Discrete Cepstrum Coefficients as Perceptual Features

Wim D'haes [a,b] * Xavier Rodet [a] †
[a] IRCAM – 1, place Igor-Stravinsky · 75004 Paris · France
[b] Visionlab – University of Antwerp (UA) – Groenenborgerlaan 171 · 2020 Antwerp · Belgium

Abstract

Cepstrum coefficients are widely used as features for both speech and music. In this paper, the use of discrete cepstrum coefficients is considered, which are computed from sinusoidal peaks in the short time spectrum. These coefficients are very interesting as features for pattern recognition applications since they allow to represent spectra by points in a multidimensional vector space. A new Mel frequency warping method is proposed that allows to compute the spectral envelope on the Mel scale which, by contrast to current estimation techniques, does not rely on manually set parameters. Furthermore, the robustness and perceptual relevance of the coefficients are studied and improved.

1 Introduction

In its elementary form, the real cepstrum of a signal is defined as the inverse fourier transform of the log magnitude spectrum. In practical recognition applications however, they are rarely used as features in this form. In the case of speech recognition for example, a filter bank is applied of which the center frequency of each bank is scaled according to the Mel scale. This scale takes into account the frequency resolution properties of the human ear. The inverse fourier transform of the log output of this filter bank yields the *Mel Frequency Cepstrum Coefficients (MFCC)*. Various other cepstrum like coefficients have been proposed and it is believed that further improvement in the front-end of a speech recognition system, i.e. the feature extraction, can be achieved (Molau, Pitz, Schlüter, and Ney 2001; Gu and Rose 2001).

Also in the music domain, cepstrum coefficients have been extensively used in numerous applications such as the retrieval of similar audio tracks (Aucouturier and Pachet 2002), instrument identification (Brown 1999), content based audio retrieval (Foote 1997; Spevak 2002), synthesis (Schwarz and Rodet 1999), and they are currently investigated for automated estimation of control parameters for musical synthesis algorithms (D'haes and Rodet 2001; D'haes and Rodet 2003).

In this work, the characterization of the *spectral envelope* of a nearly periodic sound is studied. The spectral envelope is a function of the frequency that matches the amplitudes of the individual partials in the spectrum. This captures an important aspect of the timbre since it is generally accepted that the relative strength of the amplitudes of the partials allows to distinguish musical instruments and spoken language vowels. However, a strong abstraction is still made and not all perceptually relevant features of the timbre are captured. For example, the noise component is not taken into account and the roughness is often diminished when the analysis window is taken too large. Furthermore, the estimation of the partials is often not accurate at transients.

Different representations of the spectral envelopes have been proposed such as *linear prediction coefficients (LPC)*, the cepstrum and the *discrete cepstrum*. The discrete cepstrum was originally proposed by Gallas and Rodet (Galas and Rodet 1990; Galas and Rodet 1991) and later, a regularized version was developed by Cappé and Oudot (Cappé, Oudot, and Moulines 1997; Campedel-Oudot, Cappé, and Moulines 2001). In the work of Schwarz (Schwarz and Rodet 1999), different spectral envelope representations were studied and compared. There, it was shown that the discrete cepstrum is more suitable for the representation of nearly periodic sounds than LPC or the cepstrum.

2 Discrete Cepstrum Coefficients

2.1 Definition and Computation

P discrete cepstrum coefficients c_p, with $p = 0, \ldots, P-1$ define a magnitude envelope $|H(\omega)|$ of the form

$$|H(\omega)| = \exp\left(c_0 + 2\sum_{p=1}^{P-1} c_p \cos(p\omega)\right) \quad (1)$$

$$c_p = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log(|H(\omega)|)e^{i\omega p}d\omega \quad (2)$$

Since the inverse fourier transform of the log amplitude yields again the coefficients c_p, this definition corresponds with the classic cepstrum definition. Contrary to the classic cepstrum which is computed directly from the spectrum, the discrete coefficients are matched with the individual peaks in the spectrum obtained from an additive analysis (Rodet 1997). A spectrum of this form can be described by a set of partials at frequencies ω_k with amplitudes $\hat{X}_k$ ($k = 1, \ldots K$). This can be written as

$$X(\omega) = \sum_{k=1}^{K} \hat{X}_k \delta(\omega - \omega_k) \quad (3)$$

where $\delta(\omega)$ denotes the Dirac delta distribution. The estimation of the coefficients c_p is realized by minimizing the square difference of the log amplitude envelopes $|H(\omega)|$ and $|X(\omega)|$. This equation

*wdhaes@ruca.ua.ac.be Wim D'haes is financially supported by the Flemish Institute for the Promotion of Innovation by Science and Technology (IWT).

