Discrimination between speech and music based on a low frequency modulation feature

Stefan Karnebäck
Dept. of Speech, Music and Hearing, KTH and College of Engineering, KTH-Visby, Sweden
stefan@speech.kth.se and stefan@visby.kth.se

Abstract
The possibility to discriminate between speech and music signals by using a feature based on low frequency modulation has been investigated.

Three different low frequency modulation parameters have been extracted and tested concerning the ability of discrimination.

The low frequency modulation amplitudes calculated over 20 critical bands and their standard deviations were found to be good features for this discrimination task even with VQ models. They were also found to be less sensitive to channel quality and model size than MFCC features.

1. Background and review
In preprocessors for speech recognition there is often a need to classify and segment the signal before the transcription. Many attempts have been made to discriminate between speech and music or other sound classes [1-4]. Most of these are based on knowledge of the speech production and perception or a combination of the two. The relation between these is described by Greenberg [5] where he in particular points out the important aspects of

- the spectral energy distribution
- the sound pressure
- rapid changes in spectral energy over 100 ms interval
- micro- and macro-modulations

These findings are explained from a production point of view by Stevens [6], showing that the larger organs used in the speech production, i.e. the soft palate, the tongue, the jaws and lips need approximately 50-300 ms to complete their extreme movements (from one extreme point to another and back again). This corresponds to a maximum frequency of 3-20 Hz.

Typically the spectral characteristics are used in different kinds of cepstral coefficient analysis and the added information in the rapid change is used in the delta and delta-delta features. These are used in many speech and speaker recognisers and therefore even in the preprocessing part performing the segmentation [1,2].

Some of these aspects can even be found in the time domain, used by Saunders [3] for speech/music discrimination. To catch the shape of the spectrum he used the Zero Crossing Rate, suggested by B. Kedem [7], rather than the cepstral features.

The micro-modulations are used in different pitch tracking algorithms [8] and the macro-modulation can be found in the low frequency modulation amplitude. (It will be referred to as the LF modulation or LF component.)

Greenberg uses this feature as a preprocessor for a speech recogniser [9]. This technique filters out the low and high frequency noise since the speech information is generally following the syllable rhythm of approximately 4 Hz. This fact is also used as a channel equalisation and noise reduction method, the RASTA processing [10]. Bacon and Viermeister have shown that normal-hearing persons are sensitive to LF modulation, especially in the range of 2-15 Hz [11]. Compare the findings of Stevens [6] above.

This LF modulation of speech, among other features, was used by Scheirer & Slaney in a speech/music discriminator used on broadcasting recordings [4]. They found that the normalised 4 Hz component was specifically higher in speech than in music. The component for each of 40 perceptual channels were calculated and added. This feature was found to be a good discriminator. In the same report they also used the fact that music has a beat or rhythm that follows all the frequency bands synchronously. A score for synchronous events in the different bands over 5 seconds was calculated. These features alone were two of the best discriminators in that report.

2. The low frequency modulation
The aim of this study has been to examine this LF modulation component. This includes the extraction of a feature representing this aspect and examination of its ability to discriminate between speech and music by itself or in combination with conventional cepstral features.

Discriminating between speech and music by the LF component causes problems, however, since music also has a strong LF component in all frequency bands that could be very close to 4 Hz. For example, the rhythm of 60 per second will produce a 4 Hz component on the 16th notes.

The difference between speech and music is that in music the modulation is more synchronous over all bands or a wider frequency range, while it differs more in speech. However, there could be some bands, even close to each other, that do not follow the same modulation pattern in music.

We found a difference in low frequency modulation behaviour on the following issues:

- Speech had a high correlation, especially between adjacent bands, while music had a rather low correlation between all bands, see example in Fig 1.

