A Novel Speech Reconstruction Algorithm for DSR Back-end

Jiang Wenbin, Ying Rendong and Liu Peilin
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, China

***Abstract*—In this paper, a novel speech reconstruction algorithm for DSR back-end is proposed. The algorithm is based on the classic least-squares estimate, inverse short-time Fourier transform magnitude (LSE-ISTFTM) algorithm. Unlike the classic LSE-ISTFTM algorithm, initializing speech waveform with white noise, the proposed method reconstructs voiced and unvoiced speech waveform separately, initializing with a specific signal. Furthermore, the magnitude spectrum is inversed from MFCC with Moore-Penrose pseudo-inverse by Mel-scale weighting functions. The algorithm evaluation results show that the proposed Extended LSE-ISTFTM algorithm converges faster and more stable than the classic algorithm. The speech reconstruction results demonstrate that PESQ score of the proposed algorithm is higher than the classic LSE-ISTFTM algorithm and the DSR back-end method.**

Keywords—speech reconstruction; speech synthesis; MFCC; minimum phase; cepstral analysis

I. INTRODUCTION

Speech synthesis/reconstruction is an important issue in speech codecs, Test-To-Speech (TTS), speech recognition and other speech processing systems. In recent year, Distributed Speech Recognition (DSR) [1] system has been widely used for mobile devices. DSR system transmits feature vectors, such as Mel-frequency cepstral coefficients (MFCC), to the back-end. Based on the speech signal sinusoidal model, the DSR back-end implemented a speech reconstruction algorithm with MFCC and pitch information [2, 3].

An algorithm for speech reconstruction solely from MFCC vectors is proposed in [4]. The algorithm predicts pitch and voicing by exploiting correlation between the fundamental frequency and the spectral envelope, and reconstructs speech waveform with the method of DSR back-end. The prediction 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, while the error of speaker-independent predictor is large. Besides, without objective quality measures or subjective tests, it is difficult to assure the quality of reconstructed speech.

Speech reconstruction from MFCC is a challenging task, since much information is lost by discarding the phase spectrum and Mel-scale weighting functions. An inversion method is proposed in [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 frames [6]. In addition, a low bit-rate speech coding scheme through quantization of MFCC was presented in [7]. The speech reconstruction progresses are the same as [5]. By means of Vector Quantization (VQ), the authors implemented 600-4800bps speech codecs. The test results of Perceptual Evaluation of Speech Quality (PESQ) [8, 9] show that the MFCC-based codec exceeds the MELPe codec [10]. Unfortunately, the speech waveform estimation algorithm, named LSE-ISTFTM [6], initializes speech with white noise, which leads to an unreliable speech reconstruction.

In this paper, we propose a novel speech reconstruction scheme for DSR back-end. The scheme firstly inverse the magnitude spectrum from MFCC with Moore-Penrose pseudo-inverse by Mel-scale weighting functions. And then, the voiced and unvoiced speech frames are reconstructed separately with the LSE-ISTFTM algorithm. Unlike the classic LSE-ISTFTM algorithm, initializing speech waveform with white noise, the proposed method initializes with a specific signal. For unvoiced frame, the algorithm initializes with minimum phase signal. For voiced frame, the algorithm initializes with a combined phase signal.

The organization of this paper is as follows: in Section II an overview of the proposed speech reconstruction scheme is introduced. In Section III, the inversion progresses of magnitude spectrum from MFCC are discussed. Section IV presents the proposed Extended LSE-ISTFTM algorithm in detail. Section V is the evaluation of the proposed algorithm and the speech reconstruction results. Finally, we conclude the paper in Section VI.

II. OVERVIEW OF THE SPEECH RECONSTRUTION ALGORITHUM

The proposed speech reconstruction algorithm is based on the classic inverse short-time Fourier transform magnitude (LSE-ISTFTM) algorithm. The algorithm estimates speech waveform from the short-time magnitude spectrum iteratively, in the absence of short-time phase spectrum. While, there are two challenging tasks with the classic algorithm for DSR back-

This paper was supported by National Natural Science Founding of China (NSFC) (61171171).

end speech reconstruction. Firstly, the DSR back-end receives MFCC vectors, rather than magnitude spectrum. There must be a conversion from MFCC to magnitude spectrum. Secondly, the reconstruction result of the classic LSE-ISTFTM algorithm is unstable, when the iteration times in an acceptable range, i.e. less than 200.

