EVALUATION OF MFCC ESTIMATION TECHNIQUES FOR MUSIC SIMILARITY

Jesper Højvang Jensen[1], *Mads Græsbøll Christensen*[1], *Manohar N. Murthi*[2], *and Søren Holdt Jensen*[1]

[1] Department of Communication Technology, Aalborg University
Fredrik Bajers Vej 7A-3, DK-9220 Aalborg, Denmark
email: {jhj, mgc, shj}@kom.aau.dk
[2] Department of Electrical and Computer Engineering, University of Miami,
1251 Memorial Dr., Coral Gables, FL. 33124-0640 USA
email: mmurthi@miami.edu

ABSTRACT

Spectral envelope parameters in the form of mel-frequency cepstral coefficients are often used for capturing timbral information of music signals in connection with genre classification applications. In this paper, we evaluate mel-frequency cepstral coefficient (MFCC) estimation techniques, namely the classical FFT and linear prediction based implementations and an implementation based on the more recent MVDR spectral estimator. The performance of these methods are evaluated in genre classification using a probabilistic classifier based on Gaussian Mixture models. MFCCs based on fixed order, signal independent linear prediction and MVDR spectral estimators did not exhibit any statistically significant improvement over MFCCs based on the simpler FFT.

1. INTRODUCTION

Recently, the field of music similarity has received much attention. As people convert their music collections to mp3 and similar formats, and store thousands of songs on their personal computers, efficient tools for navigating these collections have become necessary. Most navigation tools are based on metadata, such as artist, album, title, etc. However, there is an increasing desire to browse audio collections in a more flexible way. A suitable distance measure based on the sampled audio signal would allow one to go beyond the limitations of human-provided metadata. A suitable distance measure should ideally capture instrumentation, vocal, melody, rhythm, etc. Since it is a non-trivial task to identify and quantify the instrumentation and vocal, a popular alternative is to capture the timbre [1, 2, 3]. Timbre is defined as "the auditory sensation in terms of which a listener can judge that two sounds with same loudness and pitch are dissimilar" [4]. The timbre is expected to depend heavily on the instrumentation and the vocals. In many cases, the timbre can be accurately characterized by the spectral envelope. Extracting the timbre is therefore similar to the problem of extracting the vocal tract transfer function in speech recognition. In both cases, the spectral envelope is to be estimated while minimizing the influence of indivual sinusoids.

In speech recognition, mel-frequency cepstral coefficients (MFCCs) are a widespread method for describing the vocal tract transfer function [5]. Since timbre similarity and estimating the vocal tract transfer function are closely related, it is no surprise that MFCCs have also proven successful in the field of music similarity [1, 2, 3, 6]. In calculating the MFCCs, it is necessary to estimate the magnitude spectrum of an audio frame. In the speech recognition community, it has been customary to use either fast Fourier transform (FFT) or linear prediction (LP) analysis to estimate the frequency spectrum. However, both methods do have some drawbacks. Minimum variance distortionless response (MVDR) spectral estimation has been proposed as an alternative to FFT and LP analysis [7, 8]. According to [9, 10], this increases speech recognition rates.

This research was supported by the Intelligent Sound project, Danish Technical Research Council grant no. 26–02–0092.

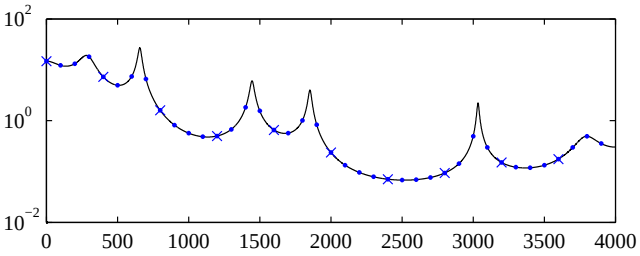

Figure 1: Spectrum of the signal that is excited by impulse trains in Figure 3. Dots denote multiples of 100 Hz, and crosses denote multiples of 400 Hz.

