FREQUENCY-DOMAIN TECHNIQUES FOR HIGH-QUALITY VOICE MODIFICATION

Jean Laroche

Creative Advanced Technology Center
1500 Green Hills Road
Scotts Valley, CA 95067
jeanl@atc.creative.com

ABSTRACT

This paper presents new frequency-domain voice modification techniques that combine the high-quality usually obtained by time-domain techniques such as TD-PSOLA with the flexibility provided by the frequency-domain representation. The technique only works for monophonic sources (single-speaker), and relies on a (possibly online) pitch detection. Based on the pitch, and according to the desired pitch and formant modifications, individual harmonics are selected and shifted to new locations in the spectrum. The harmonic phases are updated according to a pitch-based method that aims to achieve time-domain shape-invariance, thereby reducing or eliminating the usual artifacts associated with frequency-domain and sinusoidal-based voice modification techniques. The result is a fairly inexpensive, flexible algorithm which is able to match the quality of time-domain techniques, but provides vastly improved flexibility in the array of available modifications.

1. INTRODUCTION

The frequency-domain technique presented in this paper is an extension of the algorithm presented in [1] which achieved arbitrary frequency modifications in the short-time Fourier transform domain. The new technique attempts to achieve a sound quality comparable to TD-PSOLA (Time-Domain Pitch Synchronous OverLap Add) [2], [3], while providing the flexibility offered by the frequency-domain representation. The algorithm uses a pitch-estimation stage (which can be nicely combined with the short-time Fourier analysis) and makes use of the knowledge of the harmonic locations to achieve arbitrary pitch and formant modifications.

2. ALGORITHM

2.1. The fundamental technique

The new algorithm is based on the technique described in [1], which is now briefly outlined: The algorithm works in the short-time Fourier transform (STFT) domain, where $X(u, \Omega_k)$ the STFT at frame u and FFT bin Ω_k. After calculating the magnitude of the STFT $|X(u, \Omega_k)|$ a very coarse peak-detection stage is performed to identify ”sinusoids” in the signal (we use quotes because there is no strong assumption that the signal be purely sinusoidal). According to the desired (and possibly non-linear) pitch-modification, each peak and the FFT bins around it are *translated* (i.e. copied, shifted in frequency and pasted) to a new target frequency. The phases of the peak and surrounding bins are simply rotated by an amount that reflects the cumulative phase-increment caused by the change in frequency. The technique is both simple and efficient in terms of computations, and offers a quasi unlimited range of modifications. Voice modification, however, poses an additional problem in that a better control of the formant structure is required to preserve the naturalness of the voice. It is possible to add a spectral-envelope estimation stage to the technique outlined above, and modify the amplitude of the pitch-modified spectral peaks to preserve that envelope, but the resulting voice modifications are of poor quality, especially when the pitch is shifted downward while the formant remain at their original locations. The most likely cause for the artifacts that arise (noise bursts, loss of clarity) is the fact that some of the frequency areas (where the spectral envelope is of low amplitude) must be severely amplified to preserve the formant structure, which results in unacceptable noise-amplification. The improved frequency-domain technique presented in this paper was designed to solve that problem.

2.2. The pitch-based algorithm

The new algorithm uses a preliminary frequency-domain pitch estimation to locate the harmonics, and uses a specific scheme to select which harmonic will be cut-and-pasted to a specific area in the output spectrum to achieve a desired pitch and formant modification.

2.2.1. Frequency-domain pitch estimation

Any pitch estimation can be used at this point, but the simple STFT-based scheme presented below has the advantage to fit very nicely with the current framework. The basic idea consists of cross-correlating a *magnitude-compressed, zero-mean* version of the spectrum with a series of combs corresponding to various candidate pitches (e.g., from 60Hz to 500Hz every 2Hz). An arbitrary compression function $F(x)$ is applied to $|X(u, \Omega_k)|$ to prevent lower-amplitude higher-frequency harmonics from being overridden by stronger low-frequency ones. $F(x) = x^{1/2}$ or $F(x) = asinh(x)$ are appropriate choices. The mean (over all frequencies) of the result is subtracted, which is required to not bias the cross-correlation toward low-pitches. Finally, the cross-correlation is calculated for each candidate pitch, and only requires a few adds, because of the sparsity of the combs. The result is a pitch-dependent cross-correlation $C(\omega_o^m)$ which exhibits a large peak at or near the true pitch, and smaller peaks at multiples and submultiples of it, as shown in Fig. (1). The maximum of $C(\omega_o^m)$ indicates the most likely pitch for that frame. This simple single-frame pitch estimation scheme is quite efficient, and is almost

