LETTER

Perceptually Weighted Mel–Cepstrum Analysis of Speech based on Psychoacoustic Model

Hongwu YANG[†*], *Student Member*, Dezhi HUANG[††**], *and* Lianhong CAI[†††], *Nonmembers*

SUMMARY This letter proposes a novel approach for mel-cepstral analysis based on the psychoacoustic model of MPEG. A perceptual weighting function is developed by applying cubic spline interpolation on the signal-to-mask ratios (*SMRs*) which are obtained from the psychoacoustic model. Experiments on speaker identification and speech re-synthesis showed that the proposed method not only improved the speaker recognition performance, but also improved the speech quality of the re-synthesized speech.

key words: *weighted mel-cepstral analysis; auditory properties;*

1. Introduction

In this letter, we discuss a novel perceptually weighted mel-cepstral analysis method, which estimates the mel-cepstral coefficients (*MCCs*) through a frequency-dependent weighting technique[1]. Since this kind of method corresponds to human perception, it is better than the mel-cepstrum analysis method proposed by Tokuda et al.[2], which is widely used in HMM based speech synthesis[3]. The inadequacy of their method is that it may shift and widen the band of formants as well as enhance the energy of formants.

To obtain a set of weighted mel-cepstral coefficients (*WMCCs*), we first have to calculate its perceptual weighting function. This is obtained by applying the cubic spline interpolation on the signal-to-mask ratios (*SMRs*) obtained from the psychoacoustic model of MPEG[4]. We then use the Newton-Raphson method [6]to apply the weighting function on the same set of speech frequencies used in estimating *MCCS* to calculate the *WMCCs*. Experiments on speaker identification demonstrate that the *WMCCs* outperform both the mel-frequency cepstral coefficients (*MFCCs*) and the *MCCs*. The objective and subjective experiments on speech re-synthesis also indicate that using the *WMCCs* not only reduces the spectral distortion, but also improves speech quality for the re-synthesized speech. Therefore, the *WMCCs* may model spectral structure more accurately than that of the *MCCs*.

2. Mel-Cepstrum Analysis

The mel-cepstral analysis represents the spectrum based on the unbiased estimation log spectrum (UELS) method[2], [5]. In mel-cepstral analysis, optimal cepstral coefficients are estimated from the short-time spectrum of speech signals based on the nonlinear perceptual prosperities of human auditory sensation. Given a frame of speech signal, the modified periodogram is defined as follows:

$$I_n(\omega) = \frac{\left|\sum_{n=0}^{N-1} w(n)x(n)e^{-j\omega n}\right|^2}{\sum_{n=0}^{N-1} w^2(n)} \tag{1}$$

where $w(n)$ is the window function.

The estimation of mel-cepstrum $H(\omega)$ is defined as Eq. 2:

$$\begin{aligned} H(\omega) &= \left| e^{\sum_{m=0}^{M} c_m \tilde{\omega}^{-m}} \right| \\ \tilde{\omega} &= \frac{1-\alpha e^{-j\omega}}{e^{-j\omega}-\alpha}, |\alpha| < 1 \end{aligned} \tag{2}$$

where α is the mel-frequency transform coefficient, M is the order of estimation for mel-cepstral , and $\{c_m\}$ is the mel-cepstral coefficients.

The estimation error function is defined as Eq. 3:

$$\varepsilon(\mathbf{c}) = \frac{1}{2\pi}\int_0^{2\pi}\left(\frac{I_N(\omega)}{H(\omega)^2} - \log I_N(\omega) + 2\log H(\omega) - 1\right)d\omega \tag{3}$$

where $\mathbf{c}$ is the vector of *MCCs*.

The optimal mel-cepstral coefficients are estimated by solving Eq. 3 as a minimization problem. From Eq. 3 we can see that the proportion of each frequency's estimation error is completely the same.

Manuscript received April 26, 2006.
Manuscript revised August 9, 2006.
Final manuscript received January 1, 2007.

[†] The author was with the College of Physics and Electronics Engineer, Northwest Normal University, Lanzhou 730070,China.

