Spectro-Temporal Properties of the Acoustic Speech Signal used for Speech/Music Discrimination

Stefan Karnebäck

Licentiate Thesis
Stockholm 2004
Akademisk avhandling som med tillstånd av Kungliga Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologielicentiatexamen torsdagen den 4 november 2004 kl. 15.00 i Seminarsalen Fantum vid Institutionen för Tal, Musik och Hörsel, Kungliga Tekniska högskolan, Lindstedtstvägen 24, Stockholm.

ISBN 91-7283-872-8
TRITA TMH 2004:5
ISSN 1104-5787

© Stefan Karnebäck, oktober 2004
Abstract

This thesis presents work that contributes to the speech/music discrimination (SMD) task. Two new methods of discriminating between speech and music are presented. Both are inspired by knowledge of the acoustic speech signal. They originate from the spectro-temporal properties of the acoustic speech signal in two different time intervals.

The first aspect explored and used is the low frequency modulation of the speech, with a peak around 4 Hz. A new feature, 4 Hz ASD (Amplitude and Standard Deviation) is presented and evaluated in the SMD task. The feature reflects the modulation in separate Bark scaled filter banks and takes the dimension of 40. Preparatory experiments showed that the correlation in the 4 Hz Amplitude between the different frequency bands for speech and instrumental music respectively differ. This is explained by the fact that a speech signal is modulated synchronously since it only has one sound source, while music contains several sound sources. The switching between voiced and unvoiced speech, typically high frequency fricatives, shows a negative correlation. The 4 Hz ASD is less model sensitive than Mel Frequency scaled Cepstral Coefficients (MFCC), and together with MFCC, the SMD result is improved. This feature is generalized and the Low Frequency Modulation Amplitude and Deviation (LFMAD) feature is presented and investigated. The best result was achieved when exactly one period of the explored modulation frequency was used as the analysis window size. LFMAD is also explored in noisy environments and is found to be less sensitive to noisy channels than MFCC.

The second aspect is the regular switching between quasi-stationary states where the most significant one is between voiced and unvoiced speech. Hidden Markov Models (HMMs) are good representatives of this switching. Several features and feature combinations were extracted and evaluated in SMD experiments. The results show that an HMM-based SMD system performs very well. Comparative studies were performed varying mainly the number of symbols and the number of states in the models. Investigations presented suggest that there might be an optimal size of the HMM in SMD tasks.
Acknowledgement

There are several persons I would like to thank for their contributions and help to the progress of this work. First of all my mentor Rolf Carlson for supporting me and showing patience with my sometimes very split situation. My assistant advisor Mats Blomberg deserves all my thanks for patient reading, valuable ideas, remarks and fruitful discussions.

I would also like to thank all the other colleagues at TMH for creating a warm, friendly and inspiring atmosphere at the institution.

Sheri Hunnicutt deserves a special thanks for her thorough and patient reading in order to make this thesis understandable in the English language.

Ingrid, Fríða, Oskar and Josef – please forgive me for being absent-minded sometimes and not giving the distinct answers you deserved when addressing me.

This work was partly carried out at the Center for Speech Technology (CTT), supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating partners, of which Flextronics International, Sweden, specifically supported a part of my work. Special support was also given by the School of Engineering at KTH, campus Visby.

Stefan Karnebäck, October 2004
Contents
1 Acronym List, AL ........................................................................................................... 4
2 Included papers ............................................................................................................. 5
3 Introduction .................................................................................................................. 6
4 Segmentation and classification in audio data retrieval systems .................................. 8
5 Information coding of human speech in the acoustic domain ..................................... 12
6 Speech/Music discrimination systems ......................................................................... 17
6.1 Background and application .................................................................................... 17
6.2 The Speech/Music Discrimination task .................................................................... 17
7 Features generally used for Speech/Music Discrimination ......................................... 19
7.1 Short time features or primary acoustic features ..................................................... 20
7.2 Dynamic features or secondary features .................................................................. 20
8 Model and classifier .................................................................................................... 23
8.1 Databases ............................................................................................................... 25
9 Contribution ................................................................................................................. 26
10 Summary of papers .................................................................................................... 27
10.1 Paper 1. Discrimination between Speech and Music based on a Low Frequency Modulation Feature ......................................................................................... 27
10.2 Paper 2. Expanded Examinations of a Low Frequency Modulation Feature for Speech/Music Discrimination .................................................................................... 30
11 Future work ............................................................................................................... 33
12 References ................................................................................................................. 34
13 Appendix, 4 Hz ASD feature, plot examples ............................................................ 37
13.1 Example 1. Speech signal ....................................................................................... 38
13.2 Example 2. Music, Piano concerto ......................................................................... 39
13.3 Example 3. Music, Violin concerto ......................................................................... 40
13.4 Example 4. Techno music ...................................................................................... 41
13.5 Example 5. Rhythmic speech ................................................................................ 42
# 1 Acronym List, AL

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Hz ASD</td>
<td>4 Hz Amplitude and Standard Deviation</td>
</tr>
<tr>
<td>ACF</td>
<td>AutoCorrelation Function</td>
</tr>
<tr>
<td>ADP</td>
<td>Automatic Data Processing (=ADB in Swedish)</td>
</tr>
<tr>
<td>AM</td>
<td>Amplitude Modulation</td>
</tr>
<tr>
<td>AMDF</td>
<td>Average Magnitude Difference Function</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neuron Network</td>
</tr>
<tr>
<td>BCF</td>
<td>Block cepstrum Flux (Average value of Cepstrum Flux)</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transform</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>FM</td>
<td>Frequency Modulation</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>LF</td>
<td>Low Frequency (in speech processing approximately in the range of 1-20 Hz)</td>
</tr>
<tr>
<td>LFCC</td>
<td>Linear Frequency scaled Cepstral Coefficients</td>
</tr>
<tr>
<td>LFMAD</td>
<td>Low Frequency Modulation Amplitude and Deviation</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear Predictive Coding (or Coefficients)</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel Frequency scaled Cepstral Coefficients</td>
</tr>
<tr>
<td>RASTA</td>
<td>RelAtive SpecTrA processing method</td>
</tr>
<tr>
<td></td>
<td>(A filter processing method where LF modulation frequencies in each spectral band are band pass filtered, to perform channel equalisation.)</td>
</tr>
<tr>
<td>SMD</td>
<td>Speech/Music Discrimination (or Discriminator)</td>
</tr>
<tr>
<td>VQ</td>
<td>Vector Quantisation</td>
</tr>
<tr>
<td>ZCR</td>
<td>Zero Crossing Rate</td>
</tr>
</tbody>
</table>
2 Included papers


3 Introduction

Information Technology (IT) was established as a concept in the early 1990's. IT was preceded by the concept ADP (Automatic Data Processing, ADB in Swedish). This older term described a single computer, processing incoming data and producing results as an output stream. IT as a concept, comprises more distributed data processing, i.e. technology for supplying information. Focus was moved from the processing itself, where it took place and in what way, to the information retrieval and distribution.

Nowadays the information flow is larger than ever (the Internet is one major reason). For a recipient, the necessary information processing consists of rejecting, accepting and searching for information. On the World Wide Web (WWW), several search engines help to retrieve the desired information. In structured databases, special search tools or query languages are available.

For use in audio databases or archives, data retrieval tools are also being developed, though they are more difficult to design than their text-based equivalents. If the audio is non-speech sound, there is no natural way to organize data, since there is no alphabetical sort order. Much research has been carried out and more is ongoing, to index music and other sounds in order to enable retrieval of a specific melody or sound type (Feiten and Güntzel, 1994; Pfeiffer et al., 1996; Wold et al., 1996; Foote, 1997; Zhang and Kuo, 1999; Allegro et al., 2001; Guo and Li, 2003). Similar work focuses on scene analysis. For example Liu et al. (1997) use background sound for video scene classification. Algorithms for environment classification in hearing aids are proposed by Nordqvist & Leijon (2002).

If the audio is speech from different speakers or a mix of sounds where the speech is searched for and transcription or indexing should be accomplished, a segmentation and classification procedure is performed first (Hain and Woodland, 1998; Chen and Gopalakrishnan, 1998; Woodland et al., 1998; Lu et al., 2001; Zhou and Hansen, 2002; Ajmera et al., 2003; Panagiotakis and Tziritas, 2004). This procedure labels the audio with different classes, such as pure speech, a specific speaker, telephone speech, music, speech on music, noise, etc. These segments will be processed for either a complete transcription or just a keyword index. When this is accomplished, the search can take place as in any database. The segmentation process scans the audio stream for boundaries between different sound types or sources. Each segment is then classified and a clustering procedure might follow. The segmentation and classification are performed in order to avoid transcription of sound that is not speech which would produce peculiar results. If the two dominant sound sources are speech and music, the
classification task is a speech/music discrimination (SMD) task. The SMD itself can also be a part of the segmentation process. As a convention, the term “classification” is used when more than two sound classes are present and “discrimination” when only two sound sources exist.

To reduce problems with transcription of non-speech segments, a good segmentation system is of great importance. A good SMD system can, as mentioned, be a part of such a segmentation. Some of the work mentioned above describe SMD systems (Aarts and Dekkers, 1999; Ajmera et al., 2003; Panagiotakis and Tziritas, 2004; Pinquier et al., 2002a and 2002b; Saunders, 1996; Scheirer and Slaney, 1997) or segmentation and classification systems (Hain and Woodland, 1998; Montacié and Caraty, 1998; Samouelian et al., 1998; Siegler et al., 1997).

There are still unexamined features and models that could help to improve the SMD task. The papers included in this thesis focus on some dynamic behaviors in the speech signal to be used in SMD systems.

The next section describes how an SMD system can be implemented in the segmentation process. Sections 5-7 describe the origin and motive of the ideas behind the presented papers. Section 8 discusses the model and classifier issue, which explains the decision thresholds used in the papers. Following these discussions, Section 9 clarifies the contribution that the included papers give to the SMD task.
4 Segmentation and classification in audio data retrieval systems

One of the applications for SMD systems is in the context of audio data retrieval, normally as one of several components. For extraction of speech or music in a sound file or database, segmentation is normally performed first.

In Figure 1 is shown a general example of a three-step segmentation, classification and merging procedure. The example is constructed by concatenating several sound signals for pedagogical purposes. The labeling is performed by hand in this example. The software Wavesurfer (Sjölander & Beskow, 2000) is used for presentation. The signals are first segmented into acoustically similar segments. Thus, the acoustic change is also used in the segmentation procedure. The segmentation can be performed by a distance measure between the left and right side of a time instant (a border candidate). A border is found at a local maximum. The measure can be a relative cross entropy (Sieglar et al., 1997), a Bayesian Information Criterion (BIC) (Chen and Gopalakrishnan, 1998), a combination of BIC and $T^2$-statistics (Zhou and Hansen, 2000) or entropy of probability dynamism (Ajmera et al., 2003), for example. The size of these segments depend on the algorithm and the knowledge of the content. After segmentation, the signals are assigned one of five classes; Speech (S), Music (M), Silence (Z), Noise (N) and Others (O). Other classes could be speech on music, telephone speech, etc. The merging procedure is performed in order to achieve consecutive Speech, Music and Others segments where smaller parts of Silence is embedded. Others also includes the Noise segments. The result of the merging at 4 seconds is one Speech segment, where the Silence is embedded. The Silence, Others, Silence, Noise and Silence segments between 5.3 and 14.4 seconds are clustered into one segment named Others. A similar rule-based post-filtering procedure is used by Chou and Gu (2001).
Figure 1. A pedagogical example of the segmentation and classification procedures. The example contains concatenated segments with sound sources at the top. Results are shown after segmentation (top), classification (middle) and merging (bottom), see text.

If only two sound sources, Speech and Music, are involved, the segmentation and classification can use the Speech/Music discriminator. Assuming that every Speech or Music segment should be longer than 2.5 seconds, the segmentation, classification and clustering procedure can be designed as described in the next section. This example is also constructed for pedagogical purposes by concatenating speech and music files surrounded by silence.
A Segmentation and classification

A first classification is carried out using the SMD system in short segments. The segment length should be shorter than 2.5 seconds, and in this example 0.5 seconds is used. Depending on the type of SMD system, either a shorter or longer segment size has its advantages. Factors to consider are time consumption depending on the algorithm used, accuracy and appropriate decision, and analysis window sizes. The segments could also be overlapping followed by a smoothing procedure. Either two classes, Speech and Music, are used, or three classes, adding U for Unclassified segments. The U segments have score values on the border between Speech and Music.

A separate model for Silence or Others can also be useful, classifying all these segments as U until later. See Figure 2 for the result.

Figure 2. Result after segmentation and classification into three classes of a mixed speech/music file.

B Merging step 1

All consecutive segments identically classified are merged (Pinquier et al., 2002b).

C Merging step 2

Segments shorter than 2.5 seconds are merged and embedded into surrounding segments, see Figure 3.
D Refinement

A finer segmentation limit can be defined within border segments, either in a U segment or by adjusting the border between S and M (see segments at 4-4.5 seconds and 10.5-11 seconds). An alternative could be to include the U segments with the S segments if Speech is to be detected, or to include them in M segments if Music is to be detected, to avoid false rejection.

Once the segmentation is accomplished, a clustering procedure could be applied to identify several appearances of the same speaker or the same gender, if desired. This enables the use of separate speaker models for transcription.

The techniques described here assume an off-line system, partly because the signal is processed in several stages and partly due to reasons concerning processor time consumption. However, processors are constantly becoming faster and more efficient. Tritescler & Gopinath (1999) have developed a segmentation, clustering and speaker identification system using an improved BIC measure that works in real time.
5 Information coding of human speech in the acoustic domain

Before deciding on ways to discriminate between speech and music, their acoustic signals should be examined. It is very common that scientists in the speech community concentrate on knowledge of the speech and either consider music as different or perform a brief investigation of the music signal. This section sheds light on some aspects of the speech signal and its production.