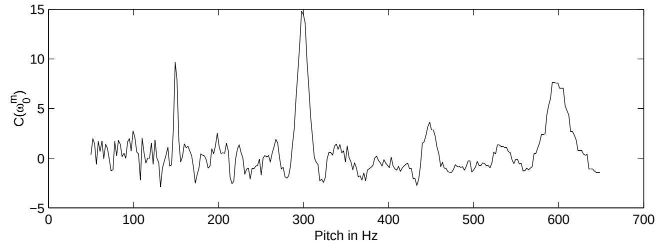

Figure 1: Cross correlation $C(\omega_o^m)$ as a function of the pitch candidate ω_o^m for a male voice.

completely free of octave-errors. A simple voiced/unvoiced decision can be derived by comparing the maximum of $C(\omega_o^m)$ to a predefined threshold. In the present version of the algorithm, frames that are non voiced are not further modified.

2.3. A new technique for formant-preserving pitch-modification

Harmonic assignment: Given the pitch-estimate ω_o at the current frame, individual harmonics are easily located at multiples of the pitch. As in [1], the frequency axis is divided into adjacent *harmonic regions* located around harmonic peaks, and extending half-way in between consecutive harmonics. To achieve formant-preserving pitch-modification (i.e., a modification of the pitch that leaves the spectral envelope constant), we will copy and paste individual *input* harmonic regions as in the algorithm described in [1], the difference being which input harmonic is selected to be pasted in a given location. Assuming a pitch modification of factor α, our goal is to create *output* harmonics at multiples of $\alpha\omega_o$. To create the ith output harmonic of $\alpha\omega_o$, at frequency $i\alpha\omega_o$, we will *select the input harmonic in the original spectrum that is closest to that frequency* and paste it in the output spectrum at the desired frequency $i\alpha\omega_o$. The rationale behind this choice is that the amplitude of the output harmonic will be close to the input spectral envelope at that frequency, thereby achieving the desired formant-preservation. This will become clear in the example below. Since the frequency of the i-th output harmonic is $i\alpha\omega_o$, denoting $j(i)$ the selected input harmonic, of frequency $j(i)\omega_o$, we must have

$$j(i)\omega_o \approx i\alpha\omega_o \tag{1}$$

Denoting $y = \text{round}(x) \overset{\triangle}{=} \text{floor}(x + 0.5)$ the integer y closest to the real number x, this yields

$$j(i) = \text{round}(i\alpha) \tag{2}$$

This does not define a one-to-one mapping, and the same input harmonic may be used to generate two or more output harmonics. This is illustrated in Fig. (2). The vertical dashed lines indicate the target frequencies of the output harmonics, for a pitch modification factor $\alpha = 0.82$. The arrows indicate which input harmonic is chosen to generate each output harmonic. The second input harmonic is used to generate both the second and third output harmonics.

Harmonic generation: The output spectrum is generated by copying and pasting the input harmonics into the output spectrum, as described in [1]. To generate the i-th output harmonic, input harmonic $j(i)$ will be shifted from its original frequency $j(i)\omega_o$ to the output frequency $i\alpha\omega_o$. Care must be taken to properly interpolate the spectral values if the amount of shift is not an integer number of bins. Refer to [1] for details on how this interpolation can be done, and how the phases of the bins around the output harmonic should be modified to account for the frequency shift. Fig. (3) presents the result of the pitch-modification for the same signal as above. Note that the second and third output harmonics have the same amplitude, because they were both obtained from the second input harmonic.