[††] The author was with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.

[†††] The author is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.

[*] Presently, the author is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.

[**] Presently, the author is with the France Telecom R & D, Beijing 100084, China.

3. Weighted Mel-Cepstral Analysis

3.1 Weighted Estimation Error

Different frequencies of speech signal have different functions in human auditory perception. In fact, only part of frequencies are sensitive for auditory perception. For the mel-cepstral analysis, perceptually sensitive frequencies should be estimated more accurately than perceptually non-sensitive frequencies. Therefore, we proposed a weighted estimation error function as Eq. 4 to simulate the nonlinear characteristic of auditory perception.

$$\tilde{\varepsilon}(\tilde{\mathbf{c}}) = \frac{1}{2\pi}\int_0^{2\pi}\left(\frac{I_N(\omega)}{H(\omega)^2} - \log I_N(\omega) + 2\log H(\omega) - 1\right)W(\omega)d\omega \quad (4)$$

where $W(\omega)$ is a non-negative weighting function. If the weighting function has the same value for each frequency, the weighted mel-cepstral analysis is equivalent to the mel-cepstral analysis. From Eq. 4 we can get a set of new mel-cepstral coefficients $\tilde{\mathbf{c}}$, which we call the weighted mel-cepstral coefficients (*WMCCs*).

3.2 Perceptual Weighting Function

The psychoacoustic model of the MPEG outputs a set of signal-to-mask ratios (*SMRs*) that flag frequency components with amplitude below the masking level, so the *SMRs* can be used to represent the contributions of the various frequencies on the acoustic perception. Therefore, we adopt the *SMRs* as the perceptual weighting function. However, the *SMRs* obtained from the psychoauditory model are discrete values. In order to get the continuous perceptual weighting function, the cubic spline interpolation between the discrete *SMRs* are performed for all frequencies so that the initial continuous weighting function, i.e. $L(\omega), \omega \in [0, \pi]$ is obtained. Therefore, the perceptual weighting function is defined as Eq. 4. Fig. 1 shows an example of continuous weighting function $W(\omega)$, which is obtained by applying cubic spline interpolation on the discrete *SMRs* of syllable /a:/.

$$W(\omega) = \begin{cases} \dfrac{L(\omega)}{2\int_0^{\pi} L(\omega)d\omega} & \omega \in [0, \pi], \\[2ex] \dfrac{L(2\pi-\omega)}{2\int_0^{\pi} L(\omega)d\omega} & \omega \in [\pi, 2\pi]. \end{cases} \quad (5)$$

3.3 Solving the Minimization Problem

Given the modified periodogram and the weighting function, the calculation of the *WMCCs* is equivalent to minimizing the estimation error function given in Eq. 4. By expanding Eq. 4, we get:

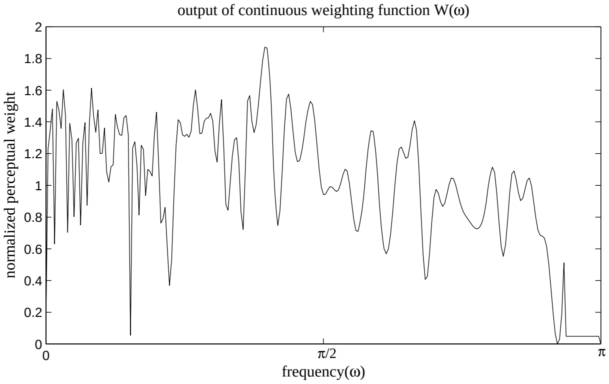

Fig. 1 The continuous weighting function $W(\omega)$ for syllable /a:/.