Humans communicate in several ways. Speech is a very common way in a face-to-face situation and, of course, when using telephones. Human speech carries information mainly formulated in utterances, containing words, consisting of phonemes. This information is transmitted in the air: the speech production apparatus is the transmitter and the ears and the rest of the auditory system constitute the receiver. The information can be considered as acoustically modulating a carrier. The carrier has its origin in the air stream created by the alveolar pressure. It is produced either by the vocal folds, vibrating in the air stream, or by any constriction in the vocal tract. In the former case the fundamental frequency and its harmonics can be considered as the carrier while in the latter, a broad-band noisy (not white) signal constitutes the carrier. Wherever the carrier may be created, however, the result when speech is produced is a modulated acoustic signal, where the information is embedded in the modulation of the carrier.

In analogue radio transmission, the message signal modulates a carrier of higher frequency than the frequency content of the message signal. Either the frequency (FM) or the amplitude (AM) of the carrier is modulated in the transmitter. The signal can be denoted

\[ s(t) = A \cdot \cos(2\pi ft) \]  

where either \( A \) or \( f \) is time variant for amplitude or frequency modulation.

The acoustic speech signal exhibits both amplitude and frequency modulation, but in a more complex combination. The frequency modulation (FM) can be observed in a varying fundamental frequency, F0, or pitch. Tones, in some languages, are used to determine the identity of a word. Tones are different pitch patterns used within a syllable. In most European languages, however, pitch is used by the speaker for phonetic highlighting alone, such as for stress and accent, often together with a prolonged duration. The amplitude modulation (AM) in the speech signal follows the syllable rhythm and is produced by the speech production organs: the lips, jaw and tongue. It can be considered as operating on a
multi-channel transmitter. All these carriers are created by the vocal folds, and the shape of their waveform defines the spectrum of the carriers (Fant, 1979; Stevens, 1998). In unvoiced speech, however, the carrier energy is more continuously distributed.

![Diagram of vocal tract and carriers](image)

**Figure 4. The production of low frequency modulation in voiced speech, see text.**

The production of low frequency (LF) modulation in voiced speech is shown in Figure 4. The example shows production of a vowel, where no constriction is present in the supraglottal area. The carrier is thus created by the vocal folds and can be described as the sum of the pitch and a number of its harmonics. The carrier signal can be denoted (in the time domain) as

\[
c(t) = \sum_{n=0}^{N} c_n \cdot \cos(n \cdot \omega_0 t + \varphi_n)^{1,2}
\]

1 c is used for carrier, since modulation theory is referred to, while in source-filter theory, \( S \) is used for the source signal.

2 The bias of the acoustic signal, \( \zeta_0 \), is not desired for speech analysis. If this bias is present in the microphone signal, it will normally be removed early in the signal processing.
The fundamental frequency, F0, is related to $\omega_h$,

$$F0 = \frac{\omega_h}{2\pi}$$

$\varphi$ is often disregarded in speech analysis.

The output spectrum of the speech signal is the product in the frequency domain of the carrier, $C(f)$, and the transfer function, $T(f)$.

$$A(f) = C(f) \cdot T(f)$$  (3)

where $C(f)$ is the Fourier transform of $c(t)$. The transfer function $T(f)$ has a distribution over frequency depending on the shape of the vocal tract and nasal cavity, position of tongue, lips and jaw and position of the velum. The resulting power density spectrum of the signal can be seen in Figure 4. The power is normally presented on a log scale and thus

$$|P| = 20 \cdot \log(\text{abs}(A))$$  (4)

is expressed in dB. The resulting power of the acoustic signal has a frequency distribution and can be denoted

$$P = \sum_{i=1}^{N_f} AM(f_i)$$  (5)

where $P$ is used instead of $|P|$ for simplicity. Frequencies are summed over intervals. $AM(f_i)$ is the average power in the i-th frequency band and derives from the product of $C$ and $T$

$$AM(f_i) = 20 \cdot \log(\text{abs}((CT)_i))$$  (6)
where \((CT)_i\) is the sum of the products of \(C\) and \(T\) in the \(i\)th frequency interval. It should be noted that (5) uses a set of frequencies, or rather frequency intervals, which do not correspond to the harmonics of the fundamental frequency, \(F_0\). In practice, the frequency distribution of the carrier is also time variant, since it is produced either by the vocal folds or some other constriction in the vocal tract. Thus, the energy is more or less continuously distributed (harmonics, however, can be observed in voiced speech) over the frequency interval, and can be observed in frequency bands rather than in discrete frequency components. This is especially valid for unvoiced speech.

Henceforth, we leave the specific voiced speech, and observe a general spectrum of the speech signal. The transfer function, \(T(\beta)\), varies when the lips, jaw and tongue are moving. The carrier distribution also varies for several reasons: the glottal pulse shape varies, constrictions in the air flow introduce other sound sources and the vocal folds are open in unvoiced speech. Thus, \(AM\) is time variant. This variation is the LF amplitude modulation, often called only LF modulation. The major contribution to the LF modulation originates from consonant-vowel (CV) switching, which can be observed in the wave signal. Its frequency distribution depends on the shape of and movements in the vocal tract. The magnitude of the acoustic power can thus be written

\[ P(t) = \sum_{i=1}^{N_i} AM_i(t) \]  

(7)

where \(AM_i(t)=AM(f_{st},t)\) has a two-dimensional distribution over time and frequency, which can be denoted

\[ AM_i(t) = \sum_{j=0}^{N_{low}} AM_{\beta} \cdot \cos(2\pi f_{\beta} t) \]  

(8)

where \(N_{low}\) denotes the highest low frequency band with a magnitude above 0. Note that the index \(j\) is used (in 8) for summation over modulation frequencies, starting at DC, and the index \(i\) is used (in 7 and 8), for carrier frequency bands. Using (8), equation (7) can be written

\[ P(t) = \sum_{i=1}^{N_i} \sum_{j=0}^{N_{low}} AM_{\beta} \cdot \cos(2\pi f_{\beta} t) \]  

(9)

In natural fluent speech, the spectral shape is constantly changing. However there are relatively stationary periods (ca. 50-150 ms), separated by Auditory Events or Avents (Morgan et al. 1995). In these periods each phone has its own characteristic spectral shape, coloured by the context. Besides the trill, tap and stop consonants, the duration of a phone can be prolonged, giving a stationary
shape of the spectrum, making $AM_j > 0$ only for $j=0$. We consider those states as quasi-stationary. This also means that the matrix $AM_j$ is not constant but varies over time.

Pitch, intensity and duration are known as the prosodic features. The prosody reflects mainly semantic information, either concerning the utterance or concerning the speaker's mental state.

The switching between the above mentioned quasi-stationary states, emphasized by the prosody, creates the characteristic shape of the acoustic speech signal. When exploring the wave signal, the amplitude modulation can be clearly observed. The main modulation rate is the same as the syllable rate and is normally in the range of 4-5 Hz.

Speech recognition and speaker recognition or verification systems mostly detect and demodulate the amplitude modulated information. This modulation accomplishes a temporal variation of the spectral shape. The shape itself is typically described by the cepstrum coefficients or by the use of filter banks, from which the temporal behavior is extracted using several different methods.
6 Speech/Music discrimination systems

6.1 Background and application
As already mentioned, SMD systems can be used for several different purposes and real-time and off-line systems may use different techniques. Saunders (1996) presented a paper in which the objective was to select the music in an FM radio receiver and avoid the speech, which contained commercials. Since a real-time system was desired, time domain features were used. In radio broadcasting, where the speech and music signals are processed in different manners, a real-time SMD system would be useful. In addition audio coding can be improved by decreasing the bit-rate for silence segments, if a silence detector is included. The papers presented in this work focus on off-line situations.

6.2 The Speech/Music Discrimination task
When designing an SMD system there are several components or algorithms to investigate and define. A generic SMD test system is presented in Figure 5. This test system presumes a preceding training phase to define the models.

![Speech/Music discrimination test](image)

*Figure 5. A generic block diagram of an SMD test system.*
Also technical data such as analysis frame size, bandwidth of the incoming signal, sample rate and frame rate have to be established. For evaluating speaker and speech verification systems, several dedicated databases for comparative studies are available. However, this is not the case for SMD evaluations. Work is going on to establish databases for segmentation and classification purposes (Cost 278 BN), but not for SMD tasks specifically, so each researcher has to design his own.

When designing an SMD system, either the speech signal, music signal or both can be explored and its characteristics chosen to constitute the basis for the system. If no aspect of either the speech or the music signal is taken into account, a general learning system can be designed based on statistical methods only. A majority of the attempts in the speech community explore the characteristics of the speech signal and more or less assume that music is of another character. Specific characteristics are searched for primarily in the speech signal and then compared with the music signal. However, some researchers have investigated both the music and speech signal before designing the SMD system, for example Saunders (1996) and Scheirer & Slaney (1997). A combination of knowledge-based and statistical methods is generally applied.
7 Features generally used for Speech/Music Discrimination

The features used for SMD tasks can be classified into mainly two categories, either as time-domain vs frequency-domain (spectral) features or as stationary vs dynamic features. The extraction of features from the time-domain is often chosen due to a requirement for short execution time. Stationary features reflect a quasi-stationary spectral shape, and the dynamic features, a temporal behavior in this spectral shape. The dynamic features are often extracted from the stationary features in some way and thus are secondary features while the stationary features are considered as primary.

Another difference among systems is the decision window size. A question one needs to ask is whether it is possible or necessary to make a decision on very short segments. Since the speech and music signals mainly share the same frequency range and partly show the same spectral distribution, experiments using only short decision windows will get error rates larger than those with longer decision windows. The longer segments are, the more information they provide and the more likely one is to make a correct decision. This longer segment could be used either to make statistical calculations to improve the likelihood, or to derive secondary features. However, a remark is in order concerning the decision window size: even if the decision is made with a certain rate, giving a virtual window size as the inverse of the rate, the actual analysis window size is larger than that, using frames outside this window.

One of the most characteristic properties of human speech is the regular switching between voiced and unvoiced segments. This property affects the energy contour as well as the temporal variation of the spectrum. The same temporal variation is not normally found in the music signal. Most of the features extracted for SMD purposes reflect this switching in some way.

The frequency range of the pitch in voiced speech is only about three octaves (much less for one single speaker), while the corresponding range for music can be up to six octaves (Saunders, 1996).

The fact that speech normally originates from a single sound source (assuming a single speaker) whereas music usually derives from a combination of several sources, can also be used when extracting features.

Two different analysis window sizes can generally be defined, besides the time interval of one frame, 5-20 ms, which can be called stationary. The short time variation of 50-250 ms where phoneme transitions and vibrato in both voice and music, (the voice can a singer or a speaker) are found. This interval also includes what is called quasi-stationary segments of ca. 50-150 ms. The larger interval,
ranging from 150 up to 1000 ms, is motivated by the syllable duration in speech and the rhythm in music. Speech has an average amplitude peak rate at four to five Hz, normally. However, it may vary a great deal depending on speaker and situation (Caroline et al., 1993; Greenberg, 1995). Harb et al. (2001) found values of 5-10 energy drops per second below a silence threshold, giving time intervals of 100-200 ms. The amplitude peak rate in music varies more.

### 7.1 Short time features or primary acoustic features

Short-time features are supposed to be stationary or at least quasi-stationary, meaning that the spectral content of the signal is almost constant for a short time period. To derive the spectral envelope, a Fast Fourier Transformation (FFT) of the acoustic time signal is often applied. The shape is normally represented by a Discrete Cosine Transform (DCT) of the log magnitude spectrum, giving the cepstrum coefficients. The frequency scale is often transformed according to the Mel scale before DCT calculation, giving the Mel scaled Frequency Cepstral Coefficients (MFCC). They reflect some properties of the auditory system, with higher resolution in the lower frequency bands. The Linear scaled Frequency Cepstral Coefficients (LFCC) are also used. The use of filter banks is an alternative description of the spectrum, giving the energy in separate frequency bands (Carey et al., 1999).

Other ways to describe the spectrum are the gravity and spectral roll-off-point (95% energy) (Scheirer and Slaney, 1997).

Zero-Crossing Rate (ZCR) is a time-domain feature that captures the dominant frequency (Kedem, 1986), used by Saunders (1996) and Scheirer and Slaney (1997). Combined with filtering, it can give a brief description of the spectrum shape, with the cost of calculation resources. Since a reason for using the time domain is to avoid too much calculation and reduce execution time, too many filters may reduce the time gain.

The amplitude itself, or the energy, as a static feature, is not very useful alone. Secondary features need to be calculated.

### 7.2 Dynamic features or secondary features

When extracting secondary features, the temporal variation of some static features is explored. Two approaches can be defined for this purpose, a knowledge-based approach and a mathematical/statistical one. In the first case, the knowledge of specific sounds is used to extract features reflecting specific properties. In the
latter case, a statistical or mathematical computation is performed from a time series of primary features. These methods partly overlap.

The most commonly used secondary features are the first and second order time differentials of the cepstrum coefficients, often named \textit{delta cepstrum}. These features are commonly used in speaker and speech recognition. Other temporal features are \textit{spectral flux} (Scheirer & Slaney, 1997), which is the 2-norm of the spectral vectors of two consecutive frames. \textit{Cepstrum Flux} is the 2-norm of two consecutive cepstral vectors, and the time average of the Cepstrum Flux is called the \textit{Block Cepstrum Flux} (BCF). Takeuchi et al. (2001) refines the BCF in two different ways into \textit{Weighted Block Cepstrum Flux} (WBCF) or \textit{Block Weighted Cepstrum Flux} (BWCF). Linear Predictive Coding (LPC) cepstrum vectors are used in their work. Ezzaidi & Rouat (2002) calculate a distance measure on delta cepstrum, \textit{DAMFCC}, which is 12-dimensional, based on the 2-norm, and could be called "delta cepstrum flux".