- Both the LF modulation amplitude and its standard deviation varied more for speech than for music, especially in bands 5-10 (or 12), 400 – 1300 (1700) Hz, corresponding to the movements of the first and partly the second formant. This behaviour can be noticed in Fig 2.
The first observation could be explained by the fact that almost all frequencies are modulated by the same signal (or movement) in speech while music contains several modulation sources, i.e. many instruments.

3. The LF modulation features

3.1. 4 Hz amplitude and standard deviation

A new feature was extracted from the LF modulation amplitude and its standard deviation for 20 critical bands, as mentioned above, giving a 40-dimensional vector, referred to as 4 Hz ASD. This feature was extracted in the way it was suggested by Greenberg [5]. This means using a critical-band filter bank (20 bands), half wave rectifying, lowpass filtering at 28 Hz, normalising by long-term average and finally extracting the log power of the wanted low frequency component, calculated by FFT. In this first attempt the 4 Hz component was chosen. The used window size was 250 ms. The standard deviation of each amplitude was calculated using 20 overlapping windows with a frame interval of 12.5 ms, giving a total size of 500 ms.

3.2. 4 Hz normalised amplitude

The 4 Hz normalised feature as described by Scheirer & Slaney [4] was also calculated for comparison. It was, however, not extracted in the same way. Since the LF components in each band were already available, the ratio between the 4 Hz component and the sum of all components from 1-14 Hz were calculated and added over the 20 critical bands. A smoothing procedure was finally performed. This gives just one single value for each frame, a one-dimensional vector. In this case a window size of 1 second was used.

3.3. 2-4 Hz normalised amplitude

In order to increase the bandwidth of the LF component, the 2, 3 and 4 Hz components were added and divided by the sum of all as before. This gives us also a one-dimensional feature vector, also on a 1 second window.

4. Models and data

4.1. Models

GMM’s (Gaussian Mixture Models) were trained as suggested by Reynolds and Rose [12] with 16 and 32 component mixtures. Vector quantization models with 16, 32 and 64 code words were also trained using the generalised Lloyd algorithm [13].

These models are referred to as GMM-16, VQ-16 etc. When testing only the LF features, even lower dimensions of these models were used.

4.2. Data

Clean speech data from the Waxholm data base [14] containing 68 speakers (51 male and 17 female) were used as training and EER data. For model training 49 speakers (12 female and 37 male) were used. The remaining 19 speakers were used for equal error rate (EER) adjustment, see below. The total training data length was approximately 17 minutes and the EER data length approximately 8 minutes.

As music training data, clean music from CD recordings of different kinds was used (pop, rock, jazz, country, world music, classical music and with many different instruments). Each one contained 15 –25 seconds of instrumental music, with no singing. The 53 segments for the training session and
23 for EER calculation corresponded to approximately 16 and 7 minutes respectively.

The test data were collected from the Swedish broadcast using a standard FM receiver. The 48 segments of speech with equal distribution between male and female speakers and 48 segments of music were collected during January and February 2001. The speech contained a variety of speaking styles and the music represented different styles such as pop, rock, country classical music etc. The test data base contained approximately 15 minutes each of speech and music.

All data were sampled at 16 kHz with 16 bits in mono.

5. Results from tests discriminating between speech and music

In order to find out how well the LF features act as discriminators between speech and music, some comparative tests on our data base were performed. In the first phase only the 3 LF features were compared. In the second phase the best LF feature was compared with conventional cepstrum features and a new mixed feature.

Mel frequency cepstrum coefficients, MFCC, were calculated with a 32 ms Hamming window using 39 (3 x 13) and 78 (3x 26) coefficients. This includes the delta and delta-delta coefficients calculated with linear regression over 100 ms segments. They will be referred to as 39-MFCC and 78-MFCC respectively.

5.1. Scoring and decision

As a scoring meter the log likelihood for GMM's and the distortion for the closest symbol in the VQ model were used, both adjusted for EER [15]. A weighting factor for VQ models and a threshold for the GMM's were calculated to get Equal Error Rate for speech and music on this 'EER data base'. Note that this EER is only calculated and valid on a frame-by-frame basis. Using other sizes of the decision window makes the result diverge.

