SOURCE-FILTER MODELS FOR TIME-SCALE PITCH-SCALE MODIFICATION OF SPEECH

Alex Acero

Microsoft Research
One Microsoft Way
Redmond, Washington 98052, USA

ABSTRACT

This paper presents two time-scale pitch-scale modification techniques to be used in speech synthesis systems. They have been applied to Microsoft's Whistler system, which is based on concatenative synthesis. Both methods are based on a source-filter model, one of them using LPC parameters and the other one using cepstral parameters. The proposed methods achieve high quality prosody modification, retain the characteristics of the donor speaker, allow for spectral manipulation (to reduce spectral discontinuities at unit boundaries), yield compact acoustic inventories and improved voiced fricatives.

1. INTRODUCTION

Although Text-to-Speech (TTS) systems today have achieved a high level of intelligibility, their unnatural prosody and synthesis voice quality still prevent them from being widely deployed in man-machine communication. In addition, the process of building a new synthesis voice often is highly labor-intensive.

There are two main methods used for speech synthesis: formant synthesis [1] and concatenative synthesis [9]. Generally, formant synthesizers use a set of rules to generate speech. While these systems can achieve high intelligibility, their naturalness is typically low, since it is very difficult to accurately describe the process of speech generation in a set of rules. In recent years, data-driven approaches such as concatenative synthesis [9][14] have achieved a higher degree of naturalness. While these speech units are often tediously extracted by human experts, there are some automatic ways of generating them [2][3][5]. In most cases, the speech units have to be synthesized with a different prosody than that of the original database.

A very popular technique of doing prosodic modification of a speech unit is the so-called Time-Domain Pitch-Synchronous Overlap-and-Add (TD-PSOLA) [9]. This approach can perform prosody modification on a speech segment with excellent quality, and the original speaker's characteristics are retained. On the other hand, it cannot do any spectral modification, which is often needed to smooth out spectral discontinuities at unit boundaries, because it operates in the time domain. Moreover, for many practical applications the acoustic inventory needs to be compressed, which can lead to a degradation in the final quality.

Another popular technique uses a source-filter model [14], 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 for spectral smoothing across unit boundaries by smoothing the LPC coefficients -- actually equivalent parameters sets that can be manipulated better such as reflection coefficients, or Line Spectral Pairs (LSP). However, since the pulse generator has not been estimated from the original speech unit, the generated speech does not resemble the original speaker.

The objective of this paper is to derive a method that can (a) retain the characteristics of original speaker after prosodic manipulation, (b) allow for spectral manipulation (c) be compact and (d) generate improved voiced fricatives. In this paper we describe some of the techniques we have experimented with to improve the speech synthesis quality of Microsoft's *Whistler* (Whisper Highly Intelligent Stochastic TaLkER). An early implementation of the Whistler TTS system [3] can be downloaded from Microsoft Research's web site [8].

This paper is organized as follows. In Section 2 we discuss source-filter models for speech synthesis. In Section 3 we then describe how to extract pitch and epochs from the input speech. Section 4 deals with our proposed LPC-based source-filter and Section 5 presents another source-filter model that is based on cepstrum. Finally we summarize our major findings and outline our future work.

2. SOURCE-FILTER MODELS

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).

To obtain natural sounding speech, we need to estimate both the excitation and the filter from input speech. In synthesis, we can modify the prosody by changing the spacing between impulses.

To estimate the excitation from input speech, we need to determine first the regions where the signal is voiced or unvoiced, and the *epochs* or exact location of the impulses for the voiced regions (see Section 3). Section 4 will describe how to estimate the time-varying filter such that the resynthesized signal is as close as possible to the original signal.

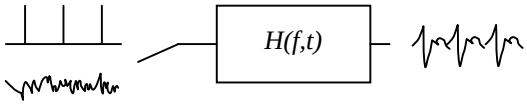

Figure 1. Basic source-filter speech production model. An impulse train is used as the source for voiced sounds and white Gaussian random noise as source for unvoiced sounds, both followed by a time-varying filter.

