PRESERVING THE CHARACTER OF PERTURBATIONS IN SCALED PITCH CONTOURS*

Thomas Baran, Nicolas Malyska, Thomas F. Quatieri

MIT Lincoln Laboratory
tbaran@mit.edu, [nmalyska,quatieri]@ll.mit.edu

ABSTRACT

The global and fine dynamic components of a pitch contour in voice production, as in the speaking and singing voice, are important for both the meaning and character of an utterance. In speech, for example, slow pitch inflections, rapid pitch accents, and irregular regions all comprise the pitch contour. In applications where all components of a pitch contour are stretched or compressed in the same way, as for example in time-scale modification, an unnatural scaled contour may result. In this paper, we develop a framework for scaling pitch contours, motivated by the goal of maintaining naturalness in time-scale modification of voice. Specifically, we develop a multi-band algorithm to independently modify the slow trajectory and fast perturbation components of a contour for a more natural synthesis, and we present examples where pitch contours representative of speaking and singing voice are lengthened. In the speaking voice, the frequency content of flutter or irregularity is maintained, while slow pitch inflection is simply stretched or compressed. In the singing voice, rapid vibrato is preserved while slower note-to-note variation is scaled as desired.

Index Terms— Time-scale modification, pitch perturbation.

1. INTRODUCTION

In performing time-scale modification of speech, standard approaches involve obtaining various prosodic contours, followed by resampling of the contours, and synthesis of a time-scaled signal from the modified contours. For example, in pitch-synchronous overlap-add (PSOLA) [1], the sinewave-transformation system (STS) [2-3], and the phase vocoder [4-5], a pitch contour is either directly or indirectly resampled as a basis for time-scale modification. Resampling such contours, however, may result in a time-scaled contour that gives a perceptual voice character that is different from the original, beyond a simple change in articulation rate. For example, a pitch contour typically contains smooth trajectory segments on which are imposed higher-frequency perturbations such as jitter or flutter in a speaking voice and vibrato in a singing voice. Although a time-scaling of the smooth components is consistent with a natural physiological movement in rate change, such scaling for the perturbation component results in an unnatural synthesis. The purpose of this paper is to introduce an approach to independently modify the slow trajectory and fast perturbation components of a contour for a more natural synthesis.

As a further motivation, Fig. 1 depicts two pitch contours that have been resampled by a factor of 2: one with flutter and the other with vibrato. In each case, the perturbation component is stretched along with the underlying smooth contour, resulting in a modified contour having a perturbation component that takes on an unnatural character: excessively slow flutter and vibrato. One approach to a more natural contour invokes three steps: (1) separate the pitch components; (2) modify the smooth component by resampling and modify the perturbation component by time-scale modification (as we would time-scale modify any signal, thus preserving the spectrum); and (3) sum the modified components. As we will show in this paper, such a methodology can work when the perturbation components fall in distinct low- and high-frequency bands. Components, however, that are scattered throughout a signal's bandwidth and/or are overlapping in frequency pose a more challenging modification task for which we propose a generalization of the two-band methodology.

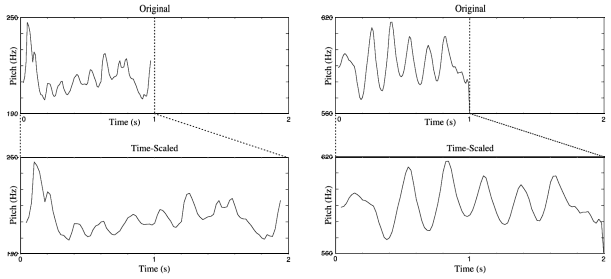

Figure 1. Actual pitch contours corresponding to (a) speech with pitch flutter and (b) a singer's vibrato, before and after upsampling by 2.

