Speech-Music Classifier

Group30

Abstract—**Given a audio sample we will attempt to design a system that will give an output that will classify whether the sample contains speech or music. Audio sample should purely consists of speech or music. We will analyse the audio sample to come up with a criteria for classification.**

I. Overview of Step In Designing Classifier

feature extraction

In this step we extract the features of audio signal like Intensity variations,zero crossing rate,short time energy ,short time energy variance,short time auto-correlation etc

feature analysis

In this step we will analyze these features for known music and voice samples . This is the training phase.In the initial stages of building the classifier we will analyse all the different features independently and construct independent decision tree classifiers to observe if these features are able to classify the unseen signal.
After initial analysis we construct a decision tree taking into account all the feature.
Testing phase:-
In the testing phase,we provide a know audio signal and observer if the decision tree classifier performs accurately.Else we will again train the classifier including more number of samples or use additional features.
Acutal classification Task: Now we provide unseen audio sample and classifier will provide whether given sample as speech or music.

II. Feature Extraction

Audio signal's are non stationary signals.Conventional techniques of signal analysis source of signal is Linear and Time invariant.So that entire system as a whole is visualized as a LTI system.

Speech systems are produced by our vocal tract which is a time varying system with time varying excitation.Hence we cannot apply tools available to process LTI systems directly on speech signals as they violate underlying assumptions.

To overcome this solution,we divide the speech signal into small intervals of 30ms-50ms and we make a reasonable assumption that source producing signal is stationary and LTI for such a short duration of time.and speech processing system can be visualised as cascade LTI system each acting on a stationary part of the signal.

In our current application we have assumed that source is stationary for a time interval of 50ms.We divide the audio signal's to be processed into time intervals of 50ms.Each section may be called as a frame.

We know that audio signals are highly correlated .Each frame is part of whose signal and cannot be treated as independent entity .If they are treated independently then we have effectively introduced discontinuities at the instants the frames are joined.We will observer spectral leakage and time domain aliasing effect due to this.

One way to avoid this to divide the signal into overlapping frames weighted by suitable window to reduce the effects of spectral leakage.Such discontinuities also lead to phase discontinuities which can be resolved by proper alignment of the blocks.

In all the methods of feature extraction ,we will extract the features of individual frames and take the mean values of features of individual frames as feature for the given audio signal,Since current application is designed to analyse the entire audio signal and classify the content of the audio signal.

III. VARIANCE OF SHORT TIME ENERGY

Short-Time energy is a simple short-time speech measurement. It is defined as:

$$E(k) = \Sigma_N |X(k)|^2$$

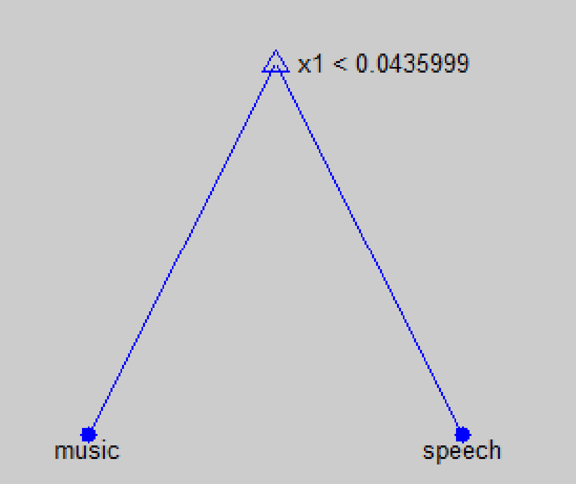
where X(k) is the FFT of a overlap window frame of duration 50ms .
we calculate the energy for each frame and calculate the variance of energy in the frames
This variance parameter is a feature for the given audio signal.
Speech signal will have more silent frames compared to audio signal which is the nature of speech signal.Thus we expect a higher value of variance for speech than music .We use decision tree classifier and arrive at below classification criteria.
We find the mean values of variance of short time energy for various know audio samples. and try to build a classifier based on this criteria as :
if $variance < 0.0435999$ then class = music else class = speech

We can see in the parameter table provided that the data is relatively well clustered for the available training data.

TABLE I
SHORT TIME ENERGY VARIANCE TRAINING DATA

Class	Parameter
1	0.0891
1	0.0896
1	0.0661
1	0.0542
1	0.0531
1	0.0340
1	0.0454
2	0.0406
2	0.0294
2	0.0291
2	0.0133
2	0.0265
2	0.0741
2	0.0418

Fig. 1. classifier for short time energy



IV. SHORT TIME AUTO-CORRELATION

Short time auto-correlation is a time domain approach.Music in general will be more correlated than speech as it contains less number of silent frames and is slowly varying as compared to speech signal. It computer the auto correlation parameter of various frames.short time auto correlation parameter is the mean values of parameter over all the frame.Again we are taking the average values as this parameter will represent the entire audio signal.