Methods using mean and standard deviation calculated on some acoustic features are used by Saunders (1996), Scheirer & Slaney (1997) and the author in Paper 1 in this thesis.

Ratios of different kinds such as percent low-energy frames are used by Scheirer & Slaney (1997). Panagiotakis & Tziritas (2004) use a similarity measure calculated from a $\chi^2$-distribution fit of a histogram over RMS values for speech and music respectively.

A specific feature analyzing the time domain signal has also been used. Aarts & Dekkers (1999) extract speech and music features from a time series which are processed by a fuzzy combiner to discriminate between speech or music. This feature is based on the switching in the acoustic signal. They primarily use the ratio of the energy content in the intervals $\{70 \text{ Hz} < f < 700 \text{ Hz}\}$ and $\{f < 130 \text{ Hz}, f > 1200 \text{ Hz}\}$. For speech signals this ratio should vary rapidly and often, compared with music signals. A difference between consecutive amplitude maxima and minima in the acoustic signal were used by Samouelian et al. (1998). Scheirer & Slaney (1997) use the "percentage of low energy frames" as a feature. A high value indicates a speech signal. Harb et al. (2003) use the Silence Cross Ratio (SCR), which detects and counts the energy dip below a silence threshold, as a speech detector. All these features attempt to catch the switching between voiced and unvoiced speech.

Temporal variations in ZCR are used by Saunders (1996) with the intention of finding speech/music discrimination properties such as tonality, bandwidth distribution, excitation patterns, tonal duration and energy sequence.

Where a fundamental frequency, F0, exists, there is an implication of voiced speech or music, while an absence implies unvoiced speech, silence, pause or
noisy segments. Harmonics are often present together with F0. The ratio of harmonic to unharmonic segments is used as a measure of tonality (Saunders, 1996; Allegro et al., 2001), and a high rate is usually valid for music. F0 can be searched for by ZCR, autocorrelation techniques, average magnitude difference function (AMDF) or other algorithms (Rabiner et al., 1976). AMDF can be used just to detect voiced segments (pitch detection) (Montacié and Caraty, 1998). Carey et al. (1999) also use the delta pitch, which together with the pitch value, discriminates well between speech and song.

Phone recognizers are used for secondary feature extraction, such as probability dynamism and entropy (Ajmera et al., 2003; Williams and Ellis, 1999).

The 4 Hz modulation energy deserves special attention since the syllable rate for human speech normally has a peak value at 4-5 Hz and this value corresponds to the response rate of the neurons in the auditory cortex (Greenberg, 1995). These findings are used in the RASTA filtering technique (Hermansky and Morgan, 1994) for improving speech recognition. Scheirer and Slaney (1997) and Pinquier et al. (2002a) extract one-dimensional 4 Hz energy modulation features in a similar manner. A 4 Hz Modulation Harmonic Coefficient is used by Chou & Gu (2001). It is calculated from a combination of temporal and spectral autocorrelation functions.

Music also has strong components in the 4-5 Hz range. This fact complicates the possibility of using the amplitude modulation in SMD tasks. One of the main differences between speech and music mentioned above, is that speech is produced by one single sound source, while music normally consists of several sound sources. The many combination possibilities in music also makes it an acoustically wider class than speech. This fact can be used when deriving secondary features from 4 Hz features.

The 4 Hz feature was examined and extended by the author and used in two of the papers, described in more detail below.
8 Model and classifier

A crucial part of the discrimination or classification task is to compute decision parameters or classifiers, using the extracted features, and to apply a decision algorithm, as described in Figure 5. This procedure can be straightforward, a binary tree model or a recursive algorithm when used in segmentation tasks. Gaussian Mixture Models (GMM) or Vector Quantifiers (VQ) are commonly used. Artificial Neural Networks (ANN) and Hidden Markov Models (HMM) are used separately or in combination with each other.

A classification, or in a 2-class situation, a discrimination, is to be done for a certain window size, depending on the application. The longer time for a decision, the higher probability of making a correct decision, in general. Decisions can be made frame by frame, meaning approximately every 5-20 ms, followed by a majority decision using a cost function; or a classifier can be calculated over a larger window size. A cost function can also be implemented as (Duda & Hart, 1973)

\[ C = C_{FCM} \cdot P_{M} \cdot P(\hat{S} | M) + C_{FCS} \cdot P_{S} \cdot P(\hat{M} | S) \]  

(10)

where \( C_{FCM} \) (Cost for Falsely Classified Music segments) and \( C_{FCS} \) (Cost for Falsely Classified Speech segments) are the costs of classifying a signal falsely as speech or music, respectively. A zero cost is assigned to correct classification. The probabilities \( p_{M} \) and \( p_{S} \) are the a priori probabilities and \( P(\hat{S} | M) \) and \( P(\hat{M} | S) \) are the probabilities that the signals will be falsely classified. The decision should be made so as to minimize the cost function in Eq. (10). For this purpose the likelihood ratio, LR, is used

\[ LR = \frac{P(x | M)}{P(x | S)} \]  

(11)

where \( P(x | M) \) and \( P(x | S) \) are the probabilities that the signal, \( x \), is produced by a Music or Speech source respectively. This likelihood ratio should be compared with a threshold for decision. The threshold that minimizes Eq. (10) is

\[ \Theta = \frac{C_{FCS} \cdot P_{S}}{C_{FCM} \cdot P_{M}} \]  

(12)
and the decision algorithm will be

\[
\frac{P(x|M)}{P(x|S)} \quad \begin{cases} \text{Music} \\ \text{Speech} \end{cases} \quad \Theta \quad (13)
\]

To evaluate the system, separate training and test databases are normally used. The less similar they are, the better the test of the robustness can be performed. A separate development database can be used to define the threshold for the decision algorithm. The threshold can, for example, be defined to achieve an equal error rate (EER) on speech and music data. Other values can be chosen depending on the application, often calculated from the cost function in Eq. (10) including a priori probabilities.

Using log likelihood in (11), the decision rule will be

\[
\begin{align*}
\log P(x|M) - \log P(x|S) &> \eta \\
\log P(x|M) - \log P(x|S) &< \eta
\end{align*}
\quad (14)
\]

where \( \eta \) is the resulting threshold

\[
\eta = \log C_{FCS} - \log C_{FCM} + \log p_S - \log p_M
\quad (15)
\]

Assuming equal a priori probabilities, the cost functions remain to be decided. With a requirement of missing as few speech segments as possible, \( \eta \) should increase, given by a higher value for \( C_{FCS} \) than for \( C_{FCM} \). A special case is obtained with equal costs and equal a priori probabilities, giving a threshold value of zero: in this case the class with largest a posteriori probability is chosen.

If the probability density functions of the two classes have the same deviations and are symmetrical, but only differ on mean values, this threshold also gives the EER. If they, which is most common, have different standard deviations or different distributions, another method is needed to find the EER value. The EER threshold could then be defined empirically by varying the threshold value and calculating the errors on the speech and music signals respectively. The threshold would then be assigned the value at which speech and music errors are equal. For this purpose, separate training and development databases are normally used. The models are trained on training data and a threshold for EER is
calculated on development data. This threshold is applied on the test database. The less similar the training/development and test databases are, the better the test of the robustness could be performed.

8.1 Databases

A special remark is in order concerning the choice of database. Evaluations conducted on SMD systems suffer from a lack of standardized public databases. Speech and speaker verification systems are provided with databases for this evaluation purpose. Unfortunately no such possibility exists for SMD tasks; thus every researcher has to construct his own database. The difficulty on the SMD task is very much dependent on the type of music included, especially since music is a wider class than speech, meaning that the feature space is larger for music data than for speech data. Vocal music increases the difficulty, and the amount of song and type of music background generally affect the discrimination results.
9 Contribution

As described above, there are already many SMD systems investigated and reported. This thesis mainly contributes with two ideas to improve SMD performance using two different dynamic aspects of the speech signal, the 4 Hz modulation and the more rapid phoneme switching.

The 4 Hz modulation has been used in SMD systems earlier, but only as a few dimensional feature. The investigations reported on in Paper 1 and Paper 2 show that there is more information in the acoustic speech signal that can be used for SMD purposes.

In the many SMD systems reported, no one uses HMM as a phoneme classifier. Paper 3 in this report is the author's contribution to the SMD problem, inspired by the idea of speech as an acoustic state machine.

HMMs are commonly used in speech and speaker recognition systems. Each state in the model is supposed to correspond to a phoneme or a phoneme class and the transitions originate from the switching between phonemes or phoneme classes or small variations within a phoneme or class. These changes follow a rhythm in the speech. If music is modeled, each state would represent the timbre and the transitions in the same way, the rhythm in the music. If only few states, three or four, are used, they represent phoneme classes in the speech. If discrete models are used, even the observation symbols would represent a phoneme or a phoneme class. One suggestion would be that when using around 40 observation symbols, they would represent phonemes, while the states, when only few, represent phoneme classes. This implies that HMM would be a good model of the speech signal and probably also of the music signal. Thus the model can be used in an SMD task. The question of the model size remains. Is a phoneme class model satisfactory, or do we need a more detailed model on a phoneme level? The result from paper 3 indicates preliminarily that there might be an optimal size of the HMM for the SMD purpose on a phoneme level.

The main objective of paper 3 is to verify the hypothesis that HMM is a good model in SMD systems. However, the investigation of good discriminating features has only begun. Traditional LFCC is used together with ZCR, autocorrelation function (ACF), energy and spectral gravity in different combinations. The preliminary investigations indicate that these features add information and improve the discrimination ability.
10 Summary of papers

The papers presented in this thesis are the results of the author's own work. They all address the dynamic aspect of the acoustic speech signal, applied in the SMD task. In the first two papers, the dynamic aspect is captured by the extracted features, and in the last paper, this aspect appears mainly in the used model, i.e. the HMM.

10.1 Paper 1. Discrimination between Speech and Music based on a Low Frequency Modulation Feature.

This paper presents a new dynamic feature, based on the extraction of the amplitude modulation (AM) in the acoustic speech signal. The interpretation of this modulated information is not as straightforward as for the pitch. It originates from the syllables in the speech, and the information is spread out over all frequency bands of the carrier. The syllable rate is normally in the interval of 4-5 Hz. The new feature 4 Hz ASD (Amplitude and Standard Deviation) is defined and some preliminary results presented on the SMD task. This feature uses strictly 250 ms analysis window size. The shortest decision window size was, however, 500 ms since 20 frames were used in the SD calculation.

10.1.1 The Low Frequency Modulation feature, 4 Hz ASD

Figure 6 describes the extraction of the 4 Hz ASD feature. the first, the 4 Hz component of the power spectrum is chosen, assigning j=4 in Eq. (9). In the experiments, time consecutive samples of $AM_i$ is calculated on each frequency band. The Bark scale is used for the frequency distribution. For analysis window $m$, the feature can be denoted

$$4Hz ASD_m = \left\{ AM_i(m), \left\{ \text{Std}(AM_i(i, k)) \right\} \right\}$$

(16)

where $AM_{i,m}$ is the 4 Hz component of critical band $i$ in analysis window $m$. With a frame rate of 80 Hz, the standard deviation was thus calculated over 250 ms, a 500 ms segment being needed for the first calculation.
Introductory investigations showed that the vector elements representing the amplitude modulation 4HzASDm(1:20), Eq. (16), appeared to exhibit different behaviour for speech and music respectively, as well as the standard deviation elements 4HzASDm(21:40). When correlations between amplitude vector elements were investigated in small time intervals, they also indicated that speech and music behave differently.

Summing up the situation for speech and music concerning the 4 Hz ASD feature gives the following:

- variations in amplitude are more evenly distributed over the frequencies in music than in speech.

Figure 6. Extracting the low frequency modulation feature, 4 Hz ASD. The incoming signal is segmented into 250 ms frames, with 80 Hz frame rate. Each frame is band pass filtered into 20 critical bands (Bark scaled), center frequencies ranging from 50 to 5800 Hz. Each BP filtered band is AM demodulated by rectifying, LP filtering, normalizing and Fourier transforming the signal. Finally the log power is extracted. 20 consecutive frames were used to calculate a standard deviation for each filter bank, thus establishing a 40-dimensional vector.
• the variation in amplitude and standard deviation elements are larger in the mid frequency bands than in the lower and upper frequency bands in the speech signal
• speech is more regular than music, regarding correlations
• adjacent frequency bands are more correlated than distant frequencies in a speech signal
• when a single instrument, such as a piano, produces the signal, a strong correlation can be possible between many bands, assuming many tones at a time, or chords are played
• when an orchestra accompanies a solo instrument, the correlations are weak

Thus, besides the 4 Hz modulation itself, both its correlations between frequency bands and the standard deviation in each band showed different behaviours when comparing speech and music signals. However, only the standard deviations together with the amplitude values were used in the experiments. Tests with the correlation would probably also yield good results. However, this will acquire a more thorough investigation of what correlation coefficients to use among the 210 possible for only the amplitude elements.

The suggested interpretation of this supra-temporal behavior is that the speech signal originates from one sound source, thus the temporal changes are synchronized and highly correlated in adjacent frequency bands, while the correlation is decreasing for more separated bands and could even be negative for the highest frequency bands. Music signals mostly consist of a mix of several sources. Although they might show a synchronous behavior, they exhibit a more varied correlation matrix. This can be understood in the frequency domain as several spectral shapes interacting with each other, more or less correlated. Some examples are presented in an appendix.

