Speech Synthesis using Mel-Cepstral Coefficient Feature

By

Lu Wang

Senior Thesis in Electrical Engineering

University of Illinois at Urbana-Champaign

Advisor: Professor Mark Hasegawa-Johnson

May 2018

Abstract

This thesis presents a method to improve quality of synthesized speech by reducing the vocoded effect. The synthesis model takes mel-cepstral coefficients and spectrum envelopes as features of the original speech waveform. Mel-cepstral coefficients could be used to generate natural sounding voice and reduce the artificial effect. Compared to regular linear predictive coding (LPC) coefficient which is also widely used in speech synthesis, mel-cepstral coefficient could resemble the human voice more closely by providing the synthesized speech with more details in the low frequency band. The model uses synthesis filter to estimate log spectrum including both zeros and poles in the transfer function, along with the mixed excitation technique which could divide speech signals into multiple frequency bands to better approximate natural speech production.

Subject Keywords: Speech Synthesis; Cepstrum Analysis

Contents

1. Introduction

In text-to-speech or image-to-speech synthesis process, the Back-end part involves the use of excitation vocoder to generate speech waveform. In human speech production, the excitation is generated with vocal folds and air flow in the lungs. Excitation models usually imitate human speech production process to have synthesized speech with natural tone. This thesis mainly explores the use of mel-scale in synthesizing natural sounding speech.

In the past years, LPC vocoder was broadly used in speech production [1]. However, it has some drawbacks in that it synthesizes "vocoded" speech. This thesis explores a model that generates more accurate outputs. The mixed excitation model based on MELP vocoder structure is proposed to enhance the naturalness of the speech. The model easily fits into modern HMM based text-to-speech systems [2]. Improvements provides by this model are due to the use of mel cepstrum which mimics human audio perception [3]. This method analyzes the log spectrum of speech in mel scale. The analysis of the original speech signal extracts mel cepstral coefficients to approximate the non-linear transfer function of mel cepstrum with reasonable time complexity [6].

This thesis compares and analyzes the performance of LPC coefficients and mel cepstral coefficients as extracted feature in speech synthesis. The experiments comprise two parts: extracting excitation based on mixed excitation methods and linear prediction method using human voice waveform as sources, and synthesizing back the speech waves corresponding speech parameters.

2. Background

2.1 Linear Predictive Coding

Linear predictive coding is broadly used in speech coding, speech recognition and speech synthesis. The LPC model uses linear combinations of previous values to predict the next value, and the transfer function of the LPC model contains only poles [5].

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + e[n] \tag{2.1}$$

Linear prediction coding provides coefficients to the all-pole filter. Since poles are generally more important than zeros with respect to the human auditory system, LPC performs well on detecting important features of the speech signal. The filter coefficients a_k represent important spectral envelope information. Different formulations, such as covariance method and autocorrelation method, have been developed in the past and can be used to obtain LPC coefficients.

2.2 Mel Cepstral Analysis

Cestrum is defined as the inverse Fourier transform of the logarithm of the spectrum, and offers the advantages of low spectral distortion, low sensitivity to noise and efficiency in representing log spectral envelop. In mel cepstral analysis, the log spectrum is non-uniform spaced in frequency scale.

Mel cepstral coefficients can be derived from LPC coefficients but with non-linear transfer function. Unlike the LPC model cares only filter poles, the mel cepstral model includes both poles and zeros.

2.2.1 Mel Scale Approximation

Mel cepstral analysis uses logarithmic spectrum on mel frequency scale to represent spectral envelopes and provide extra accuracy [4]. Mel frequency scale is particularly useful in speech synthesis. The mel frequency scale has a characteristic that it will expand the low frequency part and squeeze the high frequency part of the signal. Human ears have non-linear perception of frequency of sound, and are more sensitive to low frequency than to high frequency; therefore, mel frequency scale is more effective than linear frequency scale.

