A SPEECH-MUSIC DISCRIMINATOR USING HILN MODEL BASED FEATURES

Balaji Thoshkahna, Sudha.V and K.R.Ramakrishnan

Music and Audio Group(MAG),
Learning Systems and Multimedia Labs(LSML),
Department of Electrical Engineering,
Indian Institute of Science

ABSTRACT

We propose a simple speech music discriminator that uses features based on HILN(**H**armonics, **I**ndividual **L**ines and **N**oise) model. We have been able to test the strength of the feature set on a standard database of 66 files and get an accuracy of around 97%. We also have tested on sung queries and polyphonic music and have got very good results. The current algorithm is being used to discriminate between sung queries and played (using an instrument like flute) queries for a Query by Humming(QBH) system currently under development in the lab.

1. INTRODUCTION AND PREVIOUS WORK

Speech Music discrimination is an important task for Music Information Retrieval(MIR), highlight extraction and also advertisement detection. For the purpose of discriminating between programs and advertisements, Saunders[1] designed a realtime speech/music discriminator(SMD) using zerocrossing rates as a simple feature. Scheirer et al[2] designed a SMD using 13 temporal and spectral features followed by a GMM based classifier. A novel SMD was designed by Jarina et al[3] that used rhythm based features of polyphonic music. Maleh et al[4] describe a frame level SMD that works on 20 millisec frames to declare them as speech/music.

A much simpler form of SMD was designed by Karneback[5] using the 4Hz modulation energy feature. A PR based SMD that could detect sung phrases was designed by Chou et al[6]. A fast SMD with low complexity was designed by Wang et al[7] using a modified energy ratio feature followed by a Bayesian classifier.Our work is aimed at developing an SMD that can distinguish between singing and music also along with speech versus music.

Most of the algorithms described above use GMMs, Neural networks etc for classification purposes and use a large number of features. These are usually compute intensive and slow for real-time purposes. In this work it is desired to incorporate the capabilities of Chou's algorithm along with low complexity in our algorithm using the HILN model to derive features that separate speech against music. The HILN model is based on the Sine+Noise model for musical analysis made popular by Rodet[8]. We are currently using the speech-music discriminator for the purpose of distinguishing between sung queries and instrumental queries in a Query By Humming(QBH) system being developed in the lab.

2. SPEECH MUSIC DISCRIMINATOR ALGORITHM

The QBH system mentioned above allows for users to hum a query or to play it using an instrument such as the flute, violin etc. Since sung queries and instrumental music are processed separately by the system, we would like to discriminate between the sung queries and instrumental queries. Figure.1 shows the block diagram of our speech-music discriminator. All input signals are resampled to 16kHz and normalized to an average intensity of 70dB. We then pass the processed signal through a HILN model based feature extraction block. This block extracts 4 features that are described as below.

2.1. The HILN model

The HILN model has been proposed as an extension to the basic sine model of Quatieri and Macaulay[9]. MPEG4 audio codec is based on the HILN model. This model uses harmonics of sinusoids, individual sinusoids and noise to code audio.For more details about HILN model and its applications to audio coding, refer[10]. We use HILN model based features that are found for speech, music(both monophonic and polyphonic) and also human singing. In the following section we show the discriminating capabilities of the features we derive through a set of experiments and plots.

2.2. HILN model based features

The sine+noise model analyzes the signal and picks peaks in the frequency spectrum. We have used Dan Ellis' implementation of the sine+noise model[1]. The signal is split into 20