$$\tilde{\varepsilon}(\tilde{\mathbf{c}}) = \frac{1}{2\pi}\int_0^{2\pi}\left(\frac{I_N(\omega)}{H(\omega)^2} - \log I_N(\omega)\right)W(\omega)d\omega + \frac{1}{2\pi}\int_0^{2\pi}(2\log H(\omega) - 1)W(\omega)d\omega \quad (6)$$

where $\tilde{\mathbf{c}} = (c_0, c_1, c_2, \cdot, c_M)^T$ is the *WMCCs* that need to be solved, M is the order of the weighted mel-cepstral estimation. After replacing the $H(\omega)$ in Eq. 6 with Eq. 2, we can get:

$$\tilde{\varepsilon}(\tilde{\mathbf{c}}) = \tilde{\varphi}(\tilde{\mathbf{c}}) + E$$

$$\tilde{\varphi}(\tilde{\mathbf{c}}) = \frac{1}{2\pi}\int_0^{2\pi}\left(\frac{I_N(\omega)}{e^{2\sum_{m=0}^{M}\tilde{c}_m Re(\tilde{\omega}^m)}} + 2\sum_{m=0}^{M}\tilde{c}_m Re(\tilde{\omega}^m)\right) W(\omega)d\omega$$

$$E = -\frac{1}{2\pi}\int_0^{2\pi}(\log I_N(\omega) + 1)\,W(\omega)d\omega \quad (7)$$

Since E is shown to be unrelated to $\tilde{\mathbf{c}}$ in Eq. 7, it can be considered as a constant in solving the optimal estimation problem. The minimization problem of $\tilde{\varepsilon}(\tilde{\mathbf{c}})$ is thus equivalent to the minimization problem of $\tilde{\varphi}(\tilde{\mathbf{c}})$, i.e. the nonlinear equation set can be shown as follows:

$$\frac{\partial\tilde{\varphi}(\tilde{\mathbf{c}})}{\partial\tilde{c}_m} = \frac{1}{\pi}\int_0^{2\pi}\left[W(\omega) - \frac{I_N(\omega)W(\omega)}{e^{2\sum_{m=0}^{M}\tilde{c}_m Re(\tilde{\omega}^m)}}\right]Re(\tilde{\omega}^m)d\omega = 0, m = 0, 1, \cdots, M \quad (8)$$

Since $\tilde{\varphi}(\tilde{\mathbf{c}})$ is a concave function, solution using the Newton-Raphson method is stable, and can be applied to solve Eq. 8. The increment of $\tilde{\mathbf{c}}$, i.e. $\Delta\tilde{\mathbf{c}}$ can be solved as follows:

$$\mathbf{H}\Delta\tilde{\mathbf{c}} = -\Delta\tilde{\varphi} \quad (9)$$

where $\Delta\tilde{\varphi} = \left(\frac{\partial\tilde{\varphi}}{\partial\tilde{c}_0}, \frac{\partial\tilde{\varphi}}{\partial\tilde{c}_1}, \cdots, \frac{\partial\tilde{\varphi}}{\partial\tilde{c}_M}\right)^T$. $\mathbf{H}$ is a Hessian matrix, and is further defined as:

$$\mathbf{H} = \frac{\partial^2 \tilde{\varphi}}{\partial \tilde{\mathbf{c}} \partial \tilde{\mathbf{c}}^T} = \begin{pmatrix} \frac{\partial^2 \tilde{\varphi}}{\partial \tilde{c}_0 \partial \tilde{c}_0} & \frac{\partial^2 \tilde{\varphi}}{\partial \tilde{c}_0 \partial \tilde{c}_1} & \cdots & \frac{\partial^2 \tilde{\varphi}}{\partial \tilde{c}_0 \partial \tilde{c}_M} \\ \frac{\partial^2 \tilde{\varphi}}{\partial \tilde{c}_1 \partial \tilde{c}_0} & \frac{\partial^2 \tilde{\varphi}}{\partial \tilde{c}_1 \partial \tilde{c}_1} & \cdots & \frac{\partial^2 \tilde{\varphi}}{\partial \tilde{c}_1 \partial \tilde{c}_M} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \tilde{\varphi}}{\partial \tilde{c}_M \partial \tilde{c}_0} & \frac{\partial^2 \tilde{\varphi}}{\partial \tilde{c}_M \partial \tilde{c}_1} & \cdots & \frac{\partial^2 \tilde{\varphi}}{\partial \tilde{c}_M \partial \tilde{c}_M} \end{pmatrix} \quad (10)$$