Specifically, our solution is to first decompose the pitch contour into several components representative of the perturbations and underlying smooth pitch components. Following this decomposition, we lengthen each component's duration using a possibly different technique. These lengthened signals are then combined to create the final output. We address the problem in detail for the case where components exist in several different frequency bands of the pitch contour signal, and we present a novel algorithm for implementing the separation and scaling on a continuum of bands. We emphasize that this paper focuses on an algorithm for manipulating the pitch contour and that the algorithm in turn forms the basis of ongoing work in developing a complete analysis/modification/synthesis system.

2. APPROACH

Our approach for addressing contour modification in a general sense is depicted in Fig. 2. A pitch contour is first decomposed into the various components of the contour, followed by a set of

*This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

systems used to change the duration of the components, possibly each in a fundamentally different way, and a combination of the resulting modified components into the final scaled pitch contour. The techniques for changing the duration of the component signals might include, for example, linear resampling, inserting zeros between key events so as to separate them in time, or using time-scale modification techniques that have been conventionally applied to voice signals [1-5] to stretch the component while roughly preserving its frequency spectrum.

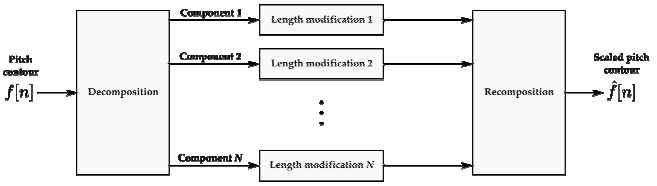

Figure 2. Approach for contour modification.

In many cases, it is natural to interpret a pitch contour as a signal whose low-frequency component represents the underlying intonation and whose high-frequency component represents additive perturbations on the slowly-varying component. Fig. 3 depicts the decomposition of the speech pitch contour in Fig. 1(a) into two such linearly-separable components, also in addition to a third component that represents a transition region. As is evident in the decomposition, a natural way to extend the duration of the components while preserving their character might be to upsample the slowly-varying component and process the high-frequency perturbation component using a time-scale modification algorithm, thereby roughly preserving its spectrum. Modification of the third component could be performed in different ways, but we consider extending its length by performing a combination of upsampling and time-scale modification. In the next section, we arrive at an algorithm that generalizes this approach, allowing modification of a pitch contour with components in a continuum of bands.

3. MODIFICATION ALGORITHM

As is indicated by the example in Fig. 3, a natural framework for modifying the duration of a pitch contour in a way that preserves the character of the underlying components involves separating the contour linearly into non-overlapping bands through a bank of linear time-invariant (LTI) filters, followed by resampling and/or time-scale modification of the pitch contour components in each band. Fig. 4 depicts a system that can be used to perform modifications of this type. In each band, the component may be resampled, shifted in frequency (i.e., the positive frequencies in its discrete-time Fourier transform may be shifted while maintaining conjugate symmetry), and time-scaled by a pre-specified amount, allowing for a broad set of modifications. Using this structure, a continuous, invertible, piecewise-linear mapping from input frequencies to output frequencies is realized using filters with a bandpass response, setting the resampling factors in the structure to the reciprocal of the slopes of the line segments in the frequency mapping, and adjusting the frequency shifts in the structure so that the line segments join to form a continuous function.

In particular, the LTI filters $G_k(e^{j\omega})$ are chosen to be ideal nonoverlapping frequency-selective bandpass filters that cover the entire range of frequencies, i.e.

$$\sum_{k=1}^{N} G_k\left(e^{j\omega}\right) = 1\,. \tag{1}$$

The product of the resampling factor and the time-scale modification factor in each band is constrained to be the same value c, resulting in an overall scaling of the duration of the pitch contour by a factor of c. The variable c is therefore used to denote the overall scaling factor.

Our proposed algorithm approximates the behavior of the system in Fig. 4 as the number of bands N becomes infinite, resulting in an arbitrary invertible, continuous frequency mapping. This allows for a very general set of manipulations of perturbation components in an arbitrary number of bands.

