PRODUCTION BASED PITCH MODIFICATION OF VOICED SPEECH

Yinglong Jiang, Peter Murphy

Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland

ABSTRACT

Previous research has shown that the voice source is strongly correlated with speech quality [1][2][3]. However in many existing pitch modification algorithms only the impulse train excitation is modified, while the voice source is normally included in vocal tract transfer function and remains unchanged during modification. We present a production based pitch-scale modification scheme, which modifies the voice source waveform extracted from the speech signal. Experiment on synthesized voice shows that compared to previous methods, better quality can be achieved for a wider pitch modification range.

1. INTRODUCTION

In many applications, it is useful to be able to implement a pitch-scale modification of the speech waveform. For example, in concatenative speech synthesis systems, pitch is modified to a certain contour in a sentence for intonation.

A pitch modification process normally includes three stages:

- Analysis stage: transfers the speech signal into another representation that facilitates the intended modification.
- Modification stage: modifies the representation corresponding to the new pitch. The spectral contour remains the same so that the formant frequencies and bandwidths are unchanged following modification.
- Synthesis stage: combines the modified part and unmodified spectral envelope, concatenates consecutive segments and generates the modified speech waveform.

The linear predictive coding (LPC) based method is probably the most widely used representation for pitch-scale modification [4]. It decomposes the speech signal into a slowly time-varying spectral envelope and an excitation signal. The spectral envelope is represented by a set of coefficients of an all-pole linear filter. By keeping the filter constant during the modification stage, the formant frequency and bandwidth remains unchanged following modification. Ideally, the excitation signal is a series of impulses spaced at the pitch period. During the modification, the impulses are scaled to a new pitch period.

More recently, a sinusoidal model based method was also developed for representing speech signals [5]. It represents the speech signal as a sum of sine-wave components. Each sine-wave has a time varying amplitude, frequency and phase. The amplitude is the product of the excitation signal amplitude and the vocal tract system amplitude. The phase is the sum of the excitation phase and the vocal tract system phase. In the modification stage, the model modifies the excitation signal frequency and phase to a new frequency and phase corresponding to the new pitch.

It has been shown that these pitch modification methods can achieve high efficiency and reasonable modification quality. However problems arise when altering the pitch over a large scale, especially for increasing pitch. For example, it is reported that when using the sinusoidal model, when pitch is changed by a factor greater than 20%, hoarseness is presented in the reconstruction [5]. Similar problems occur with other methods.

2. VOICE SOURCE ANALYSIS FOR PITCH MODIFICATION

The modification methods mentioned above have one characteristic in common, i.e. essentially, during the modification stage, the part to be modified has an impulse train shape in the time domain (as in the LPC model) or a flat spectrum in frequency domain (as in the sinusoidal model). A question arises as to whether this is the adequate part of the signal to be modified.

Referring to the speech production source-filter model:

$$s(n) = e(n) * g(n) * v(n) + r(n)$$

(1)

In this model, the speech signal is seen to be a convolution of an impulse train $e(t)$ (as excitation to the system), a glottal shaping filter $g(n)$, a vocal tract transfer function $v(n)$ and a lip radiation function $r(n)$. The glottal shaping filter reflects the voice source which excites the vocal tract. It can be derived from the speech signal by glottal inverse filtering.
Fig. 1. Illustration of the glottal air flow during impulse train scaling in pitch modification

techniques such as Iterative Adaptive Inverse Filtering (IAIF) [6]. We mark the pitch period before modification as \( T \), the modification methods mentioned previously modify an excitation signal similar to \( e(t) \) (in Eq. (1)) to a new pitch period \( T' \). The glottal shaping filter \( g(n) \) is included as part of the vocal tract transfer function and remains un-changed during modification.

To understand the effects of the impulse train scaling during pitch modification, we consider the impulses spaced at the pitch period convolved with the glottal flow waveform. The upper panel in Fig. 1 shows the voice source waveform. The non-zero parts of the waveform correspond to glottal vibration in the open phase, during which there is airflow through the glottis. The zero parts between correspond to the closed phase, during which there is no airflow through the glottis.

After the pitch-scale modification, the new voice source is obtained by convolving the impulse train spaced at the new pitch period with the glottal flow waveform. If the glottal flow waveform remains unchanged during modification, the convolution results in a new voice source that has the same open phase, but the duration of the closed phase has been altered, as shown in the mid panel in Fig. 1. In the case of modification with increased pitch, the pitch period is shortened to the open phase duration, when the closed phase has just disappeared. Further increase in pitch causes overlap between the consecutive open phases, as shown in the lower panel in Fig. 1.

Research on the relationship between the voice source and speech quality shows there is a high degree of correlation between the glottal flow waveform shape and voice quality, with a shorter closed phase relating to a pressed [1] or falsetto sound [2]. In addition in the limited available experiment data, it has been shown that the open phase and close phase are not strongly correlated with pitch [3].

In our previous work [7], we used a formant synthesizer to simulate the pitch modification by impulse train scaling and by voice source scaling. It is shown that when pitch increases, a distorted, unnatural sound appears in the synthesized speech that simulates modification based on impulse train scaling. The point the distortion appears is the point that the voice source open phase starts to overlap.

Both the simulation using the synthesized speech and the observation from real speech suggests that the voice source waveform, instead of impulse train excitation, should be modified in the pitch-scale modification schemes.

