Speech and Music Discrimination based on Signal Modulation Spectrum.

Pavel Balabko

June 24, 1999

1 Introduction.

This work is devoted to the problem of automatic speech and music discrimination. As we will see here, speech and music signals have quite distinctive features. However, the efficient distinction between speech and music is still an open problem. This problem arises when it is necessary to extract speech information from the data containing both speech and music. The typical use of such segmentation is the extraction of speech segments from broadcast news for further processing by an Automatic Speech Recognition System (ASR). This work proposes a simple and quite effective solution to this problem based on the analysis of the speech and music modulation spectrum.

The organisation of this paper is as follows. Section 2 discusses the main problems and possible solutions for the speech and music discrimination problem. Section 3 presents some results from the study of human speech perception that have been used in this work. Section 4 outlines our approach for music and speech discrimination. Experiment results of the method proposed here and result analysis are presented in Section 5. Finally, the conclusion will be given in Section 6. A short description of function implementing approach proposed here is presented in Appendix 1.

2 Different Approaches to the Speech and Music Discrimination.

There could be a lot of different approaches to the problem of music and speech discrimination. Even just looking at the spectrogram one can see the big difference between speech and music. See Figure 1 as a typical example of such spectrogram. Here the music part is up to the 2-nd second and speech part is represented from 2 to 4 seconds. The spectrogram for the different types of speech always has some common features - relatively high energy values in the low part of the spectrum (below 1 kHz) that typically correspond to formants. In contrast the spectrogram for each type of music can be extremely different. See Figure 2 below. It could be very similar to the speech like in Figure 2.a, especially if the music is accompanied with a voice (a song). But even in this case it could also be extremely different from the speech spectrogram as we can see in Figure 2.b and 2.c.

However, although these different features, the effective discrimination between speech and music is still an open problem. Those examples illustrate possible problems in the speech an music differentiation. In order to clearly differentiate speech and music we need to include temporal characteristics of the signal. From the spectrograms presented here we can see that for some type of music (see for example figure 2.a) we can make a conclusion just only after analysing several seconds of the signal (1, 2 or even more). That could be accomplished by several different methods for segmenting acoustic patterns. For example, autoregressive and autoregressive moving average models (ARMA) (See [1]) or the 'segment evaluation function' (See [2]) can be used for solving this problem.

This work pays more attention to the rhythmical properties of the signal. The main problem here is how to exactly define the border between speech and music. We can consider a song like a music but in the same time we should interpret a voice on the music background (usually in headlines in the beginning of the news flash) like speech. This work analyses the rhythmical property of the signal by means of computing the modulation spectrum for some subband. The results show that those rhythmical properties of the signal are quite different for speech and music. This fact was used in the method that is described below.

3 Speech and Music Recognition by Humans.

A central result from the study of human speech perception is the importance of slow changes in the speech spectrum. These changes appear as low-frequency amplitude modulations with rates of below 16 Hz in subband signals following spectral analysis. The first evidence for this perspective emerged from the development of the channel vocoder in the early 1930s (Dudley, 1939). Direct perceptual experiments have shown that modulations at rates above 16 Hz are not required, and that significant intelligibility remains even if modulations at rates 6 Hz and below are the only preserved. It is interesting that the human auditory system is most sensitive to modulation frequencies around 4 Hz, that correspond to the average syllable rate. Now this property is widely used for improving the quality of ASR systems. For example the robustness of ASR systems could be enhanced by using long-time information, both at the level of the front-end speech representation, and at the level of phonetic classification [3]. The syllable-based recogniser can be built using modulation spectrogram features for the front-end speech representation.

4 Method Description.

4.1 Method.

The method developed here to perform speech and music discrimination is based on the following general ideas:

1. The regular spectral analysis based on 30 ms window, shifted by 10 ms;
2. The computation of a long-time average modulation spectrum for speech and music;
3. Gaussian estimation for the components of modulation spectrum for speech and music;
4. Making a choice between speech and music for the test data by means of computing the closest Gaussian (for speech or music) to the modulation spectrum of speech or music.

