AN EXCITATION MODEL BASED ON INVERSE FILTERING FOR SPEECH ANALYSIS AND SYNTHESIS

Zhengqi Wen, Jianhua Tao

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{zqwen,jhtao}@nlpr.ia.ac.cn

ABSTRACT

Speech Synthesized in LPC-like vocoders suffered from a typical buzz problem. It is mostly due to the fact that the excitation is either a pulse train or a white Gaussian noise. In this paper, a new excitation model is proposed to reconstruct residual signal derived from inverse filtering. A residual frame of two-pitch periods length is intercepted to do spectrum analysis in every speech frame. Amplitude spectrum of only half of pitch period length is preserved in synthesis stage and zero-phase criterion is used to synthesize the excitation frame. Then the excitation signal is constructed by pitch-synchronous overlapping method (PSOLA). Speech synthesized by this excitation model can give a CMOS of 1.56 compared to the traditional excitation model. After that Mel Generalization Cepstrum (MGC) and LBG algorithm are adopted to manipulate the amplitude spectrum of proposed excitation model. MSE distortion and listening test showed that LBG algorithm is better than MGC to compress the amplitude spectrum.

Index Terms—excitation model, inverse filtering, PSOLA, MGC, LBG

1. INTRODUCTION

The parametric representation of speech signal is an essential operation in many speech processing tasks, e.g., speech synthesis, speech recognition, speech enhancement and so on. Statistical parametric speech synthesizers have shown their ability to produce natural-sounding voices [1] which is mostly due to the accurate models used to manipulate the speech signal.

During the past three decades, several models were proposed to manipulate speech signal. In source-filter model, speech is comprised of a source originating from the vocal cords and a shaping filter imitating the effect of the vocal tract. In sinusoidal model [2], speech is represented as a sum of sinusoids and phase and amplitude information are preserved to synthesis speech. In this model, only harmonic structure is preserved in synthesized speech while aperiodic information is ignored. In Harmonic plus Noise Model [3], speech is treated as a linear combination of a number of harmonically related sinusoids and a high-passed white Gaussian noise. STRAIGHT [4] proposed by Kawahara uses pitch-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency domain, and an excitation source model based on group delay manipulation. All these methods have showed their ability to manipulate the speech signal while maintaining high reproduction quality.

In this paper, we devote our attention to excitation constructing which is relaying on inverse filtering technique in source-filter model. In LPC-like vocoders [5], spectral coefficients are used to manipulate the effects of the source, vocal tract and lip radiation. Due to the complexity of speech signal, some spectral information is lost in these spectral coefficients and residual derived from inverse filtering does not have a flat spectrum. So our proposed excitation model is committed to reconstruct the residual signal and conserve the lost spectral information. In this model, a residual frame of two-pitch periods length is intercepted to do spectrum analysis. Amplitude spectrum of only half of pitch period length is preserved in synthesis stage. An excitation frame of two-pitch periods length is synthesized by IDFT of the preserved amplitude spectrum with zero-phase criterion. And the excitation signal is constructed by pitch-synchronous overlapping method (PSOLA, [6]). Speech synthesized by this excitation model can give a Comparative Mean Opinion Score (CMOS, [7]) of 1.56 compared to the traditional excitation model. For preserved spectrums having different length, we normalize these spectrums to a const length. And then MGC [8] and LBG algorithm [9] are adopted to model the normalized spectrums. MSE distortion and listening test showed that LBG algorithm is better than MGC to compress the spectrum.

The remaining part of this paper is organized as follows. Section 2 gives a detail description of the excitation model proposed in this paper. Section 3 illustrates a vocoder based
on this excitation model. In Section 4 MGC and LBG algorithm are adopted to model the normalized amplitude spectrum. Then listening tests are conducted and results are given in Section 5. Finally in Section 6, conclusions and future work are summarized.

2. EXCITATION MODEL BASED ON INVERSE FILTERING

In the frequency domain, the speech production model can be represented by

\[ S(\omega) = D(\omega)G(\omega)V(\omega)R(\omega) \]  

where \( D(\omega) \) is the Fourier Transform (FT) of an impulse train, \( G(\omega) \) is the FT of a glottal pulse, \( V(\omega) \) is the vocal tract transfer function and \( R(\omega) \) is the radiation characteristic.

In LPC-like vocoders, this model can be simplified to

\[ S(\omega) = D(\omega)H(\omega) \]  

where \( H(\omega) \) represents the spectral envelope of the speech signal which combines the effects of the source, vocal tract and lip radiation.

In [4], Kawahara mentioned that this approach does not provide reliable estimates for natural speech because this method assumes that the autocorrelation function of the periodic source is a regular pulse train and the spectrum model only represents the auto-regressive components of natural speech. In order to preserve the speech flexibility, we adopt the inverse filtering technique and reconstruct the residual as the excitation.

2.1. Inverse Filter

Inverse filtering is used to cancel the effects of the vocal tract transfer function and recover the glottal flow [10].