TABLE II
SHORT TIME AUTO CORRELATION TRAINING DATA

Class	Parameter
1	208.5398
1	152.7837
1	222.6496
1	222.6496
1	6.0494
1	193.7587
1	-107.8561
2	249.8945
2	240.8344
2	212.2958
2	281.5736
2	111.3175
2	169.0144
2	870.8536

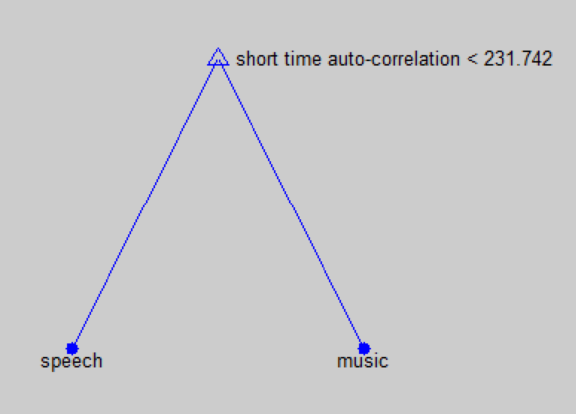
we can observer the parameter matrix and see that parameters are not clustered.

Decision criteria is :

if short time auto-correlation$<$ 231.742 then class = speech else class = music

This parameter will not provide correct classification alone,hence it is suitable to use with other parameters.

Fig. 2. classifier for short time auto correlation



V. Zero Crossing Detector

VI. Decision Tree Classifier

A decision tree is a class discriminator that recursively partitions the training set until each partition consists entirely or dominantly of examples from one class. Each non-leaf node of the tree contains a split point that is a test on one or more parameters or features which determine which class the given belongs to. The test condition at any node of tree are binary test condition. We construct a 3-dimensional(one for each feature vector) vector for each audio signal .Decision tree algorithm provides to us the optimized criteria for classification.We can observer that this will only depend in the short time energy variance criteria,since for the criteria this parameter will give error,the Short time auto correlation also gives error.Hence optimized criteria will only depend on the energy variance
If more data is available ,the classifier will be more robust and come up with a more fine tuned classification criteria. we can observe scatter plot for short time variance is relatively well clustered compared to scatter plot for short time auto correlation.Hence short time variance is a more dominant and accurate is used for classification in classier where all parameters are provided as input to the learning and classification algorithms.

Fig. 3. scatter plot for short time auto correlation

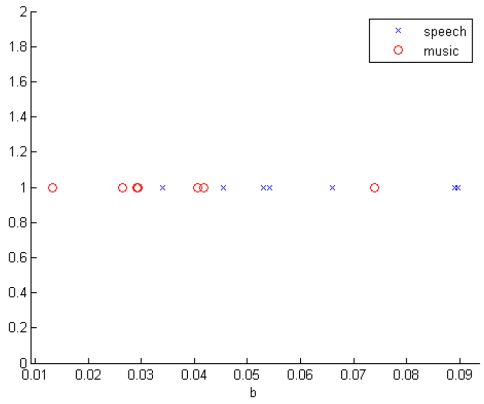


Fig. 4. scatter plot for short time auto correlation

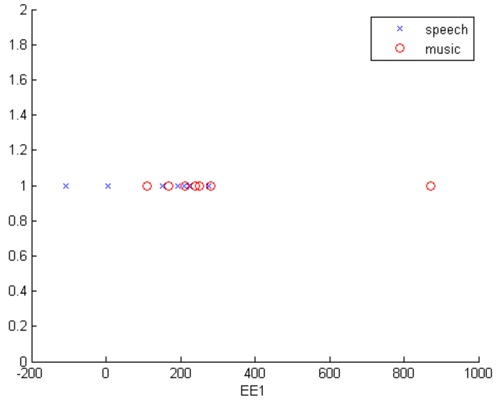


TABLE III
FILES USED FOR TESTING

Class	Parameter
train1.m	training function for short time energy variance parameter
train2.m	training function for short time autocorrelation parameter
train.m	training function including all the parameters
acorr11	function to calculate short time auto correlation
test1.m	testing function for short time energy variance
test2.m	testing function for short time auto correlation

VII. FILES INCLUDED

VIII. CONCLUSION

We are successfully able to classify the audio signals as music and speech for the given test data files based on short time features of audio signal like short time energy variance and short time auto correlation.This a design for initial approach for a classifier.As we extract more feature and if more training samples are available we will be able to build a robust classifier.