The mixed excitation model of Fig. 2 improves on the previous model, since it is well known that voiced fricatives contain some

amount of aspiration noise in addition to the voiced component. This model will be used in Section 5.

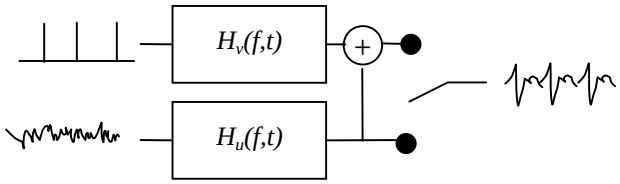

Figure 2. Mixed excitation speech production model. For unvoiced sounds, Gaussian random noise is filtered by a time-varying filter. For voiced sounds, the signal is the sum of a voiced and an unvoiced components.

Prosody modification implies pitch-scale and time-scale modification of the segment simultaneously. The time varying filter $H(f,t)$ is estimated at times 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 t_i' different than that of analysis. This involves computing a mapping $t_i' = f(t_i)$, that assigns an analysis epoch to a synthesis epoch [10], and involves repeating or removing a filter $H(f,t_i)$ for some pitch periods.

3. EPOCH EXTRACTION

Epoch extraction refers to the process of computing the glottal closure instants (GCI). Traditionally, this is done by analyzing the speech signal with a pitch tracker, which also classify different regions of the input signal as either voiced and unvoiced. In addition, we also investigated obtaining the epochs from a laryngograph signal, which ended up being superior.

We implemented a pitch tracker similar to that described in [15], and modified it to compute the epochs in addition to the pitch period. This pitch tracker had a 10% voiced/unvoiced classification error. We observed that the quality of the pitch-scale time-scale modified signal using the methods described in Sections 4 and 5, could be significantly degraded because of those epoch errors. There are several types of possible errors: epoch deletions and insertions and epoch inaccuracy. Epoch deletions *(i.e.* a voiced region being classified as unvoiced) resulted in rough speech when prosody modification was done. Epoch insertions also resulted in severe distortions (*i.e.* when an additional epoch was inserted in a voiced region, when an unvoiced frame was repeated since it is being considered voiced). Epoch inaccuracies, resulted in jitter and rough speech as well when prosody modification was done.

In order to obtain a more accurate epoch sequence, we investigated epoch extraction through an electroglotograph (EGG) signal [7], also called laryngograph signal. This signal can be recorded in one channel with the speech signal being in the other channel of a stereo recording. Briefly, a laryngograph signal can be obtained by placing two electrodes on the subject's larynx to capture the opening and closing of the vocal cords. Using a laryngograph signal (see Fig. 3) to detect voicing is much simpler because for unvoiced sounds this signal contains almost no energy. Also, since the signal has an almost constant power spectrum, it is much easier to detect periodicity.

3.1 Epoch Extraction from Laryngograph Signal

High quality epoch extraction can be achieved by performing peak picking on the derivative of the laryngograph signal. In practice, the derivative operation is accomplished by a first order pre-emphasis filter $H[z] = 1 - \alpha z^{-1}$ with α being near 1 (0.95 is a good choice).

We pre-processed the signal to filter out the low frequencies (lower than 100 Hz) and high frequencies (higher than 4kHz). This can be done with rectangular window filters that are quite efficient and easy to implement. There is a significant amount of energy outside this band that does not contribute to epoch detection, yet it can complicate the process, as can be seen in Fig. 3, so this band-pass filtering is quite important.