In this paper, we compare MVDR to FFT and LP analysis in the context of music similarity. For each song in a collection, MFCCs are computed and a Gaussian mixture model is trained. The models are used to estimate the genre of each song, assuming that similar songs share the same genre. We perform this for different spectrum estimators and evaluate their performance by the computed genre classification accuracies.

The outline of this paper is as follows. In Section 2, we summarize how MFCCs are calculated, what the shortcomings of the FFT and LP analysis as spectral estimators are, the idea of MVDR spectral estimation, and the advantage of prewarping. Section 3 describes how genre classification is used to evaluate the spectral estimation techniques. In Section 4, we present the results, and in Section 5, the conclusion is stated.

2. SPECTRAL ESTIMATION TECHNIQUES

In the following descriptions of spectrum estimators, the spectral envelope in Figure 1 is taken as starting point. When a signal with this spectrum is excited by an impulse train, the spectrum becomes a line spectrum that is non-zero only at multiples of the fundamental frequency. The problem

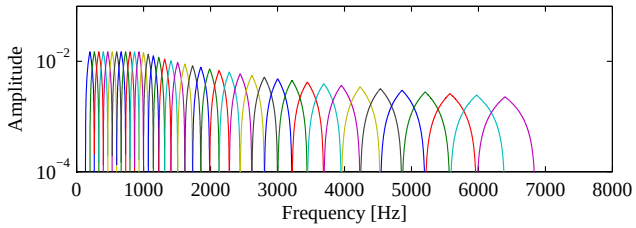

Figure 2: Mel bands

is to estimate the spectral envelope from the observed line spectrum. Before looking at spectrum estimation techniques, we briefly describe the application, i.e. estimation of mel-frequency cepstral coefficients.

2.1 Mel-Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients attempt to capture the perceptually most important parts of the spectral envelope of audio signals. They are calculated in the following way [11]:

1. Calculate the frequency spectrum
2. Split the magnitude spectrum into a number of bandpass filters (40 bands are often used) according to the mel-scale, such that low frequencies are given more weight than high frequencies. In Figure 2, the bandpass filters that are used in [11] are shown. We have used the same filters.
3. Sum the frequency contents of each band.
4. Take the logarithm of each sum.
5. Compute the discrete cosine transform (DCT) of the logarithms.

The first step reflects that the ear is fairly insensitive to phase information. The averaging in the second and third steps reflect the frequency selectivity of the human ear, and the fourth step simulates the perception of loudness. Unlike the other steps, the fifth step is not directly related to human sound perception, since its purpose is to decorrelate the inputs and reduce the dimensionality.

2.2 Fast Fourier Transform

The fast Fourier transform (FFT) is the Swiss army knife of digital signal processing. In the context of speech recognition, its caveat is that it does not attempt to suppress the effect of the fundamental frequency and the harmonics. In Figure 3, the magnitude of the FFT of a line spectrum based on the spectral envelope in Figure 1 is shown. The problem is most apparent for high fundamental frequencies.

2.3 Linear Prediction Analysis

LP analysis finds the spectral envelope under the assumption that the excitation signal is white. For voiced speech with a high fundamental frequency, this is not a good approximation. Assume that $w(n)$ is white, wide sense stationary noise with unity variance that excites a filter having impulse response $h(n)$. Let $x(n)$ be the observed outcome of the process, i.e. $x(n) = w(n) * h(n)$ where $*$ denotes the convolution operator, and let a_1, a_2, …, a_K be the coefficients of the optimal least squares prediction filter. The prediction error, $y(n)$, is then given by

$$y(n) = x(n) - \sum_{k=1}^{K} a_k x(n-k). \tag{1}$$

The orthogonality principle says that $y(n)$ will be uncorrelated with the inputs of the prediction filter, $x(n-K), \ldots, x(n-1)$. If furthermore K is chosen so large that $\mathrm{E}\left[x(n)x(n-k')\right] \approx 0$ for $k' > K$ (it is often assumed that $h(n)$ is an FIR filter), then this implies that $\mathrm{E}\left[y(n)y(n+m)\right] \approx 0$ for $m \neq 0$, since $y(n)$ is a linear combination of the inputs. This means that $y(n)$ will be white. Now, let $A(f)$ be the transfer function of the filter that produces $y(n)$ from $x(n)$, i.e.,

