APPLYING SOURCE-FILTER MODEL IN CHINESE SPEECH SYNTHESIS

YI LIFU  TIAN JING  SUN JINGCHENG
Chinese Academy of Sciences, Institute of Acoustics Beijing

ABSTRACT
This paper presents a novel algorithm for the signal modification component of concatenative text-to-speech systems. The algorithm described here is based on the LPC analysis/synthesis framework, and achieves prosodic modification by time-domain processing of the LPC residual (LF-4 model). This method is based on a source-filter model. The proposed method achieves high quality prosody modification, retains the characteristics of the donor speaker, allows for spectral manipulation, and yields compact acoustic inventories and improved voiced fricatives.

Key Words: Speech synthesis  Linear prediction  Derivative glottal flow  Prosody control

1. INTRODUCTION
Although mandarin corpus-based concatenative text-to-speech (TTS) systems today have achieved a high level of naturalness, their unit selection approaches and the large speech corpus still prevent them from being widely deployed in mobile man-machine communications. A very popular technique of doing prosodic modification of a speech unit is TD-PSOLA for these systems[4]. This approach can perform prosody modification on a speech segment with excellent quality, but it can’t do any spectral modification. So they don’t allow “personalization” of text-to-speech synthesizer, nor can they smooth out spectral discontinuities at unit boundaries.

Another popular technique uses a source-filter model, where the filter is an LPC filter estimated from the speech unit, and the excitation is a parameterized pulse generator. This source-filter approach allows spectral smoothing across unit boundaries by smoothing the LPC coefficients (using LSP parameters actually).

Linear predictive coding (LPC) is one of the explicit implementations of the source-filter theory for processing speech signals. The excitation for linear prediction (LP) is substituted by the LF derivative glottal flow. On the basis of LF-4 derivative glottal flow, a simplification algorithm of the LF-4 derivative model is proposed. This simplification not only improves the computation efficiency, but also can be conveniently used to manipulate the voice source parameters as well.

The objective of this paper is to derive a source-filter model that can (1) retain the characteristics of original speaker after prosody modification (2) allows for spectral manipulation and voice conversion (3) generates improved voiced fricatives. In this paper we describe some techniques on LF model and speech modifications.

2. SOURCE-FILTER MODEL
The traditional source-filter model consists of an excitation followed by a linear time-varying filter. The excitation can be white Gaussian noise for unvoiced sounds or an impulse train for voiced sounds (See Fig. 1).

The view put forward in the source filter model is that speech sounds are produced by the action of a filter, the vocal tract, on a sound source, either the glottis or some other constrictions within the vocal tract. Fundamental of the model is the assumption that these are independent - that is the properties of the filter can be modified without changing the properties of the source and vice versa. This assumption might not be strictly true followed in all cases but in practical terms it provides us with a very useful and largely accurate model of speech production.

Figure 1. Basic source-filter speech production model. An impulse train is used as the source for voiced sounds and white Gaussian random noise as source for unvoiced sounds, both followed by a time-varying filter.

There are two acoustic sources in speech production corresponding to voiced and unvoiced speech sounds.

2.1 Voiced Speech
The source in voiced speech is the vibration of the vocal folds in response to airflow from the lungs. This vibration is periodic and if could be examined independently of the properties of the vocal tract, it would be seen to consist of a series of broad spikes.

The spectrum of the glottal source is made up of a number of frequency spikes corresponding to the harmonics of the fundamental frequency of the vibration of vocal folds. The spectrum decreases in amplitude with increasing frequency at a rate of around -12dB per octave -- that is, for each doubling in frequency, the amplitude of the spectrum decreases by around 12dB.

The effect of increasing the frequency of the vocal folds’ vibration is to widen the gap between the frequency spikes in the glottal source spectrum, since these spikes occur at multiples of
the base or fundamental frequency. The overall shape of the spectrum remains unchanged.

2.2 Unvoiced Speech

In unvoiced speech the sound source is not a regular vibration but rather vibrations caused by turbulent airflow due to a constriction in vocal tract. The various positions at which vocal tract can constrict have been discussed in the earlier part of the course.