In order to deal with these tasks, we propose a novel method. The method obtains magnitude spectrum from MFCC with a series of inverse progress, reconstructs voiced and unvoiced speech frames separately. The Block diagram of the proposed method is presented in Figure 1.

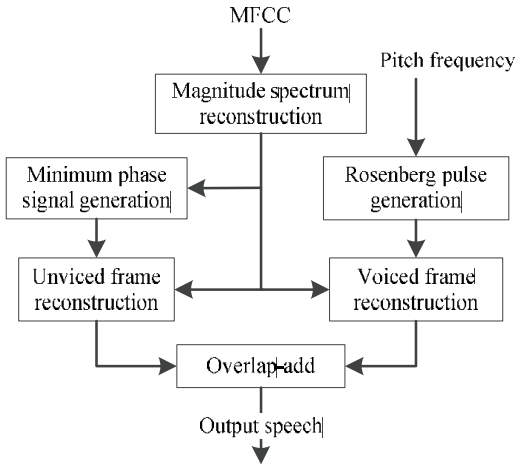


Fig. 1. Block diagram of the proposed speech reconstruction scheme for DSR back-end

The algorithm comprises following steps:

1) Magnitude spectrum reconstruction

The magnitude spectrum reconstruction is the inversion progresses of MFCC calculation. We apply Moore-Penrose pseudo-inverse to remove the Mel weighting function, and then utilize exponential and Inverse Discrete Cosine Transform (IDCT) operations to obtain magnitude spectrum.

2) Unvoiced frame reconstruction

In this step, we get the minimum phase signal with the magnitude spectrum obtained in step 1. Then we utilize minimum phase signal initialized LSE-ISTFTM algorithm (i.e. Extended LSE-ISTFTM for unvoiced) to reconstruct the unvoiced speech frame.

3) Voiced frame reconstruction

In this step, we generate Rosenberg pulse with pitch frequency. Then we synthesize a signal for the LSE-ISTFTM algorithm, whose phase spectrum is combine with Rosenberg pulse and the minimum phase signal, whose magnitude spectrum is obtained in step 1. The voiced speech frame is reconstructed with the synthetic signal initialized LSE-ISTFTM algorithm (i.e. Extended LSE-ISTFTM for voiced).

Finally, the complete speech waveform is reconstructed by overlap-add of the voiced and unvoiced frames. The detail of each step will be described in following sections.

III. MAGNITUDE SPECTRUM RECONSTRUCTION FROM MFCC

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 [7]. The Mel-cepstrum M of t^{th} frame speech $s_t(n)$ is computed as (the subscript t is dropped to simplify notation)

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

where w_m is the Mel-scale weighting function, and $S(\omega)$ is spectrum of $s(n)$.

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

$$|S(\omega)|^2 \approx w_m{}^{\dagger} w_m |S(\omega^{mel})|^2 \quad (2)$$

where $w_m{}^{\dagger} = (w_m{}^{\mathrm{T}} w_m)^{-1} w_m{}^{\mathrm{T}}$ is the Moore-Penrose pseudo inverse of matrix w_m, $S(\omega^{mel})$ is the inversed Mel-domain spectrum.

With the result of (2), the cepstrum of the reconstructed magnitude spectrum $|S(\omega)|$ can be calculated by

$$s(\tau) = F^{-1}\{\log|S(\omega)|\} \quad (3)$$

IV. SPEECH RECONSTRUCTION ALGORITHUM

The well-known LSE-ISTFTM algorithm can estimate speech frame from magnitude spectrum [6]. However, the algorithm initializes speech with white noise. As it will be shown in Section V, the algorithm is unreliable due to the quality of estimated speech is unstable. Furthermore, the digital model for speech [11] signal shows that: for voiced speech, excitation signal is generated by a pseudo-periodic sequence of impulses exciting a glottal filter; for unvoiced speech, excitation is generated by random noise. That is to say, voiced and unvoiced speech frames are generated with different way. In this section, we propose an Extended LSE-ISTFTM algorithm, which reconstructs voiced and unvoiced speech frames separately, initializes speech with a specific signal rather than white noise.