$$A(f) = 1 - \sum_{k=1}^{K} a_k e^{-i2\pi fk}. \tag{2}$$

Moreover, let $H(f)$ be the Fourier transform of $h(n)$, and let $S_x(f)$ and $S_y(f)$ be the power spectra of $x(n)$ and $y(n)$, respectively. Since $y(n)$ is approximately white with variance σ_y^2, i.e. $S_y(f) = \sigma_y^2$, it follows that

$$\begin{aligned} S_y(f) = \sigma_y^2 &= S_x(f)|A(f)|^2 \\ &= S_w(f)|H(f)|^2|A(f)|^2. \end{aligned} \tag{3}$$

Rearranging this, we get

$$\frac{\sigma_y^2}{|A(f)|^2} = S_w(f)|H(f)|^2. \tag{4}$$

The variables on the left side of Equation (4) can all be computed from the autocorrelation function. Thus, when the excitation signal is white with unity variance, i.e. $S_w(f) = 1$, linear prediction (LP) analysis can be used to estimate the transfer function. Unfortunately, the excitation signal is often closer to an impulse train than to white noise. An impulse train with time period T has a spectrum which is an impulse train with period $1/T$. If the fundamental frequency is low, the assumption of a white excitation signal is good, because the impulses are closely spaced in the frequency domain. However, if the fundamental frequency is high, the linear predictor will tend to place zeros such that individual frequencies are nulled, instead of approximating the inverse of the autoregressive filter $h(n)$. This is illustrated in Figure 3, where two spectra with different fundamental frequencies have been estimated by LP analysis.

2.4 Minimum Variance Distortionless Response

Minimum variance distortionless response (MVDR) spectrum estimation has its roots in array processing [7, 8]. Conceptually, the idea is to design a filter $g(n)$ that minimizes the output power under the constraint that a specific frequency has unity gain. Let $\boldsymbol{R}_x$ be the autocorrelation matrix of a stochastic signal $x(n)$, and let $\mathbf{g}$ be a vector representation of $g(n)$. The expected output power of $x(n) * g(n)$ is then equal to $\mathbf{g}^{\mathrm{H}}\boldsymbol{R}_x\mathbf{g}$. Let f be the frequency at which we wish to estimate the power spectrum. Define a steering vector $\mathbf{b}$ as

$$\mathbf{b} = \begin{bmatrix} 1 & e^{-2\pi if} & \ldots & e^{-2\pi iKf} \end{bmatrix}^{\mathrm{T}}. \tag{5}$$

