

Music and Speech Detection System Based on Hidden Markov Models and Gaussian Mixture Models

A public White-Paper

by

Compure

www.compure.com

<http://www.compure.com/contactus.htm>

1 Introduction

More and more audio and multimedia information is being recorded and consequently stored in archives. In order to use such big amounts of stored audio data sophisticated technologies to automatically classify these huge amounts of data are needed. One of the first basic classifications is the classification of music and speech.

This White Paper describes a general approach to classify music and speech in audio data. The approach is implemented in commercially available products like ACTNow from Compure.

The music and speech detection system analyses an incoming stream of audio data and detects if either speech or non-speech or music or non-music can be found.

The speech vs. non-speech detection is completely separated from the music vs. non-music detection. A straight combination of both independent results can be used to get a continuous classification of an audio stream into the following classes: music, speech, speech with music, neither music nor speech.

Special requirements for the music and speech detection are the Signal Dominance Requirement (SDR). The Signal Dominance Requirement means that for example speech should only be detected if speech is the dominant signal. It should not be detected if speech is just a background signal. The same is required for the music detection and the speech with background music detection.

The described detection system uses very specialized and optimized feature extraction algorithms and parameterization in combination with

a statistically trained Hidden Markov Model (HMM) using Gaussian Mixture Models (GMMs) as emission probability models.

Due to its statistic nature the system can be trained to match any input stream. The more data used to train the system the better the system will perform in general.

The outputs of the system are of the type start- and end-timestamp and probabilities of each detection item.

As a first step the detection system extracts the feature vectors using different specialized analysis methods to model the characteristics of music and speech respectively.

In a next step the feature vectors are used to do the actual classification step where music vs. non-music, speech with background music vs. no-speech with background music and speech vs. non-speech will be evaluated.

Finally the result contains the start and end time of each classification result including a probability which can be used to prioritize the results.

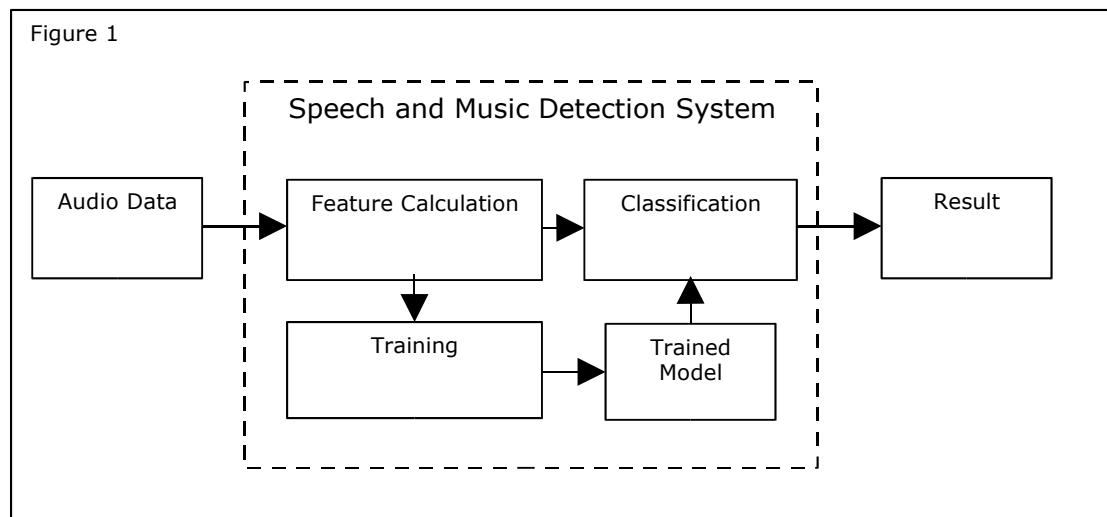
2 Detection System Description

Figure 1 shows the general system architecture of the Speech/Music Detection System.

Audio data is fed as input into the detection system. Theoretically the audio source can be anything. Practically typically a telephone, microphone, broadcasted music / TV or stored files are audio sources. Due to the bandwidth limitations of

2.1 Audio Data Input

Samples from an audio source (e.g. microphone, telephone, broadcast, audio files...) are called audio data. The data should be uncompressed in PCM (Pulse Code Modulation) format. Data from telephony systems are typically 8 kHz samples with 16 Bit amplitudes. 8 Bit data should be expanded to 16 Bit by applying A-law



telephones (8kHz sampling rates) a special parameterization of the detection system is needed.

or μ -law. Audio sources which produce higher quality samples (e.g. 32 kHz) or 22kHz) can be converted to 16kHz data.

