the classification results on the training and test data sets. The correct classification rate at the frame level, for the training data was 95.5%, while for the test data it ranged from 60.4% to 84.5%, depending on the TV program type. At the segment level, the correct recognition rate and accuracy on the train data were 100% and 95.1%, respectively while for the test data the %correct ranged from 80% to 100% and %accuracy ranged from 64.7% to 100%.

1. INTRODUCTION

There is currently a huge growth of the number of hours stored at different digital media archives all over the world. A big part of these collections is made up of broadcast TV material stored for future reference or retrieval. In order to have efficient access to such collections, good indexing is needed. The indexing cannot be made manually due to the large number of audio processing hours; therefore the need for automated indexing techniques.

There has been some previous work in this area. Some of this work concentrated on discrimination between speech and music. John Saunders [2] uses a discrimination technique based on statistics of the energy contour and the zero-crossing rate. Eric Scheirer and Malcolm Slaney use various combinations of 13 features [3]. Other researchers discrimination between speech, music, silence and other sounds. Pfeiffer *et al* [4] use perceptual criteria by matching characteristics such as amplitude, pitch and frequency. Jonathan Foote uses a supervised tree-based vector quantizer trained to maximize mutual information (MMI) [5], [6].

This paper concentrates on the indexing of music, speech, silence and noise segments in an audio file. We are interested in labelling any video program into segments of speech, segments of music and segments of noise. The timing accuracy of the segments for such labelling is not of critical importance. We need more accuracy in "what" is in the track then "when" it occurs.

This basic indexing will enable us to determine the percentage of speech and music contained in each video track and will allow efficient browsing of the video archive. We will be able, with systems such as ACSys Film Reserachers Archival Navigation Kit [7] to view only the relevant segments that contain only speech or only music.

The proposed system [8], [9] uses information theory to construct a decision tree from several different TV programs. During the training phase, the feature extraction framework [10] extracts features from the continuous audio signal on a frame by frame basis, and the C4.5 induction system [11], [12] uses these features to train a decision tree. The classification is performed at the frame level, using an inference engine to execute the decision tree and classify the firing of the rules. The classified frames are then searched using a simple Viterbi algorithm, with fixed cost penalty, to identify the best path through the audio track.

The decision tree can serve two purposes. It can be used to classify a set of unlabelled attributes, and it can also provide an insight into the reason for that decision or classification [5]. The proposed approach provides also a simple method

for analysing the many acoustic-phonetic theories using *real* audio data.

This paper is organised as follows. In section 2 we introduce the training and classification strategy. In section 3 we present the classification results, at the frame and segment levels, on the train and test database. Performance evaluation of the classifier is covered in section 4. A summary of our finding is presented in section 5 and section 6 concludes the paper.

2. TRAINING AND CLASSIFICATION STRATEGY

2.1 Database

The hand labelled training data consisted of one TV broadcast audio file of about 3 minutes in duration (164 sec). The test data consisted of three different TV broadcast files, each of about 1 minute duration. Only one test file, B2 was extracted from the same program as the training data, but it did not form part of the training data.. Thus we have not generated all the test files from a subset of the training. This makes the task of classification more difficult but also more realistic from the point of view of real application. The test files were also hand labelled to allow for direct comparison with the labels generated during the recognition stage. The composition of training and test data is shown in Table 1.

Name	Train/Test	Duration (s)	Comments
B0	Train	164	Documentary A
B1	Test	42	Documentary B
B2	Test	46	Subset of Doc. A
B3	Test	32	Documentary C

Table 1. Composition of training and test data

The task chosen was to classify a subset of 4 broad classes (Speech, Music, Silence, Other) by reducing the original 7 classes as shown in Table 2.

2.2 Feature Extraction

Seven (7) features were selected, five (5) time domain features and two (2) others. The audio signal was sampled at 8 kHz and each frame was 256 samples long and was shifted forward by 128 samples (16 ms). The samples were normalized and then pre-emphasized using a first-order filter:

7 classes	**4 classes**
VoMale	Speech
VoFemale	Speech
MuInst	Music
MuOther	Music
MuPerc	Music
Sil	Silence
NoiseOther	Other

Table 2. Training and testing subsets of the database

$$H(z) = 1 - 0.95\, z^{-1}$$

The samples were then multiplied by a Hamming window, defined as:

$$W_H(n) = 0.45 - 0.46\cos\left[\frac{2n\pi}{N-1}\right]$$

where N is the number of samples per frame. For each audio frame, the following time domain features were extracted:

1. Root Mean square (RMS). The logged sum of the squared samples
2. Zero crossing rate (ZCR). The number of time-domain zero-crossings per second.
3. Envelope. This is the maximum amplitude.
4. Difference in amplitude between maximum peak and previous minimum peak within an audio frame (Local_diff_p)
5. Difference in amplitude between maximum peak and following minimum peak within an audio frame (Local_diff_n)

The features Local_diff_p and Local_diff_n were extracted to try and track the dynamics of the time domain signal and it is illustrated in Figure 1. The other two features were extracted from ESPS signal processing package and consisted of:

1. Fundamental frequency (F0) or pitch.
2. Voiced/Unvoiced.

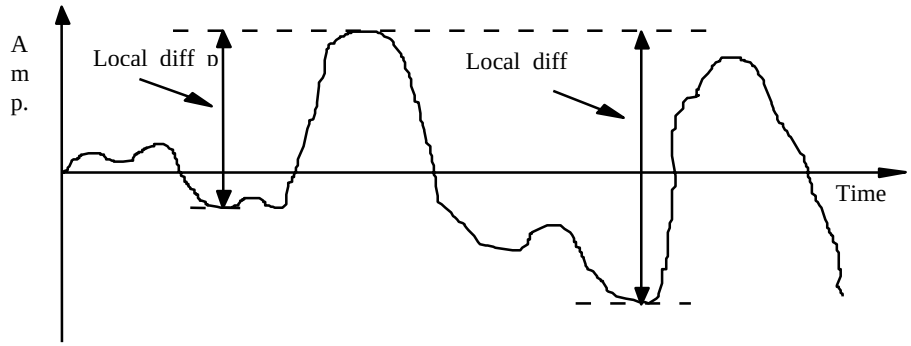

Figure 1. Method of extraction of Local_diff_p and Local_diff_n from the audio signal

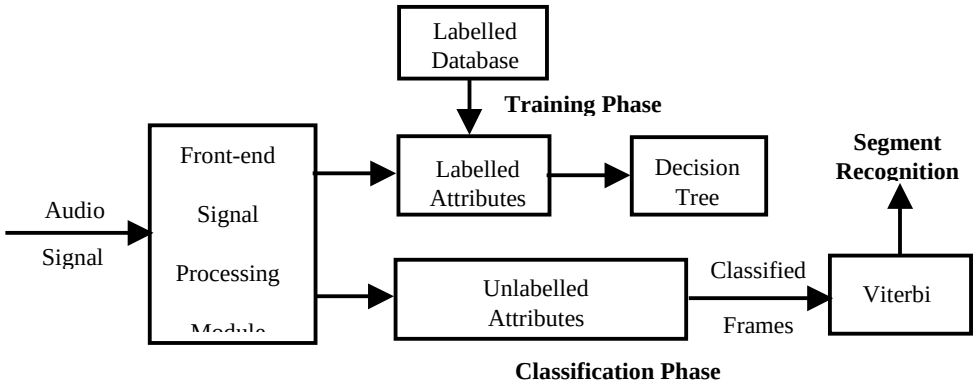

Figure 2. Block schematic of training and classification strategy

2.3 Training

A block schematic of the training and classification strategy is shown in Figure 2. During the training phase, the feature extraction framework extracted the features from the continuous speech signal on a frame by frame basis. The time aligned labelled files were then used to associate each frame with its corresponding label and generate a training data file. The data file contained labelled examples in the form *(X,a)*, where *X* is the feature vector and *a* is the corresponding class. The C4.5 program then used this file to generate a decision tree.

2.4 Classification and Recognition

The classification was performed at the frame level and the performance was evaluated by comparing each classified frame against the reference frame derived from the labelled data. This procedure allowed the correct identification of substitutions and insertions per frame.

In HMM, at each audio frame a probability of observation and transitional probability is generated. Using decision trees for classification, we could only extract a probability of observation in the form of confidence factor, for each class. Since we had no transitional probability, we used a fixed cost factor in the viterbi algorithm to find the best segmentation path through the audio track. To reduce the number of insertions and deletions, a minimum duration constraint was imposed for each class using the average duration from the training data. The output of the Viterbi identified class segment boundaries and generated a label file. This label file was then time aligned, using dynamic programming against the reference label file, to produce the final %correct and %accuracy.

3. CLASSIFICATION RESULTS

The Classification technique was tested on real audio data extracted from TV programs. Three different programs were used and the training data was selected from only one of these (B2). Table 3 shows frame level classification results for train and test audio data. Tables 4 shows the classification results at the segment level after alignment of the reference and test label files, using dynamic programming. Table 5 shows the number of insertions and deletions.

File	Type	% Error
B0	Train	4.5
B1	Test	15.5
B2	Test	26.9
B3	Test	39.6

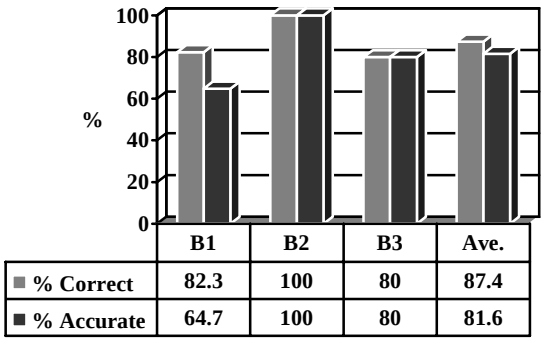

Table 3. Frame level classification results

Table 4. Segment level recognition results

File	%Corr.	%Subs.	%Del.	%Ins.	%Acc.
B0	100.0	0.0	0.0	4.9	95.1
B1	82.4	17.6	0.0	17.6	64.7
B2	100.0	0.0	0.0	0.0	100.0
B3	80.0	6.7	13.3	0.0	80.0

Table 5. Segment level performance results

4. PERFORMANCE EVALUATION

The results presented in the previous section cannot be directly compared with other published results in the open literature, since as yet there is no standard audio database available, similar to TIMIT, Wall Street Journal or Switchboard database which are used by speech recognition researchers.