†rod@ircam.fr

defines $|X(\omega)|$ only at the peak frequencies ω_k from which the following square error function $\chi(\mathbf{c})$ can be derived in function of the cepstrum coefficients $\mathbf{c}$.

$$\chi(\mathbf{c}) \quad = \quad \sum_{k=1}^{K} \left(\log(|H(\omega_k)|) - \log(\hat{X}_k)\right)^2 \qquad (4)$$

This is solved easily using a least mean squares procedure which results in a set of linear equations from which the coefficients can be computed.

2.2 Overfitting and Adapting the Order

Since the cepstrum coefficients are computed from a set of linear equations, the computation of P coefficients requires at least an equal number of detected peaks. As can be seen from Fig. 1, *overfitting* occurs when the number of coefficients equals the number of peaks. This can easily be avoided by lowering the number of coefficients. However, when too few coefficients are used, a low pass filtered envelope is obtained that fails to match the peaks accurately.

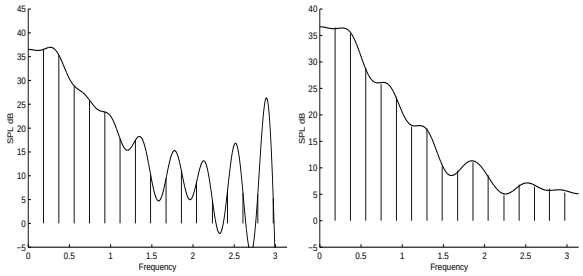

Figure 1: Spectral envelope estimations over a range of 15.000 Hz for a trumpet sound with $f_0 = 886Hz$ using 17 and 14 cepstrum coefficients respectively.

Obviously, for a sound with a lower pitch, more peaks will be detected in the same frequency interval, and as a consequence more coefficients are needed to match them accurately. Note that when the peaks are positioned exactly at multiples of $\frac{\pi}{K}$, with K being the number of peaks, the estimation of the cepstrum coefficients is equivalent to a discrete inverse fourier transform which implies no information loss. Therefore, the number of cepstrum coefficients is scaled with the number of peaks. In addition, two extra control point were added at the interval bounds as was proposed previously in (Galas and Rodet 1991). This yielded a high quality synthesis while overfitting was avoided successfully.

2.3 Why the Discrete Cepstrum ?

Comparing spectral envelopes is very interesting since it is related to the timbral similarity between two short time spectra in a trivial way. The fact that the perceived loudness of a human listener is approximately logarithmic with the signal amplitude suggests that the square difference between the log magnitude spectra can be used to express the perceived similarity. This difference, computed for two spectral envelopes $|H_1(\omega)|$ and $|H_2(\omega)|$ defined by two vectors of cepstrum coefficients $\mathbf{c}_1$ and $\mathbf{c}_2$ respectively, is equivalent to the Euclidean distance between these vectors.

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \left(\log(|H_1(\omega)|) - \log(|H_2(\omega)|)\right)^2 d\omega$$
$$= \quad (\mathbf{c}_1 - \mathbf{c}_2)^T(\mathbf{c}_1 - \mathbf{c}_2) \qquad (5)$$

This shows that the spectral envelopes defined by the discrete cepstrum can be represented by points in a multidimensional vector space where each axis corresponds with a cepstrum coefficient. This is particularly interesting for pattern recognition applications and allows for example the use of K nearest neighbor classification.

A second important property is that the spectral envelope of the sound is relatively independent of its fundamental frequency. This is not the case for other spectral envelope representations which tend to follow the individual peaks (Schwarz and Rodet 1999).

Thirdly, the spectral envelope allows, in combination with the frequencies and phases, the resynthesis of the sound. This plays an important role in a recognition system since it allows to verify to what extent the features are actually representative for the sound. It is an important advantage compared to other features that are frequently used as sound descriptors (Peeters, McAdams, and Herrera 2000). The importance of the avoidance of overfitting should not be underestimated since very similar spectra can produce very different feature vectors because of it.