The generalized mel cepstrum is calculated on the warped frequency scale, which is approximated as $\beta_\alpha(\Omega)$ [3].

$$\beta_\alpha(\Omega) = \tan^{-1}\frac{(1-\alpha^2)\sin\Omega}{(1+\alpha^2)\cos\Omega - 2\alpha} \tag{2.2}$$

Different parameter α gives a different spectrum characteristics, Varying α to the appropriate value around 0.35 can better approximate the phase response $\beta_\alpha(\Omega)$ of the auditory system.

2.2.2 Mel Cepstral Adaptive Method

The desired mel cepstral coefficients minimize the spectral criterion estimating log spectrum in the mean square sense. The quality of the synthesized signal is optimized by minimizing the value of the unbiased estimator of log spectrum. Newton-Raphson method is used here for solving this minimization problem. Mel cepstrum is derived from LPC, and Mel cepstral coefficients could be calculated from LPC coefficients with recursive method. Minimizing the spectral criterion corresponds to minimizing the linear prediction error $\mathrm{e}(n)$ [3, 5].

The system is best represented with a mel log spectrum approximation(MLSA) Filter because of its low sensitives to noise and fine quantization characteristics [6]. MLSA filter is an adaptive IIR

filter which has filter coefficients obtained from cepstral analysis. MLSA filter has an all-pass transfer function as follows:

$$\mathrm{H(z)} = \exp\big(F_\alpha(z)\big) = \exp\sum_{m=0}^{M} c_\alpha(m)\tilde{z}^{-m} \tag{2.3}$$

$$\tilde{z} = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}} = e^{-j\beta_\alpha(\Omega)} \tag{2.4}$$

Because of the nonlinearity of the transfer function, the system cannot be directly implemented with a regular filter. Estimation of log spectrum involves non-linearity which could be solved by the iterative algorithm. The basic filter $F_\alpha(z)$ has adaptive implementation in IIR form. To obtain the minimum phase of the MLSA filter, the basic filter must be stable [4].

From cepstrum $\mathrm{c}(m)$, the filter parameter $\mathrm{b}(m)$ used in the basic filter implementation can be derived. The analysis model solves a set of linear equations and updates the gradient to find convergence of filter coefficients. $\mathrm{b}(m)$ is updated as follows, where μ is unit matrix:

$$b^{(i+1)} = b^{(i)} - \mu\nabla\varepsilon \tag{2.5}$$

Since the model spectrum also involves the exponential function which is not realizable, the exponential function is approximated with the cascaded form of a rational function. The adaptive analysis method typically needs a few iterations to have converged coefficients, which is computationally efficient [4].

2.3 Excitation Model

The most basic excitation model uses periodic pulses with fundamental frequency F_0 to represent voiced speech signal, and white noise to represent unvoiced signal. The voiced and unvoiced are differentiated by the short term energy frame to frame. The excitation model is improved with adding more features; one of the useful models is the mixed excitation model [2], with voiced and unvoiced decision set according to different passband.

2.3.1 Mixed Excitation

In order to limit the vocoded effect on synthesized speech, improvements are made in the mixed-excitation linear prediction (MLEP) vocoder [7]. Mixed excitation model is based on MELP and requires more spectral parameters to easily incorporate into the HMM-based TTS system because its parameters are all trainable. In the analysis process, beside pitch periods and white noises, the mixed excitation model also includes the mel cepstral coefficients presented previously as static features and its delta coefficients as dynamic features [2].

In the synthesizing stage, instead of directly applying the inverse synthesis filter to the sum of the white noise and pulse train, the periodic pulse train represented voiced signal and Gaussian white noise represented voiced signal are filtered with a bandpass filter to determine the frequency band of voiced and unvoiced signal [7]. The voiced and unvoiced decision of the passband is determined by the voicing strength which is estimated with correlation function. The vocoder uses aperiodic pulses in the transition from voiced to unvoiced, and periodic pulses for elsewhere in voiced speech. The entire frequency band of the signal is evenly divided into four frequency passbands from 0 to 8000 Hz. The synthesized speech is obtained by applying inverse synthesis filter to the mix of filtered pulses and noise excitation. The synthesis filter for this model has the structure of the MLSA filter.