[1] www.ee.columbia.edu/∼dpwe/resources/matlab/sinemodel/

1-4244-0469-X/06/$20.00 ©2006 IEEE

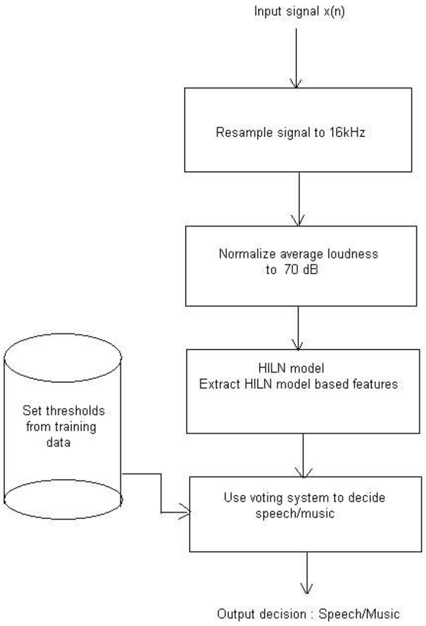
Input signal x(n)
Resample signal to 16kHz
Normalize average loudness to 70 dB
HILN model
Extract HILN model based features
Set thresholds from training data
Use voting system to decide speech/music
Output decision : Speech/Music

Fig. 1. Block diagram of the speech-music discriminator

millisecond frames with an overlap of 10 milliseconds. A high resolution 4096 point STFT(Short time Fourier transform) is taken and the peaks are picked in the frequency domain using an energy threshold(i.e all peaks within 90% of the maximum peak are picked). A continuity analysis is performed in the frequency domain to pick only sinusoids that are stable[11]. Only sinusoids that are continuous for atleast 15 frames are retained. This way, spurious / discontinous peaks are eliminated from further processing.

Now, we search for sinusoids and their harmonics in every frame. If a frame doesn't contain any sinusoids, it is labelled "unvoiced" and removed from further processing.If a frame contains atleast 1 sinusoid,it is labelled "voiced". For every "voiced" frame, we search for the harmonics of a given sinusoid in that frame. A sinusoid of value S Hz in the k^{th} frame is searched for its harmonics. Sinusoids that are off by +/- 10Hz of the multiples of S are also considered as its harmonics. This way the harmonics of every sinusoid in the k^{th} frame is calculated.

The sinusoids that are not the harmonic partials of any of the sinusoids found in the frame are labelled "individual lines"(See eqn.1). This way we calculate the "average number of sinusoids per voiced frame"(S_{avg}), the "average number of harmonics per voiced frame"(H_{avg}) and the "average number of individual lines per voiced frame"(I_{avg}).

$$S(k) = H(k) + I(k) \qquad (1)$$

where $S(k)$ is the total number of sinusoids in the k^{th} frame, and similarly $H(k)$ and $I(k)$ are the total number of harmonics and individual lines in the k^{th} frame respectively.

Each frame in the original signal is classified as "silence" or "sound " based on a simple threshold. The noise residual is calculated for the frames classified as "sound" as follows. For every frame, we calculate the energy in the sinusoids(if they are present) picked from the previous steps and subtract it from the energy of the frame. Before subtraction, the sinusoids are smoothened by the windowing signal as shown by the following steps.

Let the original signal be $x(n)$ and the DT-STFT(Discrete Time - STFT) be $X(n, k)$. Let the windowing function be $w(n)$ and the Fourier transform of $w(n)$ be $W(w)$. Then;

$$X(n, k) = \sum_{k=n}^{n+N-1} x(m).w(n-m).e^{-2j\pi mk/N} \qquad (2)$$

In the frequency domain, the multiplication in eqn.(2) corresponds to convolution of the Fourier spectrum of the signal with the Fourier spectrum of the window. Let the signal representing the harmonics picked, as mentioned in the previous paragraph be $H(n, k)$. $H(n, k)$ now contains only impulses at the samples corresponding to the frequencies identified by the peak picking algorithm. We convolve the signal $H(n, k)$ with $W(w)$. This leads to a smoothened version of the harmonic spectrum $H(n, k)$.

$$H_{smooth}(n, k) = H(n, k) * W(w) \qquad (3)$$

This $H_{smooth}(n, k)$ is the Fourier spectrum of the windowed version of the harmonics in the previous section. We subtract the $H_{smooth}(n, k)$ from X(n,k) to get the residual function $R(n, k)$. We find the average energy of $R(n, k)$ and call it the "average residual energy per voiced frame".