10.1.2 4 Hz ASD in the SMD task

Results from SMD tests using the new feature, 4 Hz ASD, are presented in this paper. In the initial experiment, this feature was first compared with features similar to the 4 Hz energy modulation presented by Scheirer & Slaney (1997), which was a one-dimensional feature. The 40-dimensional feature out-performed the one-dimensional one. The new feature was then examined and compared with MFCC features. The tests were performed using both VQ and GMM. The new feature proved to be less sensitive to model choice, while the best single performance was achieved with MFCC features. However, a combination of 4 Hz ASD and MFCC was found to yield best performance. The results on a 2.5-
second segment, however, resulted in only 93.6 % correctly classified frames, which is less than Saunders’ (1996) 98 %. Since different databases were used, the results are not quite comparable.


This paper describes further investigations and extensions of the 4 HZ ASD feature, proposed by the author. The feature is more generalized and called LFMAD (Low Frequency Modulation, Amplitude and Deviation). This feature was evaluated in the interval 37.5 - 1000 ms. The paper consists three main parts which will be commented upon.

10.2.1 LFMAD dependency of analysis window size and chosen LF component

There was a very small difference found in performance in the best range between 37.5 and 250 ms, calculated on EER_{test} (Equal Error Rate on test database, explained in more detail in paper 3), assuming that exactly one period of the analyzed frequency is used. It was found in these experiments that the SMD performance decreased almost linearly as a function of the number of periods used in the analysis. This was not an expected result since normal behavior is that an increased analysis window size yields increased performance. One explanation could be that the variations found in the standard deviation will be less when using a larger window – they will be smoothed. Since these deviations were one of the differences between speech and music signals, the performance decreases with a prolonged analysis window size.

10.2.2 Several music classes

It is well known that vocals in music complicates the speech/music discrimination task. In this paper some classification tests on vocal music, instrumental music and speech are carried out. In general, it was found that the SMD performance was reduced by approximately 30% when going from pure instrumental music to vocal music. However, this can only be used as an indication, since the difficulty depends on the number of vocals and the balance between the vocals and the accompanying instruments.

Since no general test database is available, it had to be produced locally.
Two classification tests were carried out with three and four classes. To achieve a more comparable evaluation result than the hit rate, a Kappa (Carletta, 1996) coefficient was calculated. The Kappa coefficient reduces the influence of chance.

10.2.3 Noisy signals

The results by Greenberg (1997) when using RASTA processing technique indicates that LFMAD would perform better than conventional MFCCs when used on noisy environment. SMD tests confirmed this assumption. White Gaussian noise was added to the test signals and tested against the models trained on clean signals. At SNR values between 3 and 10 (less obvious at 20) dB, LFMAD outperforms the MFCCs. At 0 dB the classification becomes guessing, with an error rate of 50%.


This paper describes an SMD system where the dynamic spectral behavior is taken care of by the HMM. The test system uses discrete ergodic HMM. It should be noted that the objective of this work was not to build an SMD system with the best performance possible. The purpose was to verify that the switching between some quasi-stationary acoustic states found in the speech signal does not show the same behavior as for the music signal and that this difference can be observed using HMM. A second purpose was to design a tool for feature evaluation in the future to search those features that best reflect this switching behavior. The focus of this paper was finding an optimal size for the HMM. Both number of states and number of symbols were varied and the performance in an SMD task was measured. These tests indicated that 20-24 states and 48-54 symbols are the best choice. A hypothesis is that the optimal number of states is close to the number of phones in the language and that the optimal number of symbols would be larger, perhaps twice as large. Each state would then be represented by one or more symbols, each symbol representing either a state or a transition between two states.

Good results were achieved with these combinations (20-24 states and 48-54 symbols) and the best results were achieved with the 'All' feature combination. This feature was designed with \(3 \times 13\) LFCC + ZCR (4 kinds) + Acf + Spectral Gravity + Binary energy + Voiced/Unvoiced, giving a 47-dimensional vector. The test results indicate very clearly that this system outperforms the system in the two first papers, where GMM classifiers were used. Results for 1-second analysis
window size yielded between 2% and 3% using HMM and between 9% and 10% using GMM with majority decision over all frames in a 1-second window.

One question arising in the paper is what amount of information is carried in the first and second order time differentials of the cepstrum features. It was difficult to obtain good results with those features. The difficulty might be that the dynamic behavior of the speech signal is already covered in the HMM, so that the delta cepstrum features would not add much information.

The aspect of degree of similarity between training and test databases was also illuminated. With mismatched databases, the models easily became over-trained. The value of the EER threshold and the EER_{test} value were not relevant for the test database, since it's score values showed a different distribution than for the development database. The test results improved when using matched databases, reaching error rates around 1%.
11 Future work

The papers presented in this thesis add knowledge to the topic of Speech/Music Discrimination. Indications that discrete hidden Markov models are useful in this task have been found. The investigations also indicate what sizes to choose for the HMM. The earlier papers present a new feature, LFMAD, that adds information and thus can be combined with the traditional cepstrum features. The last paper shows that ZCR, Acf and Gravity also add information and improve the detection performance together with the cepstrum features. However, more features and feature combinations are still to be investigated, some of which are already presented in similar situations and some not yet proposed. The evaluation tool can be the presented system.

Different model sizes for speech and music respectively is another issue to investigate in the presented system, as tested by Pinquier et al. (2002b).

More training data is needed to verify the implication of optimal model size found in paper 3. Unfortunately there is not yet a generally available speech and music database for evaluation of SMD tasks for comparison, such as can be found for speaker recognition (Petrovska et al. 1998) or speech recognition (Speechdat) evaluation. Such a database would be of great importance.

It is the author's hope that the findings in these papers would be part of a future segmentation-classification-transcription system for audio data retrieval. Before such a system can be designed, still more research must be done, especially concerning a good feature combination and fast and reliable algorithms for segmentation and labeling. Another issue for research would be automatic or semi-automatic adaptation of the models to actual data. Such an algorithm could be designed on confidence measure during classification work.
12 References


COST 278 BN. COST action 278, Spoken Language Interaction in Telecommunication, Special Interest Group on Broadcast News.


Tritschler, A. and Gopintath, R., 1999. Improved Speaker Segmentation and Segments
   Clustering using the Bayesian Information Criterion. Eurospeech, pp. 679-682.
Williams, G. and Ellis, D., 1959. Speech/Music discrimination Based on Posterior Probability
Wold, E., Blum, T., Keislar, D. and Wheaton, J., 1996. Content-Based Classification, Search
Zhang, T. and Kuo, J., 1999. Hierarchical Classification of Audio Data for Archiving and
   the Bayesian Information Criterion. Internat. Conf. Spoken Language Processing, pp 714-717.
   Retrieval System for Historical Audio Archives. Internat. Conf. Spoken Language
13 Appendix, 4 Hz ASD feature, plot examples

Five examples concerning the 4 Hz ASD feature for speech and music respectively are presented below. Each example comprises the spectrum and wave signals, the 4 Hz Amplitude (20-dimensional), the 4 Hz SD (20-dimensional) and the correlation of the 4 Hz amplitudes measured over 3.5 seconds.

The examples represent a connected speech example, a violin concerto, a piano concerto and a piece of techno music. These are all part of the database. Finally a speech signal is presented where the phoneme combination /sa/ is repeated with a frequency very close to four hertz. These examples are primarily presented to display the different patterns of the correlation matrices. The speech signals show a more regular pattern than music signals. Adjacent bands are highly correlated while some of the higher frequency bands have a very low or negative correlation to the lower frequency bands. Since the modulation is the result of movements and resonances in the vocal tract, this effect can be understood. The negative correlation could be interpreted as a switching between vowels and unvoiced fricatives with high frequency content, such as /s/. The center frequencies for critical bands 18 to 20, are 4000 Hz, 4800 Hz and 5800 Hz respectively. This effect is emphasized in the constructed last example.

The patterns of the correlation matrix for music signals vary rather much depending on type of music.
13.1 Example 1. Speech signal

The speech example is a sample from a fast speaker in the Waxholm database used in all papers, FP2001PR06, "Fortkörning är värre än mord, sa konstapel Törnhjort". The average syllable rate is 5.4 Hz.

Figure 7. A speech signal: spectrogram and waveform.

Figure 8. 4 Hz Amplitude and Standard Deviation calculated on the speech file in Figure 7.

Figure 9. Correlations between the 4 Hz Amplitude components shown in Figure 8.
13.2 Example 2. Music, Piano concerto

This music example is a small part of the first movement of Piano Concerto No. 1 by Pjotr Tchaikovsky in B minor. The piano is playing solo in this part.

Figure 10. A music (piano) signal: spectrogram and waveform.

Figure 11. 4 Hz Amplitude and Standard Deviation calculated on the music file in Figure 10.

Figure 12. Correlations between the 4 Hz Amplitude components shown in Figure 11.
13.3 Example 3. Music, Violin concerto

This is a segment of the first movement of the Violin concerto by Felix Mendelssohn-Bartholdy, Op. 64 in E minor. The orchestra is accompanying the solo violin in this part.

Figure 13. A music (violin concerto) signal: spectrogram and waveform.

Figure 14. 4 Hz Amplitude and Standard Deviation calculated on the music file in Figure 13.

Figure 15. Correlations between the 4 Hz Amplitude components shown in Figure 14.
13.4 Example 4. Techno music

This piece of music is called Parallel Universe, played by the group Red Hot Chili Peppers. All frequency bands are very synchronous, resulting in a strong amplitude correlation.

**Figure 16. A music (Techno) signal: spectrogram and waveform.**

**Figure 17. 4 Hz Amplitude and Standard Deviation calculated on the music file in Figure 16.**

**Figure 18. Correlations between the 4 Hz Amplitude components shown in Figure 17.**
13.5 Example 5. Rhythmic speech

The phoneme combination /sa/ is repeated with a frequency very close to 4 Hz.

Figure 19. A speech signal (/sasa/): spectrogram and waveform.

Figure 20. 4 Hz Amplitude and Standard Deviation calculated on the speech file in Figure 19.

Figure 21. Correlations between the 4 Hz Amplitude components shown in Figure 20.
Paper 1
Discrimination between speech and music based on a low frequency modulation feature

Stefan Karnehäck
Dept. of Speech, Music and Hearing, KTH and College of Engineering, KTH-Visby, Sweden
stefans@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/musical 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 Schreit & 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 minute 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.

![Correlation matrices for the 4 Hz amplitude, extracted as described in 3.1, over 20 critical bands for a 3.5 second speech signal (top) and music signal (below). The area of the bubble represents the value. White is a negative value.](image1)

![Codebook for a 4 Hz ASD feature (see 3.1) from VQ-3 for speech (top) and music training data (below) (see 4.2). The first 20 elements are the amplitudes and the last 20 their standard deviation.](image2)

These differences point out that there is probably more information to extract when using all 20 bands and their standard deviation rather than only one summed scalar.

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

GMMs (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 (3 x 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 86.8%](image)

![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%](image)

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%](image)

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%.
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 mode 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

Paper 2
Expanded Examinations of a Low Frequency Modulation Feature for Speech/Music Discrimination

Stefan Karnebäck
Centre for Speech Technology (CTT), KTH and Flextronics International, Sweden
stefan@speech.kth.se and stefan.karneback@se.flextronics.com

Abstract

A low frequency modulation feature, LFMA, was examined under several conditions with regard to its robustness on speech/music discrimination. The feature was tested on LF components from 2 Hz to 27 Hz and with different analysis window sizes. This feature performs best when using an analysis window containing only one period of the LF component to be used. When the music contained much vocals, the error rate increased compared with only instrumental music in the speech/music discrimination task. This effect was found in LFMA as well as in the MFCC feature, which was used for comparison. Tests were also carried out with signals in additive noise from 30 dB to 0 dB SNR. LFMA performed better than MFCC in these tests. The error rate was higher for speech signals. There was a bias towards classifying data as music when the test conditions diverged from those of the training condition. This effect is less obvious for LFMA than for MFCC. The best results in this study were obtained when combining the two features LFMA and MFCC into a mixed feature. This seems to be a more robust feature regarding the speech/music discrimination ability and could be recommended when scanning databases of unknown quality for speech events.

1. Background and report layout

The speech/music discrimination task has been examined by several authors [1, 2, 3, 4] with different approaches. The LF modulation of speech and music respectively show different characteristics. An LF modulation feature can be used to discriminate between speech and music [1]. The used feature, 4 Hz ASD, is now being further examined on several aspects.

The following investigations are presented in this paper:

- Comparison between the different LFMA-n features (Low Frequency Modulation Amplitude and Deviation, n Hz) and some combinations of them. This includes a study of different window sizes.
- The effect of vocals in the music
- The effect of additive noise in the test data

2. Data base, features and model

The data base used in these experiments was the same as in the previous paper [1] extended with vocal music and choir music.

<p>| Table 1. Sound database overview. Nr means number of speakers or number of music pieces. |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>Sound Class</th>
<th>Training Minutes</th>
<th>Test Minutes</th>
<th>Training Nr</th>
<th>Test Nr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>17 49 8 19 15 48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instrumental</td>
<td>16 53 7 23 15 48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocal</td>
<td>19 41 10 23 14 32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choir</td>
<td>14 9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>16 51 8 (25) 11 38</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Speech training and EER data are samples from the Washburn data base [5]. EER data is used to calculate a threshold for the score, giving Equal Error Rate on EER data when using the models from training data. This threshold is also used on test data whenever a decision on only two classes is to be made. The music database consists of the four parts instrumental, vocal, choir and mixed music. All music training and EER data are samples from CD recordings. Mixed music contains both instrumental and vocal music pieces. The choir music is a cappella music. The mix of male/female speakers was approximately 75%/25% for training and EER data and 50% each for test data.