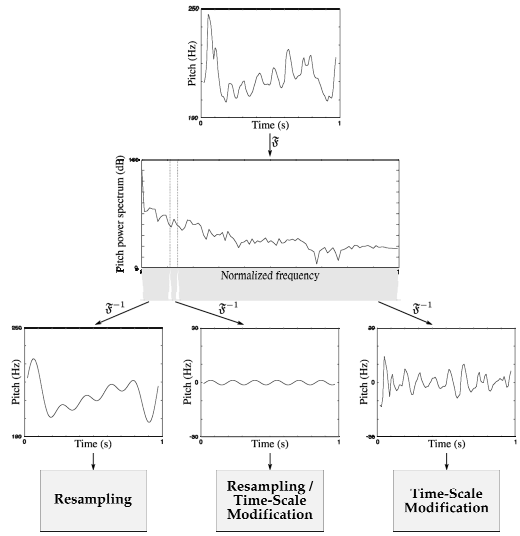

Figure 3. Example linear decomposition of pitch contour. $\mathfrak{F}$ and $\mathfrak{F}^{-1}$ denote the forward and inverse Fourier transform, respectively.

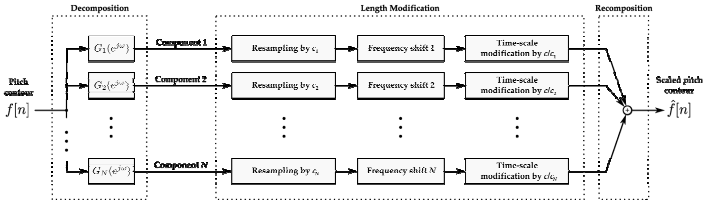

Figure 4. Modification framework for linearly-separable components.

3.1. Input-output characterization

In developing an algorithmic realization of the system in Fig. 4, we first characterize the system from an input-output perspective, and then design an algorithm that operates on a continuum of bands, approximating this characterization.

There are many ways to characterize the salient behavior of a system that involves time-scale modification of a component in a pitch contour, since the majority of existing time-scale modification techniques, traditionally applied to speech or other acoustic signals, are nonlinear. However, a common theme among existing frequency-domain time-scale modification algorithms, and the property that we rely on within our context, is that time-scale modification techniques for acoustic signals tend to scale the slowly-varying envelopes of modulated sinusoids while preserving their frequencies. It is for this reason that we characterize the system in Fig. 4 using modulated sinusoids of the form

$$f[n] = a[n]\cos(\omega n + \phi)\,, \tag{2}$$

where the envelope $a[n]$ is slowly-varying. Applying this signal to the system in Fig. 4 results in the output signal

$$\hat{f}[n] = \hat{a}[n]\cos(\hat{\omega}n + \hat{\phi}) . \quad (3)$$

Specifically, within our context we are interested in the behavior of the system in Fig. 4 with respect to a pitch contour composed of a superposition of modulated sinusoids, where for example a collection of high-frequency sinusoids may correspond to perturbation components, their envelopes to the location of the components in time, and where the envelope of a low-frequency collection of sinusoids may correspond to the underlying intonation of a pitch contour. In many time-scale modification algorithms for acoustic signals, a superposition of sufficiently few input modulated sinusoids that are sufficiently separated in frequency results, at least approximately, in a superposition of output modulated sinusoids, and the extent to which this is true is an important criterion in selecting a well-suited time-scale modification technique for use in our overall proposed algorithm.

