Voice source analysis for pitch-scale modification of speech signals

Yinglong Jiang, Peter Murphy***

Department of Electronic and Computer Engineering,
University of Limerick, Limerick, Ireland
`*yinglong.jiang@ul.ie`
`**peter.murphy@ul.ie`

Abstract

Much research has shown that the voice source has strong influence on the quality of speech processing [4][5][6]. But in most of the existing speech modification algorithms, the effect of the voice source variation is neglected. This work explains why the existing modification scheme can't truly reflect the voice source variation during pitch modification. We use synthesized voiced speech sound to compare an existing pitch modification scheme with our proposed voice source scaling based modification scheme. Results show that voice source scaling based pitch modification can be used for wider range pitch modification.

Key word: speech pitch modification, voice source, formant synthesis.

1. Analysis of the voice source effects in speech modification

1.1. Pitch modification framework

Speech modification plays an important role in many aspects of speech processing, for example: text-to-speech synthesis, speech recognition, speaker recognition, speech conversion etc. Much research has been done in time-scale/pitch-scale modification. Efficient speech synthesis and modification methods like Pitch Synchronized OverLap Add (PSOLA) are widely used in many systems [1]. Recently other speech modification models such as sinusoid model [2], or the harmonic plus noise model (HNS) [3] have also been presented. All of these methods based on speech production source-filter model. The source-filter model consists of a source that generates a sequence of glottal pulses, act as the input to a filter that models the vocal tract system, and a differentiation operator that models the radiation at the lips, it can be expressed as a convolution as in Eq.1:

$$s(t) = g(t) * v(t) * r(t) \tag{1}$$

in which *g(t)* is the excitation signal to the vocal tract, it corresponding to the glottal air flow that is injected into vocal tract, *v(t)* is vocal tract transfer function and *r(t)* is radiation at the mouth. In most of the speech modification algorithm, the radiation part and the vocal tract part are inter-changed so the input to the vocal tract transfer function is a differentiated voice source waveform.

Pitch modification normally includes the following steps:

1. Apply a window on continuous speech signal to get a short time framed signal.
2. Perform a source-filter decomposition of the framed signal to get the source signal, for voiced speech, the voice source would be a series of impulses and have a flat spectrum.
3. Modify the voice source, result in a new impulse train spaced at required pitch period.
4. Apply the vocal tract filter on the modified voice source, the output is the desired pitch-scaled signal.

During the modification, the vocal tract transfer function *v(t)* remains unchanged, the input to the vocal tract system is modified to a new excitation signal, and then the excitation signal is convolved with vocal tract transfer function to get the modified signal. In this manner, the overall contour of the speech in the frequency domain is unchanged, and pitch can be modified independently, e.g. during the pitch changing, the speech duration remains unchanged. Because the modified speech has the same spectral envelope as the original speech, it retains the intelligibility of the original speech.

1.2. Analysis of pitch modification by impulse train scaling

The speech modification methods we mentioned above have achieved high efficiency, but are short of naturalness and sometimes cause distortion in modified speech. An important aspect missed in them is the voice source effect under different conditions. During modification, the vocal tract transfer function is obtained by estimating the linear prediction coefficients *a(i)* over the whole frame, which is about 2-4 pitch period's duration. Then the residual signal is extracted from the speech waveform by the inverse filtering as voice source is represented in Eq. 2:

$$e(n) = s(n) - \sum_{i=1}^{p} a(i)s(n-i) \tag{2}$$

In which *a(i)* is a set of linear prediction coefficients, *a(i)* characterizes the spectral envelope of the speech signal. *s(n)* is the speech signal. e(t) is the difference between current sample and estimated value from previous *p* samples. The result is the excitation signal which has a nearly impulse train shape and a flattened spectrum. An example is illustrated in Fig. (1) and (2).