The simplified case of the speech production model has been represented by Equ.2 where only an excitation \( D(\omega) \) and a filter \( H(\omega) \) are considered. Inverse filtering is conducted in Equ.3. Due to the filter \( H(\omega) \) in the LPC-like vocoders only represents the auto-regressive components of natural speech, residual \( D(\omega) \) got from inverse filtering contains the spectral information which can be fund in Figure 2.

\[ D(\omega) = S(\omega) / H(\omega) \]  

Figure 2: The amplitude spectrum of a two-pitch length residual frame.

Figure 3: The preserved amplitude spectrum of a two-pitch length residual frame.

2.2. Spectral Analysis

Spectral analysis is affected by the window type and window length. In [11], Harris made a concise review of data windows and their effect on the detection of harmonic signals in the presence of broad-band noise, and in the presence of nearby strong harmonic interference. In [4], Kawahara concluded that if the signal is purely periodic and the period is an integer multiple of the sampling period, a pitch-synchronous analysis can perfectly eliminate temporal variations by using a rectangular window which length is an integer multiple of the fundamental period in samples.

In this paper, we assume that the signal is purely periodic and adopt a rectangular window whose length is two-pitch periods in sample. Let \( s(k), \ k = 1,2,\cdots,K \) be a residual frame of two-pitch periods length with corresponding discrete Fourier transform (DFT) \( S(n), n = 1,2,\cdots,N \). It is clear that the odd lines and even lines of \( S(n) \) contain periodic components and aperiodic components of \( s(k) \), respectively. In Figure 2, we
can find that aperiodic components of \( s(k) \) are nearly zero and only periodic components preserve the spectral information. Beside this, spectrum \( S(n) \) in Figure 2 shows a symmetrical characteristic. So length of half of pitch period \( \Delta T \) as expressed in Figure 3 is enough to conserve the spectral information.

**2.3. Excitation Construct**

In excitation constructing stage, we adopt the inverse discrete Fourier transform (IDFT) technique and consider the amplitude spectrum with zero-phase criterion. The result of IDFT is complex numbers of two-pitch periods length and we take the real part of these complex numbers as the excitation. Figure 4 showed a frame of the constructed excitation.

![Excitation Wave](image)

Figure 4: An excitation frame of two-pitch periods length from IDFT of the amplitude spectrum with zero-phase.

### 3. VOCODER WITH PROPOSED EXCITATION MODEL

A workflow summarizing the vocoder based on the excitation model proposed in previous section can be found in Figure 5. In voiced region, an excitation frame of two-pitch periods length is synthesized from the pitch and amplitude spectrum and these frames added together by applying PSOLA [12]. In unvoiced region, white Gaussian noise is used to model the unvoiced residual. And then these two components add together so as to generate the excitation signal. Finally the excitation signal is the input of the filter to engender the synthesized speech.

![Workflow](image)

Figure 5: The workflow of vocoder with the proposed excitation model.

### 4. AMPLITUDE SPECTRUM COMPRESS

Vocoder illustrated in previous section has an extra part of amplitude spectrum comparing to traditional excitation model. This part is a compensation for the vocal tract filter \( H(\omega) \).

In Section 2, the length of amplitude spectrum is half of pitch period. Before compressing, these spectrum frames should be normalized to a const length. In [13], the pitch value \( F_s' \) for normalization is such that:

\[
F_s' \leq \frac{F_s}{F_{\text{max}}} \quad (4)
\]

In Equ.4, \( F_s \) and \( F_{\text{max}} \) respectively denote the maximum frequency of the deterministic part, the Nyquist frequency and the minimum pitch value of the considered speaker. In [14], the speaker’s pitch histogram \( P(F_s) \) is analyzed and the normalized pitch value \( F_s' \) satisfies:

\[
\int_{F_s'} P(F_s)dF_s = 0.8 \quad (5)
\]

In this paper, we normalize the amplitude spectrum to a length of 127 points. There are two reasons for this. Firstly, the frequency range of amplitude spectrum preserved by such length is 60Hz–16000Hz. It is enough for reserving a male or a female frequency range. Secondly, such length of spectrum is fitted to do spectrum compressing in the next step.

After normalization, MGC [8] and LBG algorithm [9] are adopted to moderate the amplitude spectrum.

#### 4.1. Mel-Generalization Cepstrum

In Mel-Generalization Cepstrum analysis, a speech spectrum is modeled by the MGC coefficients in Equ.6.

\[
H(z) = \begin{cases} 1 + \gamma \sum_{m=0}^{N} c(m)z^{-m} \gamma, & -1 \leq \gamma < 0 \\ \exp \sum_{m=0}^{N} c(m)z^{-m}, & \gamma = 0 \end{cases} \quad (6)
\]

where \( z \) is an all-pass transfer function defined by

\[
\bar{z}^{-1} = \left(1 - \alpha \bar{z}^{-1}\right)^{-1} \mid a \mid < 1 \quad (7)
\]

And the parameters \( \alpha \) and \( \gamma \) control the frequency warping and the weight for pole/zero representation, respectively.

MGC has shown its ability to moderate the spectrum envelope in [15] where 40 Mel-Cepstral coefficients including the zeroth coefficient were extracted from the STRAIGHT spectrum envelope.