A. Reconstruction of the Unvoiced Frame

For unvoiced speech frame, the vocal system can be assumed to be minimum phase model. Minimum phase signal, whose rational z-transform has no poles or zeros outside the unit circle, can be represented by the real parts of their Fourier transforms [12]. Thus, we should be able to represent the

complex cepstrum of minimum phase signals by logarithm of the magnitude of the Fourier transform alone. The real part of the Fourier transforms is the Fourier transform of the even part of the sequence, the relationship of real cepstrum $c(\tau)$ and complex cepstrum $s(\tau)$ can be expressed as

$$c(\tau) = \frac{s(\tau) + s(-\tau)}{2} \tag{4}$$

For a minimum phase signal, the rational z-transform has no poles or zeros outside the unit circle. Then, the complex cepstrum

$$s(\tau) = 0, \quad \tau < 0 \tag{5}$$

Using (4) and (5), it is easily shown that

$$s(\tau) = \begin{cases} 0, & \tau < 0 \\ c(\tau), & \tau = 0 \\ 2c(\tau), & \tau > 0 \end{cases} \tag{6}$$

Thus, the complex cepstrum can be obtained by real cepstrum using (6), and the minimum phase signal $s_{\min}(n)$ can be computed by the complex cepstrum $s(\tau)$ with following equitation

$$s_{\min}(n) = F^{-1}\{\exp F[s(\tau)]\} \tag{7}$$

The proposed Extended LSE-ISTFTM algorithm for unvoiced speech frame initializes speech frame with minimum phase signal, and then iteratively estimates the phase spectrum and couples this to the given magnitude spectrum. The algorithm is halting when magnitude error e is smaller than the given threshold value or the iteration counter i reaches the maximum. The algorithm comprises following steps:

1) Initialize speech signal with the minimum phase signal $\tilde{s}_0 = s_{\min}$, initialize iteration counter i and magnitude error e

2) Compute the DFT of the speech signal, $\tilde{S}_i = \mathrm{DFT}[\tilde{s}_i]$, where the subscript i is the iteration counter

3) Calculate the magnitude error $e = \sum(|S| - |\tilde{S}_i|)^2$, where $|S|$ is the magnitude spectrum that reconstructed from MFCC

4) Couple the estimated phase to the given magnitude to get the estimated spectrum $S = |S| e^{j\arg[\tilde{S}_i]}$

5) Compute the IDFT of the estimated spectrum to get the estimated signal $\tilde{s}_i = \mathrm{IDFT}[S]$

6) Check the iteration counter i and magnitude error e, go to step 2 or halt.

B. Reconstruction of the Voiced Frame

For voiced speech frame, the vocal system function cannot be assumed to be minimum phase, because there is a glottal filter in the excitation. By studying of the effect of glottal pulse on speech quality, Rosenberg found that the natural glottal pulse waveform could be replace by a synthetic pulse waveform [13], which is called Rosenberg pulse:

$$g_R[n] = \begin{cases} 0.5(1 - \cos(\pi n / N_1)), & 0 \le n \le N_1 \\ \cos(\pi(n - N_1)/(2N_2)), & N_1 \le n \le N_1 + N_2 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where N_1 and N_2 are the parameters for the glottal opening and glottal closing durations. By varying N_1 and N_2, different glottal pulse duty cycles can be modeled and approximated.

Denote $R(\omega)$ and $\varphi_R(\omega)$ to the magnitude and phase spectra of Rosenberg pulse $g_R[n]$ respectively. The effect of glottal filter $G(z)$ can be removed by dividing the magnitude spectrum $S(\omega)$ by $R(\omega)$. The divided spectrum $S^v_{\min}(\omega) = S(\omega) / R(\omega)$ can be assumed to be minimum phase. The minimum phase signal and corresponding phase spectrum $\varphi^v_{\min}(\omega)$ can be obtained by the method in previous subsection. The synthetic phase $\varphi_{syn}(\omega)$ is obtained by adding Rosenberg pulse phase $\varphi_R(\omega)$ to minimum phase signal phase $\varphi^v_{\min}(\omega)$.