The system in Fig. 4 may be characterized by the mappings from $a[n]$, ω and ϕ to $\hat{a}[n]$, $\hat{\omega}$ and $\hat{\phi}$ for input and output signals of the form of Eqns. (2-3). It can be shown that under the previously-mentioned constraints on the LTI filters, resampling factors, frequency shifts, and time-scale modification factors that resulted in a continuous frequency mapping, the system in Fig. 4 maps an input sinusoid of the form of Eq. (2) to an output sinusoid that is approximated by the form of Eq. (3), where

$$\hat{a}[n] = \text{resampling-by-}c\{a[n]\} , \quad (4)$$

$$\hat{\omega} = \beta(\omega) , \quad (5)$$

and

$$\hat{\phi} = \phi . \quad (6)$$

Here $\beta(\omega)$ is an arbitrary invertible, piecewise-linear frequency mapping, an example of which we depict in Fig. 8. The input parameters to the algorithm are therefore c and $\beta(\omega)$.

3.2. Algorithm formulation

In extending the system in Fig. 4 to the continuum case, we wish to design a system that implements the mappings prescribed by Eqns. (4-6), where the mapping $\beta(\omega)$ is now continuous and invertible but is otherwise arbitrary. In the proposed algorithm, this is achieved by operating on the pitch contour $f[n]$ in two stages. The first stage is a linear, *frequency-dependent resampling* stage that implements the frequency mapping in Eq. (5) and produces a single output pitch contour, denoted $\bar{f}[n]$, that is analogous to the set of inputs to the time-scale modification subsystems in Fig. 4. However, this stage also resamples the envelope of a particular modulated sinusoid by a factor that depends on its frequency, as opposed to the desired constant scaling. The second stage of the algorithm therefore corrects this by processing $\bar{f}[n]$ using a *frequency-dependent time-scale modification* algorithm, resulting in an output signal $\hat{f}[n]$ that achieves the desired mappings. The overall algorithm is illustrated in Fig. 5.

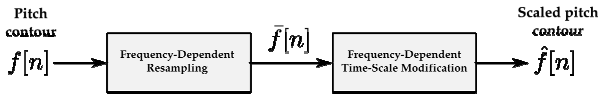

Figure 5. Overall contour scaling algorithm.

Frequency-dependent resampling: In designing an algorithm to implement the first stage in Fig. 5, which realizes the frequency mapping $\beta(\omega)$, we use a structure that is reminiscent of the frequency-warping structures presented previously in [6] and [7]. Fig. 6 depicts our implementation of this structure, which achieves the arbitrary invertible, continuous frequency mapping $\beta(\omega)$ specified in Eq. (5). The basic structure consists of a series of linear allpass filter elements, with the flat magnitude response of the elements maintaining the amplitude of the various components and the group delay function of the elements implementing a frequency-dependent time delay. Intuitively, the behavior of the structure can be explained by considering as its input a narrowband energy packet. In this interpretation, the chain of allpass filters functions as a frequency-dependent delay line. The speed at which the packet propagates down the chain therefore depends on its center frequency, and a signal that is composed of sampled outputs from the allpass filters, each sampled at a single time $n = 0$, contains a resampled and possibly modulated version of the original energy packet. It is for this reason that we refer to the structure as a frequency-dependent resampler.

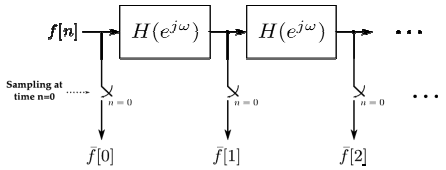

Figure 6. First stage of presented algorithm. Each filter $H(e^{j\omega})$ in the chain is a linear allpass filter.

Although the structure in Fig. 6 is related to the sampling and frequency-warping networks described in [6] and [7], in that work the constraint of causality on the allpass filters limits the set of achievable frequency mappings. In contrast, the allpass filters used here are allowed to be noncausal and, as can be shown, consequently result in a system that is in theory capable of implementing an arbitrary continuous frequency mapping. In this context, the allpass filter $H(e^{j\omega})$ is designed to approximate

$$H\left(e^{j\omega}\right) = e^{j\beta(\omega)} , \quad (7)$$

which for an input modulated sinusoid as in Eq. (2) can be shown to produce the output

$$\bar{f}[n] = \left(\text{resampling-by-}\left(-1/\tau_g(\omega)\right)\{a[n]\}\right)\cos\left(\beta(\omega)n + \phi\right) , \quad (8)$$

where $\tau_g(\omega) = -d\beta(\omega)/d\omega$ is the group delay of $H(e^{j\omega})$. Eq. (8) indicates that the frequency mapping specified by Eq. (5) is achieved by this stage of the algorithm, while the envelope relationship specified by Eq. (4) is not. The following stage in the algorithm therefore focuses on achieving the desired envelope relationship.