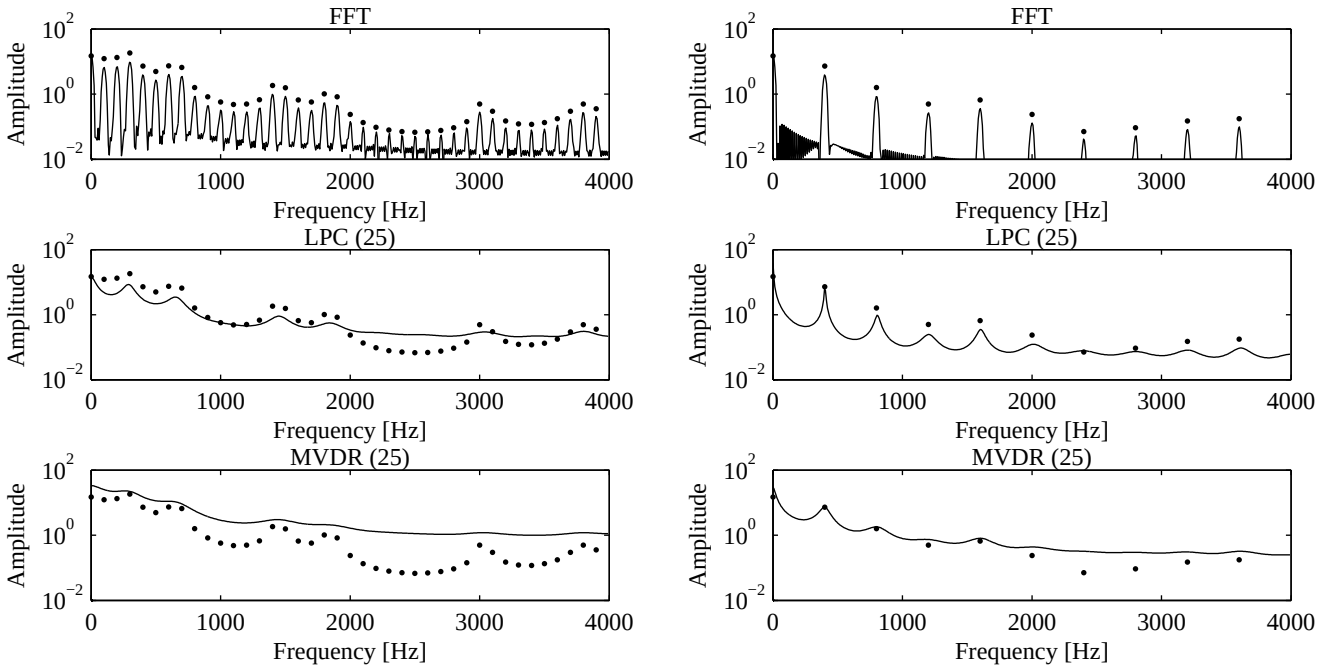

Figure 3: Three different spectral estimators. The dots denote the line spectres that can be observed from the input data. To the left, the fundamental frequency is 100 Hz, and to the right it is 400 Hz.

Compute $\mathbf{g}$ such that the power is minimized under the constraint that $\mathbf{g}$ has unity gain at the frequency f:

$$\mathbf{g} = \arg\min_{\mathbf{g}} \mathbf{g}^{\mathrm{H}} \boldsymbol{R}_x \mathbf{g} \quad \text{s.t.} \quad \mathbf{b}^{\mathrm{H}} \mathbf{g} = 1. \tag{6}$$

The estimated spectral contents, $\widehat{S}_x(f)$, is then given by the output power of $x(n) * g(n)$:

$$\widehat{S}_x(f) = \mathbf{g}^{\mathrm{H}} \boldsymbol{R}_x \mathbf{g}. \tag{7}$$

It turns out that (7) can be reduced to the following expression [7, 8]:

$$\widehat{S}_x(f) = \frac{1}{\mathbf{b}^{\mathrm{H}} \boldsymbol{R}_x^{-1} \mathbf{b}}, \tag{8}$$

In Figure 3, the spectral envelope is estimated using the MVDR technique. Compared to LP analysis with the same model order, the MVDR spectral estimate will be much smoother [12]. In MVDR spectrum estimation, the model order should ideally be chosen such that the filter is able to cancel all but one sinusoid. If the model order is significantly higher, the valleys between the harmonics will start to appear, and if the model order is lower, the bias will be higher [12]. It was reported in [10] that improvements in speech recognition had been obtained by using variable order MVDR. Since it is non-trivial to adapt their approach to music, and since [10] and [13] also have reported improvements with a fixed model order, we use a fixed model order in this work. Using a variable model order with music is a topic of current research.

2.5 Prewarping

All the three spectral estimators described above have in common that they operate on a linear frequency scale. The mel-scale, however, is approximately linear at low frequencies and logarithmic at high frequencies. This means that the mel-scale has much higher frequency resolution at low frequencies than at high frequencies. Prewarping is a technique for approximating a logarithmic frequency scale. It works by replacing all delay elements $z^{-1} = e^{-2\pi i f}$ by the all-pass filter

$$\tilde{z}^{-1} = \frac{e^{-2\pi i f} - \alpha}{1 - \alpha e^{-2\pi i f}}. \tag{9}$$

For a warping parameter $\alpha = 0$, the all-pass filter reduces to an ordinary delay. If α is chosen appropriately, then the warped frequency axis can be a fair approximation to the mel-scale [9, 10]. Prewarping can be applied to both LP analysis and MVDR spectral estimation [9, 10].

