IEICE TRANS. INF. & SYST., VOL.E101–D, NO.2 FEBRUARY 2018

LETTER

Pitch Estimation and Voicing Classification Using Reconstructed Spectrum from MFCC

JianFeng WU[†], **HuiBin QIN**[†], **YongZhu HUA**[†a)], ***Nonmembers***, ***and*** **LingYan FAN**[††], ***Member***

SUMMARY In this paper, a novel method for pitch estimation and voicing classification is proposed using reconstructed spectrum from Mel-frequency cepstral coefficients (MFCC). The proposed algorithm reconstructs spectrum from MFCC with Moore-Penrose pseudo-inverse by Mel-scale weighting functions. The reconstructed spectrum is compressed and filtered in log-frequency. Pitch estimation is achieved by modeling the joint density of pitch frequency and the filter spectrum with Gaussian Mixture Model (GMM). Voicing classification is also achieved by GMM-based model, and the test results show that over 99% frames can be correctly classified. The results of pitch estimation demonstrate that the proposed GMM-based pitch estimator has high accuracy, and the relative error is 6.68% on TIMIT database.

key words: *pitch estimation, voicing classification, MFCC, GMM*

1. Introduction

Mel-Frequency Cepstral Coefficients (MFCC) is widely used in speech recognition, speaker identification and other speech processing systems. In recent years, there is an emerging method that predicts fundamental frequency and voicing from MFCC vectors, which enables speech signal to be reconstructed solely from a stream of MFCC vectors in Distributed Speech Recognition (DSR) back-end [1]–[3]. The algorithm predicts fundamental frequency by modeling the joint density of fundamental frequency and MFCC. The method is based on a Gaussian Mixture Model (GMM), and utilizes Hidden Markov Model (HMM) to link together a series of state-dependent GMMs. The speaker-dependent HMM-GMM predictor shows good results. However, the error of speaker-independent predictor is large. Besides, the HMM-GMM predictor requires a set of monophone-based HMMs and a set of state-specific GMMs training. The trained predictor is specified to a certain language, and must be re-trained in other languages.

A low bit-rate speech coding scheme through quantization of MFCC is presented in [4], [5], which reconstructs speech waveform from MFCC without pitch and energy. The reconstruction progresses recover magnitude spectrum from MFCC by Moore-Penrose pseudo-inverse, and then utilize least-squares estimate, inverse short-time Fourier Transform Magnitude algorithm to reconstruct speech frame. In addition, some novel pitch detection methods are proposed in recent years. An algorithm named Pitch Estimation Filter with Amplitude Compression (PEFAC), utilizes non-linear amplitude compression to attenuate narrow-band noise components, and a comb-filter to attenuate smoothly varying noise components in the log-frequency power spectral domain [6]. Motivated by the study of [4], [5] and the novel method of [6], we estimate pitch from cepstrum/spectrum, which is reconstructed by MFCC with inversion operation.

The proposed algorithm reconstructs cepstrum/spectrum from MFCC with Moore-Penrose pseudo-inverse by Mel-scale weighting functions. With the reconstructed cepstrum, pitch can be estimated from the peak directly, while there may be large error due to the reconstruction distortion. A method that combines non-linear amplitude compression and a log-frequency power spectral domain filter is presented to reduce error. Making use of the correlation of filtered power spectral and pitch, a GMM-based pitch estimation method is proposed to get more reliable pitch. Moreover, a voicing classification method is also presented. The main advantage of the proposed method is the feature extraction, which uses the reconstructed and filtered magnitude spectrum, rather than the original MFCC vectors. Compared with previous work [1]–[3], the proposed method is a speaker/language independent predictor.

The organization of this paper is as follows: in Sect. 2 a brief introduction of spectrum reconstruction from MFCC is given. In Sect. 3 we discuss the pitch estimation method with the reconstructed spectrum/cepstrum, and a GMM-based voiced/unvoiced classification method is also presented. Section 4 is the experimental results of voiced/unvoiced classification and pitch estimation. Finally, we conclude the paper in Sect. 5.