The test data (except for the choir data) were collected from the Swedish broadcast during January and February 2001 using a standard FM receiver. The speech contained a variety of speaking styles and the music represented different styles such as pop, rock, country classical music etc. The choir music data were all collected from CD recordings.

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

As a standard feature for comparison was used the 39-MFCC features, i.e. 13 Mel frequency cepstrum coefficients using a Hanning window of 32 ms with their delta and delta-delta coefficients calculated over 100 ms with linear regression. This gives a 39-dimensional vector. A 78-MFCC feature uses 26 MFCC features, giving a 78-dimensional vector. The LFMA feature is described in section 3. Also the earlier reported [1] mixed feature, combined of 39-MFCC and LFMA, were used in the tests.

Some of the best of the LFMA features, namely the LFMA-4 (the same as 4 Hz ASD) and LFMA-3 were used. All tests were performed using a GMM-32 (Gaussian Mixture Model with 32 mixture components).

3. Choice of LF component

3.1 Feature extraction

The features were extracted from the LF modulation amplitude and its standard deviation for 20 critical bands, giving a 40-dimensional vector, referred to as LFMA, in the same way as the 4 Hz ASD [1]. This means using a critical-band filter bank (20 bands), rectifying, low pass filtering at 28
Hz, normalising by long-term average and finally extracting the log power of the desired low frequency component, calculated by FFT. The size of the analysis window was varied from 37.5 ms for the 27 Hz feature up to 1 second. The standard deviation of each amplitude was calculated using 20 overlapping windows with an increment of 12.5 ms, giving a decision window size that varies from approximately 290 to 1250 ms.

The influence of the analysis window size was examined by computing LPFMD-n, with n=2 to 27 and testing the discrimination ability. Some of them were also combined to see if this could improve the results.

3.2 Results

In Fig. 1 and Table 2 the error rate is presented vs. analysis window size in ms as well as in number of periods of the LF modulation component. This means that the 4 Hz component calculated with a window size of 1 second contains 4 periods, while 250 ms contains 1 period.

![Figure 1. Error rate, for test and EER data for speech/music discrimination with LPFMD features with different analysis window sizes. The results are presented as number of LF periods (top) and in ms (bottom) for those with one LF period. Linear regression lines are embedded in the top graph. EER and test data are almost parallel with a distance of 9.3%.](image)

When performing a t-test with samples-in-pair a significant difference is found between long and short analysis window size for the same LF component. It is also significant that the error rate in speech data is higher than in music data. From these results it does not seem to be critical what LF component to use as long as the analysis window size is one period. The differences between test data results in the interval between 37.5 and 250 ms are small, not linear and not statistically significant. For the used test data, the 4 and 5 Hz together with the 16 and 20 Hz performs best. The investigations in section 4 and 5 are using the 4 Hz and 5 Hz features with one period analysis window, LPFMD-4 and LPFMD-5.

The LF component, especially for speech, is not static and the larger the window, the more it varies. This explains why this feature performs better for smaller windows. From this investigation it is found that the whole interval from 37.5 ms to at least 250 ms can be used. The longer durations probably correspond to the syllable rate, 4-5 Hz, while the shorter ones could correspond to movements in the lips and glottis. According to Stevens [6] these durations can be as short as 5-10 ms for the lips and 80 – 150 ms for glottis.

Combining the LF components did not improve the results. It is probably due to the need of a larger analysis window size that an expected improvement from the combination did not have an effect.

### Table 1. Error rate for test data for different Low Frequency Modulation features when discriminating between speech and music. The best are 4, 5, 16 and 20 Hz, with no significant difference.

<table>
<thead>
<tr>
<th>LF comp (Hz)</th>
<th>Analysis window size (ms)</th>
<th>Periods</th>
<th>Equal Error Rate (%)</th>
<th>Error in Test data (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>500</td>
<td>1</td>
<td>10.6</td>
<td>20.0</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>2</td>
<td>12.1</td>
<td>21.7</td>
</tr>
<tr>
<td>3</td>
<td>320</td>
<td>1</td>
<td>10.2</td>
<td>20.8</td>
</tr>
<tr>
<td>3</td>
<td>1000</td>
<td>3</td>
<td>12.9</td>
<td>20.7</td>
</tr>
<tr>
<td>4</td>
<td>250</td>
<td>1</td>
<td>10.1</td>
<td>18.7</td>
</tr>
<tr>
<td>4</td>
<td>500</td>
<td>2</td>
<td>13.3</td>
<td>20.6</td>
</tr>
<tr>
<td>4</td>
<td>1000</td>
<td>4</td>
<td>17.5</td>
<td>24.9</td>
</tr>
<tr>
<td>5</td>
<td>200</td>
<td>1</td>
<td>9.9</td>
<td>18.7</td>
</tr>
<tr>
<td>5</td>
<td>1000</td>
<td>5</td>
<td>16.6</td>
<td>24.5</td>
</tr>
<tr>
<td>6</td>
<td>162.5</td>
<td>1</td>
<td>9.8</td>
<td>19.1</td>
</tr>
<tr>
<td>8</td>
<td>125</td>
<td>1</td>
<td>9.8</td>
<td>19.5</td>
</tr>
<tr>
<td>8</td>
<td>500</td>
<td>4</td>
<td>18.3</td>
<td>25.8</td>
</tr>
<tr>
<td>8</td>
<td>1000</td>
<td>8</td>
<td>18.4</td>
<td>28.2</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>1</td>
<td>9.5</td>
<td>18.3</td>
</tr>
<tr>
<td>16</td>
<td>62.5</td>
<td>1</td>
<td>9.1</td>
<td>18.7</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>1</td>
<td>9.2</td>
<td>18.6</td>
</tr>
<tr>
<td>27</td>
<td>37.5</td>
<td>1</td>
<td>9.4</td>
<td>19.3</td>
</tr>
</tbody>
</table>

### 4. Several music classes

To investigate the effect of vocal vs. instrumental music, a comparison was performed with different kind of music classes. The database used is presented above. First a vocal music model was trained in the same way as the instrumental music models above. The discriminating ability was examined for some features and compared with the results from speech/instrumental music discrimination.

Also models for choirs music a cappella were built. Confusion tests with several music classes and one speech class were performed. Note that these tests could not be performed with the calculated EER threshold.

#### 4.1 Results

##### 4.1.1 Vocal vs. Instrumental music

Tests were performed with speech vs. two different music classes. One class contained only instrumental music, the same as above, and the other consisted of the vocal music class. The amount of song in each piece was 75 – 90%. The result, presented in Table 2, shows that the error rate for vocal music is larger than for instrumental music in all cases, and more for MFCC than for LPFMD. MFCC is designed for recognising the timbre in the sound while LPFMD recognises the rhythm (LF modulation) and it's distribution over frequency bands. This can explain the smaller increase in
error rate for LFMAD than for MFCC. The mixed feature seems to take advantage of both, since it performs best. There is a bias towards classifying data as music. Keep in mind that the test data are sampled from other conditions than the training data.

Table 2. Frame by frame error rate (%) for discriminating between speech and music. The average value is calculated between speech and music test data.

<table>
<thead>
<tr>
<th>Frame type</th>
<th>Instrumental</th>
<th>Vocal</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>LFMAD-4</td>
<td>10.3</td>
<td>12.8</td>
<td>12.4</td>
</tr>
<tr>
<td>Speech</td>
<td>10.3</td>
<td>10.2</td>
<td>10.6</td>
</tr>
<tr>
<td>Mixed music</td>
<td>4.6</td>
<td>4.6</td>
<td>4.6</td>
</tr>
<tr>
<td>Music</td>
<td>4.6</td>
<td>4.6</td>
<td>4.6</td>
</tr>
<tr>
<td>MFCC</td>
<td>17.7</td>
<td>5.9</td>
<td>11.8</td>
</tr>
<tr>
<td>Speech</td>
<td>17.7</td>
<td>5.9</td>
<td>11.8</td>
</tr>
<tr>
<td>Mixed music</td>
<td>5.9</td>
<td>5.9</td>
<td>5.9</td>
</tr>
<tr>
<td>Music</td>
<td>5.9</td>
<td>5.9</td>
<td>5.9</td>
</tr>
</tbody>
</table>

4.1.2. Confusion matrices for several classes

Classification tests were carried out on 3 and 4 different classes. The music was divided into one choir music class and either one mixed music class or one vocal and one instrumental music class. The number of correctly classified frames for the choir music class remains almost constant when splitting the mixed music class into one vocal and one instrumental class. It is also relatively well classified which could depend on the fact that both training and test data were collected from CD recordings. Overall there was a small difference between MFCC and LFMAD to the advantage of MFCC, but again the mixed feature performed best. Only part of the result is presented in Tables 3 and 4. Three features are presented from the 3-class classifying task and only the best, the mixed feature, from the 4-class classifying task.

It can be seen that speech and choir music are seldom confused with each other and music, either two or three classes, takes the mis-classified frames. This is partly due to the common tendency to classify more frames as music and here this is reinforced due to the absence of an IER threshold in the classification procedure.

The results from 78-MFCC and LFMAD-4 show that a total agreement of 70% and 68% respectively was obtained. A better measure could be the Kappa coefficient [7, 8], which compensates for the chance probability. It is defined as

\[ K = \frac{P(A) - P(E)}{1 - P(E)} \]  \hspace{1cm} (1)

\[ P(A) = \sum \left( \frac{N_{i,j}}{N} \right) \] \hspace{1cm} (2)

\[ P(E) = \sum \left( \frac{n_i}{N} \times \frac{n_j}{N} \right) \] \hspace{1cm} (3)

where \( P(A) \) is the overall agreement and \( P(E) \) is the probability that agreement should occur by chance.

Table 3. Confusion matrix for 3 features and 3 classes.

<table>
<thead>
<tr>
<th></th>
<th>Speech</th>
<th>Mixed music</th>
<th>Choir music</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>59.8</td>
<td>27.7</td>
<td>12.5</td>
</tr>
<tr>
<td>Mixed music</td>
<td>6.6</td>
<td>75.3</td>
<td>18.1</td>
</tr>
<tr>
<td>Choir music</td>
<td>1.4</td>
<td>22.3</td>
<td>76.3</td>
</tr>
</tbody>
</table>

Table 4. Confusion matrix for the mixed feature and 4 classes. Test input in rows and classifications in columns.

<table>
<thead>
<tr>
<th></th>
<th>Speech</th>
<th>Mixed</th>
<th>Choir</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>66.4</td>
<td>27.1</td>
<td>6.6</td>
</tr>
<tr>
<td>Mixed</td>
<td>11.8</td>
<td>62.2</td>
<td>26.1</td>
</tr>
<tr>
<td>Choir</td>
<td>3.5</td>
<td>21.0</td>
<td>75.5</td>
</tr>
</tbody>
</table>

Figure 2. Result from classification test. % correct (left) and false (right) classified frames for the classes Speech (S), Vocal music (V), Instrumental music (I), mixed music (M) and Choir music (C) respectively. The left graphs for three classes and the right for four classes. The features presented are the mixed feature (top) and LFMAD-4 (bottom).

Table 5. Results from confusion tests. See text.

<table>
<thead>
<tr>
<th></th>
<th>3 Classes</th>
<th>4 Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P(A)</td>
<td>Kappa</td>
</tr>
<tr>
<td>39-MFCC</td>
<td>0.71</td>
<td>0.56</td>
</tr>
<tr>
<td>78-MFCC</td>
<td>0.70</td>
<td>0.59</td>
</tr>
<tr>
<td>LFMAD-4</td>
<td>0.68</td>
<td>0.52</td>
</tr>
<tr>
<td>Mixed</td>
<td>0.73</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Kappa is normally used to measure agreement between two or more annotators or judges. In our case, one of the judges is the correct classification and the other is the classifier to be tested. Kappa values for the tests performed
are presented in Table 5. A Kappa value around 0.5 is only considered as ‘moderate’.

5. Noisy signals

In order to examine the robustness against noise for different features, while Gaussian noise was added to the test data. The noise was controlled to achieve a certain SNR, measured over one second. The noise was added assuming that the original test data was clean.

5.1 Features

LFMA-4 and LFMA-5 were chosen as LF features since they were among the best in the tests in section 3. Also 39-MFCC and 78-MFCC together with the mixed feature was used in the test.

5.2 Data base

The test database for speech and instrumental music described above, with additive white noise, was used.

5.3 Results

The results of these investigations, presented in Fig. 3, show that the LFMA features do not degrade as fast as the conventional MFCC. This was expected, since when going from training (and EER) data to test data with a different environment (or channel) the error rate for MFCC increased more than for LFMA. However, when adding noise, the error rate for speech increases more than for music, as noticed in earlier tests too. This can be explained simply by the fact that the noise itself can be considered as another sound source and music already contains more than one source. This was found for both MFCC and LFMA. The LFMA-4 and LFMA-5 performed almost equally but only LFMA-5 is presented.

Also note that the mixed feature takes advantage of both the MFCC and the LFMA, and actually performs best.

6. Discussion and concluding remarks

This paper shows that the LF modulation feature LFMA is more robust than MFCC in a speech/music discrimination task. It does not degrade as much as MFCC when the test conditions diverge from those of the training situation. However, the conventional MFCC features perform better under circumstances with only instrumental music. Thus, LFMA or even better, the mixed feature, could be recommended when scanning data bases of unknown quality for speech events in segments with sizes of a couple of seconds, choosing the threshold in such a way that speech segments will not be lost in the scan.