3 Mel Scaled Discrete Cepstrum

Since the goal of the features is to define a perceptual distance between two envelopes, it is more appropriate to express the envelope on the Mel scale. The monotone and invertible Mel scale warping function $g(\omega) : [0, \pi] \rightarrow [0, \pi]$, converting a linear scale frequency ω to a Mel scale frequency $\bar{\omega}$ is given by

$$g(\omega) = \frac{\pi}{\log(1 + \frac{f_s}{2 \cdot 700Hz})} \log\left(1 + \frac{\omega f_s}{2\pi 700Hz}\right) \qquad (6)$$

according to (Molau, Pitz, Schlüter, and Ney 2001) where f_s denotes the sampling frequency.

3.1 Regularization

Analogue to the MFCC's used in speech, Galas and Rodet proposed the *discrete MFCC's* (Galas and Rodet 1990; Galas and Rodet 1991) which are computed by first warping the peaks on the Mel scale and then computing the envelope over these peaks. The disadvantage of this technique is that after the warping, all high frequency peaks are positioned closer to one another than the low frequency peaks. As a result, the high frequency peaks predominate the estimation resulting consistently in overfitted envelopes. The solution that was proposed consisted of introducing to each observation a cluster of neighboring points which yields satisfying results in many cases but increases the numerical complexity and depends on the initial choice and number of points. Cappé and Oudot proposed to cope with the ill-posed nature of the problem by adding a penalty function (Cappé, Oudot, and Moulines 1997)

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \left[\frac{\partial}{\partial\bar{\omega}} \log(|H(\bar{\omega})|)\right]^2 d\bar{\omega} \qquad (7)$$

to the error criterion given in Eq. (4). This penalty function is multiplied by a regularization parameter λ controlling the relative importance of the smoothness of the envelope versus the exactness of the

envelope fit. Regularization and cloud smoothing were also combined to obtain smooth envelopes which can be controlled locally by adding additional points (Schwarz and Rodet 1999).

3.2 Posterior Warping

The techniques described in the previous subsection convert the peaks to the Mel frequency scale before the envelope is estimated, what we named *prior warping*. In addition, the envelope depends on parameters that need to be set manually by the user which have a large influence on the exactness and smoothness of the fit. As stated in section 2.2, it is rather easy to obtain a spectral envelope on the linear scale that is at the same time accurate and smooth by automatically adapting the number of used cepstrum coefficients to the number of peaks. This led to the idea of first estimating the envelope on the linear scale and computing the warping from the linear scale cepstrum coefficients a posteriori.

A spectral envelope on the Mel scale $\bar{\omega}$ defined by the Mel scaled cepstrum coefficients $\mathbf{d}$ is given by

$$|G(\bar{\omega})| = \exp\left(d_0 + 2\sum_{p=1}^{P-1} d_p \cos(p\bar{\omega})\right) \tag{8}$$

We show that Mel scale coefficients $\mathbf{d}$ can be computed directly from the linear scale coefficients $\mathbf{c}$ defining an envelope $|H(\omega)|$ on the linear frequency scale (see Eq. (1)). The computation of the coefficients $\mathbf{d}$ from the warped linear envelope, given by $|H(\bar{g}^{-1}(\bar{\omega})|$, results in

$$\begin{aligned} d_k &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \log(|H(g^{-1}(\bar{\omega})|)e^{j\bar{\omega}k}d\bar{\omega} \\ &= \sum_{p=0}^{P-1} c_p \frac{2-\delta_{0p}}{\pi}\int_0^{\pi} \cos\left(pg^{-1}(\bar{\omega})\right)\,\cos(\bar{\omega}k)d\bar{\omega} \end{aligned} \tag{9}$$

where δ_{0p} denotes the Kronecker symbol. Note that this is equivalent to the minimization of the error between the Mel scale log envelopes in function of $\mathbf{d}$ given by

$$\chi(\mathbf{d}) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\log(H(g^{-1}(\bar{\omega}))) - \log(G(\bar{\omega}))\right]^2 d\bar{\omega} \tag{10}$$

Eq. (9) shows that a Mel scale coefficient can be computed from a linear combination of linear scale coefficients what can be written as a matrix multiplication

$$\mathbf{d} = \mathbf{A}\mathbf{c} \tag{11}$$

with

$$\begin{aligned} A_{k+1,l+1} &= \frac{2-\delta_{0l}}{\pi}\int_0^{\pi}\cos\left(lg^{-1}(\bar{\omega})\right)\,\cos(k\bar{\omega})d\bar{\omega} \\ &\cong \frac{2-\delta_{0l}}{N}\sum_{n=0}^{N-1}\cos\left(lg^{-1}\left(\frac{\pi n}{N}\right)\right)\cos\left(\frac{\pi nk}{N}\right) \end{aligned} \tag{12}$$