3. Experiments and Results

3.1 Overall Experiment Procedure

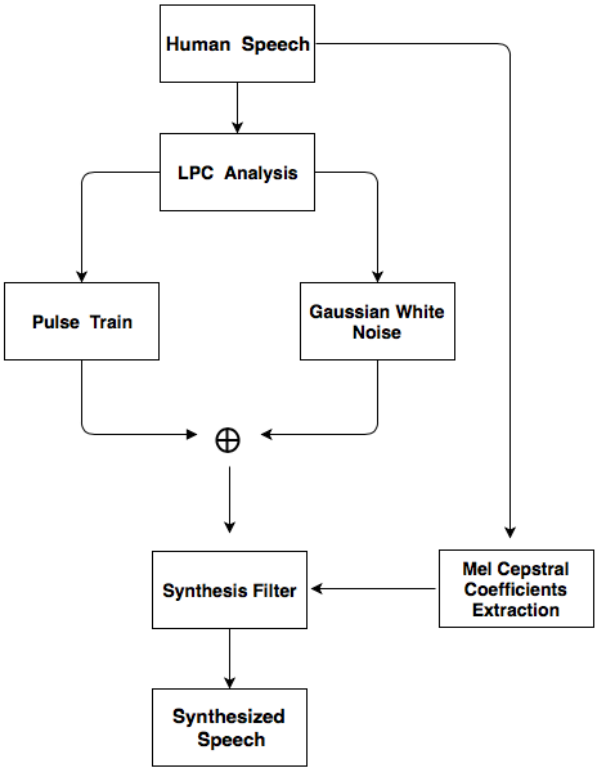

Fig. 3.1 Synthesis Procedure

Figure 3.1 shows the overall procedure to test the efficiency of mel cepstral coefficients for speech synthesis. The pulse train and voiced and unvoiced decision information is obtained with linear prediction method. Synthesis filter with MLSA structure and mel cepstral coefficients synthesize the excitation signal provided by the LPC analysis part [5, 8]. For comparison purpose, the system also synthesize speech with linear prediction coefficients which are obtained in the analysis of the original speech. The source speech waveforms are from TIMIT speech corpus, which contains human-pronounced short sentences that are sampled at 16 kHz.

3.2 Excitation Parameters Extraction

In order to produce natural sounding speech at low bit rate, the parameters representing speech information effectively need to be extracted from source files. The excitation signal usually requests fundamental frequency F0, spectral envelopes information and voiced and unvoiced decision parameters. LP coefficients and mel-cepstral coefficients are extracted for synthesis.

3.2.1 Pitch Detection

The source signal is sampled at 16k Hz rate, and is converted to frames each of length 30 ms. Each frames is applied with hamming window $w(n)$. Pitch period is selected from 50 to 350 with the one minimizing the error between synthesized speech energy and original speech energy. The error criterion is calculated over the range of pitch period. The equation is given below where $S_W(\omega)$ is original speech spectrum, and $\hat{S}_W(\omega)$ is synthesized speech spectrum.

$$\varepsilon = \frac{\int |S_W(\omega) - \hat{S}_W(\omega)|^2 d\omega}{(1 - P\sum_{-\infty}^{\infty} w^4(n)) \int |S_W(\omega)|^2 d\omega} \tag{3.1}$$

3.2.2 Voiced and Unvoiced Decision

Most of spoken language contains vowels and consonants. Vowels are composed with purely voiced signal and consonants with both voiced and unvoiced signal. Therefore, separating voiced and unvoiced segments of speech signal is an effective way to get excitation. U/V binary decision parameter is derived from analysis of the energy of the speech segments [8]. The equation for normalized error to distinguish the voiced or unvoiced band is given below:

$$\xi_m = \frac{\varepsilon_m}{\frac{1}{2\pi}\int_{a_m}^{b_m} |S_W(\omega)|^2} \tag{3.2}$$

where ε_m is the error criterion, and $S_W(\omega)$ is the windowed speech segments [8]. Set the bands with normalized error below threshold (around 0.2) to unvoiced and above the threshold to voiced.