When the decision window was set to 1 second the results from all frames within 1 second were added and a majority decision was performed. This means that, for example, the 4 Hz normalised feature only contained one value while the 4 Hz ASD feature contained 42 vectors of dimension 40 and the 39-MFCC contained 62 39-dimensional feature vectors.

5.2. LF results

The results from the LF comparing tests are shown in Figs 3 and 4. The new 4 Hz ASD feature is found to give the best results both for EER data and test data.

It is important to keep in mind that the comparison is performed between features of different size, a 40-dimensional vector and a scalar. Therefore even lower dimensions of the VQ models were tested.

5.3. Comparing MFCC and LF

The 4 Hz ASD was used in a comparison test with MFCC features. 39-MFCC and 4 Hz ASD has approximately the same vector dimension (39 and 40 respectively). The sampling windows were synchronised so that the MFCC's were calculated in the centre of the 4 Hz ASD windows since the latter is larger (232 and 500 ms respectively).

Figure 3. Percent correct classified segments on average for speech and music over 1 second for 3 different LF component features from EER data. The best yield for the 4 Hz ASD with VQ-64 is 96.8%.

Figure 4. Percent correct classified segments on average for speech and music over 1 second for 3 different LF component features for test data. The best yield for the 4 Hz ASD with VQ-8 is 86.4%.

The discrimination ability for, on one hand, an increased number of cepstrum parameters and, on the other, the addition of the 4 Hz ASD feature were also compared to see if it would give a better yield. In this case 78-MFCC and 39-MFCC + 40 4 Hz ASD parameters were used.

Figure 5. Percent correct classified frames for EER data. The best yield for 78-MFCC with GMM-32 is 96.6%.

The results are shown in Figs 5-7 both as EER and the average error rate between speech and music tests for test data. When going from training and EER data to test data, the error rate on speech data increases while the music error rate decreases, showing a tendency in which the sound becomes more 'music-like'. This could be explained by the fact that music is a 'wider' class than speech, so when getting test data outside the training corpus they would more easily fit the music class.

For a 2.5 second decision window on test data the mixed feature gave the best result with 93.6 %, almost the same as the 4 Hz ASD with 93.2 % and 78-MFCC with 93.1 %.
Figure 6. Percent correct classified frames for test data. The best yield for the mixed feature with GMM-32 is 84.5%.

Figure 7. Percent correct classified segments over 1 second for test data. The best yield for the mixed feature with GMM-32 is 90.9%.

6. Discussion and concluding remarks

The results of this report show that this new 4 Hz ASD feature contains more information than the 4 Hz normalised and the 2-4 Hz normalised features. It also performs better than MFCC for VQ models, but keep in mind that the delta and delta-delta features are hardly used in the VQ models since they are so small compared to the coefficients, while the 4 Hz ASD components all happen to have the same order of magnitude.

It can also be seen that the 4 Hz ASD is rather model independent while the result from MFCC increases with larger models.

It is interesting that the 4 Hz ASD also seems more independent of the sound environment. Since our test data were of a totally different quality than our training data, the MFCC did much worse on test data than on EER data, while the results from the 4 Hz ASD did not decrease that much. Using a mean normalisation method or RASTA processing together with the MFCC calculation would probably increase these results.

The mixed feature is even less sensitive to model type, taking advantage of both the 4 Hz ASD for VQ model and MFCC for GMM.

We believe that there are other and probably better ways to extract a LF feature for discriminating between speech and music or other ways of combining cepstral methods and LF methods, and we shall continue in our search for these.

7. Acknowledgements

The author would like to thank Mats Blomberg and Rolf Carlson for their help and expertise during this study. He would also like to thank Peter Nordqvist in particular for making the Matlab® version of the VQ algorithm available. and Carl Welin for GMM-routines in the C language.

8. References