2.3. Using the HILN model based features

The above 4 features were calculated for a set of 84 training files that contained 45 music files (monophonic trumpet, piano, flute and violin of average 4 seconds length), 25 speech files(of average length 4 seconds) and 14 singing files(of average 15 seconds length). Figures 2, 3, 4, and 5 show a histogram of values of S_{avg}, H_{avg}, I_{avg} and R_{avg} for music,speech and singing in 3 separate subplots. As can be seen, simple thresholds can separate the speech signals from music signals. Also we can very easily see that the "average number of individual lines per voiced frame" I_{avg} is a feature that doesnt have any discrimination between speech, music and singing. We thus neglect this feature based on the histogram plot shown in Figure.3. Thresholds set for S_{avg}, H_{avg} and R_{avg} are 1.5, 2 and 0.075 respectively. These thresholds are set to minimize the misclassification error (Though we have not used Bayesian classification thresholds, it is very simple from our formulation to set those thresholds too). The experiments and results on test data and the database information are given in the next section.

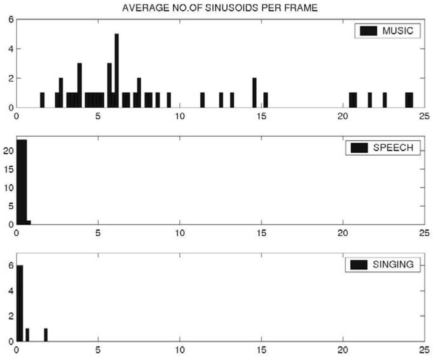

Fig. 2. Average number of sinusoids per voiced frame

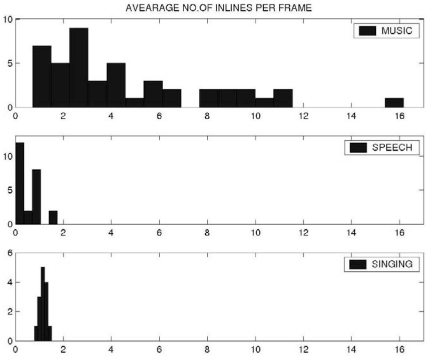

Fig. 3. Average number of individual lines per voiced frame.As can be seen from the plot, the I_{avg} feature is not of much use. So we neglected this feature.

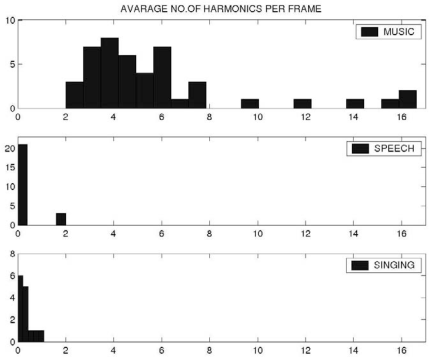

Fig. 4. Average number of harmonics per voiced frame

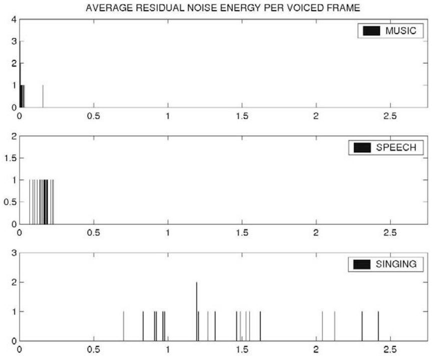

Fig. 5. Average residual noise energy per voiced frame.In all the above histograms, the x-axis corresponds to the value of the feature and the y-axis corresponds to the number of times the feature value occurs in the experiment.

3. EXPERIMENTS AND RESULTS

For the experiments, we used Slaney and Scheirer's database [2] for speech music discriminators.The database has around 150 files belonging to various categories like music, vocals, non-vocal sounds etc. We also used high quality singing files from our QBH database. We split the above database into training and testing data with 60% of the database going into training and the remaing 40% of files into testing. As described in the algorithm, the training phase was used to estimate the thresholds for the various features and their discriminating capabilities.