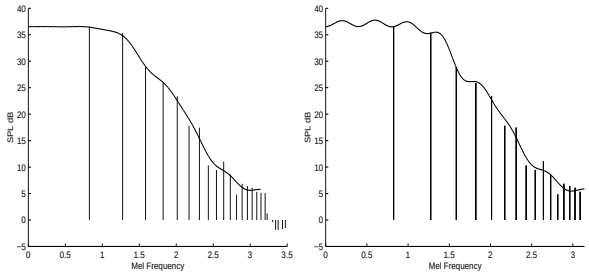

Figure 2: Regularized discrete cepstrum using 40 cepstrum coefficients with $\lambda = 0.1$ and $\lambda = 0.01$ respectively.

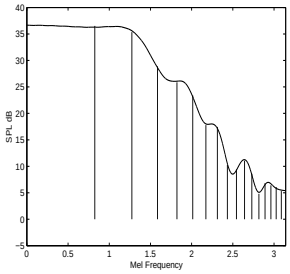

Figure 3: Discrete cepstrum using 40 cepstrum coefficients computed from 14 cepstrum coefficients on the linear scale using posterior warping.

This was named *posterior warping*, since the warping is computed after the estimation of the linear scale coefficients. Evidently, the approximation of the integral by the sum series introduces an error which approximates zero when N is large. The analytic solution of $\mathbf{A}$ was also computed and resulted in a sum of complex incomplete gamma functions. This derivation is omitted due to space limitation.

In Fig.1, a linear frequency scale envelope is shown which is at the same time accurate and smooth. Fig. 2 shows Mel scale envelopes of the same spectrum computed by the regularized discrete cepstrum. These figures show that in the case that $\lambda = 0.1$, a smooth envelope is obtained but it fails to match the peaks accurately in the high frequency band. Decreasing λ does not seem to solve the matching accuracy and introduces overfitting in the lower frequency band of the envelope. However, it is known that the resolution of the human ear is less accurate at these frequencies. The posterior warped version shown in Fig. 3, is at the same time very smooth and very accurate. In addition, no extra parameters must be determined manually.

4 Stability and Perceptual Relevance

When the cepstrum coefficients of consecutive frames were plot in time, considerable variations in these coefficients were observed although the perceived timbre remained constant. The cause of this problem is clarified in Fig. 4. On the left hand side of the figure it is shown that envelope in the lower frequency band is very stable over consecutive frames while considerable differences are shown in the high frequency band. These differences come from very small am-

plitude variations which are amplified enormously by the log function which approaches $-\infty$ when the amplitude approaches zero. The absolute threshold in quiet, represented by the dashed line suggests that these variations are not perceived by a human listener. The cepstrum coefficients on the right hand side of the image are clearly influenced by the variation in the high frequency band. From this it must be concluded that the variation in the cepstrum coefficients does not correspond with the perceived variation in timbre although these variations are actually present in the sound. This compromises the cepstrum based distance metric that was proposed previously.

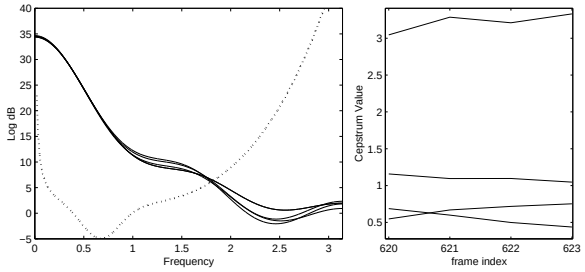

Figure 4: Spectral envelopes and cepstrum coefficients for consecutive time frames.

The stability of the cepstrum coefficients was improved by using a lower bound threshold on the amplitude of the peaks. By replacing amplitudes that were below the threshold with the threshold itself, the influence of these noisy low amplitude partials was significantly reduced. Since most of these partials are situated in the high frequency band, a second method consists in estimating the envelope over a limited spectral band. However, when only the frequency band in synthesized, the perceived quality deteriorates significantly. Fig. 5 shows that the linear scale discrete cepstrum coefficients are very noisy what makes them difficult to interpret. When the lower bound threshold is applied, the features become much more stable. One can clearly observe the silence at the beginning and end of the excerpt, the onsets between different notes and a tremolo (as a result of vibrato) from frame 1000 to 1100.