#### 4.2. LBG Algorithm
Vector quantization (VQ) is an important and powerful technique for data compression. LBG is a classical algorithm for VQ which is a finite sequence of steps. At every step, a new quantizer with a total distortion less or equal to the previous one is produced. A workflow of LBG algorithm is showed in Figure 6.

5. EXPERIMENTS AND RESULTS

Our experiments are divided into two parts. We will firstly evaluate the effectiveness of proposed excitation model and then evaluate the effectiveness of these two compressing methods. While in the second part, we will calculate the deviation between the original spectrum and the compressed spectrum and also assess the quality of synthesized speech after compressed through a listening test.

The database used in these experiments was utilized in the speech synthesis system. In this database, a female Chinese speaker was asked to read sentences in a quiet environment as a recording studio. And the pitch contours of these sentences are manual annotated by the labeling staffs. When annotating, the labeling staffs were asked to listen the synthesized speech by using the annotated pitch contour and then modulated the pitch contour. In this way, more smooth pitch contours could be obtained. The spectrum coefficient used in the spectral analysis is linear predictive coefficient and the order is 39.

5.1. Effectiveness of Proposed Excitation Model

Two versions of sentences including speech synthesized by the proposed excitation model and speech synthesized by the traditional excitation model are given. These sentences are divided into three groups based on the different length. Ten people are participated in this listening test and they are asked to give a CMOS (see Table 1) with one decimal precision to every sentence.

Table 1: Grades in the COMS scale

<table>
<thead>
<tr>
<th>CMOS</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Much Better</td>
</tr>
<tr>
<td>2</td>
<td>Better</td>
</tr>
<tr>
<td>1</td>
<td>Slightly better</td>
</tr>
<tr>
<td>0</td>
<td>About the same</td>
</tr>
<tr>
<td>-1</td>
<td>Slightly worse</td>
</tr>
<tr>
<td>-2</td>
<td>Worse</td>
</tr>
<tr>
<td>-3</td>
<td>Much worse</td>
</tr>
</tbody>
</table>

Figure 7: Average CMOS in advantage of the proposed excitation model over traditional excitation model.

The results showed in Figure 7 illustrate that the excitation model proposed in this paper has significantly improved the quality of synthesized speech comparing to the traditional excitation model. This model can give an average CMOS of 1.56 over traditional excitation model. Difference score between three groups indicates that difference rhythm information can influence on the effectiveness of proposed excitation model.

5.2. Effectiveness of Compressing Methods

The amplitude spectrum is normalized into a const length of 257 samples in Section 2. And in Section 4, MGC and LBG algorithm are used to compress the normalized spectrum. In compressing stage, the order of MGC is 10 and the size of codebook used in LBG algorithm is 32768. The compressed spectrum of these two methods and original spectrum are showed in Figure 8. And then the mean square error (MSE) distortion is used to measure the difference between compressed spectrum and original spectrum.

Figure 8: Two ways to compress the amplitude spectrum. Top: the solid line is the original spectrum, the dashed line
is spectrum got from MGC and the dotted line is spectrum got from LBG algorithm. Down: The MSE distortion of MGC spectrum (solid line) and LBG spectrum (dotted line).

MSE distortion showed in Figure 8 demonstrates that the LBG algorithm is better than MGC to compress the amplitude spectrum. The LBG algorithm preserves more spectrum information while the spectrum got form MGC is too smoothing.

A listening test is carried out to evaluate the effectiveness of these compressed methods. Four versions of speech are used in this test. They are original speech, speech synthesized from original spectrum and speech synthesized from the spectrum compressed by LBG algorithm and speech synthesized from the spectrum compressed by MGC. Ten participants are asked to listen to 30 sentences and give a mean opinion score (MOS) to every sentence. These sentences divide into three groups according to different length.

![Figure 9: The average MOS of speech synthesized by the original spectrum, spectrum compressed by LBG algorithm and spectrum compressed by MGC.](image)

The results of listening test can be viewed in Figure 9. The participants preferred the speech synthesized from the spectrum compressed by LBG algorithm to MGC. These results are consistent with the conclusion got from Figure 8. Figure 7 showed that the effectiveness of proposed excitation model can be influenced by the rhythm information. But this phenomenon can not be found in Figure 9 and the proposed excitation model can give an average MOS of 4.2.

6. CONCLUSION AND FUTURE WORKS

In this paper, a new excitation model based on inverse filtering is proposed to reconstruct the residual. A listening test is carried out to evaluate the effectiveness of the proposed excitation model. The result showed that this model can give a CMOS of 1.56 over traditional excitation model. After that, two compressing methods are used to modulate the amplitude spectrum. The results of MSE distortion and listening test showed that the LBG algorithm is better than MGC to compress the amplitude spectrum.

However, we did not take the phase information and aperiodic components into consideration in this excitation model. So in the future work, we will manipulate the phase information and then add the aperiodic components to improve the naturalness of synthesized speech. And also we will integrate this excitation model into speech synthesis system.

7. REFERENCES