where

$$\frac{\partial^2 \tilde{\varphi}}{\partial \tilde{c}_m \partial \tilde{c}_n} = \frac{2}{\pi} \int_0^{2\pi} \left[\frac{I_N(\omega) W(\omega)}{e^{2 \sum_{m=0}^{M} \tilde{c}_m Re(\tilde{\omega}^m)}} \right] Re\,(\tilde{\omega}^m) Re\,(\tilde{\omega}^n) d\omega$$

Since the Hessian matrix **H** in Eq. 10 is equal to a Toeplitz-plus-Hankel matrix, [6] leads to the conclusion that the time complexity of solving equation set 8 is $O(M \log^2 M)$. Hence the time complexity of the weighted mel-cepstral analysis is $O(K \times M \log^2 M)$, where K is the iteration time, which is dependent on the precision of the search. Because of the quadratic convergence property of the Newton-Raphson method, K generally has a small value and therefore the weighted mel-cepstral coefficients can be usually solved in real-time.

4. Experiments

4.1 Speaker Identification

In the experiment, we used all 49 speakers in the DR1 dialect region of TIMIT to perform evaluation. We chose 72-order *MFCC*s, 72-order *MCCs* and 72-order *WMCCs* as the speaker features to model the speaker's acoustic characteristic, where the first 26-order parameters of each set were constructed from the original coefficients, with the rest coming from their first- and second-order differential coefficients. The speech signals were emphasized with a coefficient setting of 0.97. Meanwhile, the blocking operation used a 25ms Blackman window to scale the frame signal, and a 10ms hopping size. We used the 7 longest (out of 10) utterances of each speaker as the training data and the remaining 3 utterances as the testing data.

For each speaker, three 16 components Gaussian mixtures model (*GMM*) each using one of the 3 features (*MFCCs*, *MCCs* and *WMCCs*)were trained as the speaker models. Table 1 shows their recognition rate and *WMCCs* are shown to outperform both *MFCCs*and *MCCs*. The result indicates that the *WMCCs* were able to characterize the vocal tract accurately to correctly identify a speaker.

Table 1 *Speaker recognition rate using the 3 features (%).*

Features	*MFCC*	*MCC*	*WMCC*
Correct recognition rate (%)	96.2	98.3	99.8

4.2 Speech Re-Synthesis

In the experiment, we used both a male (indicated with M) and a female (indicated with F) corpus collected by our laboratory. Both corpora include 1268 Chinese tone syllables and 50 utterances. Each utterance includes about 20 syllables. 26-order *MCCs* and *WMCCs* were extracted from each syllable and utterance respectively with a 25 ms Blackman window and 10 ms hopping. All syllables and utterances were then re-synthesized with the mel log spectrum approximation digital filter (*MLSADF*)[7], using *MCCs* and *WMCCs* respectively.

4.2.1 objective test

We use modified Bark spectral distortion (*MBSD*)[8] as the distortion criteria between the source speech and the re-synthesized speech. The *MBSD* just manages the perceptible spectral distortion.

Table 2 shows the results of the objective test on the syllables and utterances. The *MBSD* between the source speech and the re-synthetic speech with *MCCs* or with *WMCCs* is calculated. From Table 2 we can see that the *MBSD* is smaller for *WMCCs* than for *MCCs*, which means that the spectral distortion of the re-synthesized speech with the *WMCCs* is smaller than that with the *MCCs*. Therefore, the quality of the re-synthesized speech is higher for *WMCCs*.