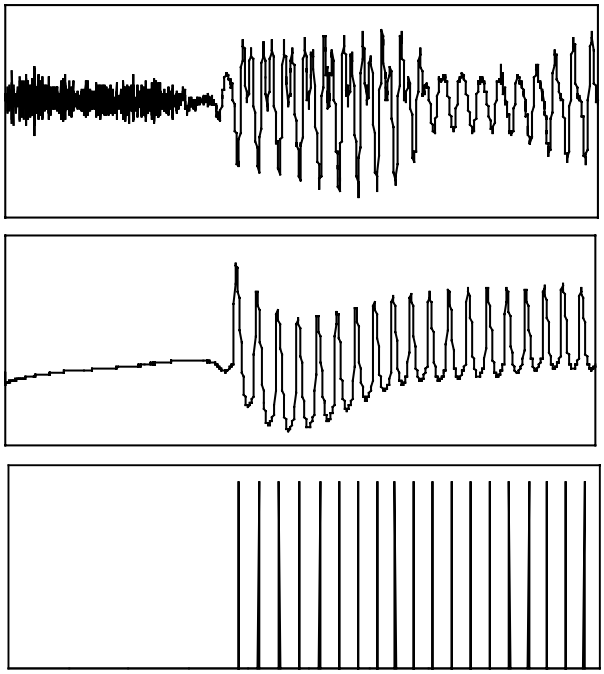

Figure 3. Speech signal, laryngograph signal, and extracted epochs corresponding to the word "city".

The pre-emphasized signal exhibits peaks that are found by thresholding. The quality of this epoch detector has been evaluated on recordings from 2 female and 4 male speakers and the voiced/unvoiced decision errors are lower than 1%. This is definitely acceptable for our prosody modification algorithms. The quality of prosody modification with the epochs computed by this method vastly exceeded the quality achieved when the standard pitch tracker was used on the original speech signal.

4. LPC-BASED FILTER MODEL

Analysis/resynthesis experiments conducted on a system like that of Fig. 1, where the filter is an LPC filter, results in speech that is unnatural and different than the original speech, especially for 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 because most LPC estimation methods have been derived under the assumption of a white noise excitation, and not 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 22kHz sampling rate, we estimated 20 LPC coefficients through the Levinson-Durbin recursion (different estimation techniques didn't produce a noticeable improvement in quality), which are transformed to LSP and quantized with 48 bits by using split-VQ similarly to [12], but with 6 codebooks of 256 entries each.

4.1 Double-Filter Model for Voiced Speech

For voiced signals we used a cascade of an *N*-tap time-varying FIR filter followed by a p^{th} order LPC filter. The LPC filter is estimated as described above for unvoiced signals (other than the window was centered at the epochs). The objective in adding the FIR filter is to retain the characteristics of the original signal. Therefore the FIR filter is estimated in three steps.

First, let's define the LPC-residual signal as $x[n]$, and a local version of it centered at time *m* by $x_m[n] = x[m+n]$, where the time instant for epoch *i* is $t_i = m$. We can create a windowed version $y_m[n]$ that can be computed as

$$y_m[n] = w_L[n] x_m[n]$$

where $w_L[n]$ can be a Hanning window

$$w_L[n] = \begin{cases} 0.5 + 0.5\cos(2\pi n / L) & |n| < L \\ 0 & |n| > L \end{cases}$$

with zero padding for $L < N$ and L being defined as

$$L = \min(t_i - t_{i-1}, t_{i+1} - t_i, N)$$

i.e. the minimum of the adjacent pitch periods and N. The use of a symmetric window makes perfect reconstruction impossible. In addition, truncation can occur for pitch periods larger than N. Nevertheless, it was empirically observed that this choice of FIR filter $y_m[n]$ resulted in no perceptual degradation in the resynthesized signal when no prosody modification was done.

Second, $y_m[n]$ is expressed in the frequency domain by taking the N-point FFT as:

$$Y_m[k] = \sum_{n=-N/2}^{N/2} y_m[n] \exp(-2\pi n k / N)$$

Finally, $Y_m[k]$ is quantized by sub-bands using delta-quantization:

$$Y_{t_i}[k] = Y_{t_{i-1}}[k] + \Delta^r_j[k] \quad l_r \le k < u_r$$

where sub-band *r* contains the frequency bins *k* between l_r and u_r, and $\Delta^r_j[k]$ is the delta contribution from codeword *j* for the frequency band *r*. We divided the spectrum into *R* sub-bands with bandwidths approximating the mel-scale.