7. Acknowledgements

This research was carried out at the Centre for Speech Technology, supported by Vinnova, KTH and participating Swedish companies and organisations. The author is also very grateful to Rolf Carlson and Mats Blomberg for their valuable help throughout this study.

8. References

Paper 3
Speech/Music Discrimination Using Discrete Hidden Markov Models

Stefan Karnebäck
stefan@speech.kth.se
Centre for Speech Technology, KTH - Royal Institute of Technology,
Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden

Abstract
A speech/music discrimination system using discrete Hidden Markov Models has been designed. The system has been evaluated using separate training, development and test databases. The discrimination ability was examined using different features and feature combinations and results are presented as error rate on the development and test databases. Features were chosen from knowledge about the speech signal. Adding the zero crossing rate, autocorrelation function and spectral gravity features to cepstrum coefficients helped to improve the discrimination result, while the cepstrum features were found to be more robust. The impact of the model size on the Speech/Music discrimination result was especially evaluated. Different compositions of the database were also explored, with and without a good match.

The best result on a mismatched situation, 2.3% error rate on test data, was achieved with 2x13 LFCC (13 Cepstrum coefficients and first time differentials), using 24 states and 96 symbols. In the matched situation, the best result on test data was achieved using a combination of all features extracted (47-dimensional), yielding 1% error rate or just below, at the optimal model size, 20-24 states and 48-54 symbols. These tests used a 1-second decision window size.

By using transcribed speech files it was also possible to compare the state assignment with the uttered phoneme for each segment. Investigating these results showed that the assigned states from the Viterbi search in a 3-state HMM, could be considered as phoneme classes.

Introduction and background
Speech/Music discrimination (SMD) systems are used for different purposes. Before transcription in mixed audio database systems, a segmentation and classification pre-processor is often used to detect the speech segments. A sound classifier that could be an (SMD) system, is a natural component in such a pre-processor. Other applications could be music detection for listening purposes or to adapt appropriate signal processing on different kinds of sound, such as speech or music, in broadcasting. The design of an SMD system, including feature extraction, can be influenced by investigating the speech and/or the music signal. In this work, knowledge about the acoustic speech signal has influenced the design of the system.

Human speech can, in one aspect, be described as concatenated syllables. As a result, the acoustic signal varies between high energetic nuclei of the syllables surrounded by weaker segments. Even the spectral shape varies with the syllable rhythm. This behaviour is clearly recognized in a spectrogram, which is composed by quasi-stationary spectral states. Speech production could thus be considered as a state machine, where the states are phoneme classes.

Ajmera et al. (2003) have shown that the probability dynamism, which captures the dynamic behaviour of the probability values, is larger for a speech signal than for a music signal, while the entropy of a speech signal is smaller than for a music signal.

Similar observations in the speech signal have been reported, for example by Greenburg (1995). He describes a number of properties to
be found in the speech signal from the perception perspective, for example the micro-modulation, in the interval of 3 to 12 ms, also considered as the pitch and the macro-modulation, 50-250 ms, associated with segments and syllables in the speech. Especially the macro-modulation supports this behaviour and it has been usec in SMD tasks by the author in earlier reports (Karnebäck, 2001 and 2002). These reports showed that a feature based on the low frequency modulation, its amplitude and deviation (LFMAD), and especially a combination of LFMAD and MFCC’s constitute robust features for SMD tasks.

Saunders (1996) used some of the characteristic features pointed out by Greenberg to successfully discriminate speech/music in FM broadcasting. Since the task was to develop a real-time system, he used only time domain features, mostly from ZCR (Zero Crossing Rate). Samouelian et al. (1998) also used some time domain features combined with two frequency features. Static spectral domain features which are used in speaker and speech recognition tasks can also be used for discriminating between speech, music and other sound sources. Hain & Woodland (1998) used MFCC (including normalised log energy and the first and second order time differentials) and Gaussian Mixture Models (GMM) separating four different sound sources (Speech, Music, Noise and Telephony speech). Gauvain et al. (1999) used a GMM system with MFCC as input features. They performed a segmentation and clustering step followed by a transcription session. Since MFCCs are often used in speech recognition systems for transcription, they were natural to use in the segmentation part too. Scheirer & Slaney (1997) pointed out and used some features for speech/music discrimination, which are closely related to the nature of human speech. Some of the features they used were the 4 Hz energy, percent of low energy frames and spectral centroid and spectral “flux” (Delta Spectrum Magnitude).

Nordqvist & Leijon (2002) used discrete HMMs to classify the acoustic environment in order to set gain and filter parameters in hearing aid equipment. By manually adjusting the constants for transitions they created a two-level system. They used 12 delta-cepsrum coefficients and modelled three sound sources (wide-band speech, traffic and telephony speech). The transitions between these sound sources were used by a second level HMM with four states. The objective was to decide whether the conversation was held on a telephone line or face-to-face.

Even though some features are already known to reflect this state switching in the speech signal, typically the energy feature, there might be other features with a strong impact from this behaviour. In order to get a tool to investigate this effect from individual features and feature combinations, a signal classification system is developed and presented in this report. The system is designed to discriminate between two signals and primarily evaluated on the SMD task. It could easily be further developed by arranging several discriminators in parallel or cascade for more complex classification tasks. The assumption in this work is that a music signal, albeit switching, does not show the same pattern as the speech signal and that it would be a good method to construct an SMD system using HMMs, since the dynamic behaviour is taken care of in the HMM. Other sound sources are also presumed to show different behaviour than speech on this matter.

HMMs are commonly used in speech recognition, speaker verification and speech recognition systems. In SMD and classification tasks, HMMs are also used in different ways. HMM/ANN hybrid systems are used by Ajmera et al. (2003) and Williams & Ellis (1999). Continuous HMMs, containing one state per sound timbre and one model per sound source are used by Zhang & Kuo (1999). Allegro et al. (2001) also use per sound source individually trained HMMs. The most commonly used features are cepstrum coefficients, or secondary features derived from these (Ajmera et al., 2003; Williams & Ellis, 1999).

This report describes a sound classification system, evaluated on the SMD task, based on discrete ergodic HMMs, using both cepstrum coefficients and other features modelling speech and music sound sources separately. This is a first step in a feature evaluation work and the objective is to evaluate if the design is useful for its purpose by performing evaluations on some features and feature combinations. As a side-effect, preliminary evaluations of a phoneme classification test, investigating the quasi-stationary segments, could also be performed. The system is
described in section 3 and the databases in section 4. In section 5, the technical data is presented together with a discussion on feature selection. The results are presented in section 6. In section 7, the result is discussed and conclusions are drawn.

Objectives and method

The objective of this work has been to design a sound classification system to be used mainly as an SMD system. The design is based on the assumption that the switching behaviour in the acoustic speech signal between phones or phone classes, is to a large degree speech specific and less found in other sound sources. The phone classes can be considered as quasi-stationary states and thus this assumption leads to the use of Hidden Markov Models (HMMs). Since HMMs are known as an effective method in automatic speech recognition tasks, it is a natural choice also for SMD tasks. The sound classification system should be used to evaluate the effect of the individual features on the discrimination or classification ability. The relation between the uttered phoneme and the states in the optimal path sequence should also be possible to investigate.

A system for training and testing was developed. The issues reported on in this paper are mainly

- feature selection
- number of observation symbols in the HMM
- number of states in the HMM
- feature evaluation
- decision window size
- matched and mismatched databases

The assumption that the states could be considered as phoneme classes was evaluated in the investigation of the assignment of each label. An argument for the use of discrete HMMs is that both the symbols and the states have some phonetic correspondence that can be evaluated. Transcriptions from the database used are available and comparison could be performed. By looking at each frame inside a label, a comparison between the assigned state, or symbol, and the uttered phoneme is performed. Only some of the feature combinations were examined on this issue.

System description

The proposed system is naturally divided into two parts, the test and the training subsystems, shown in Figs. 1 and 2 respectively. An overview is presented below and starts with the test system.

The test subsystem overview

![Speech/Music discrimination test diagram](image)

*Figure 1. Test subsystem, see text.*
The training system overview

![Diagram of the training system overview](image)

*Figure 2. Training subsystem. VQnobs is the observation symbol codebook used for both training and testing. The codebook VQnstates is only used for training, see text. The Music model training is left out for steps 2 and 3, but it is performed in the same way as the speech model.*

The system resembles a speaker verification system. The input signal is split into short-time analysis frames. From each frame is extracted a feature or feature combination. Features are normalised with mean and standard deviation derived from the training database, giving a distribution on the training database features of $N(0,1)$. These values are stored together with the codebook. Only one codebook is used and shared by both the speech and music signals. The codebook is referred to as VQn, where $n$ is the number of symbols or cells in the codebook. It would have been possible to build one separate codebook for speech signals and another for use on music signals, thus separating the two systems. Pinquier et al. (2002b) uses separately trained models for speech and music respectively. The choice in this work was, however, to use common codebooks for the two systems. In the future, it would be possible to compare the performance for the two alternatives.

Observation symbols are obtained by a search in the codebook for each feature vector. A sequence of observations, corresponding to one or a few seconds, constitutes the decision window. The observation sequence is matched by a Viterbi search to the HMM speech and
music models, respectively. One score value is calculated on each decision window at a certain frame rate.

From the Viterbi search a log probability was obtained that was used as a score in the decision algorithm. The decision rule for classification is

\[
\text{Class} = \begin{cases} 
\text{Music} : & \log P_{\text{music}} - \log P_{\text{speech}} > \eta \\
\text{Speech} : & \log P_{\text{music}} - \log P_{\text{speech}} < \eta 
\end{cases}
\]

(1)

where \( \eta \) is a threshold. This is equivalent to taking the log of the likelihood ratio and comparing with a threshold. No score normalisation was performed. The algorithm also gives the most probable path for the sequence. For a VQ3 assignment, it is possible to compare the observation sequence with the output state sequence from a three-state HMM.

The training system, Figure 2, is more complex than the test system. It can be divided into three parts: codebook training, HMM initialisation and HMM training.

**Codebook training**

Two codebooks are used. \( \text{VQ}_{\text{state}} \) corresponds to the observation symbol set and \( \text{VQ}_{\text{state}} \) to the states. The latter is only used to initialise the HMM. The codebooks were trained using the generalised Lloyd algorithm (Linde et al., 1980). The system is unsupervised, i.e. no labels are used to create the codebooks. The only manual work is to feed the system with split speech and music data files during the training phase.

**HMM initialisation**

The transition probability matrix \( \mathbf{A} \) is initialised with large numbers (0.9) in the diagonal and the rest equally distributed.

\[
\mathbf{A} = \left[ \begin{array}{c}
0.1 \\
(\theta_{\text{states}} - 1) \\
0.9
\end{array} \right]
\]

(2)

Since no element in the A-matrix is set to 0, the model is fully connected or ergodic. The observation probability matrix \( \mathbf{B} \) is initialised using the quantified training data, from two separate codebooks. The matrix \( \mathbf{B} \) is defined as

\[
b_i(k) = \frac{P(x_t = o_k | q_t = s_i)}{N_{\text{obstalist}} / N_{\text{state}}}
\]

(3)

where \( x_t \) is the observation at time \( t \), \( o_k \) is a symbol \( k \) from the observation symbol library \( \mathcal{O} \), \( q_t \) is the state at time \( t \) and \( s_i \) is state number \( i \). Since no states are defined, the codebook \( \text{VQ}_{\text{states}} \) is used to initially assign each frame to a state. Let \( V_{\text{state}}(1 : T) \) be the vector of \( T \) consecutive quantified features from training data, using codebook \( \text{VQ}_{\text{states}} \), \( V_{\text{obs}}(1 : T) \) is the training data quantified using the observation codebook \( \text{VQ}_{\text{obs}} \) and \( T \) the number of frames in the training data. Equation (4) below is then used at initialisation.

\[
b_i(k) = \frac{P(V_{\text{obs}}(t) = k | V_{\text{state}}(t) = i)}{N_{\text{obstalist}} / N_{\text{state}}}
\]

(4)

where \( N_{\text{obstalist}} \) is the number of joint events where \( V_{\text{obs}}(t) = k \) and \( V_{\text{state}}(t) = i \) and \( N_{\text{state}} \) is the number of single events where \( V_{\text{state}}(t) = i \). No element in the B-matrix is initialized to zero. If the first initialization introduced zeroes they were replaced with small numbers. When the number of symbols was the same as the number of states in the HMM, the observation probability matrix was initialized in the same way as the A-matrix.

The Baum-Welch algorithm was used for training the HMMs.

**Feature extraction**

Several basic features were studied, listed in Table 1. They were combined in many ways, giving feature vector dimensions from 1 to 47.

The energy feature switches with the syllable rate. The nucleus of the syllable contains high energy and thus this feature was assumed to be useful. The binary energy feature only fortifies this effect.

Cepstrum parameters have been used in speaker verification with success and also in SMD systems (Hain and Woodland, 1998; Gauvain et al., 1999). Mel frequency scaled cepstrum coefficients, MFCC, are often used...
on speech signals. If they are appropriate for music signals is not yet known although Logan (2000) showed that MFCC outperforms LFCC, linear scaled frequency cepstrum coefficients, in short term discrimination tasks. The higher resolution in the lower frequency bands, achieved by MFCCs, reflects the auditory system and results in a higher yield on recognition tasks. This higher yield can also be obtained by using some more coefficients from LFCC. Since they do not differ in principle, but describe the same aspect in the signal, namely the overall shape of the spectrum, the choice in this work was to use LFCC.