The modulation spectrum of the incoming signal can be obtained by the spectral analysis of the temporal trajectory of a power spectral components in the following way (see Figure 3):

- The incoming signal, sampled at 16 kHz, is analysed into the one of the critical subbands: at first the short-time Fourier transform (STFT) or spectrogram is computed. The Hamming window is used to compute FFT over 512 points (30ms) and the segment is shifted every 10 ms (with frequency 100Hz) in order to capture the dynamic properties of the signal. As a result every 10 ms. we have 256-dimensional fft magnitude vector.
- The mel-scale transformation is applied to the magnitude vector. The mel-scale transformation designed to approximate the frequency resolution of the human ear is linear up to 1000 Hz and logarithmic thereafter (for detailed description see mfcc in the Appendix 1). The output is a mel-scaled vector consisting of 40 components.

- These computations are made over approximately 30 minutes of incoming data, then one subband is chosen (in this experiment we have taken 2 subbands corresponding to 6-th/466-600Hz and 20-th/1510-1732Hz components of a mel-scaled vector). The result is a sequence of energy magnitudes for the chosen subband sampled at 100 Hz.
- The modulations of the normalised envelope signal are analysed by computing the FFT over the 256 Hamming window (that corresponds to 2.56 sec) for the sequence of energy magnitudes for a given subband. The FFT is computed every 100 msec (with a shift of 1 point). The result is a sequence of 128-dimensional modulation vectors (let's denote it m_n(i), where i=1..128, and n = sequence number). Those vectors present the modulation frequencies of the energy for the given subband.

After completing these computations we can fit a set of Gaussians to the sequence of modulation vectors where the mean and the variance for each Gaussian are given by the following formulas:

$$\mu(i) = \frac{\sum_{n=1}^{N} m_n(i)}{N} \tag{1}$$

$$\sigma(i)^2 = \frac{\sum_{n=1}^{N}(m_n(i) - \mu(i))^2}{N} \tag{2}$$

Here i=1..128 and N is the length of a training sequence. We apply this training procedure for the speech and music data. This gives us 4 vectors: $\mu_{speech}(i)$, $\sigma_{speech}(i)$, $\mu_{music}(i)$, $\sigma_{music}(i)$. On the figure 4 you can see the mean and deviation values for speech and music, computed for the subband 6 (466-600Hz).The red solid line represents the mean and the dash blue line represents the variance of the energy magnitude in the frequency domain.

music in the frequency domain. of the energy magnitude in the frequency domain). ——

These images show that there is quite a big difference in energy modulation for speech and music. The typical feature for the speech modulation spectrum is a wide peak at frequencies from 2 to 6 Hz. For the music the narrow peak with frequencies below 1 Hz is more typical. Also the experiment has shown quite a big difference for the height of these peaks. In the log10 scale it is 0.9 and 0.6 for speech and music respectively. This result makes it possible to build an automatic differentiator between speech and music.

In order to make a speech-music discrimination test for the input signal, we can compute the modulation spectrum vector y of that signal in a given time interval. Comparing this vector with the modulation spectrum of speech and music the decision could be made. We choose the closest (speech or music) spectrum vector to the computed one.

For computing the modulation spectrum vector y we have applied almost the same procedure that was described above. The only difference is that we computed the modulation spectrum average just only for 1 second:

$$y(i) = \frac{\sum_{n=1}^{10} m_n(i)}{10} \tag{3}$$

Here i=1..128 and n=1..10 that corresponds exactly to 1 second. For computing each m_n value we have used 2.56 seconds Hamming window of energy magnitudes. So we can see that this method requires 1+2.56=3.56 seconds of sound data in order to make the choice between speech and music.

Using this 1 second average value of the modulation spectrum y we can compute the probability of the given signal being music or speech in the following way:

$$p_{speech} = \prod_{i=3}^{25} N(y(i), \mu(i)^{speech}, \sigma(i)^{speech}) \tag{4}$$

$$p_{music} = \prod_{i=3}^{25} N(y(i), \mu(i)^{music}, \sigma(i)^{music}), \tag{5}$$

where:

$$N(M, \mu, \sigma) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp(-\frac{(\mu - M)^2}{2\sigma^2}), \tag{6}$$

For computing those values we have used just only 22 (i=3..25) components of modulation spectrum vectors. These components correspond to the modulation frequencies from 1 to 10 Hz, where there is a most evident difference between speech and music. The final conclusion about the nature of the data fragment is made by comparing those two probability values: P_{speech} and P_{music}.