Since in our TTS system, we need to quantize each unit independently of the others (*i.e.* we have no history), we also need to quantize the first voiced frame directly. To do that we created an equivalent sub-band quantization procedure.

For 22kHz sampling rate, a choice of *N = 512* and *R = 13* was found to be reasonable. This representation can result in a compact system.

4.2 Discussion

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 2 or more, because repeating the FIR filter ignores the fact that at high frequencies the energy is not totally periodic, and forcing it to be periodic can lead to buzzy speech.
- Repeating and deleting frames does not yield smooth waveforms and ideally one would like to interpolate filters with time. We can interpolate the LSP coefficients, but if we interpolate the time-varying FIR filter, the unvoiced component present in every voiced sound (though most prevalent in voiced fricatives) is attenuated. This results in a more "muted" speech in practice, and even in buzzy voiced fricatives.
- Because of the estimation method, the LPC vectors do not evolve smoothly with time, and neither do the FIR filter, which causes some spectral blurring if interpolation is used.

5. CEPSTRUM-BASED FILTER MODEL

To address the problem caused by both LPC and FIR filters not evolving smoothly with time, another filter technique is proposed based on the *pitch-synchronous* real cepstrum.

If we define the input signal as $x[n]$, instead of the LPC-residual signal as done in the previous section, we can proceed similarly until the computation of $Y_m[k]$, which now represents the pitch-synchronous short-time spectrum of the input signal.

We can then compute the pitch-synchronous real cepstrum $c_m[n]$ as:

$$c_m[n] = IFFT\{\log|Y_m[k]|\} \quad -N/2 < n < N/2$$

which allows us to decompose the original spectrum as

$$Y_m[k] = |Y_m[k]| Z_m[k]$$

with $|Y_m[k]|$ being computed from the real cepstrum as

$$|Y_m[k]| = \exp(FFT\{c_m[n]\})$$

and $Z_m[k]$ being the amplitude-normalized spectrum, which is indeed whiter than a standard LPC residual.

Many of the problems of the approach of Section 4.2 can be solved by using a mixed-excitation model such as that described in Fig. 2. This way we can modify the periodic and the aperiodic components of a voiced sound independently. Estimation of the voiced and unvoiced component could be done in a way inspired by Waveform Interpolation methods [6], so that $Z_m[k]$ can be decomposed as

$$Z_m[k] = S_m[k] + Q_m[k]$$

where the slowly varying component of the spectrum $S_m[k]$, obtained by low-pass filtering $Z_m[k]$ with time *m*, represents the voiced component. The rapidly changing component $Q_m[k]$ represents the unvoiced component, which can be characterized by its power spectrum.

One difference with the approach in [6] is that we use a Hanning window instead of a rectangular window, which helps reduce the variability present when pitch is changing rapidly. Another difference is that here we filter the spectrum directly, and not a length-normalized time signal. This way each frequency component can be treated differently (for example we can use different time constants for every frequency bin). We can also neglect the unvoiced component that appears at very low frequencies, which is an artifact of the window we use to estimate the pitch synchronous spectrum.

This cepstrum vector $\mathbf{c} = c_m[n]$ can then be converted into a mel-scale through the use of the bilinear transform [11]:

$$z_{new}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}$$

that has been successfully used in speech recognition. This operation is equivalent to a matrix multiplication $\mathbf{c}_w = \mathbf{L}(\alpha)\mathbf{c}$ on the cepstrum, where the matrix $\mathbf{L}(\alpha)$ depends on the warping parameter α [13]. This transformation allows us to significantly reduce the number of coefficients without a perceptual change in quality. $S_m[k]$ can also be transformed to a mel-scale by taking the inverse FFT and then applying the bilinear transform.