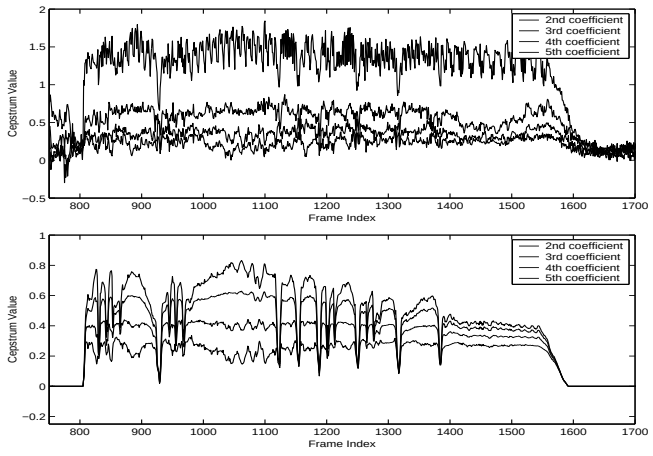

Figure 5: Linear scale cepstrum coefficients: i) Without Preprocessing, ii) Lower Bound thresholded over limited spectral band.

5 Conclusions

The use of Mel scaled discrete cepstrum coefficients as features is studied to express the perceptual similarity between two short time spectral envelopes. The observation that accurate and smooth spectral envelopes are easily obtained on the linear frequency scale resulted in the idea to compute the Mel scaled cepstrum coefficients from the linear scale coefficients. This technique was named *posterior warping* and has the advantage that no manual parameters must be set. Furthermore, it was shown that small amplitude variations are amplified enormously by the log function, compromising the perceptual relevancy of the features. This was improved by computing the envelope over a limited spectral band and applying lower bound thresholding. Since the discrete cepstrum is meant to characterize the deterministic component of the sound, not all perceptual relevant information is captured. However, a great advantage of the discrete cepstrum is that a resynthesis can be obtained from the features allowing a user to judge whether the features are representative for the original sound.

References

Aucouturier, J.-J. and F. Pachet (2002). Finding songs that sound the same. *1st IEEE Workshop on Model based Processing and Coding of Audio (MPCA)*, 91–98.

Brown, J. C. (1999). Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *Journal of the Acoustic Society of America*, 1933–1941.

Campedel-Oudot, M., O. Cappé, and E. Moulines (2001, july). Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach. *IEEE Transactions on Speech and Audio Processing 9*(5), 469–481.

Cappé, O., M. Oudot, and E. Moulines (1997, October). Spectral envelope estimation using a penalized likelihood criterion. *IEEE WASPAA*.

D'haes, W. and X. Rodet (2001). Automatic estimation of control parameters: An instance-based learning approach. *ICMC*, 199–202.

D'haes, W. and X. Rodet (2003). A new estimation technique for determining the control parameters of a physical model of a trumpet. *International Conference on Digital Audio Effects (DAFx-03)*.

Foote, J. (1997). Content-based retrieval of music and audio. *Multimedia Storage and Archiving Systems II, Proc. of SPIE 3229*, 138–147.

Galas, T. and X. Rodet (1990). An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sounds. *ICMC*, 82–84.

Galas, T. and X. Rodet (1991, september). Generalized discrete cepstral method analysis for deconvolution of source-filter systems with discrete spectra. *IEEE WASPAA*.

Gu, L. and K. Rose (2001, May). Perceptual harmonic cepstral coefficients as the front-end for speech recognition. *ICASSP*.

Molau, S., M. Pitz, R. Schlüter, and H. Ney (2001, May). Computing Mel-frequency Cepstral Coefficients on the Power Spectrum. *ICASSP*, 73–76.

Peeters, G., S. McAdams, and P. Herrera (2000, September). Instrument sound description in the context of mpeg-7. *ICMC*, 166–169.

Rodet, X. (1997, August). Musical signal analysis/synthesis sinusoidal+residual and elementary waveform models. *IEEE Time-Frequency and Time-Scale Workshop (TSTF)*.

Schwarz, D. and X. Rodet (1999). Spectral envelope estimation and representation for sound analysis-synthesis. *ICMC*, 351–354.

Spevak, C. (2002). Soundspotter - a prototype system for content based audio retrieval. *International Conference on Digital Audio Effects (DAFx02)*, 27–32.