2. Spectrum Reconstruction

MFCC is defined as special cepstrum that a set of weighting functions is applied to the power spectrum prior to the log operations and Discrete Cosine Transform (DCT). This weighting function is based on human perception of pitch and is most commonly implemented in the form of a bank of triangular filters in Mel-scale [4], [5]. The Mel-cepstrum M of t^{th} frame speech $s_t(n)$ is computed as (the subscript t is dropped to simplify notation)

Manuscript received July 24, 2017.
Manuscript revised October 9, 2017.
Manuscript publicized November 15, 2017.

[†]The authors are with the Institute of Electron Device & Application, Hangzhou Dianzi University, Hangzhou, China.
[††]The author is with the Institute of Communication, Hangzhou Dianzi University, Hangzhou, China.
a) E-mail: huayongzhu0704@163.com

DOI: 10.1587/transinf.2017EDL8162

$$M = \mathrm{DCT}\{\log(w_m |S(\omega)|^2)\} \quad (1)$$

where w_m is the Mel-weighting function, and $S(\omega)$ is Discrete Fourier Transform (DFT) of $s(n)$.

The power spectrum with Mel-weighting in (1) can be expressed in matrix form as

$$\mathbf{y} = \mathbf{W_m} |\mathbf{s}(\omega)|^2 \quad (2)$$

where $\mathbf{y}$ is a vector of $J \times 1$ (J is the number of Mel-filters), and $\mathbf{W_m}$ is the weighting matrix of $J \times L$ (L is frame length).

In (1), spectrum information is lost by applying the Mel-scale weighting, while other operations – DCT, log, and square-root are all invertible. To invert the Mel weighting, a solution of minimal Euclidean norm can be used

$$\left|\tilde{S}(\omega)\right|^2 = \mathbf{W_m}^{\dagger}\mathbf{y} = \mathbf{W_m}^{\dagger}\mathbf{W_m} |\mathbf{s}(\omega)|^2 \approx |S(\omega)|^2 \quad (3)$$

where $\mathbf{W_m}^{\dagger} = (\mathbf{W_m}^{\mathrm{T}}\mathbf{W_m})^{-1}\mathbf{W_m}^{\mathrm{T}}$ is the Moore-Penrose pseudo inverse of matrix $\mathbf{W_m}$.

3. Pitch Estimation and Voicing Classification with Reconstructed Spectrum

3.1 Spectrum Filter-Based Pitch Estimation

In [2], the authors make use of the correlation of the MFCC and fundamental frequency (i.e. pitch), predict the pitch with GMM. While, we have found that the correlation between reconstructed magnitude spectrum and pitch is higher. Table 1 shows the correlation between different vectors (MFCC, reconstructed magnitude spectrum and filtered spectrum) and pitch, which is computed using the TIMIT training subset.

The results of Table 1 show that the correlation between reconstructed magnitude spectrum and pitch is higher than that of MFCC. Comparing with the MFCC vector, the magnitude spectrum contains much more information about pitch frequency. The reconstructed magnitude spectrum is more suitable for GMM-based pitch estimation and voicing classification.

In practice, speech signal is always added or convoluted by various noises, and the harmonic peak of magnitude spectrum will be broadened by framing window. In order to reduce the error caused by reconstruction distortion, we utilize a pitch estimation method that combines non-linear amplitude compression to attenuate narrow-band noise components, and a log-frequency power spectral domain filter to attenuate smoothly varying noise components [6]. The algorithm is described as follow

1) From MFCC, obtain the reconstructed magnitude spectrum $\left|\tilde{S}(\omega)\right|$ by Inverse Discrete Cosine Transform (IDCT), exponent and (3).

2) Transform $\left|\tilde{S}(\omega)\right|$ to power spectral log-scale domain, $\tilde{S}(q) = \log\left|\tilde{S}(\omega)\right|^2$, where $q = \log\omega$

3) Compress the Power Spectral Density (PSD) by

$$\tilde{S}'(q) = \tilde{S}(q)^{a(q)} \quad (4)$$

where $a(q)$ is calculated by low-pass filtered $\tilde{S}_{filt}(q)$ and long-term average spectrum $\tilde{S}_{aver}(q)$, $a(q) = \frac{\log \tilde{S}_{aver}(q)}{\log \tilde{S}_{filt}(q)}$

4) Filter the compressed PSD with

$$h(q) = \beta - \log(\gamma - \cos(2\pi e^q)) \quad (5)$$

where β is a shift so that $\int h(q)dq = 0$, and γ controls the peak width. The output is

$$\tilde{Y}(q) = \tilde{S}'(q) * h(q) \quad (6)$$

The pitch can be estimated from the peak of $\tilde{Y}(q)$. With the compressing and filtering progresses in log-scale domain, the algorithm is able to reduce the pitch estimation error caused by spectrum reconstruction distortion. However, the position of peak will shift between original filtered and the reconstructed PSD. Therefore, a more robust method based on GMM model is used, which will be discussed below.