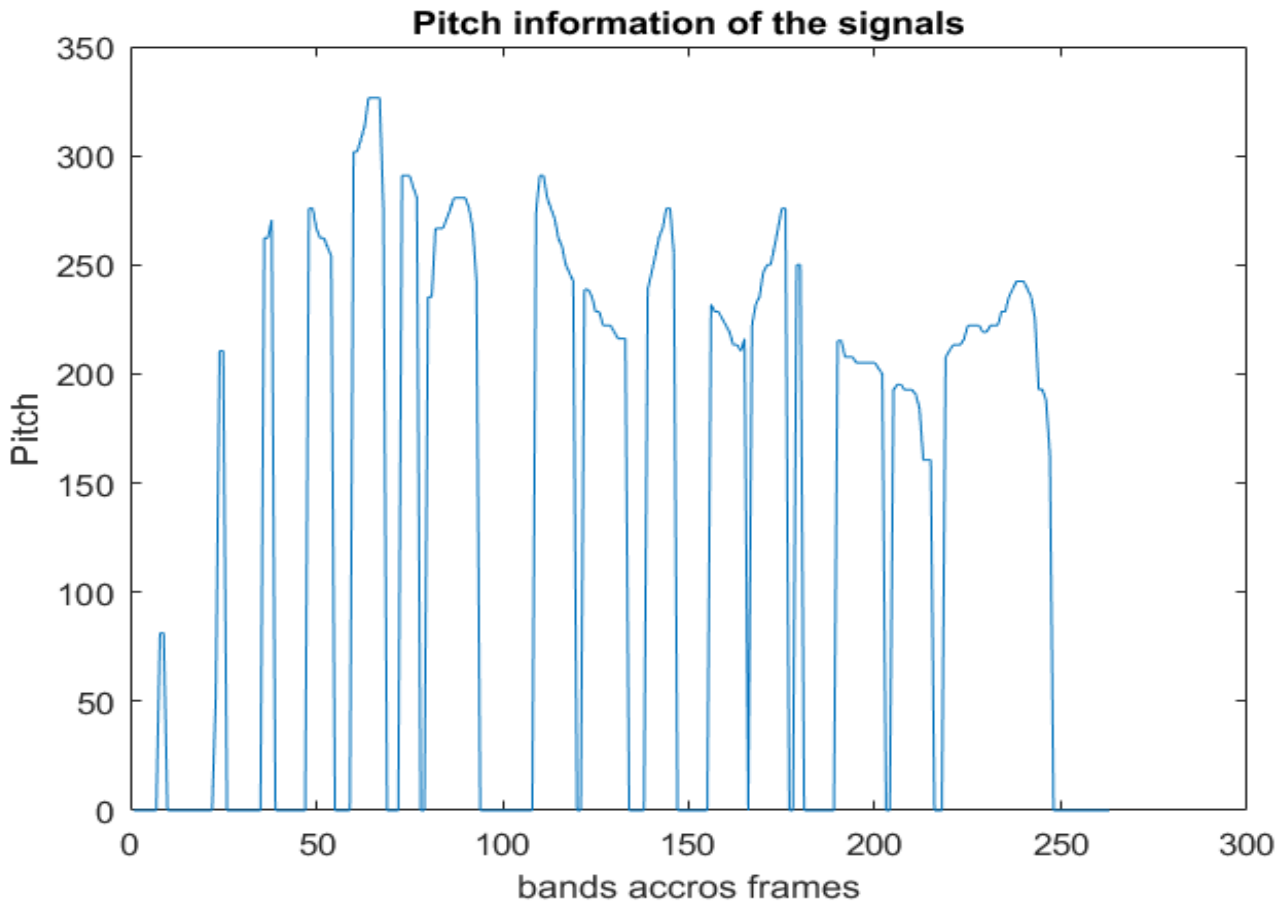

Fig. 3.3 U/V Decision of the Speech Signal

Figure 3.3 illustrates waveform's voiced and unvoiced decision for the excitation with multiband excitation technique. The results are refined with the use of pitch tracking technique to ensure adjacent voiced pitches cannot vary by large pitch periods to prevent sharpness transition in synthesized speech. The frequency bandwidths are determined based on the harmonic location of the current frames [8].

