A MIXED-EXCITATION FREQUENCY DOMAIN MODEL FOR TIME-SCALE PITCH-SCALE MODIFICATION OF SPEECH

Alex Acero

Microsoft Research
One Microsoft Way
Redmond, Washington 98052, USA

ABSTRACT

This paper presents a time-scale pitch-scale modification technique for concatenative speech synthesis. The method is based on a frequency domain source-filter model, where the source is modeled as a mixed excitation. This model is highly coupled with a compression scheme that result in compact acoustic inventories. When compared to the approach in the *Whistler* system using no mixed excitation, the new method shows improvement in voiced fricatives and over-stretched voiced sounds. In addition, it allows for spectral manipulation such as smoothing of discontinuities at unit boundaries, voice transformations or loudness equalization.

1. INTRODUCTION

In recent years, data-driven approaches, such as concatenative synthesis, have achieved a high degree of naturalness for speech synthesis. While these speech units are often tediously extracted by human experts, there are some automatic ways of generating them [3][6] [7]. While there are systems that do not modify the waveform [7], in many 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) [11]. 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 manipulation, which is often needed to smooth out spectral discontinuities at unit boundaries, because it operates in the time domain. This approach typically repeats pitch periods when a unit needs to be lengthened, which can result into buzziness, particularly for voiced fricatives. Moreover, for many practical applications the acoustic inventory needs to be compressed, which is done independently of the PSOLA algorithm, therefore leading to degradation in the output quality.

Several approaches have been proposed to address the deficiencies of TD-PSOLA. The system described in [12] uses an LPC model to smooth out transitions. LP-PSOLA [11] performs the PSOLA on the residual signal obtained after LPC filtering allowing modification of its LPC parameters. Synthesis based on the sinusoidal model [2][9] allows more control over the spectrum. The MBROLA [4] approach uses TD-PSOLA on segments that have constant pitch and phase of their harmonics, which results in an efficient implementation and allows for smoothing across unit boundaries. The Harmonic plus Noise Model (HNM) [13] has been proposed to combat buzziness in voiced fricatives in addition to allow spectral smoothing.

The objective of this paper is to derive a method that can (a) allow for spectral manipulation, (b) be compact and (c) can achieve lengthening of unvoiced and voiced fricatives without buzziness. In this paper we describe one technique we have experimented with to improve the speech synthesis quality of Microsoft's *Whistler* (Windows Highly Intelligent Stochastic TaLkER). One version of the *Whistler* TTS system [6] can be downloaded from Microsoft Research's web site as part of the Speech SDK [10].

This paper is organized as follows. Section 2 presents the baseline source-filter model, which is then enhanced in Section 3 to include mixed-excitation frequency domain processing. Section 4 deals with parameter estimation, Section 5 with acoustic compression and Section 6 with decompression and synthesis. Finally an evaluation is presented in Section 7, after which we summarize our major findings and outline future work.

2. SOURCE-FILTER MODEL

In this section we'll present a reformulation of the well known TD-PSOLA algorithm [11] for prosody modification in a framework of a source-filter model. This will let us later extend this to a more general model of mixed excitation in the frequency domain.

First, let's define the input signal as $x[n]$, and a set of time marks $\{t_m,\ m=-\infty,\ldots,\infty\}$. Let's further define a local version of $x[n]$ centered at time t_m as $x_m[n]=x[t_m+n]$. We can then define $y_m[n]$ as a short-time version of $x_m[n]$ by multiplying it by a window $w_m[n]$

$$y_m[n]=w_m[n]x_m[n] \tag{1}$$

where the window $w_m[n]$ is 0 for $|n|>N/2$, with N being the window length. Then we can define $\tilde{x}[n]$ as

$$\tilde{x}[n]=\sum_{m=-\infty}^{\infty} y_m[n-t_m] \tag{2}$$

which is an approximation of $x[n]$. We can express (2) as a convolution of an impulse train with a time-varying filter:

$$\tilde{x}[n] = \sum_{m=-\infty}^{\infty} \delta_m[n - t_m] * y_m[n] \quad (3)$$

The time-varying filter $y_m[n]$ can also be expressed in the frequency domain by taking the *N*-point FFT as:

$$Y_m[k] = \sum_{n=-N/2}^{N/2} y_m[n] \exp(-2\pi nk / N) \quad (4)$$

Fig. 1 shows (3) and (4) in a block diagram.