The sound created via a constriction is described as a noise source. It contains no dominating periodic component and has a relatively flat spectrum implying that every frequency component is represented equally. Observing the time waveform of a noise source we see only a random pattern of movement around the zero axis. The spectrum looks the same, with random peaks and valleys, but the overall trend is for it to be flat across the frequency range.

3. LPC-BASED FILTER MODEL

Linear Predictive Coding (LPC) is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate. It provides extremely accurate estimates of speech parameters, and is relatively more efficient for computation.

In order to synthesize more natural speech, it’s necessary to estimate the linear filter and the excitation. During speech synthesis, we can modify the pulse of excitation for prosodic modification.

![Figure 2. LPC based speech production model. For unvoiced sounds, Gaussian random noise is filtered by a time-varying filter. For voiced sounds, the signal is the sum of a voiced and an unvoiced component.](image)

Analysis/resynthesis experiments conducted on a system like that of Fig. 3, in which the filter is an LPC filter, results in speech that is unnatural and different from the original speech, especially voiced frames. The residual signal obtained through various LPC estimation methods, including Levinson-Durbin, has a non-white magnitude, and it can be understood without problems if listened to through speakers. This is possibly due to the fact that LPC analysis models only poles, not zeroes, and most LPC estimation methods have been derived under the assumption of a white noise excitation, rather than a periodic excitation. Therefore we decided to use the LPC filter for unvoiced segments only, for which it produces satisfactory results.

For a TTS system, one needs to store a fairly large set of speech segments and in practice these segments need to be compressed. While we could have opted for using a standard speech compression scheme to do this, we decided to investigate schemes that would compress the speech in a way that integrates well with the approach required to do prosody modification.

For 11025kHz sampling rate, we estimated LPC coefficients through the autocorrelation recursion (different estimation techniques didn’t produce a noticeable improvement in quality), which are transformed to LSP and quantized.

3.1 Prosody Modification Framework

The source-filter model consists of a source that generates a sequence of glottal pulses, acting 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:

\[ s(t) = g(t) * v(t) * r(t) \]  

in which g(t) is the excitation signal to 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 algorithms, the radiation part and the vocal tract part are interchanged so the input to the vocal tract transfer function is a differentiated voice source waveform.

Prosody 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 and time-scale signal.

Holmberg suggested that the voice source open/close quotient has no significant link to the pitch changing [5]. Other researches on voice source analysis also drew a similar conclusion. To overcome the possibility that degrades the quality of modified speech, we proposed a pitch modification scheme that scales the voice source extracted from the speech signal, a new voice source signal that is used as input to the vocal tract transfer function is obtained by scaling the real voice source waveform, and use this as a new voice source for the vocal tract transfer function to get the modified speech signal. By this way, this modification will keep the voice source characteristics, and expected to produce better voice quality is.

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 envelopes as the original speech, it retains the intelligibility of the original speech.

Prosody modification implies pitch-scale and time-scale modification of the segment simultaneously. The time varying filter $H(t_i)$ is estimated at time $t_i$, which corresponds to the input epochs for voiced speech and are arbitrary for unvoiced speech. In synthesis, re-sampling of this filter is necessary at a time sequence, it is different from that of analysis. This involves computing a mapping $t'_i = f(t_i)$ and involves repeating or removing a filter $H(t_i)$ for some pitch periods[3]. Lengthening unvoiced sounds or voiced fricatives results in buzziness. Since it is produced by repeating frames, it can cause undesired periodicity at high frequencies. Reversing the repeated frames for unvoiced sounds allows lengthening by a factor of 2. Lengthening voiced fricatives results in buzziness by creating an artificial periodicity at high frequencies. One possible resolution of this problem is to interpolate frames instead of repeating them, but this would attenuate the aspiration component.

4. VOICE SOURCE MODEL

Many researches have shown that the voice source would affect synthesized speech. Existing popular prosody modification schemes always ignored the voice source. The prosody modification scheme with our proposed voice source scaling seems to have wider prosody range. By means of inverse filtering, we can research on the voice source.