Table 1 Correlation between different vectors (MFCC, reconstructed magnitude spectrum and filtered spectrum) and pitch

Vector	MFCC	Reconstructed magnitude spectrum	Filtered spectrum
Correlation	0.863	0.898	0.925

3.2 GMM-Based Pitch Estimation

In order to estimate pitch more reliable, we make use of the correlation of filtered PSD and pitch. Denote the feature vectors Φ by

$$\Phi = [\Omega, f] \quad (7)$$

where f is the pitch frequency and Ω is a row vector constructed by of $\tilde{Y}(q)$ for q in a discrete set, which will be discussed in Sect. 4.

The feature vector Φ is modeled by GMM. From the training set of Φ, a set of K Gaussian clusters are produced using the expectation-maximization (EM) algorithm. The probability density function (PDF) of Φ is given by

$$p(\Phi) = \sum_{k=1}^{K} \pi_k N(\Phi; \mu_k, \Theta_k) \quad (8)$$

Each of the K clusters is represented by a prior probability π_k, and a Gaussian PDF $N(\Phi)$, with mean vector μ_k and covariance matrix Θ_k

$$\mu_k = [\mu_k^{\Omega}, \mu_k^{f}]\Theta_k = \begin{bmatrix} \Theta_k^{\Omega\Omega} & \Theta_k^{\Omega f} \\ \Theta_k^{f\Omega} & \Theta_k^{ff} \end{bmatrix} \quad (9)$$

where μ_k^{Ω} and μ_k^{f} is mean value of Ω and f in clusters k, respectively. $\Theta_k^{\Omega\Omega}$ is covariance matrix of the Ω, Θ_k^{ff} is covariance of f, $\Theta_k^{f\Omega}$ and $\Theta_k^{\Omega f}$ are covariance of Ω and f, respectively.

The initial cluster positions of EM training are found using the well-known Linde-Buzo-Gray (LBG) algorithm. The maximum EM clustering iteration is 100. The selection

of the mixture components K is discussed in the experimental section.

A maximum a posteriori (MAP) pitch estimation method given by [8] is used to estimate the pitch frequency $\hat{f}$ from Ω_i of i^{th} frame, which is given by

$$\hat{f}_i = \sum_{k=1}^{K} h_k(\Omega_i)\left(\mu_k^f + \Theta_k^{f\Omega}(\Theta_k^{\Omega\Omega})^{-1}(\Omega_i - \mu_k^{\Omega})\right) \tag{10}$$

where the posterior probability $h_k(\Omega_i)$ is given as

$$h_k(\Omega_i) = \frac{\pi_k p(\Omega_i|c_k^{\Omega})}{\sum_{k=1}^{K} \pi_k p(\Omega_i|c_k^{\Omega})} \tag{11}$$

and $p(\Omega_i|c_k^{\Omega})$ is the marginal distribution of Ω for the k^{th} cluster c_k^{Ω} in the GMM.

3.3 Voiced/Unvoiced Classification

Pitch estimation should only be applied to reconstruct spectrum which represents voiced speech. In [2], this is implemented by extending the HMM-GMM pitch predictor, which requires a set of monophone-based HMMs and a set of state-specific GMMs training. The trained predictor is specified to a certain language, and must be retrained in other languages.

We propose a low complexity voicing classification algorithm, which is achieved by GMM-based model. The model utilizes the correlation of the frame mean power and the peak of filtered spectrum, which comprises more latent voicing information.

The feature vectors ψ are extracted from a set of training data

$$\psi = [s, \chi] \tag{12}$$

where $s = \log \mu_{psd}, \chi = \Sigma_{pitch}/\mu_{psd}$. μ_{psd} is the mean power value of compressed PSD with (4), and Σ_{pitch} is the sum of three candidate pitch in (6).