4.2 Training.

4.2.1 Speech.

Training over the speech in this experiment was made using the data from broadcast news. 24 minutes of speech were used from the file:

rsr_news_980309_1230_1305.wav. This file was previously manually labelled into 5 categories: speech, music, music+speech, noise and pause. Only speech fragments with some noise and pause (not more than 1 second) were used for training of Gaussian parameters.

4.2.2 Music.

Training over the music was made using the data on the CD (THISL-MUSIC). First 100 files were used to compute Gaussian parameters (in the directory /data/music0000). The overall time of the training including all 100 files consisted of 25 minutes.

5 Test.

The test consisted of a lot of different experiments over the different data sets. Each experiment was the discrimination test between speech and music for the signal 3.56 seconds in length. The test interval was moved by 3 seconds for each experiment. So the test intervals were overlapped by 0.56 second. Test experiments were made separately on the music data and speech data. For each experiment a conclusion about the correctness of the algorithm was made. The results of the correctness test is presented below in the form of tables.

5.1 Discrimination test on the training data

.

Results for the discrimination test on the data that has been used for training (0000 Directory for music and rsr_news_980309_1230_1305.wav file for speech) are presented here. The number of the intervals where the test has shown correct results as well as incorrect results and its fractions are presented in the following table.

Table 1. Band: 466-600 Hz

Experiment	Correct	Incorrect
Music		
(0000 Directory)	478 (95.6%)	22 (4.4%)
Speech		
file 980309/180-840 sec/	216 (98.18%)	3 (1.36%)
file 980309/855-1674 sec/	260 (95.24%)	13 (4.76%)

5.2 Testing over the different data.

In this section we present test results for data different from those we have used for training of Gaussian parameters. For the testing experiment we have used two different subbands (466-600 Hz and 1510-1732 Hz) which can allow to compare results between them.

5.2.1 Music.

The music part was tested on the data from 4 directories on the CD (THISL-MUSIC). The data presented there consist of music of different types (rock music, pop music, classical music, hard music etc.) and are similar to the music we have used for the training. The number as well as the fraction of correctly and incorrectly recognised segments are presented here.

Table 1. Band: 1510-1732 Hz

Experiment	Correct	Incorrect
1 (0100 Directory)	468 (93.6%)	32 (6.4%)
2 (0200 Directory)	464 (92.8%)	36 (7.2%)
3 (0300 Directory)	475 (95%)	25 (5%)
4 (0400 Directory)	471 (94.2%)	29 (5.8%)

Table 1. Band: 466-600 Hz

Experiment	Correct	Incorrect
5 (0100 Directory)	471 (94.2%)	29 (5.8%)
6 (0200 Directory)	479 (95.8%)	21 (4.2%)
7 (0300 Directory)	477 (95.4%)	23 (4.6%)
8 (0400 Directory)	476 (95.2%)	24 (4.8%)

5.2.2 Speech.

The error rate for speech was tested on the data from the file rsr_news_980312_1230_1305.wav, that contains broadcast news that are similar to the data we have used for the training. For this experiments only the data consisting of speech information have been used. The number and the fraction of correctly and incorrectly recognised segments are presented here.

Table 2. Band: 1510-1732 Hz

Experiment	N. of tests	Correct	Incorrect
11. file 312 /342-1146 sec./	268	265 (98.88%)	3 (1.12%)
12. file 312 /1155-1473 sec./	106	65 (61.35%)	41(38.65%)
13 file 312 /1497-1788 sec./	97	61(62.89%)	36 (37.11%)

Table 2. Band: 466-600 Hz

Experiment	N. of tests	Correct	Incorrect
14 file: 312 /342-1146 sec./	268	266(99.25%)	2 (0.75%)
15 file: 312 /1155-1473 sec./	106	69(65.09%)	37 (34.91%)
16 file: 312 /1497-1788 sec./	97	67(69.07%)	30 (30.93%)

5.2.3 Discrimination test using two bands.

In this section we present results of the discrimination experiment using two bands simultaneously (band 6: 466-600 Hz and band20: 1510-1732 Hz). After computing $P_{speech}(Band6)$, $P_{music}(Band6)$ and $P_{speech}(Band20)$, $P_{music}(Band20)$ (see section 3) we can compute P_{speech} and P_{music} in the following way:

$$P_{speech} = P_{speech}(Band6) * P_{speech}(Band20) \tag{7}$$

$$P_{music} = P_{music}(Band6) * P_{music}(Band20) \tag{8}$$

The final conclusion about the nature of the data fragment is made by comparing those two probability values: P_{speech} and P_{music}.