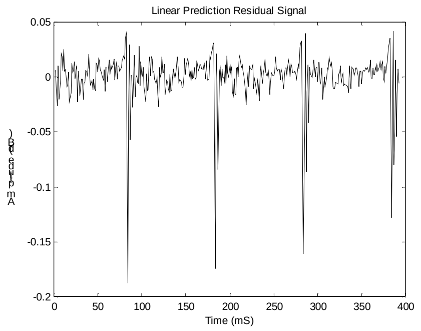

Fig 1. Excitation signal from inverse filtering

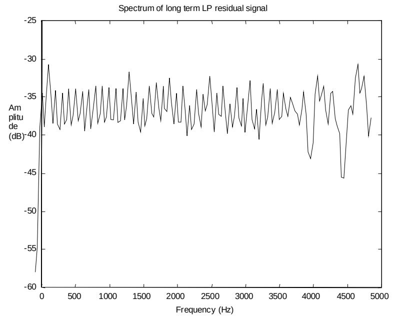

Fig. 2 Spectrum of the excitation signal from inverse filtering

Although it is treated as an excitation signal for the vocal tract transfer function or voice source, it doesn't represent the true voice source at the glottis. By this inverse filtering decomposition, glottal effects are mostly included in the vocal tract transfer function. To more better , we present the glottal filter and vocal tract transfer function separately in Eq. 3:

$$s(t) = e(t) * g_e(t) * v(t) * r(t) \tag{3}$$

Here, voice source is decomposed as a glottal system impulse response $g_e(t)$, excited by impulse train $e(t)$. We can see from Eq. 3, the residual signal we get and modify to the new excitation is actually $e(t)$ in Eq.3, while $g_e(t)$ and $v(t)$ are both included in the vocal tract transfer function.

For voiced speech sound, the vocal cords vibrate periodically, producing a periodic puff of air. So one period of the voice source signal includes open phase and close phase, a typical voice source waveform obtained by close phase inverse filtering (CPIF) [4] is shown in Fig. 3.

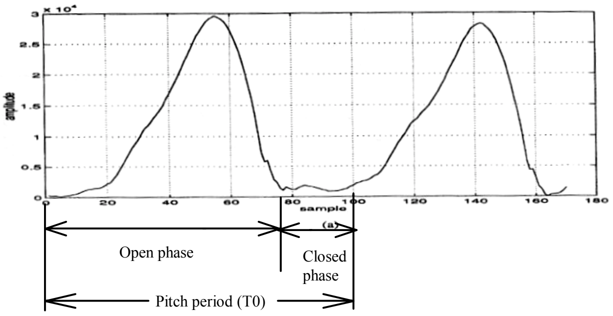

Fig. 3 A typical voice source waveform obtained by inverse filtering

From a signal processing point of view, the voice source waveform can been seen as a impulse response of glottal system, i.e. $g_e(t)$ in Eq.3. During the pitch modification, only impulse train $e(t)$ is scaled to new pitch period. Glottal system $g_e(t)$ and the vocal tract transfer function $v(t)$ are included in the filter parts and kept unchanged. We call this type of modification scheme impulse train scaling based pitch modification. Fig. 4 shows how voice source changes in this pitch modification scheme.

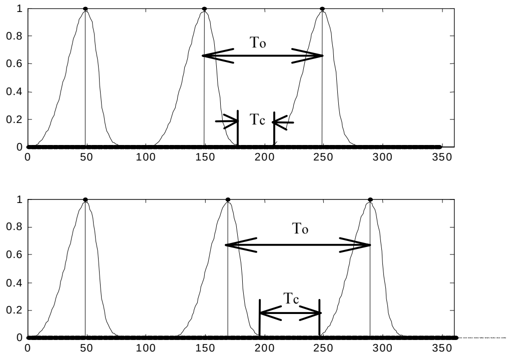

Fig. 4 Illustration of voice source in pitch modification based on modification of impulse train

Where T_0 is the pitch period, T_c is the period in which the vocal cords are totally closed.

From Fig. 4 it can be seen that because the glottal system impulse response remains unchanged, the new voice source would be open phase spaced at new pitch period, in other words, pitch modification by impulse train scaling would maintain the open phase of the voice source waveform. But the open quotient would increase (decrease) when the pitch period decreases (increases).

To present the relation between open/close phase and the pitch period, open quotient and closed quotient of voice source are defined:

- Open quotient:

$$Q_o = T_c / T_0 \tag{4}$$

- Close quotient:

$$Q_c = (T_0 - T_c) / T_0 \tag{5}$$

Researchers have shown that the change of voice source open/closed quotient can distinctly alter voice type or voice quality [5] [6]. So there is the possibility that for pitch modification based on impulse train scaling, the quality of the speech could be changed.

1.3. Pitch modification by voice source waveform scaling

Holmberg carried out research on voice source waveform analysis of recorded speech in different pitch context. She suggested that the voice source open/close quotient has no significant link to the pitch changing [5]. Other research on voice source analysis also drew a similar conclusion [6]. To overcome the possibility that degrading the quality of modified speech, we proposed a pitch modification scheme that scales the voice source extract from the speech signal instead of scales the impulse train.