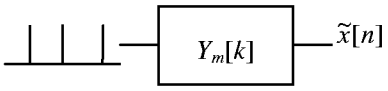

Figure 1. Approximation of *x*[*n*] as an impulse train driving a time-varying filter.

A good choice for the time marks t_m is to coincide with the epochs in the signal. Epochs are the instants of closing of the vocal folds, and indicate the periodicity of speech. For unvoiced speech, these marks could be arbitrarily placed. Epoch estimation from speech waveforms is a very difficult problem, but it can be done quite accurately using electroglotograph (EGG) signals [1].

Given the pitch synchronous time marks t_m, a good choice for $w_m[n]$ is, for example, a Hanning window

$$w_m[n] = \begin{cases} 0.5 + 0.5\cos(\pi n / L(m)) & |n| < L(m) \\ 0 & |n| > L(m) \end{cases} \quad (5)$$

with *L*(*m*) being defined as

$$L(m) = \min(t_m - t_{m-1}, t_{m+1} - t_m, N/2) \quad (6)$$

The use of a symmetric window makes perfect reconstruction impossible, unless time marks t_m are equally spaced (impossible in real speech). In addition, truncation will occur if these time marks are spaced more than *N*/2 apart (very long pitch periods). Nevertheless, it was empirically observed that this approximation $\tilde{x}[n]$ computed from (3) was perceptually indistinguishable from the original signal *x*[*n*] for real speech signals. A necessary condition for this is that the marks are not spaced more than 10ms in unvoiced regions, to preserve time resolution for stops.

The above is analysis-resynthesis but prosody modification implies pitch-scale and time-scale modification of the segment simultaneously. In synthesis, re-sampling is necessary at a time sequence t'_m different than that of analysis. This involves computing a mapping $t'_m = f(t_m)$, that assigns an analysis epoch to different synthesis epoch [11], and typically involves repeating or removing a filter $y_m[n]$ for some pitch periods.

Lengthening unvoiced sounds or voiced fricatives results in buzziness. Since it is accomplished by repeating frames, it can cause undesired periodicity at high frequencies. Reversing the repeated frame for unvoiced sounds [11] allows for lengthening by a factor of 2. Lengthening voiced fricatives results in buzziness by creating an artificial periodicity at high frequencies. One possibility suggested in [11] is to interpolate frames instead of repeating them, but this would attenuate the aspiration component.

3. MIXED EXCITATION MODEL

In this section we present a frequency-domain mixed-excitation model as a solution to the lengthening problems of the model described in Section 2. One way of removing that lengthening restriction for unvoiced frames is to generate random noise shaped by the power spectrum for that frame.. In fact, we have observed that replacing $y_m[n]$ by random noise with the power spectrum of $Y_m[k]$ doesn't result in any perceptual degradation in practice.

To address the lengthening problems for voiced fricatives we propose in Fig. 2 a mixed excitation model, which has a switch to produce purely unvoiced sounds and mixed excitation sounds. The voiced and unvoiced components of a voiced sound can then be processed independently. The goal is that if a voiced fricative needs to be lengthened, the noisy component can be independently generating random noise, which will not lead to buzziness. To lengthen voiced sounds, we don't need to repeat pitch frames, which can result in a metallic sound, rather we interpolate frames instead.

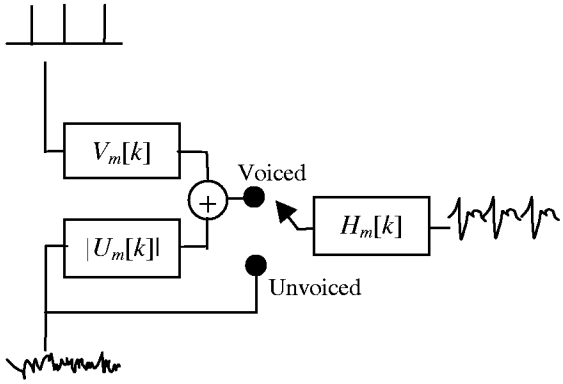

Figure 2. Mixed excitation speech production model. For unvoiced sounds, Gaussian random noise is filtered by a time-varying filter. For voiced sounds, the signal is the sum of a voiced and an unvoiced components.

To obtain natural sounding speech, we need to estimate all the time-varying filters in Fig. 2 from input speech. We should also note that while the system in Fig. 2 is not a minimal system, it's advantageous from the point of view of estimation and compression of its parameters, as will be shown in the next sections.