3. GENRE CLASSIFICATION

The considerations above are all relevant to speech recognition. Consequently, the use of MVDR for spectrum estimation has increased speech recognition rates [10, 13, 14]. However, it is not obvious whether the same considerations hold for music similarity. For instance, in speech there is only one excitation signal, while in music there may be an excitation signal and a filter for each instrument. In the following we therefore investigate whether MVDR spectrum estimation leads to an improved music similarity measure. Evaluating a music similarity measure directly involves numerous user experiments. Although other means of testing have been proposed, e.g. [15], genre classification is an easy, meaningful method for evaluating music similarity [16]. The underlying assumption is that songs from the same genre are musically similar. For the evaluation, we use the training data from the ISMIR 2004 genre classification contest [17], which contains 729 songs that are classified into 6 gen-

res: classical (320 songs, 40 artists), electronic (115 songs, 30 artists), jazz/blues (26 songs, 5 artists), metal/punk (45 songs, 8 artists), rock/pop (101 songs, 26 artists) and world (122 songs, 19 artists). Inspired by [2] and [3], we perform the following for each song:

1. Extract the MFCCs in windows of 23.2 ms with an overlap of 11.6 ms. Store the first eight coefficients.
2. Train a Gaussian mixture model with 10 mixtures and diagonal covariance matrices.
3. Compute the distance between all combinations of songs.
4. Assume the song has the same genre as the most similar song apart from itself (and optionally apart from songs by the same artist).

We now define the accuracy as the fraction of correctly classified songs. The MFCCs are calculated in many different ways. They are calculated with different spectral estimators: FFT, LP analysis, warped LP analysis, MVDR, and warped MVDR. Except for the FFT, all spectrum estimators have been computed with different model orders. Furthermore, the MFCCs have been calculated both with and without the use of a Hamming window. Before calculating MFCCs, pre-filtering is often applied. In speech processing, pre-filtering is performed to cancel a pole in the excitation signal, which is not completely white as otherwise assumed [5]. In music, a similar line of reasoning cannot be applied since the excitation signal is not as well-defined as in speech due to the diversity of musical instruments. We therefore calculate MFCCs both with and without pre-filtering.

The Gaussian mixture model (GMM) for song l is given by

$$p_l(\mathbf{x}) = \sum_{k=1}^{K} c_k \frac{1}{\sqrt{2\pi\,|\mathbf{\Sigma}_k|}} \exp\Big(-(\mathbf{x}-\mu_k)^{\mathrm{T}}\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\mu_k)\Big), \quad (10)$$

where K is the number of mixtures. The parameters of the GMM, $\mu_1,\ldots,\mu_K$ and $\mathbf{\Sigma}_1,\ldots,\mathbf{\Sigma}_K$, are computed with the k-means-algorithm. The centroids computed with the k-means-algorithm are used as means for the Gaussian mixture components, and the data in the corresponding Voronoi regions are used to compute the covariance matrices. This is often used to initialize the EM-algorithm, which then refines the parameters, but according to [15], and our own experience, there is no significant improvement by subsequent use of the EM-algorithm. The similarity between two songs is computed by comparing their Gaussian mixture models. Let $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$ be the GMMs of two songs, and let $\mathbf{x}_{11},\ldots,\mathbf{x}_{1N}$ and $\mathbf{x}_{21},\ldots,\mathbf{x}_{2N}$ be random vectors drawn from $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$, respectively. We then use the following symmetric distance measure [3]:

$$d = \sum_{n=1}^{N} \big(\log(p_1(\mathbf{x}_{1n})) + \log(p_2(\mathbf{x}_{2n})) - \log(p_1(\mathbf{x}_{2n})) - \log(p_2(\mathbf{x}_{1n}))\big). \quad (11)$$

In our case, we set $N = 200$. When generating the random vectors, we ignore mixtures with weights $c_k < 0.01$ (but not when evaluating equation (11)). This is to ensure that outliers do not influence the result too much. When classifying a song, we either find the most similar song or the most similar song by another artist. According to [2], this has great impact on the classification accuracy. When the most similar song is allowed to be of the same artist, artist identification is performed instead of genre classification.