Frequency-dependent time-scale modification: In implementing the second stage in Fig. 5, we use a modification of Flanagan's phase vocoder-based time-scale modifier [4], traditionally used to process acoustic signals. Because the phase vocoder operates on signal bands independently, it may consequently be modified in a straightforward way to perform scaling by different factors in different bands of $\bar{f}[n]$. One disadvantage to using this particular algorithm is that due to its use of the discrete Fourier transform (DFT), a discrete set of time-scale factors is required, and the overall contour modification algorithm consequently does not operate on a true continuum of bands. However, it may be possible to retrofit other existing time-scale modification techniques that do not rely on a discretized spectrum,

such as the STS-based algorithm in [3], to perform frequency-dependent time scaling as well.

In this stage of the algorithm, the time-scale factor is a function of frequency and is chosen so that for an input signal as described by Eq. (8), an output signal as described by Eqns. (3-6) is obtained. This is achieved by time-scaling signals in the vicinity of the frequency $\beta(\omega)$ by a factor of $-c\tau_g(\omega)$, thereby correcting for the envelope scaling from the previous stage.

4. RESULTS

In this section we consider three pitch contours that were chosen to illustrate the possible behavior of the algorithm presented in Section 3. Fig. 7 compares the contours before and after scaling by a factor of $c = 2$, using both traditional upsampling, typical of existing speech time-scale modification techniques, as well as the presented contour scaling algorithm. As previously mentioned, the algorithm is focused on the processing of pitch contours, as opposed to the details associated with pitch analysis and synthesis. The contours in Fig. 7(a-b) are therefore synthetic contours intended to demonstrate the behavior of the algorithm for perturbations that occur at distinct times (irregularity for speech and vibrato for singing). The contour in Fig. 7(c) is an actual pitch contour containing flutter, which was used in Figs. 1(a) and 3.

In processing the irregular speech contour in Fig. 7(c), the frequency mapping function $\beta(\omega)$ depicted in Fig. 8 was used. The allpass filters used in the frequency-dependent rate converter were 150^{th}-order noncausal filters that were designed to approximate the phase response of $\beta(\omega)$ using the technique in [8]. The phase vocoder-based time-scale modifier used a length-32 DFT, Hann analysis and synthesis windows, and a frame step of 2 samples. The small step size was chosen so as to more effectively track the envelopes of the low-frequency sinusoids composing the underlying contour. The original and processed pitch contours were sampled at 100 Hz. Similar parameters were used in processing the synthetic contours in Fig. 7(a-b). For each example, the qualitative nature of the high-frequency perturbations in the contour $f[n]$ was preserved in the lengthened contour $\hat{f}[n]$.

5. CONCLUSION AND FUTURE WORK

In this paper, we introduced a general framework for scaling pitch contours so as to preserve the naturalness of pitch perturbations, and we presented an algorithm to perform such a scaling in the case of linearly-separable perturbation components. Future work includes completing further evaluations of the algorithm within the context of a complete analysis/modification/synthesis system, using subjective and objective measures on a large set of data. Informal listening involving vowel synthesis from the various contours in Fig. 7 verified that the qualitative nature of the perturbations that was lost in the resampled contours $r[n]$ was preserved in the scaled contours $\hat{f}[n]$, and additional evaluations would give a better sense of the best algorithm parameters to use for those and other classes of pitch perturbations.