3. PRODUCTION BASED MODIFICATION

3.1. Production based modification

This production based pitch modification scheme is also a three stage method:

- **Voice source extraction stage:** voice source waveform \( g(n) \) (in Eq. (1)) is extracted from the speech signal by inverse filtering;
- **Modification stage:** modify the voice source waveform \( e(n) \) to the new waveform corresponding to the new pitch period, the vocal tract transfer function \( v(n) \) remains unchanged;
- **Synthesis stage:** synthesize the modified voice source waveform and the vocal tract transfer function \( v(n) \) to produce the modified speech waveform.

3.2. Voice source extraction

In the production based modification scheme, the voice source extraction stage is the crucial processing. A pitch synchronous iterative adaptive inverse filtering (PSIAIF) algorithm based on iterative adaptive inverse filtering (IAIF) algorithm [6] is used for voice source waveform extraction. Fig. 2 shows the flowchart of the IAIF algorithm: The PSIAIF method applies the IAIF first to obtain the glottal pulse and the pitch period for pitch synchronous analysis and then applies the IAIF again pitch synchronously.

The PSIAIF algorithm has been used in studying voice source parameters for both synthetic and natural vowel sounds. Compared to an earlier closed phase inverse filtering (CPIF) algorithm, the advantage of this algorithm is that there is no need to locate the open/closed phase for the vocal tract transfer function estimation, which has proven difficult for real speech. Also the vocal tract transfer function is estimated from not only the closed phase, so it can be used for short closed phase voiced speeches such as female speech.

3.3. Synthesized sound for modification

Synthesized speech sound was used in this study. A male British English vowel /a:/ was synthesized by a formant syn-
Fig. 2. Iterative Adaptive Inverse Filtering [6]

Fig. 3. Voice source LF model

thesizer. The voice source was specified and passed through the formant filter to generate the synthesized voice sound. The synthesized sound has a pitch of 100Hz.

The voice source is generated based on a 4 parameter LF model [8]. Fig. 3 shows the waveform of the model and its time parameters:

The value of the parameters for this experiments are: 
\[ t_p = 0.41, t_c = 0.55, t_e = 0.62, t_a = 0.005 \]. These parameters are normalized to the pitch period as 1.

3.4. Modifications

The original synthesized sound is then modified in two ways:

1. LPC model based modification: As stated in Section 1, the original synthesized speech waveform is decomposed into impulse train excitation signal and all pole filter coefficients. The excitation signal is modified.

2. Voice source scaling based modification: Voice source waveform is extracted by the PSIAIF algorithm described in Section 3.2, and then the voice source is modified.

Both methods modified speech to pitch from 80Hz to 200Hz at 10Hz steps (normal male speaker pitch range).

4. RESULTS

4.1. Modification results

The upper half of Fig. 4 shows the waveform and spectrum of voice source extracted by PSIAIF from the synthesized speech signal. The waveform has a similar pattern as the LF model shown in Fig. 3. The lower half shows the waveform and spectrum of the residual signal of LPC method. It is similar to a impulse train which has a flat spectrum.

Fig. 5 shows a segment of the modified speech waveform of the LPC based modification method (dashed line) and production based modification method (solid line) at a pitch of 120Hz. It shows that LPC based modification has an earlier onset time in the opening phase due to the shorter closed phase.
4.2. Auditory test result

Auditory tests are performed for all the synthesized speech signals. The original synthesized speech sounds somewhat buzzy, which is expected at the beginning of the research. To simplify the experiment, variables such as jitter, shimmer and additive noise, which could improve the naturalness of formant synthesized speech, were purposely not added.

For production based modification, all the pitches from 80Hz to 200Hz are successfully synthesized. The synthesized sound remains intelligible and smooth, no distortion or other audible noise presents for the whole pitch range.

For LPC based modification, there is no significant difference from the result obtained from the production based method at pitch up to 160. The intelligibility of the synthesized sound remains good between the pitch at 80Hz and 160Hz. Severe distortion appears in the modified sounds for pitch above 160Hz.

5. DISCUSSION

The auditory test shows there is a difference between the two modification methods after 160Hz. The voice source of the original speech has an open phase of 58%. For LPC based modification, when the pitch is increasing, the closed phase is shortened. When the pitch reaches: $F = \frac{F_{\text{orig}}}{T_c}$, the closed phase disappears. After this point, the open phase overlaps between consecutive cycles as shown in the bottom panel in Fig. 1. For this study, the open phase is 58% of the pitch period, and original pitch was 100Hz, so the highest pitch before open phase overlap $F = 100/62\% \approx 161\,Hz$. The overlapped part changes the input to vocal tract filter. From a speech production point of view, the disappearance of the closed phase changes the voice type and voice quality, which could be the major reason caused the distortion.

Unlike the impulse shape excitation signals, the voice source waveform has a tilt in its spectrum of typically -12 – -15 dB/oct for normal speakers (as shown in upper-right figure of Fig. 4). This tilt provides a major contribution to the overall tilt of the speech signal spectrum. When the voice source is being scaled in the time domain for modification, its spectrum is scaled in the frequency domain in a reverse way, i.e. a 30% compression in time domain corresponding to 30% expansion in frequency domain. Because the spectrum scales in a linear way during the pitch modification, it maintains the same degree of tilt as before modification.

6. CONCLUSION AND FUTURE WORK

In this work, we compared two pitch modification methods: a voice production based model and the LPC based model. Modification of the synthesized speech shows that the voice production based modification can achieve intelligible speech over a wider pitch range. The result suggests that during the pitch modification, the voice source waveform should be extracted from the vocal tract transfer function and modified. Voice source extraction does require extra calculation, but it does produce higher a quality of modification. Future studies includes identifying the best voice source extraction algorithm to achieve efficiency and robustness for real speech and for applying this production based modification scheme on real speech.

7. REFERENCES