4. RESULTS

The computed classification accuracies are shown graphically in Figure 4. When the most similar song is allowed to be of the same artist, i.e. songs of the same artist are included in the training set, accuracies are around 80%, and for the case when the same artist is excluded from the training set, accuracies are around 60%. This is consistent with [2], which used the same data set. With a confidence interval of 95%, we are not able to conclude that the fixed order MVDR and LP based methods perform better than the FFT-based methods.

In terms of complexity, the FFT is the winner in most cases. When the model order of the other methods gets high, the calculation of the autocorrelation function is done most efficiently by FFTs. Since this requires both an FFT and an inverse FFT, the LPC and MVDR methods will in most cases be computationally more complex than using the FFT for spectrum estimation. Furthermore, if the autocorrelation matrix is ill-conditioned, the standard Levinson-Durbin algorithm fails, and another approach, such as the pseudoinverse, must be used.

The experiments have been performed both with and without a preemphasis filter. When allowing the most similar song to be of the same artist, a preemphasis filter increased accuracy in 66 out of 74 cases, and it decreased performance in 7 cases. When excluding the same artist, the accuracy was increased in 70 cases and decreased in only 3 cases. However, with a 95% confidence interval, we cannot draw any conclusion.

The improvement by using a Hamming window depends on the spectral estimator. We restrict ourselves to only consider the case with a preemphasis filter, since this practically always resulted in higher accuracies. For this case, we observed that a Hamming window is beneficial in all tests with the LPC, and with most tests using MVDR. There were no significant difference when using the warped variants. Once again, however, we cannot draw any conclusion with a confidence interval of 95%.

5. CONCLUSION

With MFCCs based on fixed order, signal independent LPC, warped LPC, MVDR, or warped MVDR, genre classification tests did not exhibit any statistically significant improvements over FFT-based methods. This means that a potential difference must be minor. Since the other spectral estimators are computationally more complex than the FFT, the FFT is preferable in music similarity applications. There are at least three possible explanations why the results are not statistically significant:

1. The choice of spectral estimator is not important.
2. The test set is too small to show subtle differences.
3. The method of testing is not able to reveal the differences.

The underlying reason is probably a combination of all three. When averaging the spectral contents of each mel-band (see Figure 2), the advantage of the MVDR might be evened out. Although the test set consists of 729 songs, this does not ensure finding statistically significant results. Many of the songs are easily classifiable by all spectrum estimation methods, and some songs are impossible to classify correctly with spectral characteristics only. This might leave only a few songs that actually depend on the spectral envelope estima-