3.2.3 Linear Prediction Vocoder

Linear prediction vocoder is a popular structure used to obtain excitation signal from raw speech. Predictor coefficients contain information about vocal tracts and glottal flow [5]. In the analysis

process, predictor coefficients are extracted by the autocorrelation method from which mel cepstral coefficients can also be derived [5]. Linear prediction values obtained with applying adaptive linear predictor to excitation and the original excitation forms synthesized speech.

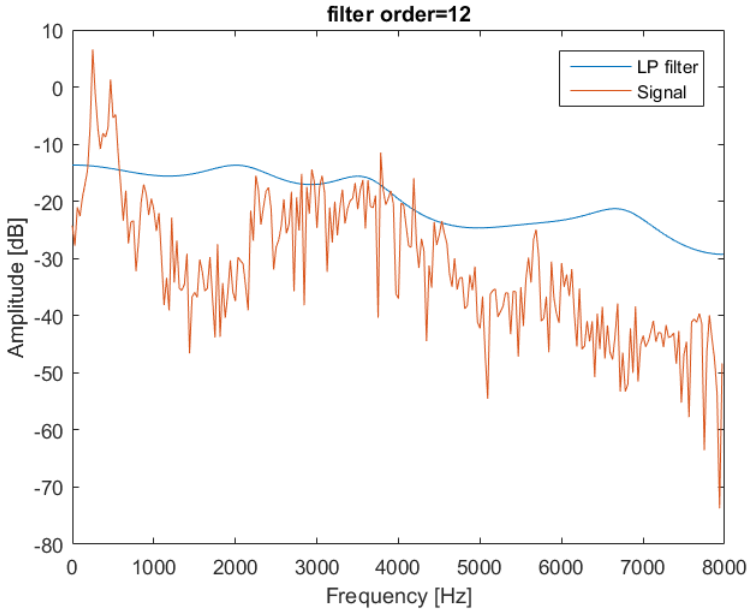

Fig. 3.2 Frequency Response of LP Filter

Figure 3.2 shows frequency response of the transfer function with linear prediction filter order = 12. Typically, an order 12 LP filter could represent the frame's information reasonably.

3.3 Synthesis Results Comparison

The synthesized waveforms are generated from mixing the voiced pulse trains and unvoiced Gaussian white noise based on U/V decision parameters and filtering with inverse LP and MLSA synthesis filter, which were introduced in the previous section. The synthesis systems process the excitation signal in time domain. Both methods could generate easily distinguishable voice outputs, but with different properties.

