A SPHINX BASED SPEECH-MUSIC SEGMENTATION FRONT-END FOR IMPROVING THE PERFORMANCE OF AN AUTOMATIC SPEECH RECOGNITION SYSTEM IN TURKISH

Cemil Demir[1,2], Erdem Unal[1], Mehmet Ugur Dogan[1]

[1]TUBITAK-UEKAE,Gebze,Kocaeli,Turkey
[2] Bogazici University, Istanbul, Turkey
cdemir|unal|mugurd@uekae.tubitak.gov.tr

ABSTRACT

In this study a system that segments an audio signal as speech and music by using posterior probability based features is proposed and implemented in Sphinx. Unlike the earlier efforts that uses Multi-Layer Perceptrons (MLP), this system uses Hidden-Markov-Model based acoustic models that are trained in Sphinx for posterior probability calculations. Acoustic Models are trained with the HMM-states that are associated with the context-independent phones. Frame likelihoods given acoustic models are computed and used to extract posterior probabilities. Entropy and Dynamism which are the main features that are selected for speech-music discrimination are constructed with respect to the calculated posterior probabilities. A two-state HMM based classifier that uses Viterbi decoding is implemented. The segmentation performance of the system is tested. Moreover, effect of speech-music segmentation to Automatic Speech Recognition (ASR) is investigated by measuring the Real Time (RT) factor and Word Error Rate (WER). Results showed that, proposed speech-music segmentation method decreases WER and increases the speed of recognition which is tested in a Turkish Automatic Speech Recognition System built in Sphinx.

1. INTRODUCTION

Locating speech and music segments in a given audio sample is called Speech-Music Segmentation (SMS). SMS has important applications in audio content analysis and automatic speech recognition (ASR). In this study, the effect of SMS over ASR is investigated, that is an SMS system will be used as a front-end processing unit for an ASR system to improve its recognition performance. Since the SMS that is proposed in this study requires features and acoustic models that are also used in the ASR, the computational complexity is decreased. It is expected that the performance of the SMS system directly affects the the ASR performance in terms of Word Error Rate (WER) and Real-Time (RT) factor.

WER will decrease with accurate segmentation because:

- without a prior music-speech discrimination, the ASR will run and decode every frame of the audio signal whether it is music or speech. This will lead unwanted word insertions at music segments.
- the music in audio can be assumed as out of vocabulary (OOV) words and as a rule of thumb an OOV word brings up on average 1.5 recognition errors so, the music segments will cause more errors [12].

The reasons why RT factor is expected to decrease with segmentation are:

- the decoder will spend time for decoding music segments and with segmentation we will save time at least as the length of the music segment.
- another reason is due to the pruning method that is used in beam search technique. In beam search technique, when the Viterbi decoding is used the paths that have probability less than some threshold will be pruned and this threshold is defined according to the path that has maximum likelihood. When we have some non-speech parts in the signal, obviously the recognizer will output some random text, however the drawback is not only limited to the unwanted text but also it also causes the computation time to increase due to competetion of many paths with each other during the decoding process and pruning rate will not be as high as the speech parts. Therefore, the extra paths that are supposed to be tracked by the recognizer causes the computation time to increase. The whole process is implemented using Sphinx code except for the Viterbi coding part, which needed to be modified in order to use the newly calculated features.

In an SMS system, there are mainly two problems that must be solved; choosing features that are discriminant for speech and music signals, and designing a robust classifier that performs accurate speech-music segmentation for a given signal. In previous works, many features that capture the temporal and spectral structure of the signals have been suggested, including zero-crossing information, energy, pitch, cepstral coefficients, line spectral frequencies, 4 Hz modulation energy, amplitude, and perceptual features like timbre and rhythm [1-5]. In this work, posterior probability based features introduced in [7] are used, namely entropy and dynamism. As shown in the rest of the paper, these features indeed exhibit efficient discriminant properties yielding high performance for speech-music segmentation.

Our main contribution in this work comes from the fact that, the calculation of the posterior probabilities is done by using MFCC based features unlike MLP that are commonly used in the literature. Since the MFCC's are already calculated by the ASR, there exists no extra computation for extraction or calculation of the features of the SMS system which makes the proposed algorithm more feasible.