The ZCR is supposed to detect the dominant frequency of the interval (Kedem, 1986), and is explored in several ways by Saunders (1996) and Scheiter & Slaney (1997). This feature obtains a high value at voiceless fricatives. When filtering in intervals, ZCR0-1, ZCR1-6 and ZCR2-7, the assumption was that the zero crossing rate should follow the dominant frequencies within the interval that could be a formant frequency in the speech signal.

Acf gives a high value for strong periodic signals and would typically obtain high values in voiced segments of the speech signal. A threshold in the Acf signal gives a binary feature, Voiced/Unvoiced. This is supposed to follow the voicing in the speech signal.

The spectral gravity reflects approximately the same character as the ZCR, only calculated in a different manner.

The features are selected from knowledge about the speech signal, assuming that the music signal, or any other sound signal, shows a different behaviour.

Special effort was dedicated to the question of feature selection, described in section 5.3.

Features were normalised with mean and standard deviation derived from the training database, giving a distribution on the training database features of N(0,1), before codebook training.

### Databases

The speech and music databases were both divided into three parts, a training, a development and a test database. Model training was performed on the training database. The development database was used...
to find the threshold giving an equal error rate for speech and music data.

The database was composed in two different ways. One with a good match and one with mismatch between training and development on one hand and test data on the other hand. The matched database were composed in 5 different combinations, derived from the same sources. On many examinations, all five were examined and average values were calculated.

Two subsets of speech data with different sound quality were used, the Waxholm database (Bertenstam et al., 1995) and Swedish FM broadcasts. The Waxholm database is recorded in a silent room using a hi-fi quality microphone with 16 kHz sample rate. Waxholm is a dialogue system where the task is to ask for boat schedule, hotel room and other facilities in the archipelago of Stockholm. It contains 68 speakers of mixed gender. The broadcasts were recorded using a standard FM receiver and a cassette tape recorder. 48 segments of speech with equal distribution between male and female speakers were collected during January and February 2001.

The music database also uses two subsets of different quality, CD recordings and FM broadcasts. The CD recordings contain a great variety of instrumental music and the broadcast recordings contain 46 segments of mixed kinds of instrumental music.

The speech databases differ more in quality than the music databases. The FM recordings contain more spontaneous speech than the short phrases in the Waxholm database. The SNRs differ approximately 12 dB on average and the transfer functions are different. The music databases are more similar, but the SNRs are higher on the CD recordings than in the radio segments. The purpose of these constructions was to evaluate the robustness of the system with a good and a bad match. The amount and mixture of data is presented in Table 2.

Initial silent parts were removed, but there is silence within the speech and at the end of the utterances, in both training and test data, while the music files were almost free from silence.

**Experiments**

**Technical data**

All signals were sampled at or resampled to 16 kHz with 16 bits in mono. The analysis window size was 20 ms and the frame rate was 100 Hz. The decision window was primarily one second. Before FFT-calculations a Hamming window was applied.

**Evaluation methods**

There are several parameters of interest to investigate in this system. As mentioned, different features and feature combinations are of interest and an initial investigation is presented in this report. However, the main focus is put on the impact of the model size on the SMD result. The database composition is also investigated. Results are presented as error rates in an SMD task.

**Error rate calculations**

The models were trained using the training databases. Evaluation was performed by calculating the equal error rate, on speech and

<table>
<thead>
<tr>
<th>Sound source</th>
<th>Mismatched database</th>
<th>Matched database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>Training</td>
<td>Development</td>
</tr>
<tr>
<td>Waxholm</td>
<td>15 min</td>
<td>8 min</td>
</tr>
<tr>
<td>FM-broadcasts</td>
<td>15 min</td>
<td>6 min</td>
</tr>
<tr>
<td>Total</td>
<td>15 min</td>
<td>8 min</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sound source</th>
<th>Mismatched database</th>
<th>Matched database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Music</td>
<td>Training</td>
<td>Development</td>
</tr>
<tr>
<td>CD-recordings</td>
<td>15 min</td>
<td>8 min</td>
</tr>
<tr>
<td>FM-broadcasts</td>
<td>15 min</td>
<td>6 min</td>
</tr>
<tr>
<td>Total</td>
<td>15 min</td>
<td>8 min</td>
</tr>
</tbody>
</table>
music test data, when testing on the development databases, by varying the threshold, \( \eta \), in Eq. 1.

This threshold is denoted \( \eta_{\text{SER}} \). The speech error rate is calculated as

\[
\text{SER} = \frac{N_{\text{err}}}{N_{\text{tot}}} \times 100\% \quad (5)
\]

where \( N_{\text{err}} \) is the number of falsely classified frames in the speech database and \( N_{\text{tot}} \) is the total number of frames in the speech database. The corresponding definition for MER is

\[
\text{MER} = \frac{N_{M_{\text{err}}}}{N_{M_{\text{tot}}}} \times 100\% \quad (6)
\]

where \( N_{M_{\text{err}}} \) is the number of falsely classified frames in the music database and \( N_{M_{\text{tot}}} \) is the total number of frames in the music database. Index is used to inform what database is used in the test. For example \( \text{SER}_{\text{dev}} \) gives the speech error rate from the development database. A total error rate (TER) was also calculated as the average of SER and MER.

\[
\text{TER} = \frac{\text{SER} + \text{MER}}{2} \quad (7)
\]

Results are presented as \( \text{EER}_{\text{dev}} \), which is the equal error rate on the development database (performed at \( \eta = \eta_{\text{SER}} \)). \( \text{TER}_{\text{dev}} \), which is the total error rate on the test database, also calculated at \( \eta = \eta_{\text{SER}} \) and finally as \( \text{EER}_{\text{dev}} \), the equal error rate achieved on the test database, normally found at a threshold \( \eta \neq \eta_{\text{SER}} \).

**Feature selection**

A method to help to select features with large impact on the discrimination ability was searched for. In many similar situations, PCA (Principal Component Analysis) is used in order to reduce the number of secondary features as input to the model-training algorithm. However, in this work the aim is also to evaluate the primary features and their impact on the discrimination ability. Fukunaga (1972) describes another possible way, also known as discriminant analysis. These methods should help to separate the sources in the static feature space. Since our system is based on the dynamic behaviour, modelled as HMM, and there is no known correlation between deviations in the static feature space and the dynamic behaviour, it is not likely that the methods would help. Besides the Vector Quantizer takes care of the separation in the static feature space. As a tool in the feature selection procedure the correlation matrices were also investigated.

**Results**

The result presentation is divided into several parts, starting with some initial investigations concerning feature correlation and increasing number of observation symbols. The main results consist of error rates as a function of either number of states or number of symbols. These tests are performed on the two differently composed databases. Finally the decision window size is prolonged in one test.

**Feature correlations**

The results when investigating combined speech and music signals showed mainly weak correlations. Table 3 presents correlations above 0.4, while the others are left out for space-saving reasons.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Ceps-1</th>
<th>Ceps-2</th>
<th>Energy</th>
<th>ZCR</th>
<th>ZCR0-1</th>
<th>ZCR1-6</th>
<th>Acf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy</td>
<td>0.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZCR</td>
<td>-0.55</td>
<td>-0.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZCR1-6</td>
<td>-0.68</td>
<td>-0.46</td>
<td>0.51</td>
<td>-0.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZCR2-7</td>
<td>-0.63</td>
<td></td>
<td>-0.40</td>
<td>0.66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acf</td>
<td>0.41</td>
<td>0.54</td>
<td>0.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voiced/Unvoiced</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gravity</td>
<td>-0.57</td>
<td>-0.42</td>
<td>0.95</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Feature correlation matrix. Both speech and music signals are included in the calculation. Only correlations larger than 0.4 (magnitude) are presented. First and second order differentials are not present in this investigation.

48
It was found that the correlations between features were much larger within speech features than within music features, see Tables 4 and 5. Similar observations, concerning differences in correlation between frequency bands, were also found when investigating the LFMaD feature described by the author (Karnbrick, 2001). The reason can be that the speech signal, when not disturbed, comes from one sound source while the music signal, when not solo, comes from several sources.

Note that in the music features no correlations between the cepstrum coefficients and ZCR, Acf and Gravity, above 0.4, were found. As expected, a large correlation was found between ZCR and Gravity in all signals.

The correlations indicate that the features from ZCR, Acf, Voiced/Unvoiced and Gravity would perform well in this task since they are used with good results in speaker recognition systems as mentioned earlier.

It was decided to perform the SMD tests on 13 LFCC (13-dimensional) as a baseline, and add either first (26-dimensional) or both first and second order time differentials (39-dimensional) of those, or ZCR, Acf and Gravity (16-dimensional). Tests were also performed on a combination of all the features presented (47-dimensional, referred to as 'All' feature) and all features except the first and second order time differentials (21-dimensional, referred to as 'All except delta'). A more thorough investigation of each feature candidate and more feature combinations has to be left for the future.

**Initial investigations on number of observation symbols**

Several speech/music discrimination (SMD) tests were initially performed on a separate small database changing primarily the number of observation symbols from 3 to 24. They showed that the error rate was dramatically reduced, from 11-15% down to 1-2%, with increasing number of symbols. This behaviour was found using three states in the HMM. Therefore it was decided to start with 24 observation symbols in the following tests.

**State duration and phoneme classification**

The average durations of the states were measured for two feature combinations on 3- and 4-states HMM, see table 6. Three and four states could correspond to some phoneme classes. The speech model was used on speech data and the music model was used on music data.
Table 6. Duration lengths on 3- and 4-states Hidden Markov Models. Durations are mean values of all states.

<table>
<thead>
<tr>
<th>Feature</th>
<th>states</th>
<th>Mean duration (over all)</th>
<th>Mean duration (over all)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Speech</td>
<td>Music</td>
</tr>
<tr>
<td>13 LFCC</td>
<td>3</td>
<td>105 ms</td>
<td>190 ms</td>
</tr>
<tr>
<td>13 LFCC</td>
<td>4</td>
<td>99 ms</td>
<td>171 ms</td>
</tr>
<tr>
<td>All except delta</td>
<td>3</td>
<td>97 ms</td>
<td>121 ms</td>
</tr>
<tr>
<td>All except delta</td>
<td>4</td>
<td>84 ms</td>
<td>122 ms</td>
</tr>
</tbody>
</table>

The differences between speech and music durations are larger for the cepstrum features. The difference itself was expected and supports our assumption that the switching is more frequent in speech than in music signals. These results are also found and used for speech/music discrimination purposes by Pinquier (2002a).

Since the Waxholm database is transcribed, the classification per state could be compared with the transcription label showing the correspondence between the states produced by the Viterbi search and the labelled phoneme. Grouping all phonemes for each state indicates the assumed correspondence between state and phoneme class. In this procedure, every first and last frame in a transcribed phoneme was omitted to avoid transition problems. Hence, the system itself assigned each frame to a class, given by a number. The most frequent state assignment for each phoneme was supposed to correspond to that phoneme class. Considering this choice, an error rate was calculated for each class and in total. The total error on three classes, using 'All except delta', was approximately 12%. The error rate on 13 LFCC was higher, approximately 17%. Note, however, that the transcription is not strictly phonetic. The transcription was automatically performed, using the uttered text, literally, with manual corrections concerning deletions and insertions but not vowel changes due to assimilations. These results were found good enough and encourage further investigations.

**SMD tests on mismatched databases**

**Effect of number of states**

Experiments on the development and test databases from the mismatched databases were performed with three up to 48 states in the HMM models using 24 symbols on six feature combinations. Due to the mismatch in the database some features performed extremely well on EERdec, but worse on TERdec and EERacc. The models became over-trained. A randomised initialisation was used for the codebook training. As a result they did not become identical if tests were performed several times with new training procedure involved. Taking this fact into consideration the only conclusion to be drawn was a small reduction in error rate above 20 states on most of the feature combinations. Figure 3 shows the result for 24 states and 24 symbols.

The overall best result on TERdec was achieved with 'All except delta', yielding 2.4% at 32 states (EERacc, yielded 0%), although the differences between the explored features were small. Tests on the test database were performed on different thresholds and not only at $\eta=\eta_{x\bar{x}}$. It was then found that the best test results could sometimes be achieved using a threshold close to the $\eta_{x\bar{x}}$ and sometimes far away. These results show the mismatch between the training and development databases on one hand and the test databases on the other hand, as expected. One explanation is that these materials are different, recorded under different circumstances giving different SNR etc. An illustration is found in Figure 5 below.

It is interesting to note that 3x13 LFCC does not show the best yield. It was also unexpected that 13 LFCC performs better than 3x13 LFCC up to 28 states. For tests on 28 states or more 3x13 LFCC performed equally well or better than 13 LFCC and 2x13 LFCC. The reason might be that the delta cepstrum force an undesirable splitting of the codebook. Other features are probably more important for this purpose, and if there were more symbols
in the codebook, the HMM would help to get a better result.

An inspection of one feature combination, 13 LFCC, showed that a large amount of the speech errors appeared in connection to a pause, speaker noise or breathing. To solve this problem a future system should be expanded with a silence detector together with the segmentation procedure.

Of great importance in a future SMD system is the robustness. The most common aspect of robustness is that there should be a small difference between the EER\textsubscript{dev} and the TER\textsubscript{test} result. However, a robust but poorly performing feature is not very useful. In general, the cepstrum parameters and especially 3x13 LFCC were found a little more robust on this aspect, as were 13 LFCC+AcF+ZCR+Gravity to some extent. The development database is half the size of the test database and thus the EER\textsubscript{dev} is expected to be a little bit lower than TER\textsubscript{test}. Since the error rates differed more than expected for all feature combinations, we cannot consider the system to be robust.