There are two GMMs for the voiced/unvoiced classifier, one is modeled for voiced with the training set of voiced vectors ψ_v, the other is for unvoiced one, ψ_u. The modeling method is the same as what is described in [8]. Probability of the frame being voiced is calculated as

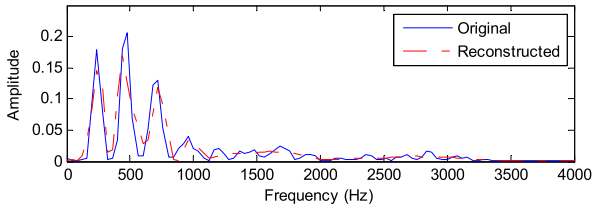
$$P(v) = (1 + \exp(p_u - p_v))^{-1} \tag{13}$$

where p_u and p_v are posterior probability of unvoiced GMM and voiced GMM, respectively.

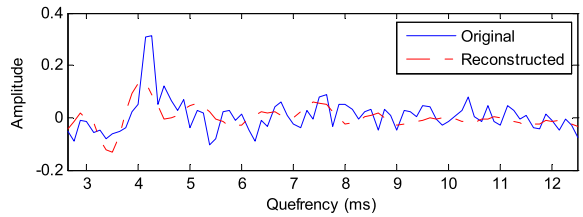
For an input feature vectors ψ_i, calculating the posterior probability of $p_v(\psi_i)$ and $p_u(\psi_i)$ with GMMs, and then the probability of being voiced is calculated by (13). If $P(v) > \varepsilon$ (ε is threshold, set to 0.5), the frame is classified as voiced, otherwise, as unvoiced.

4. Experimental Results

In this section, both the results of pitch estimation and voicing classification are evaluated. We use TIMIT database [9] for training and testing. Each sentence is about 3 seconds duration, and down-sampling to 8 kHz. The corpus is framing to 200 samples (25ms) with hamming windows, and the frame shift is 80 samples (10ms).



(a) Original spectrum V.S. reconstructed spectrum



(b) Original cepstrum V.S. reconstructed cepstrum

Fig. 1 Spectrum/cepstrum reconstruction from MFCC for a voiced frame of speech

4.1 Spectrum/Cepstrum Reconstruction Results

Inversing MFCC to spectrum is a challenging task since much information is lost by Mel-scale weighting function in (2), and (3) is only an approximate solution. It is obvious that the more the Mel-filters are, the less information of magnitude spectrum will be lost. In this paper, the number of Mel-filters is 23, just as the DSR front-end [10]. In consideration of the inversion, all of 23 MFCC are reserved, while 10 high order coefficients are discarded in DSR.

Figure 1 shows the reconstructed results of spectrum and cepstrum from MFCC. Figure 1 (a) compares the original and reconstructed spectrum, from which we can see that there is only a slight difference between the original and approximate solution of (3). Figure 1 (b) illustrates the original and reconstructed cepstrum, and the peak is the candidate pitch.

4.2 Pitch Estimation Results

Figure 2 illustrates the result of pitch estimation with the reconstructed spectrum/cepstrum from MFCC. We first estimate pitch with reconstructed cepstrum peak directly (i.e.ceps.).The result is shown in Fig. 2 with blue dash-dot line. Then, we utilize non-linear amplitude compression and a log-frequency power spectral domain filter to reduce error (i.e.filt.). The result is shown with green dotted line.

The proposed GMM-based pitch estimator making use of the correlation of filtered PSD and pitch can estimate pitch more reliable (i.e. GMM). Since the pitch frequency range is from 60 to 400 Hz and the frame length is 200, the candidate pitch is between 31 and 62 of the output of

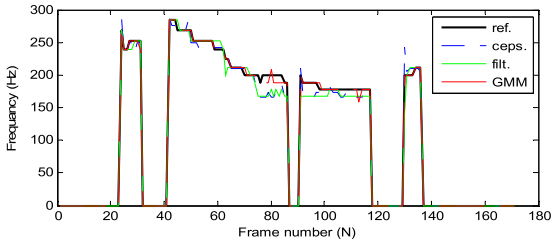


Fig. 2 Comparison of estimated and reference pitch contours

Table 2 Voiced/unvoiced classification errors E_c

GMM clusters	classification error, E_c,%
2	1.02%
4	0.98%
8	**0.96%**
16	0.96%

(6). Therefore, the dimension of Ω in (7) is 32. The result is shown in Fig. 2 with red dashed line. And the reference pitch contours (i.e. ref.) is shown with black bold solid line. In this experiment, the number of GMM clusters is 32, more details of the parameter will be discussed in following evaluation.