Another issue in the system design is the selection of a classification algorithm. Different classifiers like the Bayesian Information Criterion (BIC) [6], Gaussian likelihood ratio (GLR) [1,2,5,7], quadratic Gaussian classifier (QGC) [4], nearest neighborhood classifier [2]; and hidden Markov model (HMM) [8] have been used for this purpose.Recently, systems that uses the BIC [6] classifier for audio segmentation are most popular. The BIC technique is useful for general audio change detection, as it does not require any a priori information about the particular acoustic classes present. However,

in the case that the number and type of acoustic classes is known, it should be advantageous to explicitly incorporate this information into the design of the segmentation system. Since the type and the number of acoustic classes are known in the SMS applications, it is reasonable to use an HMM based classifier. It enables us to decrease the computation time to do segmentation.

This paper is organized as follows. Section II describes posterior-based features and training system, Section III explains the HMM based classifier. In section IV, the speech recognition system that was used to measure the effect of segmentation on the ASR performance is discussed. Section V includes information about the data and test results and lastly Section VI contains conclusions that are arrived in this study.

2. POSTERIOR BASED FEATURES AND TRAINING SYSTEM

Posterior-based features were firstly used by Williams and Ellis [7] to classify speech and music. However, in this study, we mainly use the particular posterior based features, entropy and dynamism, as defined in Ajmera's study [8]. Moreover, different from previous studies, we use HMM-based acoustic models that were used in ASR system and posterior probabilities are found using frame likelihoods. The advantage of using HMM-based acoustic models to calculate posterior probabilities is that, widely used speech recognition toolkits such as HTK and Sphinx, generally use this type of acoustic models. Therefore, we do not need an extra statistical model to estimate posterior probabilities. The acoustic model contains context-independent phones and the states of the phones are used as the acoustic model unit in order to calculate frame likelihoods. If we assumed the uniform distribution over the states, the Posterior probability, which is the probability of a state given an audio frame, can be found using the equation:

$$\mathbf{P}(\mathbf{s_k}|\mathbf{x_n}) = \frac{\mathbf{P}(\mathbf{x_n}|\mathbf{s_k})}{\sum_{\mathbf{k=1}}^{\mathbf{K}} \mathbf{P}(\mathbf{x_n}|\mathbf{s_k})} \tag{1}$$

In this equation, $\mathbf{P}(\mathbf{x_n}|\mathbf{s_k})$ represents the frame likelihood i.e.,the probability of n-th frame generated by k-th state.

2.1. Entropy

Entropy is a measure of the uncertainty or disorder in a given distribution and it can also be interpreted as the fitness measure between a model and a sample data set. In this study, entropy is a measure of fitness between an acoustic model and audio signal. Using the posterior probability defined in the previous section, the entropy can be defined for each frame as:

$$\mathbf{h(n)} = -\sum_{\mathbf{k=1}}^{\mathbf{K}} \mathbf{P}(\mathbf{s_k}|\mathbf{x_n}) \log \mathbf{P}(\mathbf{s_k}|\mathbf{x_n}) \tag{2}$$

It is advantageous to average this instantaneous entropy over a window of several frames, resulting in the averaged entropy at time n. We can find averaged entropy as

$$\mathbf{H(n)} = \sum_{\mathbf{i=n-N/2}}^{\mathbf{n+N/2}} \mathbf{h(i)} \tag{3}$$

Generally, in case of speech, the value of the posterior probability for a particular phoneme state should be much higher than other phoneme states. This means that the value of the entropy will be close to zero, indicating that little information will be gained by knowing its actual value, or, equivalently, that there is little uncertainty over the unknown segments. In the case when a music signal is passed through the primary Acoustic Model, the values of probabilities will be more uniformly distributed, resulting in a higher value for entropy. This assumption is verified with our experimental measurements and the results can be seen in Figure 1. We showed the entropy distributions using experimental data explained in Section 5.1.

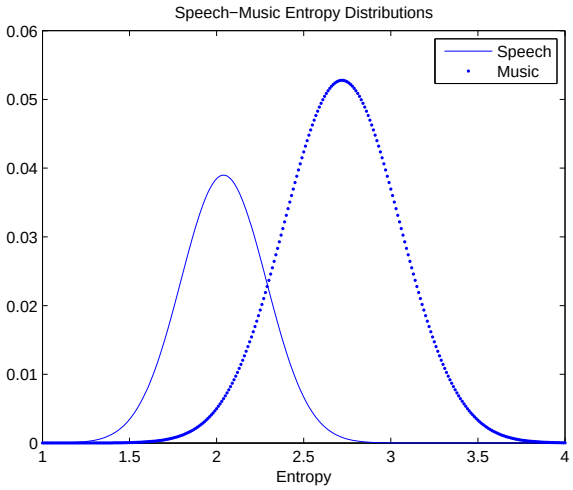

Fig. 1. Speech-Music Entropy Distribution Comparison.