**Effect of number of observation symbols**

When investigating the effect of varying the number of symbols, no obvious conclusion could be drawn, see Figure 4. A tendency that increasing number of symbols (going from 24 to 96) would improve the results, especially for 3x13 LFCC, was however found. The problem with finding other conclusions was probably due to the mismatch in the database composition. 'All except delta' seemed to be a good feature when using 24 states and 24 symbols, but not when increasing the number of symbols.

The poor result on test data for 'All except delta' with 96 symbols is probably due to a very good match between training and EER data when increasing the codebook size, while the test database is different. It recognises the training data too well and therefore it cannot perform well on the test database. It's clearly over-trained.

The overall best test result using 24 states, was achieved on 2x13 LFCC with 96 symbols, yielding 2.3 % error on TER\textsubscript{test}, while 'All except delta' yielded 2.2% on EER\textsubscript{test} using 96 symbols.

An alternative measurement of the robustness of the system and features is to calculate how many segments of music would have been falsely captured to detect at least 99% of the speech segments (the opposite could also be evaluated), see Table 7. In those tests, 'All except delta' performed best with only 3.3% falsely captured music segments. This is a bit surprising considering the high error rate, 6.8%, but this rate comprises mainly speech errors and the music errors increase slowly when speech errors decrease on increasing threshold, see Fig. 5. When comparing the TER\textsubscript{test} and EER\textsubscript{test} results for 'All except delta', this effect is also clear. The TER\textsubscript{test} is more depending on a good match between the databases than the EER\textsubscript{test} is. Thus, for feature evaluation purpose, the EER\textsubscript{test} is more relevant in a mismatched situation and the development database is thus of no use.
Figure 4. Results from SMD tests performed on 24 states as a function of number of observation symbols. $EER_{\text{dev}}$ (top), $TER_{\text{dev}}$ (middle) and $EER_{\text{test}}$ (bottom).
Table 7. Percent falsely detected music segments when detecting at least 99% of the speech segments on test data using 24 states and 96 symbols in the HMM on mismatched databases.

<table>
<thead>
<tr>
<th>Feature</th>
<th>13 LFCC</th>
<th>2x13 LFCC</th>
<th>3x13 LFCC</th>
<th>All except delta</th>
<th>13 LFCC + Acf + ZCR + Gravity</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent falsely detected segments</td>
<td>9.5</td>
<td>10</td>
<td>12</td>
<td>3.3</td>
<td>27</td>
<td>20</td>
</tr>
</tbody>
</table>

Figure 5. Speech, music and average error curves as a function of threshold value, using 'All except delta' with 24 states and 96 symbols. Note the unbalance between speech and music errors and that TER_{match} (6.8%) is much larger than EER_{test} (2.2%). EER_{test} is however not the lowest average value. The point 99% speech detect and 3.3% music error is marked in the graph, as well as EER_{delta}.

SMD tests on matched database

The system was not found to be robust against the mismatch in the databases. The EER_{delta} was extremely good while test results were only fairly good. Thus, tests were continued on a matched database. Five pseudo randomly recomposed database compositions were evaluated.

Baseline result with 24 states and 24 symbol

As a baseline and comparison with the mismatched database composition, tests were performed with 24 states and 24 symbols. Since these databases were very similar, the difference between EER_{delta} and TER_{test} became very small. As expected, TER_{test} and EER_{test} were also of the same size and much lower than in the mismatched case, see Fig. 6. When a matched situation occurs, TER_{test} is more relevant and useful for system evaluation, while EER_{test} was relevant for a mismatched situation, and only for feature evaluation purpose. The best result is now achieved with 'All' feature (1.3% on TER_{test}).

Effect of number of states

The effect of number of states were investigated on the matched database. The number of symbols were kept at 48, while the number of states were varied from 3 to 48. In Figure 7, four feature combinations are presented for one matched database composition. Due to processing time considerations, only one composition was tested with four feature combinations. The one chosen performed close to the average (slightly better) on the 'All' feature combination, which was evaluated on all 5 compositions, see Figure 8. It can be seen that ZCR, Acf and Gravity help to improve the result and that the 'All' feature
combination performs best, 0.7% for 32 states. A reduction in error rate between 12 and 32 states was found for 'All' and 'All except delta' but not seen at all for 13 LFCC and 3x13 LFCC.

A test to achieve a refined examination on number of states in the interval between 10 and 20 was performed on 13 LFCC using the same database composition. No surprise was found, maybe with one exception, a peak at 16 states on EERdev. However, TERtest and EERtest did not vary much between 10 and 20 states. It should be noted that EERdev in general varied more than TERtest and EERtest, and that 13 LFCC showed the largest variation among the features presented in Figure 7.

A small but clear error reduction is found, in Fig. 8, between 16 and 36 states. 20 or 24 states seem to be a good choice for best performance with 48 symbols, yielding 0.95 ±0.26% and 1.01 ±0.36% respectively on TERtest.

**Figure 6.** Average error rates from SMD tests performed on six feature combinations on five differently composed database compositions. 24 states and 24 symbols were used in the HMM. Results are presented together with ±0.5 SD.

**Figure 7.** SMD test results as a function of number of states on four feature combinations on one matched database composition, see text.
Figure 8. SMD test results as a function of number of states on matched databases for the ‘All’ feature combination. The average values for TERtest and EERtest from five database compositions are plotted. TERtest values are plotted together with ± 0.5 SD.

**Effect of number of observation symbols**

Also for the matched database, the effect of number of symbols were examined. Number of states were kept at 24. The first test comprises only 24 and 96 symbols for six feature combinations. Results are presented as average values for five database compositions in Figure 9. Since the difference between, EERdev, TERtest and EERtest are fairly small in this test, only the TERtest results are presented. It can be seen that the features containing only LFCC (with or without the differentials) need more symbols to perform as well as those containing also ZCR, Acf and Gravity. The best result on 96 symbols was achieved with 13 LFCC + Acf + ZCR + Gravity (1.13%).

An investigation of falsely accepted music segments when searching for speech segments was performed. The result, presented in Table 8, shows less variation than for the mismatched situation. 13 LFCC + ZCR + Acf + Gravity performs best also in this aspect.

Figure 10 presents some error curves for this well matched situation, using the same database composition as in Figure 7, which should be compared with Figure 5. However, due to a better match, EERdev, TERtest and EERtest are close to each other.

Figure 9. SMD test results for 24 and 96 symbols, 24 states, on matched databases. Average values on 5 differently composed databases. Results are presented with ± 0.5 SD.
Table 8. Percent falsely detected music segments when detecting 99% of the speech segments in test data on average for five database compositions. The HMM uses 24 states and 96 symbols.

<table>
<thead>
<tr>
<th>Feature</th>
<th>13 LFCC</th>
<th>2x13 LFCC</th>
<th>3x13 LFCC</th>
<th>All except delta</th>
<th>13 LFCC + Act + ZCR + Gravity</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent falsely detected</td>
<td>1.8 ±0.8</td>
<td>1.4 ±0.4</td>
<td>2.3 ±0.7</td>
<td>1.4 ±0.7</td>
<td>1.2 ±0.6</td>
<td>1.4 ±0.5</td>
</tr>
</tbody>
</table>

Figure 10. Speech, music and average error curves as a function of threshold value for one database composition, using ‘All except delta’ with 24 states and 96 symbols. Training, development and test data are well matched as can be seen. TER_{test} (1.56 %) is equal to EER_{dev}. The point 99% speechdetect and 2.25 % music error is marked in the graph, as well as EER_{test} (2.10 %).

Figure 11. SMD test results as a function of number of symbols. 24 states were used in the HMM. Results are presented as average values for TER_{test} and EER_{test} from 5 database compositions. TER_{test} values are plotted together with ± 0.5 SD.
The second test on varying the number of symbols in finer steps was only performed with the 'All' feature combination. The number of states were kept at 24 while the number of symbols varied from 12 to 96 in steps of 6. A clear reduction in error rate is found at 48 (TER\text{all}=1.0 ±0.36 %) and 54 (TER\text{all}=0.95 ±0.34 %) symbols. The results for TER\text{all} and EER\text{all} are very similar, which can be seen in Figure 11.

**Increased decision window size**

When implementing an SMD system in an application, it will be working on different sizes of decision windows. Normally, a segmentation has been performed in an earlier stage, using criteria based on knowledge of the data to be retrieved and a cost function. Generally the performance increases with increasing window size.

In earlier reports (Saunders, 1996; Scheirer & Slaney, 1997; Williams & Ellis, 1999), a decision window size of 2.4-2.5 seconds were used, yielding 1.3 – 2% error rate. To get an indication on how good the result can be, tests were performed with a 2.4 seconds decision window size. Due to processing time considerations only one of the matched compositions were explored. Using 24 states with 24 symbols, the error rate was reduced by approximately 50% on average, compared with a 1-second decision window size. Several feature combinations achieved error rates below 1%, for example 'All except delta' achieved TER\text{all} = 0.34% and 'All feature TER\text{all} = 0.55%'. Using an optimal model size would improve these results further.

**Feature evaluation**

The feature evaluation has only started in this work and needs to be continued. Generally, it turns out that adding more features results in better performance, assuming that large enough models are trained. Adding Acf, ZCR and/or Gravity instead of the delta cepstrum features yields better result on smaller HMM models. The cepstrum and delta cepstrum features seem to be more general and robust, while ZCR, Acf and Gravity (seen as a cluster in this work) detect more specific cues in the signal, thus increasing the discrepancy between error rate on the development and test databases, when applied on a mismatched database. Acf probably captures the voiced speech segments, which are cleaner (higher SNR) in the training and development databases than in the test database (in the mismatched database composition). However, if the training and test databases are more similar, then these features perform very well. This can be observed in Figure 6 above where 'All except delta' yields 1.4% error rate (TER\text{all} with 24 states and 24 symbols) which is better than the cepstrum features do by themselves. The best result, using 24 states and 24 symbols, was however achieved with the 'All' feature combination, yielding 1.3%.

**Discussion and conclusion**

A Speech/Music discrimination system using discrete Hidden Markov Models was designed. Several aspects of the system were investigated with focus on the HMM model size. Several different feature combinations were tested with models using up to 96 symbols and 48 states in the HMM. Different compositions of the database were tested, showing different behaviour for the features, on a good or a bad match between the training and test databases. The lowest error rate on test data, TER\text{all}, with the mismatched composition was achieved with 2x13 LFCC using 24 states and 96 symbols, yielding 2.3%. When a good match occurs the best result was achieved with the 'All' feature combination, indicating that a good match helps to get use of all the features. The error rate was just below 1%, calculated on 1-second decision windows. The results must be considered as good and tests performed on 2.4 seconds indicated a 50% reduction of the error rate, approximately.

Even though the results in this work are in the same magnitude or slightly better than earlier reports on the same task (Saunders, 1996; Scheirer & Slaney, 1997; Williams & Ellis, 1999), they cannot be compared, since the databases are different. In this database, there was no singing within the music, for example. Earlier investigations (Karnebäck, 2002) show a 30% increase in error rate when including song in the music database. Other results reported, like Ajmera et al. (2003) do not use the same decision window size, thus making a comparison difficult.

Since these results show a much better yield than earlier tests on the same database (Karnebäck, 2001), which used only static models like GMM or VQ and MFCC features
complemented with a low frequency modulation feature, LFMAD (some dynamic behavior is, however, captured in the delta cepstrum and LFMAD features), the conclusion is that discrete ergodic HMMs, perform well in SMD tasks.

Generally, it seems that adding more features results in better performance, assuming that large enough models are trained, also found by Berenzweig & Ellis (2001). However, large feature dimensions need large models and a large amount of training data to get use of the large information embedded in the feature vector. When small models are desired, it could be useful to evaluate, in advance, what features to extract. In this work, it was found that adding Acf, ZCR and Gravity was a better choice than adding the delta cepstrum to 13 LFCC, on small models, but profound investigations are needed.

The second order delta cepstrum coefficients seem to either need very large models to increase the performance or they do not add much information at all. In this work, there was no test where 3x13 LFCC performed best. However, the ‘All’ feature combination performed best, indicating that the Cepstrum features work well with other features and improve the result.

Some aspects of robustness were discussed. Which one to take into consideration depends on the application. When selecting speech segments for further transcription it is desirable to detect as many speech segments as possible with as few falsely detected music segments as possible. In this work, the ‘All except delta’ feature performed best on a mismatched situation, while several feature combinations performed almost equally well in the matched situation. The cepstrum features were less affected by the mismatched databases than feature combinations containing ZCR, Acf and Gravity.

The assumption that the phoneme classes are represented by quasi-stationary states in the HMM, could also be supported in this work. The agreement was approximately 12% and 17%, respectively, when an automatic phoneme classification for three and four classes was performed on the ‘All except delta’ feature combination. These findings have to be further investigated on larger number of states. It is likely that a large degree of agreement will be found also on 20- or 24-state models, since the discrimination performance was found to be best in that interval. If so, the states can be considered as phonemes rather than phoneme classes.

Refined and extended experiments have to verify, or discard, the implication that 24 states and 48-54 symbols are optimal sizes of the models. Is this behavior more enhanced for ‘non-cepstrum’ features? Can it be language specific? Another question to be answered is whether there are specific combinations of number of symbols and states in the HMM, that performs specifically well or bad.

The desired system for evaluating individual features or feature combinations in an SMD task, was designed. It was found useful for its purpose to investigate the impact from individual features on the SMD task. The error rates were found to be very small. Different ways to improve the performance on SMD tasks were discussed and they indicate that the system can be tuned to even better results. This tuning should be controlled by the specific application were the system should be a part. The system can also be used to investigate the agreement between the state assignment and the uttered phoneme on an individual feature basis.

Acknowledgement

This research was carried out at the Centre for Speech Technology, CTT, supported by Vinnova, KTH and participating Swedish companies and organisations.

References