Table 3. Two bands, Music.

Experiment	Correct	Incorrect
17 (0100 Directory)	474 (94.8%)	26 (5.2%)
18 (0200 Directory)	479 (95.8%)	21 (4.2%)

Table 3. Two bands, Speech.

Experiment	N. of tests	Correct	Incorrect
19 file: 312 /342-1146 sec./	268	266(99.25%)	2 (0.75%)
20 file: 312 /1155-1473 sec./	106	67(63.21%)	39 (36.79%)
21 file: 312 /1497-1788 sec./	97	61(62.89%)	36 (37.11%)

5.3 Analysis of results.

The results of experiments show that error rate for the discrimination between speech and music could vary greatly (from 98% of correct recognition to 62%). This can be explained by the big difference in speech and music test data sets.

In our experiments the quality (as well as styles) of the music for training and testing was pretty similar. The music was a high-quality studio music. As a result the recognition rate varies just a little bit somewhere around 95%. See Experiments 1-8.

In the contrast, for the training over speech we have used the data captured from the news (which means it could vary greatly depending on the reporter and the place of reporting). But mainly it consisted of the studio reporters speech (good quality and pronunciation, approximately the same speaking rate). In the experiments 11, 14 and 19 such a 'good quality' speech was used for testing. For these experiments the error rate was less than 5%.

We have several other tests (not presented in tables) on the same quality data with approximately the same error rates.

In the same time the error rate is much higher for the experiments 12-13, 15-16 and 20-21. The data for these tests was made up of the interviews from theatre and included several parts of some theatre performances. The style of speech is absolutely different from that which has been used for training. The rhythm of the speech is much slower and the style of speech sometimes is close to reading of poems. These experiments show that the method is highly sensible to the rhythm of the speech. If one would just simply slowly count 1,2,3,4,..., the method could recognise that style of speech like music.

Considering the experiments presented here we can make a conclusion about the nature of the major errors in the discrimination process:

- A noise could have a great influence on the recognition level - a noise would be more probably recognised like music.
- Rap music with fast words can be interpreted like a speech: Example: file 0217 - Music Value=0.000018 Speech Value = 0.000001
- If the percussion are too loud and it frequency is around 2-4 Hz, this type of music could be recognised like a speech. Example: file mus0217 - Music Value: 0.000003, Speech Value: 0.021271
- The different rhythm of speech could greatly increase the error rate. The system has been trained over the data taken from the news. In the example 12 the style of the speech is a completely different. That is an interview in the theatre that includes some performance parts, reading poems and some noise typical to the theatre. Let's notice that a poem or a song (even without music) is more likely music than speech in this case.

In this work we considered experiments with two different bands. The results of this experiments show that the discrimination error rate changes just a little bit from one band to the other (see Table 1 and Table 2) and even for the discrimination test where we have used two bands simultaneously the error rate is still approximately the same (see Table 1, Table 2 and Table 3).

6 Conclusion.

In this work we developed an algorithm for speech and music differentiation based on the analysis of the speech and music modulation spectrum. Speech and music modulation spectrum for given subband has been computed and analysed here. Speech modulation spectrum has a typical wide peak at frequencies from 2 to 6 Hz and the music modulation spectrum has the narrow peak with frequencies below 1 Hz. That difference has been used in the speech and music discrimination method that was presented here.

From physical point of view this difference is cased by different energy changing for speech and music data. The typical rate of speech energy changing corresponds to the average syllable rate (around 4 Hz) and the rate of music energy changing corresponds to the beat rate (around 0.7 Hz).

The method presented here has the following advantages. It is quite simple and effective (the error rate is less then 10% for clean speech and music data). The method can detect the mixture of speech and music (where the probabilities of music and speech are relatively small and equal).

In the same time the error rate greatly depends on the style of music and speech. Even for the same speaker the method can give different results depending on the speed of speech and the 'melody' or rhythm of the speech. For example, this method could recognise poem reading like a music and in the same time fast rap music could be recognised like speech.

We can see the following possible improvements for the method presented here. The number of bands can be increased - that could probably give better results. In the same time the band width and their grouping into sub-bands should also be considered. In this work we have done training just only for the good quality speech of studio reporters. Training on the data with different speech styles and speech quality can probably give some improvements. It could also be interesting to train speech parameters of the system on poems and music parameters separately on different music styles. In this case we can use Multi-Gaussians instead of Gausians with μ and σ defined for each style of speech or music.