4. PARAMETER ESTIMATION

This section deals on how to estimate the parameters in the model of Fig 2. While there are many possible ways of doing this, including through LPC and cepstral analysis [1], we estimated $H_m[k]$ as the magnitude spectrum:

$$H_m[k] = |Y_m[k]| \quad (7)$$

With the excitation then computed as

$$E_m[k] = \frac{Y_m[k]}{H_m[k]} \tag{8}$$

and then further decomposed as

$$E_m[k] = V_m[k] + U_m[k] \tag{9}$$

where $V_m[k]$ is the voiced component and $U_m[k]$ the unvoiced component.

The voiced component $V_m[k]$ can be characterized because it evolves slowly over time. Therefore, similarly to the approach taken in Waveform Interpolation coding schemes [8], it can be estimated by low-pass filtering $E_m[k]$ over the time index *m*. The unvoiced component $U_m[k]$ is then computed as $E_m[k] - V_m[k]$.

We then just keep the magnitude spectrum of $U_m[k]$ and synthesize it by:

$$U_m[k] = |U_m[k]|W[k] \tag{10}$$

where $W[k]$ is a complex random vector derived from a zero-mean unity-variance Gaussian distribution.

In [8], the same time constant is used to low pass filter all frequencies. We have observed that by doing this, we can smooth out sharp spectral transitions that often occur in natural speech, which results in distortion. Moreover, the unvoiced component in those cases is also overestimated, which leads to additional noise. To avoid this, and given the fact that typically there is very little aspiration or unvoiced component at low frequencies, we use different low-pass time constants for different frequencies. In particular, we do not low-pass filter frequency components below 3kHz at all.

5. ACOUSTIC COMPRESSION

A speech synthesis system needs to store a large number of speech units. In practice this requires the units to be compressed, which will lead to some degradation in the synthesized signal. Traditionally speech compression and prosody modification algorithms are done independently, so that each step adds distortion. By integrating the compression with the prosody modification, higher efficiency can be achieved. In this section we describe one possible way of doing such compression.

To compress $H_m[k]$, we opt for gain-shape quantization with a product code VQ [5]. To do this we divide the vector into R_h sub-bands with bandwidths approximating the Bark-scale, since this has been successfully used in audio coding [14]. The average log-energy in each band is computed as

$$G_m^i = \frac{1}{(u_i - l_i + 1)} \sum_{k=l_i}^{u_i} \ln H[k] \tag{11}$$

where l_i and u_i are the lower and upper bins for sub-band *i*. The average energy in frame *m* is the average for all sub-bands:

$$G_m = \frac{1}{R_h} \sum_{i=0}^{R_h-1} G_m^i \tag{12}$$

after which the gain-normalized $\overline{H}_m[k]$ is defined as

$$\overline{H}_m[k] = H_m[k] \exp(-G_m) \tag{13}$$

The gain G_m is scalar quantized to $\hat{G}_m$. Each sub-band *r* is then vector quantized to minimize the Euclidean distance between the logarithm of the frequency bins:

$$i_m^r = \arg\min_i \sum_{k=l_i}^{u_i} \left(\log \overline{H}_m[k] - \log C_i^r[k]\right)^2 \tag{14}$$

with $C_i^r[k]$ being the codeword *i* in codebook *r*, and i_m^r is the codeword with minimum distance, so that $\overline{H}_m[k]$ can be quantized as

$$\overline{H}_m[k] = \left\{ C_{i_m^0}^0[k] \cdots C_{i_m^{R-1}}^{R-1}[k] \right\} \tag{15}$$

The voiced component $V_m[k]$ is vector-quantized with R_v sub-bands similarly to how it is done for $\overline{H}_m[k]$, but using plain Euclidean distance. We need to note that unlike for $\overline{H}_m[k]$, we have to quantize a complex vector instead of a real one, which doubles the dimension of the codebook. We need to handle $V_m[k]$ as a complex vector because we want to retain both the magnitude and the phase of the voiced component.

We also vector-quantize the magnitude spectrum of the unvoiced component $|U_m[k]|$. We have found that using a single band is sufficient in practice, since little detail is necessary for the unvoiced spectral component.

For a 22kHz sampling rate, a choice of $N = 512$ and $R_h = R_v = 12$ was found to be a reasonable tradeoff. This representation resulted in a compact acoustic inventory.