The proposed Extended LSE-ISTFTM algorithm for voiced speech frame initializes speech frame with synthetic phase signal, and then iteratively estimates the phase spectrum and couples this to the given magnitude spectrum. With the exception of the initial signal, other steps of the voiced frame iteration are the same as the unvoiced frame. Finally, the complete speech waveform is reconstructed by overlap-add of the voiced and unvoiced frames.

V. EXPERMENTAL RESULTS

In this section, we use TIMIT database [14] for evaluation. Each sentence is about 3 seconds duration, and down-sampling to 8 kHz. The corpus is framing to 240 samples (30ms) with hamming windows, and the frame shift is 120 samples (15ms). The number of Mel-filters and Mel-cepstral coefficients is 23.

A. Evaluation of the Extended LSE-ISTFTM Algorithm

Comparing with the classic LSE-ISTFTM algorithm, the proposed Extended LSE-ISTFTM algorithm is evaluated with in three aspects, phase spectrum errors, convergence speed, and stability. It should be notice that the magnitude spectrum for the both algorithms in this subsection is the calculated with the original speech signal.

The phase spectrum errors of initial and reconstructed speech frame with two methods are computed as

$$E_\varphi = \sqrt{\frac{1}{L}\sum_{i=1}^{L}[\varphi_i - \varphi'_i]^2} \tag{9}$$

Where L is length of Fourier transform, φ_i is original phase spectrum of i^{th} bin, and φ_i' is phase spectrum of i^{th} bin to be evaluated.

TABLE I. PHASE SPECTRUM ERRORS E_φ

Method	Initial speech	Reconstructed speech
LSE-ISTFTM	2.58	2.59
Extended LSE-ISTFTM	2.34	2.56

Table 1 shows mean value of E_φ for a speech with 210 frames, from which we can see that the proposed Extended LSE-ISTFTM exceeds the classic LSE-ISTFTM on both initial speech and reconstructed speech.

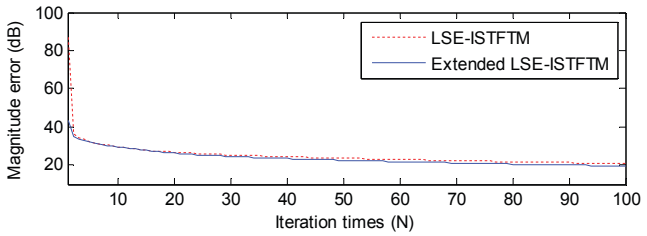


Fig. 2. Convergence of LSE-ISTFTM and Extended LSE-ISTFTM algorithm

Figure 2 shows the convergence of the proposed Extended LSE-ISTFTM and the classic LSE-ISTFTM algorithm. The proposed Extended LSE-ISTFTM algorithm (with blue solid line), starting the iteration at a lower magnitude error, converges faster than LSE-ISTFTM algorithm (with read dashed line).

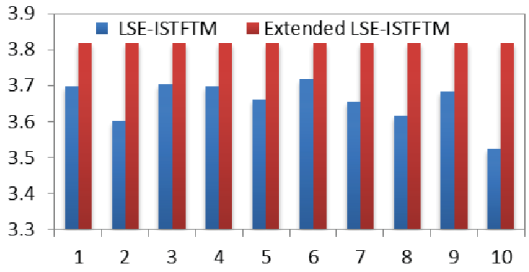
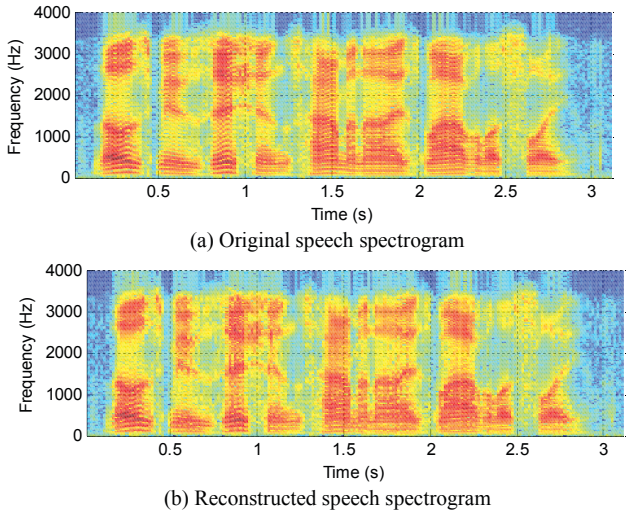