The distributions show that entropy is a discriminative feature for speech and music classes.

2.2. Dynamism

Dynamism is a measure of the rate of change of a quantity. In our approach, dynamism is defined as the rate of change of posterior probability between consecutive frames. For each frame, it can be calculated as:

$$\mathbf{d(n)} = \sum_{\mathbf{k=1}}^{\mathbf{K}} (\mathbf{P}(\mathbf{s_k}|\mathbf{x_n}) - \mathbf{P}(\mathbf{s_k}|\mathbf{x_{n+1}}))^{\mathbf{2}} \tag{4}$$

and as in the case of entropy, it is useful to average out dynamism values over frames to find averaged dynamism per frame. Therefore, average dynamism can be shown as:

$$\mathbf{D(n)} = \sum_{\mathbf{i=n-N/2}}^{\mathbf{n+N/2}} \mathbf{d(i)} \tag{5}$$

This feature captures the dynamic behavior of the probability values. As speech involves more transitions through the speech-specific primary feature space, the phoneme posteriors will exhibit more abrupt changes than other acoustic signals such as music, resulting in higher dynamism. This assumption is also verified with our experimental measurements and are shown in Figure 2. Dynamism distributions are calculated by using experimental data explained in Section 5.1. Note that in this particular system, the negative logarithm of actual dynamism values is used in order to get rid of numerical problems.

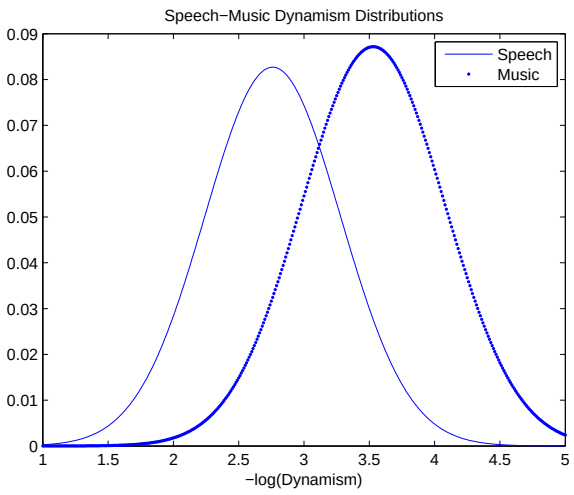

Fig. 2. Speech-Music Dynamism Distribution Comparison.

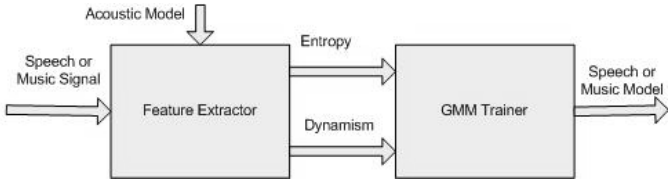

Fig. 3. Model Training System

2.3. Training System

In the previous section, it is claimed that entropy and dynamism are discriminative features for classifying speech and music signals. Therefore, a training procedure that enables us to model speech and music is needed. Gaussian Mixture Models (GMM's) are used to learn the feature distributions of these two classes. The trainer structure is shown in Figure 3 Using previously trained acoustic models and given speech or music training audio files, entropy and dynamism values for each frame are extracted and the two dimensional feature vector (Entropy & Dynamism) is formed. Using the two dimensional feature set, GMM's are trained using conventional Expectation-Maximization (EM) algorithm. Initial models for GMM training is found using the K-Means clustering method. Note that, it was sufficient to use a two Gaussian distribution to model this feature for training set.

3. HMM BASED CLASSIFIER

In this study, in order to find the segmentation for a given test audio, a 2-state HMM-based classifier is utilized. States of the HMM structure consist of previously trained music or speech GMM's. The whole segmentation system is shown in Figure 4. Feature extraction part is similar to the one in the model training. Using previously trained speech and music models and Viterbi decoding, the best path for the segmentation of the test audio is decoded. The HMM topology for the proposed system is the same as in [8]. In the case of speech-music discrimination, this HMM is a 2-state fully connected model, where a minimum duration is imposed for each state. This is achieved by simply concatenating internal states associated with the same distribution. This is enforced because it is assumed that,

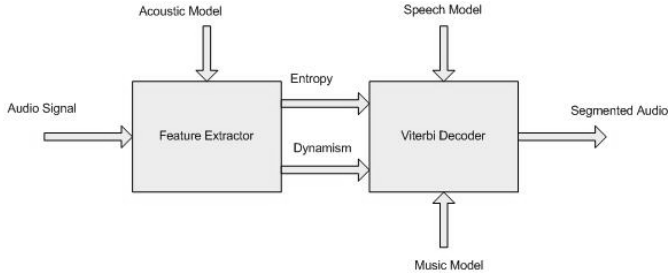

Fig. 4. Speech-Music Segmentation System.

speech and music information can not dramatically change in short periods of time.