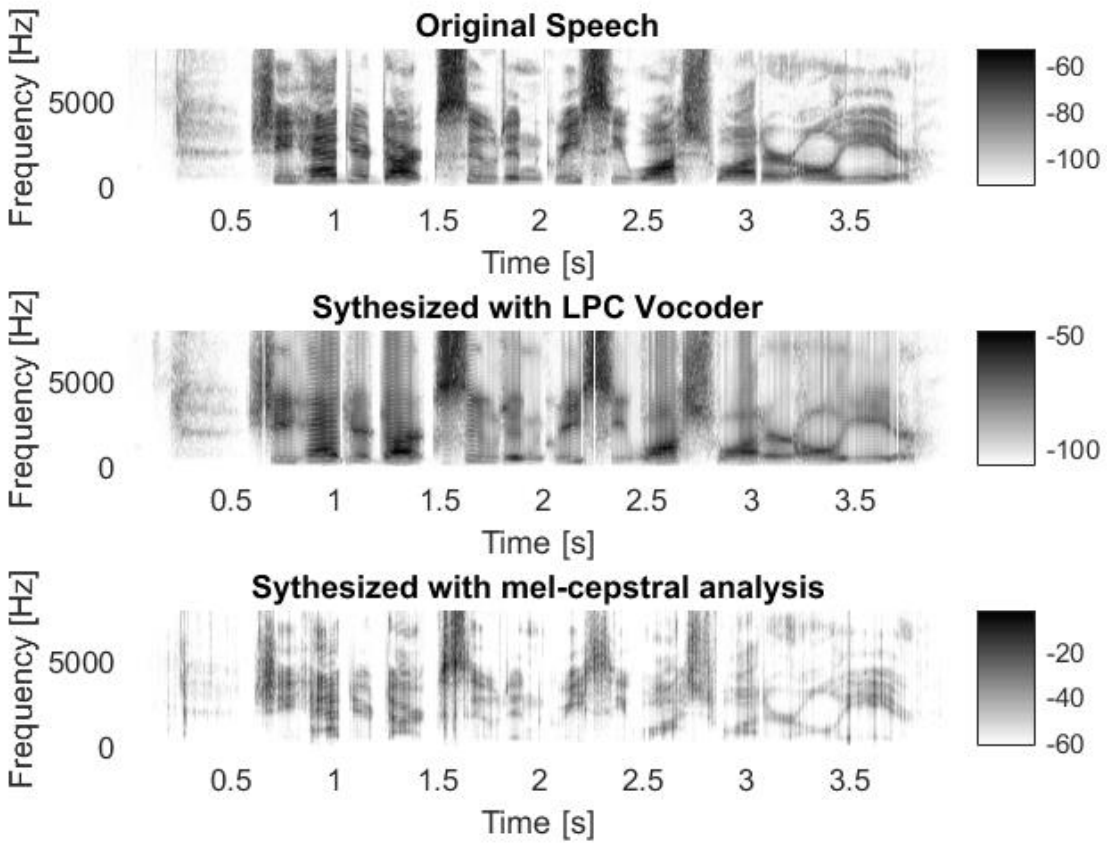

Fig. 3.4 Spectrograms for Source Speech, LP and Mel Cepstral Synthesized Speech

Figure 3.4 shows the spectrograms of source waveform and speech waveforms obtained with linear prediction and mel cepstral analysis. From the spectrogram, we can observe that the characteristics of the vocoded speech, such as fundamental frequency, are matched with the original speech. Because of the nonlinearity of the MLSA filter's transfer function, the synthesized speech includes more details on lower frequency than upper frequency as expected. Adjustment can be made to make MLSA filter have better performance than the linear prediction vocoder

4. Conclusion and Future Work

The experimental result shows a relatively smooth synthesized speech. The experiments confirm that using mel scale in speech synthesis has the advantages of insensitivity to noise and its efficiency in representing spectral information with non-linear transfer function, for it is designed to mimic the human auditory system. The algorithm converges fast and the synthesis system provides a low bit-rate coding of speech, thus the synthesis method could benefit to larger scale TTS system with higher operating speed. Improvements can be made with a more accurate pitch tracking technique such as Viterbi algorithm and includes dynamic feature, such as cepstral coefficients' delta and delta's delta to provide extra smoothness and clarity to the speech wave [2].

References

[1] B. S. Atal and S. L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” in The Journal of the Acoustical Society of America, 50:2B, pp. 637-655, 1971.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Mixed excitation for HMM-based speech Synthesis,” in Proc. of Eurospeech, pp.2259-2262, Sept. 2001.

[3] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, “Mel-generalized cepstral analysis – a unified approach to speech spectral estimation,” in Proc. ICSLP, pp. 1043-1046, 1994.

[4] T. Fukuda; K. Tokuda, T. Kobayashi, S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 137-140 vol.1, 1992.

[5] L. R. Rabiner and R. W. Schafer, Theory and Applications of Digital Speech Processing, 1st ed. Pearson/Prentice Hall, ch. pp. 137-140, 2011.

[6] S. Imai, “Cepstral analysis synthesis on the mel frequency scale,” in ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 93-96, 1983.

[7] A. V. McCree and T. P. Barnwell, “A mixed excitation LPC vocoder model for low bit rate speech coding,” in IEEE Transactions on Speech and Audio Processing, vol. 3, no. 4, pp. 242-250, Jul 1995.

[8] D. W. Griffin and J. S. Lim, “Multiband excitation vocoder,” in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 36, no. 8, pp. 1223-1235, Aug. 1988.