Accuracy of epoch placement is important for this technique. A deviation from the correct epoch location will result in a linear phase shift, which in turn will reduce the time correlation of frequency bins, especially at higher frequencies. This can result in an overestimation of the noise component $Q_m[k]$, and therefore more aspiration in the synthetic speech. Additionally, there is typically a delay between the epochs found in the laryngograph signal and the epochs in the original signal, due to the time it takes the signal to travel from the larynx to the microphone (in our recordings this averaged 20 samples at 22kHz sampling rate).

This system can also be compressed in a way similar to that described in Section 4.1: cepstral coefficients can be differentially quantized easily using split-VQ. This system allows larger prosody manipulations than the LPC-based model.

6. SUMMARY

We have presented two techniques to do time-scale pitch-scale modification for a speech synthesis systems. Both methods are based on a source-filter model, one of them using LPC parameters and the other one using cepstral parameters. The proposed methods offer high-quality prosody-modification, retain the characteristics of the donor speaker, allow for spectral manipulation (to reduce spectral discontinuities at unit boundaries) and yield compact acoustic inventories.

REFERENCES

[1] Allen J., Hunnicutt S., and Klatt D. *From text to speech: the MITalk system*. MIT Press, Cambridge, MA, 1987.

[2] Donovan R.E. and Woodland P.C. "Improvements in an HMM-Based Speech Synthesizer". *Proceedings of Eurospeech Conference*, Madrid, Spain, 1995, pp. 573-576.

[3] Huang X., Acero A., Adcock J., Hon H., Goldsmith J., Liu J., and Plumpe M. "Whistler: A Trainable Text-to-Speech System". *International Conference on Spoken Language Processing*. Philadelphia, pp. 2387-2390. Oct, 1996.

[4] Huang X., Acero A., Hon H., Ju Y., Liu J., Meredith S. and Plumpe M.. "Recent Improvements on Microsoft's Trainable Text-to-Speech System: Whistler". *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, pp. 959-962. Apr., 1997.

[5] Hunt A. J. and Black A. W. "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database" . *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, pp. 373-376. May., 1996.

[6] Kleijn W. B and Haagen J. "Transformation and Decomposition of the Speech Signal for Coding". *IEEE Signal Processing Letters*, vol. 1, no. 9, pp. 136-138, 1994.

[7] Krishnamurthy A. K. and Childers D. G. "Two-channel speech analysis". *IEEE Trans. on Acoustics, Speech and Signal processing*, Vol 34, 1986.

[8] Microsoft Research's Speech Technology Group web page: http://www.research.microsoft.com/research/srg/.

[9] Moulines E. and Charpentier F. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones". *Speech Communication*, vol. 9, no 5, pp. 453-467, 1990.

[10] Moulines E. and Verhelst W. "Time-Domain and Frequency Domain techniques for Prosodic Modification of Speech". In *Speech Coding and Synthesis*, by Kleijn et al, pp. 519-555, *Elsevier* 1995.

[11] Oppenheim A. V. and Johnson D. H. "Discrete Representation of Signals". *Proc. of the IEEE*, (33): pp. 681-691, 1972.

[12] Paliwal K.K and Atal B.S. "Efficient vector Quantization of LPC Parameters at 24 bits/frame". *IEEE Trans. on Speech and Audio Processing*, vol. 1, pp. 3-14, 1993.

[13] Shikano K. "Evaluation of LPC Spectral Matching Measures for Phonetic Unit Recognition". *Carnegie Mellon University Technical Report*, May 1986.

[14] Sproat R. and Olive J. "An Approach to Text-to-Speech Synthesis". In *Speech Coding and Synthesis*, by Kleijn et al, pp. 611-633, *Elsevier* 1995.

[15] Wise J. D., Caprio J. R. and Parks T. W. "Maximum-likelihood pitch estimation". *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 24, pp. 399-417, Oct. 1976.