After setting the thresholds, the testing was done for a simple

[2] http://www.ee.columbia.edu/∼dpwe/sounds/musp/music-speech-20051006.tgz

speech or music classification.The 3 features S_{avg}, H_{avg} and R_{avg} were calculated for each of the test file. Along each feature we classified the file as speech or music based on whether the feature was greater than or less than the set threshold . A final voting was then done to select the majority class as the class label for the audio file(i.e if the file was classified as "music" along 2 features and as "speech" along 1 feature, we classified it as a music file).

Table.1 gives the results for the test data.The first column gives the instrument class, the second column the number of files of the particular insrument/speech class and the third and fourth columns give the number of files classified as speech or music respectively for the given class of instrumens. As can be seen from the table, this simple technique gave an accuracy of 96.8% for the 32 monophonic instrumental files and 97.5% for the 41 speech and singing files we tested.

We also tested this technique on 15 polyphonic music files of average length of 7 seconds. These files were clips from various CD recordings of movie songs and various albums(Note: No training was done using polyphonic music data). 12 out of the 15 were classified as music. This technique compares well with the results from various other algorithms, but is much simpler and low on computation. With slight modifications, this can be used for realtime speech-music discrimination purposes also since the classifier is extremely simple. Compared with the various algorithms referred in this paper(See [3] for a comparative analysis of various algorithms), we get an accuracy that is good enough for an SMD with simple features. Also the peak picking in the frequency domain and continuity analysis are fast and involve no iterative techniques (unlike the other training methods mentioned in this paper).

4. CONCLUSIONS AND FUTURE WORK

The paper proposes a simple and computationally fast algorithm that discriminates between speech(spoken sentences and sung phrases) and music(monophonic intrumental performances and polyphonic sounds). This has been proposed taking into consideration that music is inherently more harmonic than speech. This algorithm can be extended to realtime applications such as advertisement detections in broadcasts, speech music segmentations in movies etc. Also the HILN model is definitely useful in various other applications than just audio coding as can be seen from this work.

5. REFERENCES

[1] John Saunders, "Real-time discrimination of broadcast speech/music," *ICASSP*, 1996.

[2] Eric Scheirer and Malcolm Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," *Proc. of High Performance Computing on the Information Superhighway*, 1997.

[3] S.Marlow R.Jarina, N.O.Connor and N.Murphy, "Rhythm detection for speech-music discrimination in mpeg compressed domain," *14th International Conference on Digital Signal Processing*, 2002.

[4] Grace Petrucci Khaled El-Maleh, Mark Klein and Peter Kabal, "Speech/music discrimination for multimedia applications," *ICASSP*, 2000.

[5] Stefan Karneback, "Discrimination between speech and music based on a low frequency modulation feature," *Eurospeech*, 2001.

[6] Wu Chou and Liang Gu, "Robust singing detection in speech/music discriminator design," *ICASSP*, 2001.

[7] W. Gao W.Q. Wang and D.W. Ying, "A fast and robust speech/music descrimination approach," *Proceedings of ICICS-PCM*, 2003.

[8] X.Rodet, "Musical sound signal analysis/synthesis:sinusoidal + residual and elementary waveform models," *TFTS-97*, 1997.

[9] R.J.Macaulay and T.F.Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on acoustics,speech and signal processing(ASSP)*, 1986.

[10] Nikolaus Meine Heiko Purnhagen, "Hiln - the mpeg-4 parametric audio coding tools," *ISCAS-2000*, 2000.

[11] Tuomas Virtanen, "Audio signal modeling with sinusoids plus noise," *MSc Thesis*, 2000.

Table 1. **Tables for the set of experiments**

Test Results			
		Class label	
Audio Class	No.of files	Speech	Music
Flute	10	0	10
Violin	10	1	9
Trumpet	6	0	6
Piano	6	0	6
Polyphonic	15	3	12
Total Music	47	4	43
TIMIT speech set	14	14	0
Slaney's speech set	7	6	1
Singing	20	20	0
Total Speech	41	40	1