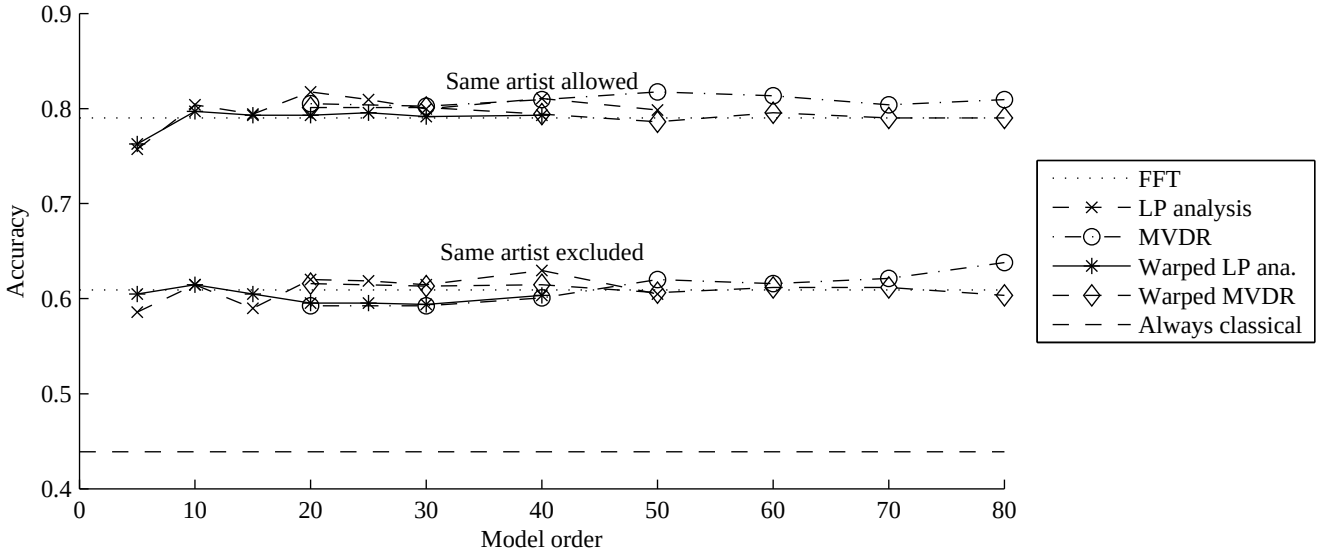

Figure 4: Classification accuracies. All methods are using preemphasis. The FFT, LP analysis and MVDR methods use a Hamming window, while the warped methods use a triangular window.

tion technique. The reason behind the third possibility is that there is not a one-to-one correspondance between timbre, spectral envelope and genre. This uncertainty might render the better spectral envelope estimates useless.

REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Processing*, vol. 10, pp. 293–301, 2002.

[2] A. Flexer, "Statistical evaluation of music information retrieval experiments," Institute of Medical Cybernetics and Artificial Intelligence, Medical University of Vienna, Tech. Rep., 2005.

[3] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: How high's the sky?" *Journal of Negative Results in Speech and Audio Sciences*, 2004.

[4] B. C. J. Moore, *An introduction to the Psychology of Hearing*, 5th ed. Elsevier Academic Press, 2004.

[5] J. John R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, 2nd ed. Wiley-IEEE Press, 1999.

[6] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *Proc. IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, 2001.

[7] M. N. Murthi and B. Rao, "Minimum variance distortionless response (MVDR) modeling of voiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Munich, Germany, April 1997.

[8] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, May 2000.

[9] M. Wölfel, J. McDonough, and A. Waibel, "Warping and scaling of the minimum variance distortionless response," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, November 2003, pp. 387 – 392.

[10] M. Wölfel and J. McDonough, "Minimum variance distortionless response spectral estimation," *IEEE Signal Processing Mag.*, vol. 22, pp. 117 – 126, Sept. 2005.

[11] M. Slaney, "Auditory toolbox version 2," Interval Research Corporation, Tech. Rep., 1998.

[12] M. N. Murthi, "All-pole spectral envelope modeling of speech," Ph.D. dissertation, University of California, San Diego, 1999.

[13] U. H. Yapanel and J. H. L. Hansen, "A new perspective on feature extraction for robust in-vehicle speech recognition," in *European Conf. on Speech Communication and Technology*, 2003.

[14] S. Dharanipragada and B. D. Rao, "MVDR-based feature extraction for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001.

[15] A. Berenzweig, B. Logan, D. Ellis, and B. Whitman, "A large-scale evaluation of acoustic and subjective music similarity measures," in *Proc. Int. Symp. on Music Information Retrieval*, 2003.

[16] T. Li and G. Tzanetakis, "Factors in automatic musical genre classificatin of audio signals," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, 2003.

[17] ISMIR 2004 audio description contest – genre/artist ID classification and artist similarity. [Online]. Available: http://ismir2004.ismir.net/genre_contest/index.htm