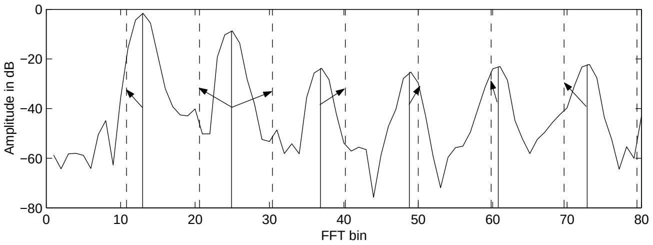

Figure 2: Assignment of input harmonic for a pitch modification factor $\alpha = 0.82$. The arrows indicate which input harmonic is used to generate the output harmonics at the vertical dashed lines.

Refining the amplitudes: Fig. (3) also displays a very simple line-segment spectral envelope (dashed-line) obtained by joining the harmonic peaks. Clearly, the amplitudes of the output harmonics do not necessarily follow exactly that spectral envelope, and this is likely to be the case no matter how the spectral envelope is defined. This may and may not be a problem in practice. In our experience, the amplitude mismatch is very rarely objectionable, although in some instances (e.g., very sharp formants), it is audible. More troublesome are the amplitude jumps that can appear from frame to frame, if two different input harmonics are selected in two consecutive frames to generate the same output harmonic. For example, still using Fig. (3), if the second output harmonic was obtained from the first input harmonic in a frame, then from the second input harmonic in the following frame, it would be given a -1dB amplitude in the first frame and a -9dB in

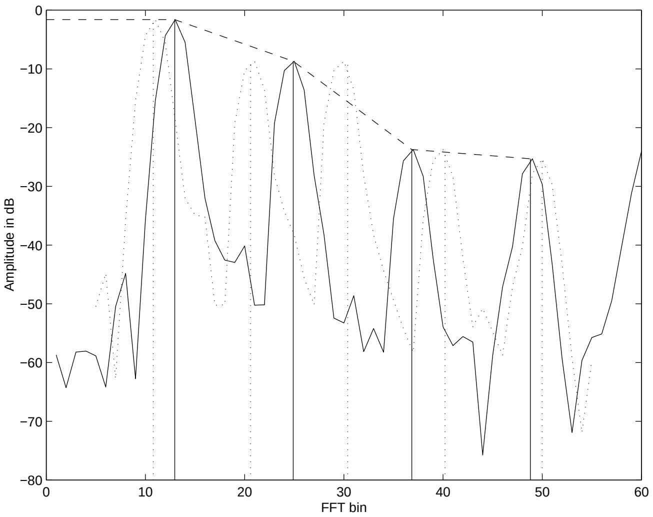

Figure 3: Input (solid line) and output (dotted line) spectra for the pitch modification factor $\alpha = 0.82$. A simple spectral envelope in shown in dashed line.

the next frame. Such amplitude ”jumps” are very audible and very objectionable. Note however, that according to Eq. (2) this only occurs if the modification factor α varies from frame to frame. In such cases, it is possible to avoid the problem by rescaling the output harmonic according to the magnitude of the spectral envelope at the target frequency, which guarantees that the output harmonic will be given the same amplitude, no matter which input harmonic was selected to generate it. Any technique to estimate the spectral envelope can be used, but the availability of the pitch makes the task much easier, see for example [4].

2.4. Joint formant-and-pitch-modification

The harmonic assignment equation Eq. (2) can easily be modified to perform formant modification in addition to pitch modification. One of the strong advantages of frequency-domain algorithms over time-domain techniques such as TD-PSOLA is the essentially unlimited range of modifications they allow. While TD-PSOLA only allows linear formant scaling [5], we can apply almost any input-output envelope mapping function. We can define a frequency-warping function $\omega' = F(\omega)$ which indicates where the input envelope frequency ω should be mapped in the output envelope. The function $F(\omega)$ can be completely arbitrary but must be invertible. To generate the i-th output harmonic, we select the input harmonic $j(i)$ of frequency $\omega = j(i)\omega_o$ which once warped through function $F(\omega)$ is close to the desired frequency of the i-th output harmonic $i\alpha\omega_o$. This can be expressed as

$$F(j(i)\omega_o) \approx i\alpha\omega_o \tag{3}$$

which yields a generalization of Eq. (2):

$$j(i) = \text{round}\left(\frac{F^{-1}(i\alpha\omega_o)}{\omega_o}\right) \tag{4}$$