Figure 2 shows that estimating pitch with reconstructed cepstrum peak directly (i.e.ceps.) can track the reference pitch contours, while there are some errors due to cepstrum reconstruction distortion. With the compressing and filtering progresses in log-scale domain (i.e.filt.), the performance is better. The pitch estimation by MAP with GMM matches the reference exactly.

Prior to the pitch estimation progress, a frame should be classified as voiced or unvoiced with the method described in Sect. 3.3. The accuracy is measured using the percentage voicing classification error E_c

$$E_c = \frac{N_{\mathrm{V}/U} + N_{U/\mathrm{V}}}{N} \times 100\% \tag{14}$$

where N is total frame number of testing set, $N_{\mathrm{V}/U}$ and $N_{U/\mathrm{V}}$ is the number of unvoiced and voiced frames that are incorrectly classified respectably.

Table 2 shows the classification error of different GMM clusters, the numbers of voiced and unvoiced GMM clusters are equal. From the results we can see that as the number of GMM clusters increased the classification error reduced. However, as the number increased to 16, the accuracy didn't improve. There may be overfit for the voiced/unvoiced classification task.

For voiced frames, pitch prediction accuracy is measured using the percentage pitch frequency error E_p

Table 3 Pitch estimation errors E_p and $E_{20\%}$

Pitch estimation	E_p,%	$E_{20\%}$,%
ceps.	15.20	5.84
filt.	12.98	3.26
GMM(16 clusters)	**7.86**	2.37
GMM(32 clusters)	**7.26**	1.42
GMM(64 clusters)	**6.68**	1.38

$$E_p = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\hat{f}_i - f_i\right|}{f_i} \times 100\% \tag{15}$$

where i is the frame index, $\hat{f}_i$ is the predicted pitch frequency, and f_i is the reference pitch frequency. As proposed in [3], pitch estimation error greater than 20% are not included in E_p, and another error measure is introduced, $E_{20\%}$.

The comparisons of pitch estimation error with the three methods are shown in Table 3. Pitch estimator with the reconstructed cepstrum peak (ceps.) performs general. With non-linear amplitude compression and a log-frequency power spectral domain filter (filt.), the accuracy is improved. The GMM-based estimator shows high accuracy and increasing the clusters number can reduce estimation error.

References

[1] X. Shao and B. Milner, "Pitch prediction from MFCC vectors for speech reconstruction," Proc. ICASSP, Montreal, QC, Canada, 2004.
[2] B. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," IEEE Trans. Audio, Speech, Lang. Process., vol.15, no.1, pp.24–33, Jan. 2007.
[3] B. Milner and J. Darch, "Robust Acoustic Speech Feature Prediction from Noisy Mel-Frequency Cepstral Coefficients," IEEE Trans. Audio, Speech, Lang. Process., vol.19, no.2, pp.338–347, Feb. 2011.
[4] L.E. Boucheron, P.L.D. Leon, and S. Sandoval, "Low Bit-Rate Speech Coding Through Quantization of Mel-Frequency Cepstral Coefficients," IEEE Trans. Audio, Speech, Lang. Process., vol.20, no.2, pp.610–619, Feb. 2012.
[5] L.E. Boucheron and P.L.D. Leon, "On the inversion of mel-frequency cepstral coefficients for speech enhancement applications," Proc. ICSES, pp.485–488, 2008.
[6] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," Proc. European Signal Processing Conference (EUSIPCO), 2011.
[7] A.M. Noll, "Cepstrum pitch determination," The journal of the acoustical society of America, vol.41, no.2, pp.293–309, 1967.
[8] B.R. Ramakrishnan, "Reconstruction of incomplete spectrograms for robust speech recognition," Ph.D. thesis, Carnegie Mellon University, 2000.
[9] [Online] available: http://nltk.googlecode.com/svn/trunk/nltk_data/packages/corpora/timit.zip, 2013.
[10] ESTI document - ES 201 108 – STQ: DSR – Front-end feature extraction algorithm; compression algorithm, 2000.