Fig. 3. PESQ scores of LSE-ISTFTM and Extended LSE-ISTFTM algorithm

PESQ is an objective measure of speech quality developed to obtain the highest correlation with subjective MOS. We utilize PESQ score to evaluate the classic LSE-ISTFTM and the proposed Extended LSE-ISTFTM algorithm. Figure 3 shows the reconstructed speech PESQ scores using two algorithms for the same speech signal. Within the ten times reconstruction results, we can see that the classic LSE-ISTFTM algorithm is unstable due to the white noise initialized speech. In comparison, the proposed extended LSE-ISTFTM algorithm is more stable, and the score is higher than the highest of the classic algorithm.

B. *Speech Reconstruction Results*

In this subsection, we evaluated the speech reconstruction result from MFCC. The magnitude spectrum is reconstructed with the method descripted in Section III, ant the speech waveform is reconstructed with the proposed Extended LSE-ISTFTM algorithm.



(a) Original speech spectrogram

(b) Reconstructed speech spectrogram

Fig. 4. Comparison of original and reconstructed spectrograms

Figure 4 illustrates the original and reconstructed spectrogram of a speech signal from TIMIT database. Only slight differences are observed between figures 4(a) and 4(b). With good performance, the proposed algorithm can be applied to the DSR back-end speech reconstruction. Table II shows the PESQ scores of reconstructed speech with the three methods: DSR back-end, the classic LSE-ISTFTM, and the proposed Extended LSE-ISTFTM. Comparing with the DSR back-end speech reconstruction method, the proposed method improves 8% in the evaluation of PESQ score.

TABLE II. PESQ SCORES OF THE RECONSTRUCTED SPEECH

Method	PESQ
DSR back-end	2.93
LSE-ISTFTM	3.05
Extended LSE-ISTFTM	3.17

VI. CONCLUSION

In this paper, a speech reconstruction method from MFCC for DSR back-end is presented. The method is based on the classic LSE-ISTFTM algorithm. Unlike the classic LSE-ISTFTM algorithm, initializing speech waveform with white noise, the proposed method initializes with a specific signal. For unvoiced frame, the algorithm initializes with minimum phase signal. For voiced frame, the algorithm initializes with a combined phase signal. The evaluation results show that the Extended LSE-ISTFTM algorithm is more stable, and the PESQ score is higher than the classic algorithm. The results of speech reconstruction from MFCC show that the proposed method is superior to the current method of the DSR back-end.

REFERENCES

[1] ESTI document - ES 201 108 – STQ: DSR – Front-end feature extraction algorithm; compression algorithm, 2000.

[2] D. Chazan, R. Hoory, G. Cohen, and M. Zibulski, "Speech reconstruction from mel frequency cepstral coefficients and pitch frequency," in Proc. ICASSP, 2000.

[3] Ramabadran, Tenkasi, et al. "The ETSI extended distributed speech recognition (DSR) standards: server-side speech reconstruction." in Proc. ICASSP, 2004.

[4] 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.

[5] L. E. Boucheron and P. L. D. Leon, "On the inversion of mel-frequency cepstral coefficients for speech enhancement applications," in Proc. ICSES, 2008, pp. 485–488.

[6] D. W. Griffin and J. S. Lim, "Signal estimation from modified short time flourier transform," IEEE Trans. Audio, Speech, Lang. Process., vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[7] L. E. Boucheron, P. L. D. Leon and Steven 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.

[8] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," in Proc. ICASSP, 2001, pp. 749–752.

[9] Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs, ITU-T Recommendation P.862, Int. Telecomm. Union, Telecomm. Standardization Sector, 2001.

[10] A. V. McCree and T. P. Bamwell, "Mixed excitation LPC vocoder model for low bit rate speech coding," IEEE Trans. Speech Audio Process., vol. 3, no. 4, pp. 242–250, Jul. 1995

[11] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[12] Institute of Electrical and Electronics Engineers. Programs for digital signal processing, Inst. of Electr. and Electronics Engineers, 1979.

[13] A. E. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels", Journal of the Acoustical Society of America, Vo1.49, No.2,

[14] [Online]. Available: http://web.mit.edu/6.863/share/data/corpora /timit/