It is easy to check that in the absence of formant-warping, $F(\omega) = \omega$, Eq. (4) collapses to Eq. (2). For a linear envelope modification in which the formants frequencies must be scaled linearly by a factor β i.e., $F(\omega) = \beta\omega$, Eq. (4) becomes $j(i) = \text{round}(i\alpha/\beta)$. Fig. (4) illustrates the results of such a linear, formant-only modification with a factor $\beta = 0.8$. The pitch is visibly unaltered, but the spectral envelope has been compressed, as desired. As in Section 2.3, it might be necessary to adjust the harmonic amplitudes so they match exactly the desired warped spectral envelope. For example, it is visible on Fig. (4) that the output spectral envelope is not exactly similar in shape to the compressed original one, in particular the second output harmonic should be of larger amplitude.

2.5. Shape-invariance

The algorithm described above performs fairly well, but as is typical with frequency-domain techniques [6] [7], the resulting speech can exhibit ”phasiness”, i.e. a lack of presence, a slight reverberant quality, as if recorded in a small room. This undesirable artifact usually plagues most frequency-domain techniques based on either the phase-vocoder or sinusoidal modeling, and has been linked to the lack of phase synchronization (or ”phase-coherence” [8]) between the various harmonics. To better understand the concept of phase-coherence and shape-invariance, it is helpful to recall a simplified model of speech production where a resonant filter (the vocal tract) is excited by a sharp excitation pulse at every pitch period. According to that model, a speaker changes the pitch of her/his voice by altering the rate at which these pulses occur. The important factor is that the shape of the time-domain signal around the pulse onset *is roughly independent of the pitch*, because it is essentially the impulse response of the vocal tract[1]. This observation is usually what is called ”shape invariance”, and it is directly related to the relative phases and amplitudes of the harmonics at the pulse onset time. The TD-PSOLA technique achieves pitch modification by extracting small snippets of signal (about 2 pitch-periods long) centered around excitation onsets, and ”pasting” them with a different onset rate. The good quality of the resulting signal can be attributed to the fact that shape-invariance is automatically achieved around excitation onsets, since the signal is manipulated in the time-domain. Shape-invariant techniques have been proposed for various analysis/modification systems for both time-scale and pitch-scale modification [9],[10],[11], and similar principles can be used in the present context. The main idea is to define pitch-synchronous input and output onset times and to reproduce at the output onset times the phase relationship observed in the original signal at the input onset times. We first define the input onset times t^i_n, and the output onset times t^o_n by the following recursion

$$t^i_n = t^i_{n-1} + \frac{2\pi}{\omega_o} \tag{5}$$

$$t^o_n = t^o_{n-1} + \frac{2\pi}{\alpha\omega_o} \tag{6}$$

with $t^o_0 = t^i_0$ (for lack of a better choice). The term $2\pi/\omega_o$ represents the pitch period. The short-time Fourier transform frame u is centered around time t^a_u, this is the time at which we are able to measure the phases of the input harmonics, and to set the phases of the output harmonics. Fig. (5) illustrates the various onset times for a pitch modification factor $\alpha = 2/3$. To calculate the phases of the output harmonics, we will use the same mapping as was used to generate the output spectrum (e.g., Eq. (2)), and we will

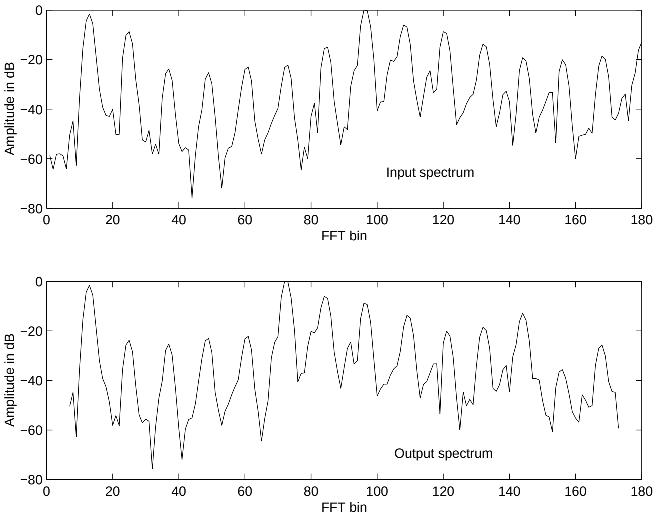

Figure 4: Input (top) and output (bottom) spectra for a formant-only modification of factor $\beta = 0.8$.