In this scheme, a new voice source signal that is used as input to vocal tract transfer function is obtained by scaling the real voice source waveform instead of the impulse train, i.e. in

Eq.3 we extract and modify both *e(t)* and *$g_e(t)$* from the original speech signal, and use this as a new voice source for the vocal tract transfer function to get the modified speech signal. This should be closer to human speech production process, and as it will, and because by this way, this modification will keep the voice source character, we would expect it could produce better voice quality.

2. Comparison of synthesized voiced sound

2.1. Analysis of synthesized voice

In order to compare how the voice source would affect the modified speech in two modification schemes described in Part 1, we use a formant synthesizer to synthesize the voiced speech sound in different pitches. A schematic diagram of formant synthesizer is shown in figure. 5.

To synthesize the voiced speech sounds, an impulse train spaced at pitch period *e(t)* is sent as input to a glottal filter *g(t)* to produce the voice source signal, then the voice source signal goes through the vocal tract transfer function to generate formants, finally a differential filter *r(t)* models radiation at the lips.

Two sets of voice sound /a:/ were synthesized to compare two synthesis schemes. First set is obtained by modifying the impulse train and leave the glottal filter *g(t)* unchanged for different pitches, this would produce the voice source signal has the same open/close quotient for different pitches, which is corresponding to pitch modification based on voice source scaling described in 1.3. Second set of speech is obtained by change the glottal filter *g(t)* to get the same duration of open phase for different pitches, but the open/close quotient of voice source are different, this is corresponding to the pitch modification by pulse train scaling method described in 1.2.

Each set of synthesized sound has pitch range from 100Hz to 200Hz, step at 10Hz. This is the pitch range for a normal male speaker use in general speech.

2.2. Voice source model and formant parameters

To produce the voice source signal, a voice source model is needed. The most widely accepted voice source model is the LF model [7]. It describe the voice waveform with a four parameter function:

$$\begin{cases} g(t) = E_0\, e^{\alpha t}\sin(\omega_g t) & 0 \le t \le t_e \qquad (6) \\ g(t) = -\dfrac{E_e}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right] & t_e \le t \le t_c \le T_0 \qquad (7) \end{cases}$$

Eq. 6 represents the voice source waveform when the vocal cords are opening, and Eq. 7 represents the other part of open quotient within one period. Figure 6 shows the voice source waveform of LF model.

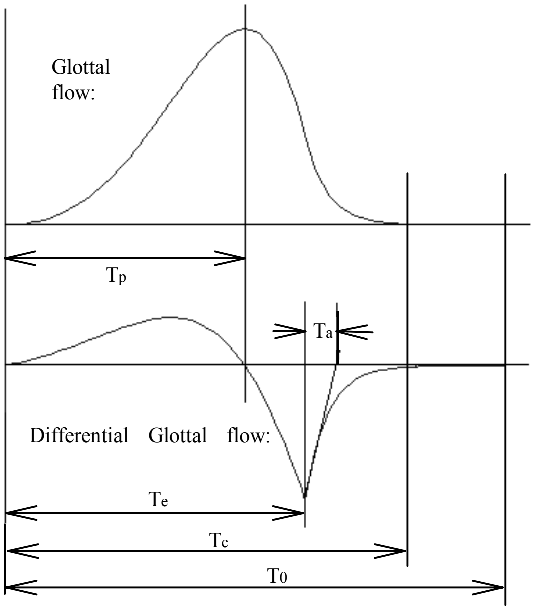

Figure 6. Waveform of voice source LF model

The modeled waveform could be specified by either the direct synthesis parameters (E0, α, ωg, ε) or by timing parameters (t_p, t_e , t_a , t_c) where a set of conditions hold in Eq.8:

$$\begin{cases} \int_0^T g(t)\,dt = 0; \qquad \omega_g = \dfrac{\pi}{t_p}; \\ \varepsilon\, t_a = 1 - e^{-\varepsilon(t_c - t_e)}; \\ E_0 = -\dfrac{E_e}{e^{\alpha t_e}\sin(\omega_g t_e)} \end{cases} \qquad (8)$$

All the timing parameters in the LF model are normalized, i.e. they are percentages of one pitch period. By altering the parameters, we can generate the voice source to meet the requirement of the present research. For this research we synthesized a normal male voice, the parameters are:

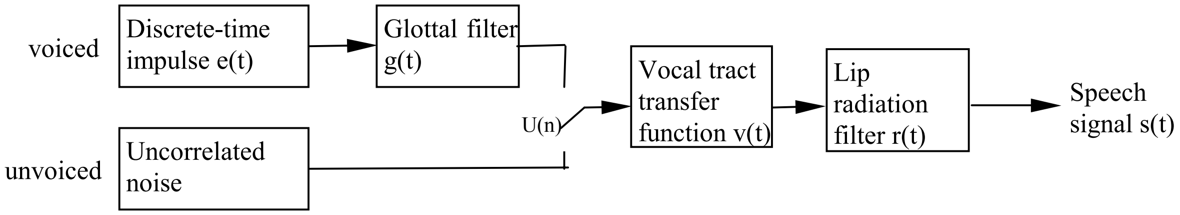
voiced
Discrete-time impulse e(t)
Glottal filter g(t)
unvoiced
Uncorrelated noise
U(n)
Vocal tract transfer function v(t)
Lip radiation filter r(t)
Speech signal s(t)

Figure 5. Schematic diagram of formant synthesizer

t_p=41%, t_e=55% , t_c= 58%, t_a=0.5%

These values are obtained from the analysis of data and suggested as standard for a normal male speaker [8].

First five formants' frequency and bandwidth of British English vowel /a:/ are:

Formant	1	2	3	4	5
Frequency(Hz)	650.3	1075	2463	3558	4631
Bandwidth(Hz)	94	91	107	198	90

Table 1 formant frequency and bandwidth for vowel /a:/ (after Rabiner [9])

3. Result

3.1. Synthesized voice

Figure 6. and Figure 7. show the waveform and spectrum of synthesized speech at pitch of 120Hz and 180Hz. Solid line represents the waveform of the speech signal synthesized by voice source scaling, dashed line represents the waveform of the speech signal synthesized by impulse train scaling only.

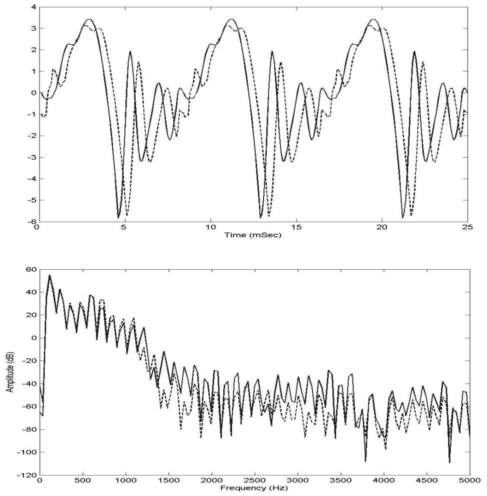

Fig 7. Waveform and spectrum of synthesized /a:/ F0=120Hz

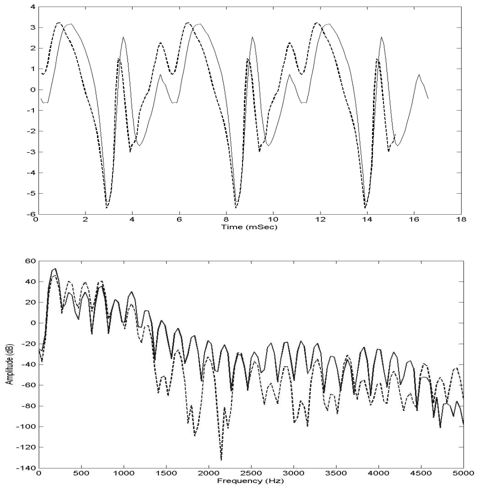

Fig 8. Waveform and spectrum of synthesized /a:/ F0=180Hz

From the waveform, it shows that the synthesized sounds by both methods are close. From the spectrum we can see for F0 at 120Hz sound, there is no significant difference between the two modification methods. For F0 at 180Hz, synthesized speech by impulse train scaling shows more high frequency decay and has deeper valleys at certain harmonics between 1000Hz and 4000Hz.

3.2. Auditory result

Auditory tests are also performed for all the synthesized speech signals. All the synthesized speech sounds buzzy, this is expected at the beginning of the research. To simplify the model, we didn't consider the many aspects like jitter (short-term variability in fundamental frequency), shimmer (short-term variability in amplitude) and additional noise, which could improve the naturalness of formant synthesized speech.