Table 2 *The* MBSD *(dB) between the source speech and the re-synthesized speech.*

	MCCs		*WMCCs*	
speaker	Syllable	Utterance	Syllable	Utterance
M	0.46	0.77	0.41	0.64
F	0.59	0.82	0.53	0.78

4.2.2 subjective test

A contrastive experiment was performed to test the quality of the re-synthesized speech in terms of user perception. 200 utterances were re-synthesized, and 5 subjects were invited to evaluate the two sets of utterances. The first set includes 100 utterances that were re-synthesized with the *MCCs* while the second set includes the same 100 utterances that were re-synthesized with the *WMCCs*. Matching utterances were identified from the 2 sets, and they were played to the subjects a pair at a time in a completely random manner. The subjects were then asked to evaluate which of the utterances sounds more pleasant and this step was repeated for all 100 pairs of utterances.

The relative perception of the speech quality are shown in Fig. 2, where the *WMCCs* utterances are considered to be more pleasant for both the male and

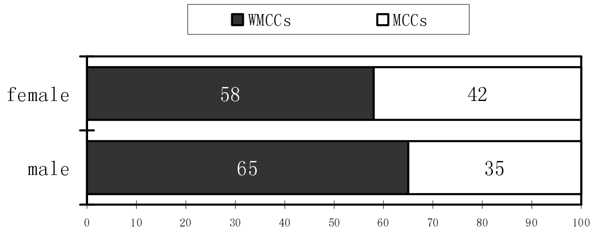

Fig. 2 The percentage of the sentences regarded as more pleasant in the two sets(%).

female utterances. This is because for *WMCA*, more spectral details of the source utterance are better preserved and the structure of formant is also better defined. Therefore, utterances re-synthesized with *WMCCs* have higher fidelity and sound more authentic and nature.

5. CONCLUSIONS

In this letter, we propose a weighted mel-cepstral analysis method. Based on the mel-cepstral analysis, we use psychoacoustic model to develop a weighting function so that different frequencies have different perceptual weights for the estimation error. Compared with the mel-cepstral analysis, the weighted mel-cepstral analysis not only characterizes the vocal tract more precisely, the structure of formants is also accurately preserved. Because the auditory properties are taken into account in the speech analysis, the *WMCCs* are in accordance with the psychoacoustic rule. This improves the quality of the re-synthesized speech.

Acknowledgments

The work is supported by the research fund from the National Science Foundation of China (NSFC) under grant No. 60433030 and the joint fund of NSFC-RGC (Research Grant Council of Hong Kong) under grant No. 60418012. We would also like to thank all reviewers as well as Mr. Andrew Lee from Human Computer Communication Laboratory, Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong for their careful review.

References

[1] H. Hermansky, "Perceptral Linear Predictive(PLP) analysis for speech," J. Acoust. Soc. Am., Vol. 87, No. 4, pp. 1738–1752, 1990.

[2] K. Tokuda, T. Kobayashi and S. Imai, "Adaptive cepstral analysis of speech," IEEE Trans. Speech and Audio Processing, Vol. 3, No. 6, pp. 481–488, 1995.

[3] J. Yamagishi, K. Onishi, T. Masuko and T. Kobayashi,"Acoustic Modeling of Speaking Styles and Emotional Expressions in HMM-Based Speech Synthesis," IEICE TRANS. INF. & SYST., VOL.E88–D, NO.3, pp502–509,March 2005.

[4] ISO/IEC. IS 13818–3. "Information technology–generic coding of moving pictures and associated audio–part 3: audio." 1994.

[5] S. Imai and C. Furnich, "Unbiased estimator of log spectrum and its application to speech signal processing," Proc. 1988 EURASIP, pp. 203–206, Sep. 1988.

[6] G. Heinig and K. Rost, "New Fast Algorithms for Toeplitz-Plus-Hankel Matrices," SIAM J. Matrix Anal. Appl., Vol. 25, No. 3, pp. 842-857, 2004.

[7] S. Imai, K. Sumita and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," Electronics and Communications in Japan, Vol. 66, No. 2, pp. 10–18, 1983.

[8] W. Yang, M. Dixon and R. Yantorno, "A modified bark spectral distortion measure which uses noise masking threshold," in Proceedings of IEEE Speech Coding Workshop. Pocono Manor, pp. 55–56, 1997.