4. SPEECH RECOGNITION SYSTEM

In this study, we developed the whole speech recognition system using CMU tools. Acoustic model training is done using 56 hours of speech that was uttered by 216 people and their transcriptions. After the training, models for 87 states of 29 phones is in Turkish is obtained. These models are trained using 13 dimensional Mel-Frequency Cepstral Coefficients (MFCC), their delta and delta-delta vectors. The acoustic model training was done by using the SphinxTrainer[9,10] and the SphinxDecoder[9,10] is used to recognize speech. In speech recognition, to decode an incoming audio, a language model is also used together with acoustic models. For this study, a text that contains 200 million words is used to train the language model and the dictionary contains the most frequent 30000 words. Language model is trained via CMU-LMTK[11] and tri-gram language model is used in the tests.

5. DATA AND EXPERIMENTS

5.1. Feature Extraction Data

In order to extract dynamism and entropy features, two hours of speech that was uttered by 10 people and 2 hours of musical instrument data was used. In this part, 13 dimensional MFCC vectors are used to find frame likelihoods and the respective the posterior probabilities are calculated using these likelihoods.

5.2. Test Data

Since we do not have any speech and music labeled data, we formed our test data synthetically. The synthetic data is formed by concatenation of randomly chosen music and speech parts from our test files. In this study, 3 speech and 2 music segments are concatenated respectively to create the test audio. However, concatenation method information is not used in segmentation part. The test data length is about 134 minutes. The ratio of speech in this test data is %65. There are 100 test files that have 90 seconds average length. The average length of the music segments and speech segments are about 14 and 17 seconds, respectively.

5.3. Test Results

For the segmentation tests, minimum duration length is defined as 3 seconds. The number of states used to impose the minimum duration constraint in the HMM was fixed to 300, thus assuming in our

case that any speech or music segment is never shorter than 3 seconds.The self-loop probabilities were set to 0.9 for the last state of each class. In order to measure the segmentation error of the system, NIST speaker segmentation evaluation method is used. The performance of the system is evaluated with respect to segmentation error and its affect on the performance improvement of the ASR in terms of WER and RT factor. To perform a comparison, MFCC's are also used to model speech and music signals and compared the results with Entropy-Dynamism approach. Firstly, the segmentation error results are presented in Table 1. In this table, ED represents Entropy-Dynamism method, MFCC-8 and MFCC-4 represent the modeling of speech and music signals using MFCC vectors with 8 and 4 gaussians respectively. Although ED method outperforms

Table 1. Segmentation Test Results

Method	SER
ED	0.14
MFCC-4	0.19
MFCC-8	0.11

MFCC-4 method, when the number of gaussians are increased to 8, MFCC method outperforms ED method. One question might be asked here is what happens if the number of gaussians is increased to model MFCC vectors of both classes? The answer is that, for this setup, it did not improve the results. In Table 2, Real Time

Table 2. Speech Recognition Test Results

Data	ST	RT	WER	PER
Correct	0	0.22	%25	%10
ED	0.05	0.31	%33	%16
MFCC-8	0.39	0.27	%32	%17
MFCC-4	0.20	0.32	%42	%28
Raw	0	0.45	%37	%19

Factor (RT) represents the ratio of recognition time to the length of the tested signal. ST shows the ratio of segmentation time to the length of the tested signal. Word Error Rate (WER) denotes the insertions, deletions and substitutions between reference text and recognized text. Phone Error Rate (PER) is the rate of the difference if the phones are treated as words. Moreover, in this table, raw results represents the ASR performance without any speech-music segmentation.

In Figure 5, the total RT factors i.e.,sum of segmentation time and recognition time, and WER's of the different methodologies to recognize a test audio are compared. ED method decreases WER with %11 and RT factor with %16 with respect to unsegmented results. While MFCC-8 decreases WER with %13, it increases RT factor with %50. Moreover, MFCC-4 increases WER with %13 and RT factor with %18. Therefore, it can be concluded that ED method outperforms MFCC methods if we consider both of WER and RT factor together to compare the performances of the methods. The reason why the speed of ED method is better than MFCC approach is that it models speech or music with two gausses and also its feature vector dimension is 2. Therefore, computation time is much lower than MFCC methods.