7 Appendix 1. Description of main functions used in this work.

This appendix presents the description of main functions used in this work and various examples. All experiments presented above were made using Matlab V5.2. under SUN Solaris 2.6. Here you can find the description of the following functions:

- test - performs the discrimination test between speech and music
- music_fft - computes mean and variance for music for a given band

- speech_fft(filename,band,start,end) - computes mean and variance for speech for a given band
- mfcc - computes: the mel-frequency cepstral coefficients (ceps), detailed fft magnitude (a signal spectrogram), the mel-scale filter bank output and the smooth frequency response;

Files:

- music_20 - contains mean and variance for the music modulation spectrum (for the frequency the band 20 - 1510-1732 Hz);
- music_6 - contains mean and variance for the music modulation spectrum (for the frequency the band 6 - 466-600 Hz);
- speech_20 - contains mean and variance for the speech modulation spectrum (for the frequency the band 20 - 466-600 Hz);
- speech_6 - contains mean and variance for the speech modulation spectrum (for the frequency the band 6 - 466-600 Hz);

7.1 mfcc

This function is a part of 'Auditory Toolbox' (see [5]) and has been used in this work for computing the mel-scale filter bank (fb) output.

[ceps,freqresp,fb,freqrecon] = mfcc(input, samplingRate).

7.1.1 Description

. Find the mel-frequency cepstral coefficients (ceps) corresponding to the input. Three other quantities are optionally returned that represent the detailed FFT magnitude (freqresp), the log 10 mel-scale filter bank output (fb), and the reconstruction of the filter bank output by inverting the cosine transform. The sequence of processing includes for each chunk of data:

- Window the data with a hamming window,
- Shift it into FFT order,
- Find the magnitude of the FFT,
- Convert the FFT data into filter bank outputs,
- Find the log base 10,

- Find the cosine transform to reduce dimensionality.

The outputs from this routine are the MFCC coefficients and several optional intermediate results and inverse results.

- freqresp the detailed fft magnitude used in MFCC calculation, 256 rows.
- fb the mel-scale filter bank output, 40 rows.
- fbrecon the filter bank output found by inverting the cepstrals with a cosine transform, 40 rows.
- freqrecon the smooth frequency response by interpolating the fb reconstruction, 256 channels to match the original freqresp.

This version is improved over the version in Release 1 in a number of ways. The discrete-cosine transform was fixed and the reconstructions have been added.

The filter bank is constructed using 13 linearly-spaced filters (133.33Hz between centre frequencies,) followed by 27 log-spaced filters (separated by a factor of 1.0711703 in frequency.)

7.1.2 Examples

Here is the result of calculating the cepstral coefficients of the ’A huge tapestry hung in her hallway’ utterance from the TIMIT database (TRAIN/DR5/FCDR1/SX106/ SX106.ADC). The utterance is 50189 samples long at 16kHz, and all pictures are sampled at 100Hz and there are 312 frames. Note, the top row of the mfcc-cepstrum, ceps(1,:), is known as C 0 and is a function of the power in the signal.Since the wave-form in our work is normalised to be between -1 and 1, the C 0 coefficients are all negative. The other coefficients, C 1 -C 12 , are generally zero-mean.

```
tap = wavread(’tapestry.wav’);
[ceps,freqresp,fb,fbrecon,freqrecon]= ... mfcc(tap,16000,100);
imagesc(ceps); colormap(1-gray);
```

After combining several FFT channels into a single mel-scale channel, the result is the filter bank output.

```
imagesc(flipud(fb));
```

7.2 test

This function performs the discrimination test between speech and music.

```
output = test(filename,band,start_sec);
```

It takes a segment of a signal from the file 'filename' starting from the point 'start_sec' and performs the discrimination test. Required segment length for distinguishing between speech and music is equal to 3.6 seconds. 'band' argument shows the band that is used to generate modulation spectrum of the signal. In our experiments we have used just only 2 different bands: number 6 (466-600 Hz) number 20 (1510-1732 Hz). This function requires 2 files - music_band_number.mat and speech_band_number.mat to be in the current directory. In our case we have files: music_20.mat, speech_20.mat, music_6.mat, speech_6.mat. The output of the function is equal to 0 if the signal is more likely music than speech and equal to 1 otherwise.