6. DECOMPRESSION AND SYNTHESIS

Decompression and resynthesis with prosody modification is accomplished as follows:

1. Decompress G_m, $\overline{H}_m[k]$, ($V_m[k]$ and $|U_m[k]|$ also for voiced frames) for the input epoch sequence t_m by doing table look-ups.
2. Compute output epoch sequence $t'_m = f(t_m)$ following [11].
3. Compute G_m, $\overline{H}_m[k]$, ($V_m[k]$ and $|U_m[k]|$ also for voiced frames) for the output epoch sequence. Instead of repeating parameters when lengthening is needed, *interpolation* is used at all times. See below for details.
4. Synthesize output frame by computing the complex spectrum $Y_m[k]$ (according to Fig. 2), taking an inverse FFT to obtain $y_m[n]$ and overlap-add according to (2).

The interpolation is step 3 is not just between the corresponding two input frames, but rather between all input frames in an N-sample window. This acts as a low-pass filter on the filter coefficients that reduces the quantization noise. Both a rectangular window and an exponential window gave satisfactory results. A time constant of less than 20ms was found to be beneficial in reducing the quantization noise, particularly for voiced sounds, without noticeably distorting the synthesized signal.

The same low-pass filtering can be done across unit boundaries to reduce the spectral discontinuities present in a concatenative synthesizer such as Whistler. A large time constant is needed to completely smooth out bad transitions, but this was observed to increase distortion in the output. A compromise in the low pass filter time constant can be achieved that reduces somewhat the discontinuity, yet doesn't increase the noise in the synthesized signal. Another possibility would be to use a longer time constant only around the concatenations if a large discontinuity was noticed, though this remains future work to be done.

Since the information is in the frequency domain, we can do other manipulations easily. For example we can equalize the signal by multiplying $\tilde{Y}_m[k]$ by another transfer function. This is useful also when implementing loudness, since soft speech tends to have greater spectral tilt. We can also simulate different vocal tract shapes by simply warping all the filters in Fig. 2. This warping can be implemented as a non-linear mapping between the input and output frequencies, and it results in realistic voices. Finally, we can increase the level of breathiness by simply increasing the gain on the unvoiced component, or adding more noise.

7. EVALUATION

We conducted an informal evaluation by doing a preference test between the new frequency-domain synthesizer and a previous one [1] based on LPC parameters with a residual and no mixed excitation. In both cases, the synthesis units were derived in an automatic way [6] and natural prosody was used. A total of 6 subjects listened to 20 utterance sets and all of them preferred the new mixed-excitation system.

The mixed-excitation speech was less noisy, particularly when listened through headphones. Voiced fricatives were more natural, and those vowels exhibiting a metallic sound (typically when several pitch periods were repeated) before were improved as well. While the distortion at the concatenation point was somewhat reduced, this wasn't significant. Since this is the largest cause of distortion, there is still a lot to improve.

8. SUMMARY

We have presented a mixed-excitation frequency domain technique to do time-scale and pitch-scale modification that improves the quality of unvoiced sounds, voiced fricatives and over-stretched sounds. This approach also reduces quantization noise by integrating the acoustic compression into the prosody modification algorithm. While frequency-domain processing makes smoothing of spectral discontinuities at unit boundaries easy, more work remains to be done in this area to bridge the gap with recorded speech.

REFERENCES

[1] Acero A. "Source-Filter Models for Time-Scale Pitch-Scale Modification of Speech". *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, USA, pp. 881-884. May, 1998.

[2] Crespo M., Velasco P., Serrano L. and Sardina J. "On the Use of a Sinusoidal Model for Speech Synthesis in text-to-Speech" in *Progress in Speech Synthesis*, pp. 57-70, *Springer*, 1996.

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

[4] Dutoit T. and Leich H., "MBR-PSOLA : Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database", *Speech Communication*, Elsevier Publisher, November 1993, vol. 13, n03-4.

[5] Gray R. M. "Vector Quantization". *IEEE ASSP Magazine*, April 1984.

[6] 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.

[7] 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.

[8] 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.

[9] Macon M. W. and Clements M. A. "Speech Concatenation and Synthesis using an Overlap-Add sinusoidal model". *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp 361--364, 1996.

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

[11] 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.

[12] 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.

[13] Stylianou Y., Laroche J and Moulines E. "High-Quality Speech Modification based on a Harmonic + Noise Model". *Proc. of Eurospeech Conference*, Madrid, Spain, pp 451-454, 1995.

[14] Veldhuis R. and Kohlrausch A. "Waveform Coding and Auditory Masking". In *Speech Coding and Synthesis*, by Kleijn et al, pp. 397-431, *Elsevier* 1995.