Compressed audio data (e.g. MP3, CELP, ADPCM...) should be converted to PCM (either 8kHz or 16kHz) prior to feeding to the Speech and Music Detection System.

Basically two models for telephony (8kHz) and other input sources (16kHz) are sufficient for most applications.

2.2 Feature Calculation

Since the feature calculation drives the complete detection system a careful selection of features is essential for the overall quality of the detection system.

Two different analyses methods are used. Basically for the speech vs. non-speech detection a cepstral analysis according to Mel scale is used. For the music vs. non-music detection mainly an analysis with spectral characteristics is used.

To obtain a high detection quality typically about hundred cepstral and spectral feature vectors need to be calculated per second.

From the vector sequences additional features specifically tailored for speech or music respectively are derived to get a complete description of all relevant input signal characteristics.

Also channel variations and quite often additive noise signals require special algorithms. Therefore enhanced feature normalization techniques are deployed to deal with channel variation and additive noise.

2.3 Trained Model

The model describes the "nature" of the signal for detection. For example the Speech model describes speech in its basic characteristics.

Speech and non-speech are modeled statistically with one or more Hidden Markov Models.

The emission probabilities are modeled as various mixtures of Gaussian probability density functions.

The modeling for the variety of duration of speech and non-speech is accomplished by the topological structure of the Hidden Markov Models and by applying optimized penalties for model changes.

Experimental tests showed us that the HMM model is by superior to heuristic smoothing methods to avoid too frequent switch in the results of the emission models.

2.4 Training

The statistical approach of the detection system allows a continued training and improvement of the detection system.

The more training accomplished the better the quality of the detection system.

It is essential that the variety of different input signals (e.g. movies, telephone, music, sport events...) is covered properly in the training sessions. The training material used consists of dozens of hours of audio data of various types and sources.

The Hidden Markov Models are trained with standard expectation-maximization techniques and up splitting of probability density functions.

Various experiments showed that an optimized HMM configuration delivers superior results to most other approaches.

2.5 Classification

Multiple Gaussian Mixture Models (GMMs) are used to calculate the emission probabilities.

The classification is made by calculation of the log likelihood for each model of speech and non-speech, music and non-music respectively.

Hidden Markov Models (HMMs) are also used to model the different duration possibilities of music or speech. Using only GMMs would result in a too frequent change between music and non-music or speech and non-speech respectively.

The Hidden Markov Models for speech and non-speech are organized in a net reflecting all allowed model transitions.

A Viterbi-Decoder processes the incoming features and determines the best path through the net. All hypotheses are continuously traced back to determine any fixed part and to provide immediate feedback with minimal delay.

2.6 Result

The detector emits multiple results during feeding audio data where as the speech/non-speech results follow consecutively. The result contains the start-time-stamp, end-time-stamp and the probability. In a final step speech and non-speech segments are combined to reflect the probability of speech. A probability of 0.99 means that a detection item (e.g. speech or music) is extremely likely. A probability of 0.1 means that the item is extremely likely not there. 0.5 means that it is not clear if the item is there not (this value carries the smallest information).

To measure the accuracy the Equal Error Rate (ERR) is used. The EER is

the error rate is the operating point where the number of False Alarms (wrong detection) is equal to the number of False Rejections (wrong missing).

Typically if broadcasted audio data consisting of news reports and music is mixed up in consecutive segments of sizes larger than one second then the EER is about 3% for speech and 6% for music.

Response time is typically between several hundred milliseconds and several seconds.

3 Conclusion

The above detection system is according to our experience and experiments one of the most accurate and sophisticated methods to detect music and speech in an audio stream.

To achieve a maximum quality the detection approach requires a combination of various mathematical statistical calculations and models.

Especially the proper choice of features is essential for the quality of the overall system.

A reasonable amount of effort should be spend on fine-tuning all parameters involved in the overall recognition process.

The statistical approach allows the system to "learn" and to improve the performance