For synthesized speech based on voice source scaling, all the pitches from 100Hz to 200Hz are successfully synthesized. The synthesized sound remains intelligible and undistorted for all pitches. For synthesized speech based on impulse train scaling, all the pitches are achieved, there is no significant difference at f0 between 100Hz and 170Hz in terms of the intelligibility and distortion. But there is severe distortion for the pitches above 170Hz, at this F0, there is no closed phase in the voice source signal for modification by impulse scaling. Because the spaces between impulses are equal or shorter than the open phase.

4. Discussion

4.1. Comparison of pitch modification schemes

From the synthesized speech waveform and the auditory test, we can see there is no significant difference between synthesized speeches based on different scaling sources at F0 between 100Hz to 170Hz.

The major different of these two modification schemes could be reflected at the changing of the voice source spectrum. The voice source spectrum normally shows a -12 dB/oct ~ -18dB/oct tilt depending on the voice source shape. For modification based on impulse train scaling, the harmonics shift to new location, the envelope of the source spectrum remains unchanged. And for modification based on voice source scaling, the relative height of the harmonics remains the same [10].

For pitch modification based on impulse train scaling, the limit of the frequency range that doesn't cause distortion depends on the open/closed quotient of original speech, i.e. distortion starts when the modification cause no close phase. For pitch increasing modification, when the space between impulses is shorter than the open phase, the spectral tilt of the voice source is changed, this causes a change in spectral envelope of the synthesized speech. In this study, the closed quotient of the original speech is 58%, so when the modified speech reaches

$$f0_{\text{modified}} = f0_{\text{original}} / 0.58 = 100/0.58 = 172\ \text{Hz}$$

The closed phase disappears, and the modification causes distortion.

4.2. Suggestion of the use of two pitch modification schemes.

From the synthesis result and the above discussion, we can see that both of the modification schemes would be capable for modification in certain ranges of pitch. Pitch modification based on voice source scaling appears to be closes to the actual human speech production process, so it can be used for wider range. But this modification scheme requires extra calculation to obtain the voice source signal.

On the other hand, pitch modification based on impulse train scaling would be suitable for pitch decreasing modification, which wouldn't reach the zero closed phase. For pitch increasing modification, this modification scheme can have only limited pitch range depending on the duration of voice source open/close phase. The advantage of this modification scheme is it only requires a relatively simple inverse filtering algorithm to obtain the modification sources.

5. Conclusion

In this work, we analyzed two pitch modification schemes, one based on voice source scaling, and the other based on impulse train scaling. A simulation of the modification has been performed by formant synthesis. Results show that voice source scaling based pitch modification can achieve the required pitch modification in a wider pitch range, while impulse train based pitch modification would only work on a more limited pitch range. It also shows that voice source analysis could help to improve speech modification schems.

6. References

[1] Moulines, E. and Laroche J., "Non-parametric techniques for pitch-scale and time-scale modification of speech", Speech Comm. 16 (1995) 175-205.

[2] George, E.B. and Smith,M.J.T., 'Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model', IEEE trans on speech and audio proc. Vol 5, No. 5, Sept. (1997), 389-406.

[3] Laroche, J. Stylianou, Y. and Moulines, E. 'HNM: A simple, efficient harmonic+noise model for speech' Proc. IEEE ICASSP-93, Minneapolis, Apr 1993.

[4] Wong, D.Y., Markel, J.D. and Gray, A.H. 'Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform', IEEE Trans. on Acoustics, Speech and Signal Porcessing, vol. ASSP-27, No. 4, Aug. 1979.

[5] Holmberg, E.B., Hillman, R.E. and Perkell, J.S. 'Glottal Air and Transglottal Air Pressure Measurements for Male and Female Speakers in Low, Normal and High Pitch.' Journal of Voice. Vol. 3, No.4 p249-305.

[6] Alku, P. and Vilkman, E. 'A comparison of Glottal Voice Source Quantification Parameters in Breathy, Normal and Pressed Phonations of Female and Male Speakers', Folia Phoniatr Logop 1996;48;240-254.

[7] Fant, G., Liljencrants, J., and Lin, Q.G. 'A four parameter model of glottal flow.' STL-QPSR, 2-3: 119-156 (1985)

[8] Childers, D.G. and Ahn, C.T. 'Moldeling th glottal volume-velocity waveform for three voice types.' J. Acoustic. Soc. Am., 97:505-519.

[9] Rabiner, L.R. and Schafer, R.W. 'Digital processing of speech signals'. Prentice Hall Press. New Jersey, 1978

[10] Murphy, P.J. 'Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis', J. Acoustic. Soc. Am. Vol. 105, Issue 5 pp. 2866-288.