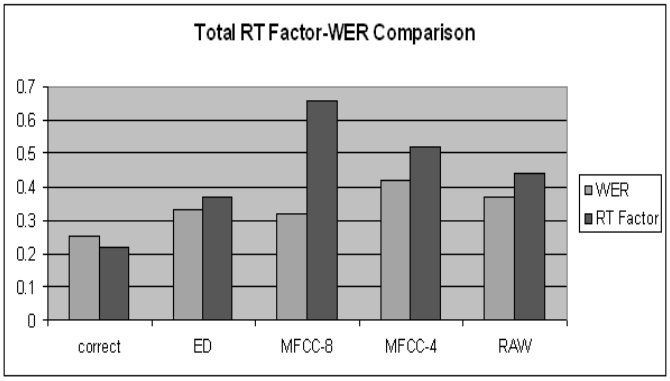

Fig. 5. Total RT Factor and WER Comparison for Applied Methods.

6. CONCLUSION

In this study, as a contribution we showed that HMM based acoustic models can be used to calculate posterior based features instead of MLP's. Another contribution of this study is to find the effect of speech-music segmentation to the ASR results, that is, we investigated the improvements of the segmentation systems to the RT factor and WER. Since we do not have an MLP system we could not compare the results of both systems. However, we compare the results with an MFCC-based feature system and showed that ED method performs as well as MFCC-8 method in terms of WER and its RT factor is %56 of MFCC-8 method. It was also found that even though using MFCC-4 method decreases the segmentation time, it's segmentation error is more than MFCC-8's by %72 and it's WER is also higher than MFCC-8's by %32. Moreover, as stated before, increasing the number of gaussians does not improve the performance in this data set. As a result, ED method can be used as a front end to an ASR system without requiring extra information and it can outperform MFCC-based methods with its computation time-efficiency. With that study we implemented a Sphinx-Based Turkish speech recognition system that uses all Sphinx tools and CMU language model toolkit. Moreover, we used Sphinx acoustic models to segment speech and music parts of an incoming signal.

7. REFERENCES

[1] Saunders, J., "Real-Time discrimination of broadcast speech/music", IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 993–996, 1996.

[2] Sheirer, E., Slaney, M., "Construction and Evaluation of a robust multifeature speech/music discriminator", IEEE Internat. Conf. Acoust., Speech, Signal Process., pp. 1331–1334, 1997.

[3] Carey, M. J., Parris, E. S., and Lloyd-Thomas, H., "A comparison of features for speech, music discrimination", IEEE Internat. Conf. Acoust., Speech, Signal Process., 1999.

[4] El-Maleh, K., Klein, M., Petrucci, G., and Kabal, G., "Speech/Music discrimination for multimedia application", IEEE Internat. Conf. Acoust., Speech, Signal Process., pp. 2445–2448, 2000.

[5] Carey, M. J., Parris, E. S., and Lloyd-Thomas, H., "Feature fusion for music detection", European. Conf. Speech Comm. Technology., 1999.

[6] Chen, S. S., Gopalkrishnan, P. S., “Speaker, environment and channel change detection and clustering via bayesian information criterion”, IBM Tech J., 1998.

[7] Williams, G., Ellis, D., “Speech/Music Discrimination based on posterior probabilities”, European. Conf. Speech Comm. Technology.,pp. 687–690,1999.

[8] Ajmera, J., McCowan, I., and Bourlard, H., “Robust HMM based speech/music segmentation”, IEEE Internat. Conf. Acoust., Speech, Signal Process., pp. 297–300,2002.

[9] Placeway, P. and Chen, S. and Eskenazi, M. and Jain, U. and Parikh, V. and Raj, B. and Ravishankar, M. and Rosenfeld, R. and Seymore, K. and Siegler, M. and others, “The 1996 hub-4 sphinx-3 system”, Proc. DARPA Speech recognition workshop, pp.85-89, 1997.

[10] www.cs.cmu.edu/ robust/Tutorial

[11] P. Clarkson and R. Rosenfeld, “Statistical language modeling using the CMU-Cambridge toolkit ” in Proceedings of European Conference on Speech Communication and Technology (ESCA, Eurospeech 97), vol. 1, Rhodes (Greece), 1997, pp. 2707-2710.

[12] I. L. Hetherington, “A characterization of the problem of new, out-of-vocabulary words in continuous-speech recognition and understanding, ” Ph.D. dissertation, Mass. Inst. of Technol., Cambridge, 1995.