Saunders [2] reports a 98% classification accuracy on commercial radio broadcasts. Scheirer and Slaney [3] report 1.4% error on a large and diverse collection of FM radio broadcasts. In our experiments, the best segment level classification was obtained on the test data (B2) that came from the same audio track as the train data, but did not form part of the train data. The other two test segments (B1, B3) that came from completely different TV programs, the average %correct and %accuracy were 81.2% and 72.4%, respectively. B1 contained segments of applause, while B3 contained cheering, Laughter and singing.

5. DISCUSSION

A decision tree generated from a relatively short (164 sec) of train data extracted from a single TV source audio track. It was sufficiently robust to be able to discriminate between speech, music, silence and noise on test data that was extracted from two different two programs, with an average duration of 44 sec. The classification performance can be further improved by including examples from these two different sources in the train data, specifically examples of applause, cheering and singing.

We could also label the train data more correctly. For example, there were segments in the train data that contained background music on top of the speech signal but were labelled as speech only. Similarly there were segments that contained clapping or laughter on top of the speech signal. These were labelled as speech. The feature set can also be further optimized. For example, the feature fundamental frequency can be represented as "present" or "absent" instead of its numerical value.

6. CONCLUSION

This paper demonstrated an application of information theory to the classification and automatic labelling of TV broadcast material into speech, music, silence and noise. We used information theory to construct a decision tree from several different TV programs. The experimental results indicate the ability of this approach to build reliable decision trees that can be used to perform the classification task on unknown audio tracks. We have demonstrated that this approach works on audio segments that are a subset of the train data and from very different audio tracks that have different program characteristics.

Unlike techniques such as HMM, this approach is computationally inexpensive with very fast classification in terms of computation cost. Another interesting feature of using decision trees is that the discriminating power of each feature can be determined by examining the structure of the tree. If a feature is never used for a split or used in only a few splits, then it can safely be ignored. This should help researchers to determine features that are important for audio indexing.

7. ACKNOWLEDGMENT

The authors would like to thank Mr. Li Jiang from Microsoft Speech Technology Group, Microsft Corporation for making his alignment program available for testing the segmentation accuracy.

8. REFERENCES

1. Wold, E., Blum, T., Keisler, D. and Wheaton, Jcontent-Based Classification, Search, and Retrieval of Audio", *IEEE Multimedia*, Vol. 3, No. 3, pp 27-36, 1996.
2. Saunders, J., "Real Time Discrimination of Broadcast Speech/Music", *Proc. ICASSP96*, pp 993-996, 1996.
3. Scheirer, E. and Slaney, M., "Construction and Evaluation of a Robust Mutifeature Speech/Music Discriminator", *Proc. ICASSP97*, pp 1331-1334, 1997.
4. Pfeiffer, S., Fischer, S. and Effelsberg, W., "Automatic Audio Content Analysis", Technical Report TR87-881, University of Mannheim, D-68131 Manheim, Germany, 1996.
5. Foote, J., "A similarity Measure for Automatic Audio Classification", *Proc. AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video and Audio Corpora*, Stanford, Palo Alto, CA 1997.
6. Foote, J., "Content-Based Retrieval of Music and Audio", *Multimedia Storage and Archiving Systems II, Proc. Of SPIE*, Vol. 3229, pp 138-147, 1997.
7. Simpson-Young, B. and Yap, K., "FRANK: Trialing a system for remote navigation of film archive", *SPIE Symp. Voice Video & Data Comm.*, Boston, 1996.
8. Samouelian, A., "Frame Level Phoneme Classification Using Inductive Learning", *Computer Speech and Language*, Vol. 11, pp 161-186, 1997.
9. Samouelian, A., "Using C4.5 for Speech processing", *International Journal of Knowledge-based Intelligent Systems*, Vol. 2, No. 2, pp 120-131, 1998.
10. Samouelian, A., "Acoustic feature Extraction Framework for Automatic Speech Recognition", *Proc. Fourth Australian Int. Conference on Speech, Science and Technology*, Brisbane, Australia, pp 629-634, 1992.
11. Quinlan, J. R., *C4.5 Programs for Machine Learning*, Morgan Kaufmann series in machine learning, Morgan Kaufmann publishers, USA, 1993.
12. Quinlan, J. R, "Improved Use of Continuous Attributes in C4.5", *Journal of Artificial Intelligence Research*, Vol. 4, pp 77-90, 1996.