4.1 LF Derivative glottal flow Model

To produce the voice source signal, a voice source model is needed. The most widely accepted voice source model is the LF model [1][2]. It describes a voice waveform with a four parameter function:

$$ E(t) = \begin{cases} 
E_1(t) = E_0 e^{\omega t} \sin(\omega t)(t < t_e) \\
E_2(t) = \frac{E_e}{\Delta t_e [e^{\omega(t_{01}-t_e)} - e^{\omega(t_e-t_1)}]}(t_e < t < t_f) \\
0(t_f < t < t_c) 
\end{cases} $$

Obviously, the waveform obtained after fitting procedure seems smoother. Auditory result shows the synthesized speech from the fitting voice source sounds better.

5. DISCUSSION

Unfortunately, things are not so simple. One reason is that there are speech sounds made with a combination of buzz and hiss sources.

Another problem is that, inevitably, any inaccuracy in the estimation of the vocal tract means that more speech information gets left in the residue. The aspects of nasal sounds that don't match the LPC model (as discussed above), for example, will end up in the residue. There are other aspects of the speech sound that don't match the LPC model; side branches introduced by the tongue positions of some consonants, and tracheal (lung) resonances are some examples.

Figure 3. Glottal flow ($U_g$) and glottal flow derivative ($dU_g$) with the parameters of the LF model: time of glottal opening ($t_0$); time ($t_1$) and value ($U_0$) of the maximum of $U_g$; time ($t_e$) and absolute value ($E_e$) of the minimum of $dU_g$; $T_a$ describes the return phase, it is the length of the time interval between $t_c$ and the projection of the tangent of $dU_g$ in $t_c$; and the time of glottal closure ($t_c$).

4.2 Estimation method of voice source

A voice source has five LF parameters $E_e$, $t_0$, $t_1$, $t_e$, and $t_a$. In our fitting estimation method these five LF parameters are estimated for each pitch period. The goal of the fitting estimation method is to determine a model which resembles the glottal pulse as much as possible.

After initializing estimation of LF parameters, we can use simplex search algorithm and Levenberg–Marquardt algorithm to give more accuracy parameters.

For the fitting procedure different nonlinear optimization techniques were used: the simplex search algorithm of Nelder and Mead (1964). Of the algorithms tested the simplex search algorithms usually came closer to the global minimum than the gradient algorithms. Owing to discontinuities in the error function, gradient algorithms are more likely to get stuck in local minima than simplex search algorithms. Therefore the best version of the simplex search algorithm is used in the second stage of our fitting estimation method. However, in the neighborhood of a minimum, the simplex algorithm may do worse. As a final optimization, the Levenberg–Marquardt algorithm is therefore used in the third stage. Result of a voice source is as follows.

Figure 4. The derivative glottal flow waveform synthesized with LF-5 parameters.
In this work, we can see that the modification scheme would be capable for modification in certain ranges of pitch and duration. The quality of the synthetic speech generated by this model is quite high, though there are several problems:

- Voiced fricatives can exhibit some buzziness when stretched by a factor of 1.7 or more.
- Repeating and deleting frames does not yield smooth waveforms and ideally one would like to interpolate filters with time.
- Because of the estimation method, the LPC vectors do not evolve smoothly with time, nor do the FIR filter, which causes some spectral blurring.

To simplify the model, we didn’t consider 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 LPC-Based synthesized speech.

6. CONCLUSION

In this work, we analyzed the prosody modification scheme based on voice source scaling. Results show that voice source scaling based pitch modification can achieve the required prosody modification in a wide range, retain the characteristics of the donor speaker, allow for spectral manipulation (to reduce spectral discontinuities at unit boundaries) and yield compact acoustic inventories. It also shows that voice source analysis could help to improve speech modification schemes.

7. REFERENCES

[1] 王聰修, 中科院聲學所博士學位論文, 《語音轉換及其相關問題的研究》, 2001.4