7.2.1 Examples

Here is the example of the discrimination test of the signal taken from the file '/tmp/rsr_news_980313_1230_1305.wav' for the band 6 (466-600 Hz). The test segment starts on the 927-th seconds and lasts 3.6 seconds.

```
test('/tmp/rsr_news_980313_1230_1305.wav',6,927)
```

```
Computing fft ...This is a music
Music: 0.003215, Speech: 0.000000
ans = 0
```

The output of the function is equal to 0 that corresponds to music. This function is also printing the probabilities of the segment being music and speech.

7.3 music_fft

This function computes mean and variance of the signal modulation spectrum for music files in the directory /mus0000 (THISL-MUSIC CD) for a given band.

```
[Mean, Deviation]=music_fft(band);
```

The output of the function presents two 256-dimensional arrays: Mean and Deviation.

7.3.1 Examples

The following example computes the mean and the deviation of signal modulation spectrum for 5 files in the directory /mus0000. The results are saved then in the file 'music_6.mat' for further using in the 'test' function. The number of files for computing the Gaussian parameters is an internal parameter of the function and by default is equal to 100 (that corresponds to 30 minutes of music).

```
[Mean,Disp] = music_fft(6);
```

```
————————- Mean Computing ——————
file 1 : Opening file ...Done.
file 2 : Opening file ...Done.
file 3 : Opening file ...Done.
file 4 : Opening file ...Done.
file 5 : Opening file ...Done.
————————- Deviation Computing ——————
file 1 : Opening file ...Done.
file 2 : Opening file ...Done.
file 3 : Opening file ...Done.
file 4 : Opening file ...Done.
file 5 : Opening file ...Done.
```

```
save music_6;
```

7.4 speech_fft

This function computes mean and variance of the signal modulation spectrum for the speech taken from the file 'filename'.

```
[Mean, Deviation]=speech_fft(filename,band,start,end);
```

'band' represents the band the modulation spectrum is computed for, 'start' and 'end' represent the staring and ending points (given in seconds) of the speech segment. The output of the function presents two 256-dimensional arrays: Mean and Deviation.

7.4.1 Examples

The following example computes the mean and the deviation of signal modulation spectrum of speech. This 30 seconds speech segment is taken from the file 'rsr_news_980313_1230_1305.wav'. The results are saved then in the file 'speech_6.mat' for further using them in the 'test' function.

```
[Mean,Disp]=speech_fft('/tmp/rsr_news_980313_1230_1305.wav',6,60,90);
```

```
——————- Mean Computing —————
1 seconds of 30:Reading data ...Computing fft ...Done .
16 seconds of 30:Reading data ...Computing fft ...Done .
——————- Deviation Computing —————
1 seconds of 30:Reading data ...Computing fft ...Done .
16 seconds of 30:Reading data ...Computing fft ...Done .
```

```
save speech_6;
```

7.5 Deviation_fft

Deviation_fft(input,band,start_point,Mean) - computes the standard deviation for the signal that is given by input.

band - the band number;
start_point - should be equal to 1;
Mean - the mean value. The deviation is computed with respect to this value.

This function was used in the music_fft.m for computing the variance of music modulation spectrum in the following way:

```
Local_Deviation(i,:) = Deviation_fft(fb,band,1,Mean);
```

References

[1] Michele Basseville, Albert Benveniste, 'Sequential Detection of Abrupt Changes in Spectral Characteristics of Digital Signals', *IEEE International Conference on Acoustic, Speech, and Signal Processing*, NO 5, 5 September 1983.

[2] John S. Bridle, Nigel C. Sedgwick 'A Method for Segmenting Acoustic Patterns, with Applications to Automatic Speech Recognition', *Transactions on Automatic Control*, May 9-11, 1977.

[3] Brian E.D. Kingsbury, Nelson Morgan, Steven Greenberg, 'Robust speech recognition using the modulation spectrogram', *Speech communications*, 25 (1998), pp. 117-132.

[4] Steven Greenberg, Brian Kingsbury, 'The modulation spectrogram: in pursuit of an invariant representation of speech', *IEEE International Conference on Acoustic, Speech, and Signal Processing, Volume III* 1997

[5] Malcolm Slaney, Auditory Toolbox (Version 2), Technical Report #1998-010, Interval Research Corporation http://web.interval.com/papers/1998-010/