[1] discounting, of course, the tail of the impulse response triggered by the previous pulse.

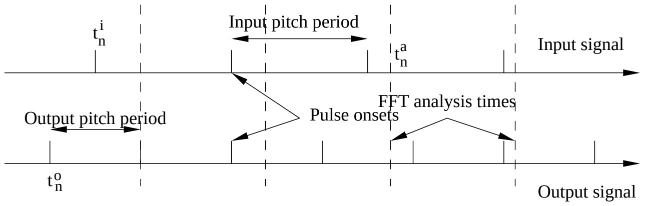

Figure 5: Input (top) and output (bottom) onset times t_n^i and t_n^o, and FFT analysis times t_u^a (vertical dashed lines).

set the phase of output harmonic i at time t_n^o to be the same as the phase of the input harmonic $j(i)$ at time t_n^i. Because we use the short-time Fourier transform, phases can only be measured and set at the short-time Fourier transform times t_u^a. We will therefore consider the input and output onset times closest to t_u^a and use our knowledge of the harmonic's instantaneous frequency to set the proper phases to the bins around harmonic i in the output spectrum. Denoting $\phi^i(t)$ and $\phi^o(t)$ the phases of the input and output harmonics at time t, we have:

$$\phi^i(t_u^a) = \phi^i(t_n^i) + \omega_i(t_u^a - t_n^i) \tag{7}$$

$$\phi^o(t_u^a) = \phi^o(t_n^o) + \omega_o(t_u^a - t_n^o) \tag{8}$$

where t_n^i is the input onset closest to t_u^a and t_n^o is the output onset closest to t_u^a. ω_i and ω_o are the frequencies of the input and output harmonics. We must ensure that $\phi^o(t_n^o) = \phi^i(t_n^i)$, which yields

$$\phi^o(t_u^a) = \phi^i(t_u^a) + \omega_o(t_u^a - t_n^o) - \omega_i(t_u^a - t_n^i) \tag{9}$$

Eq. (9) shows that the phase of the output harmonic is obtained by adding $\omega_o(t_u^a - t_n^o) - \omega_i(t_u^a - t_n^i)$ to the phase of the input harmonic, which means the harmonic bins are simply *rotated, i.e. multiplied by a complex number z*

$$z \triangleq e^{j\omega_o(t_u^a - t_n^o) - j\omega_i(t_u^a - t_n^i)} \tag{10}$$

As in [1], the spectral bins around the input harmonic are all rotated by the same complex z during the copy/paste operation, which guarantees that the fine details of the spectral peak are preserved in both amplitude and phase, which is important in the context of short-time Fourier transform modifications [6]. From a computation point of view, we can see that Eq. (10) requires minimal phase computations (no arc tangent, no phase-unwrapping/interpolation). Notice also that in the absence of pitch or formant modification, $t_n^o = t_n^i$ and $\omega_o = \omega_i$, and z becomes 1, i.e. the phases of the harmonic bins are not modified. This means that our modification algorithm guarantees *perfect reconstruction* in the absence of modification, which is usually not the case for sinusoidal analysis/synthesis [8]. Fig. (6) presents an example of pitch-modification for a male speaker. The sample rate was 44.1kHz, the FFT size was 35ms, with a 50% overlap (hop size $R = 17.5ms$), and the modification factor α was 0.75. Careful inspection of the waveforms shows great similarity between the orignal signal and the pitch-modified signal as should be expected for a shape-invariant technique. Of course, the rate at which pitch pulses occur differs between the two signals, showing the pitch has indeed been altered.

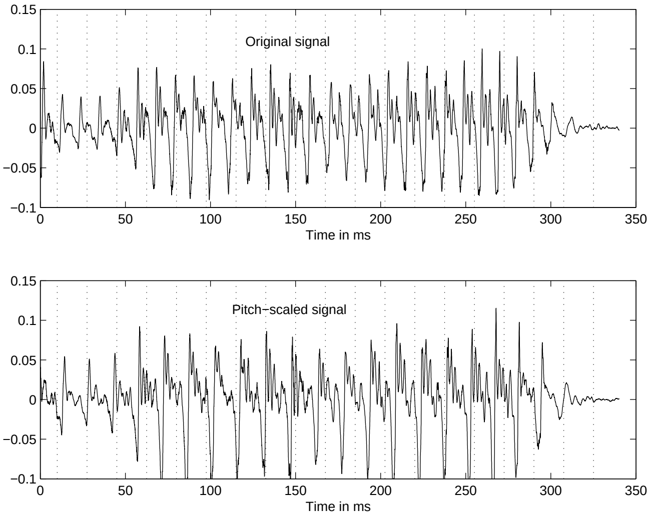

Figure 6: Speech signal from a male speaker (top) and pitch-modified version (bottom) for $\alpha = 0.75$. The vertical dotted lines indicate the analysis times t_u^a (every 17.5ms in this case).

3. RESULTS AND CONCLUSION

The voice modification technique described above was tested on a wide range of speech signals over which it performed very well. With the shape-invariant technique, the quality of the output speech is usually very good, nearly free of undesirable phasiness, similar to but still slightly inferior to the quality obtained by the TD-PSOLA technique. Because the spectral envelope can be modified in a ”non-linear” manner, for example by compressing specific areas in the spectrum, while leaving other areas unchanged, exotic vocal effects can be achieved that are out of reach of purely time-domain techniques. Using various piecewise linear frequency warping functions $F(\omega)$ in Eq. (4), we were able to impart a twang to the voice (for example, by pulling the vowel ’a’ (as in ’cast’) toward a more closed vowel ’Ç’ as in ’hot’), to dramatically accentuate the nasality of the voice, and even to increase the perceived age of the speaker. The technique lends itself well for real-time processing, although the short-time Fourier transform introduces a minimum latency equal to the size of the analysis window $h(n)$ (30 to 40ms), which may or may not be acceptable, depending on the context. From a computation point of view, the technique is relatively inexpensive. The algorithm runs at about 10x real-time for a monophonic 44.1kHz speech signal, on a 800MHz Pentium III PC (using a 35ms window, with a 75% overlap). Sound examples are available at
www.atc.creative.com/users/jeanl/SoundExamples/VoiceModif

4. REFERENCES

[1] J. Laroche and M. Dolson, “New phase-vocoder techniques for real-time pitch-shifting, chorusing, harmonizing and other exotic audio modifications,” *J. Audio Eng. Soc.*, vol. 47, no. 11, pp. 928–936, Nov 1999.

[2] F.J. Charpentier and M.G. Stella, “Diphone synthesis using an overlap-add technique for speech waveforms concatena-

tion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, Japan, 1986, pp. 2015–2018.

[3] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5/6, pp. 453–467, Dec 1990.

[4] M. Campedel-Oudot, O. Cappé, and E. Moulines, "Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 469–481, July 2001.

[5] J. Laroche, "Time and pitch scale modification of audio signals," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds. Kluwer, Norwell, MA, 1998.

[6] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, May 1999.

[7] J. Laroche and M. Dolson, "Phase-vocoder: About this phasiness business," in *Proc. IEEE ASSP Workshop on app. of sig. proc. to audio and acous.*, New Paltz, NY, 1997.

[8] T.F. Quatieri and R.J. McAulay, "Audio signal processing based on sinusoidal analysis/synthesis," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds. Kluwer, Norwell, MA, 1998.

[9] T.F. Quatieri and J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Trans. Signal Processing.*, vol. ASSP-40, no. 3, pp. 497–510, Mar 1992.

[10] D. O'Brien and A. Monaghan, "Shape invariant time-scale modification of speech using a harmonic model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Phoenix, Arizona, 1999, pp. 381–384.

[11] M. P. Pollard, B. M. G. Cheetham, C. C. Goodyear, and M. D. Edgington, "Shape-invariant pitch and time-scale modification of speech by variable order phase interpolation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Munich, Germany, 1997, pp. 919–922.