Future work may also investigate nonlinear techniques for separating the components, using for example median filtering or the bilateral filter [10], conventionally used in image smoothing. As an example motivating the use of these nonlinear techniques, note that in processing the contour in Fig. 7(c) the initial peak, which may be representative of the underlying intonation, is smeared in the scaled contour $\hat{f}[n]$. Also to that end, further investigating alternative time-scale modification techniques for use in the second stage in Fig. 5, such as the STS system [2-3] or the phase-locked phase vocoder [5], may result in algorithm improvements. More generally, future work includes extending the algorithm to process contours corresponding to energy or other parameters, as well as discontinuous pitch contours, as with voiced/unvoiced transitions or abrupt note transitions in singing.

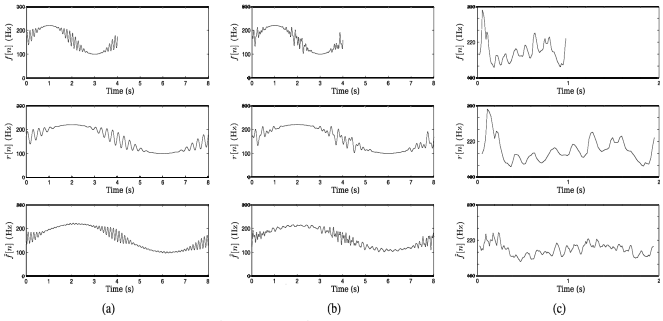

Figure 7. Algorithm performance for pitch contours $f[n]$, (a) synthesized with vibrato, (b) synthesized with flutter, modeled as in [9], and (c) actual contour from Figs. 1(a) and 3. Time expansion was performed by a factor of $c = 2$. The contour $r[n]$ (middle row) is the result of upsampling $f[n]$ by a factor of 2, as in existing time-scale modification techniques. $\hat{f}[n]$ (bottom row) is the output from the presented algorithm.

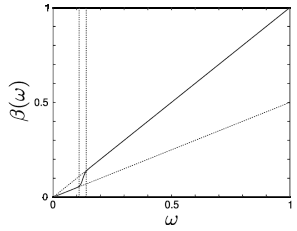

Figure 8. Frequency mapping parameter $\beta(\omega)$ used in obtaining $\hat{f}[n]$ in Fig. 7(c). The vertical dotted lines represent the band edges that were chosen for the decomposition, as in Fig. 3. Regions where the curve lies on the dotted line with a slope of ½ indicates input frequencies that are effectively upsampled by a factor of 2, and regions where the curve lies on the dotted line with a slope of 1 indicates input frequencies that are effectively time-scaled by a factor of 2.

Acknowledgment: The authors thank Robert Dunn of MIT Lincoln Laboratory for helpful discussions and for supplying examples of singer's vibrato.

6. REFERENCES

[1] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication* 9: 5-6, pp. 453-467, 1990.

[2] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE TASSP*, 34: 4, pp. 744-754, 1986.

[3] T. F. Quatieri and R.J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Trans. on Signal Processing*, 40: 3, pp. 497-510, 1997.

[4] J. Flanagan and R. Golden, "Phase Vocoder," *Bell Sys. Tech. J.* 45, pp. 1493-1509, 1966.

[5] M. Puckette, "Phase-locked vocoder," *Proc. IEEE WASPAA*, pp. 222-225, 1995.

[6] C. Braccini and A.V. Oppenheim, "Unequal bandwidth spectral analysis using digital frequency warping," *IEEE TASSP*, 22: 4, pp. 236-244, 1974.

[7] A.V. Oppenheim and D.H. Johnson, "Discrete representation of signals," *Proc. IEEE*, 60: 6, pp. 681-691, 1972.

[8] B. Yegnanarayana, "Design of recursive group-delay filters by autoregressive modeling," *IEEE Trans. on Acoustics, Speech and Signal Processing*, 30: 4, pp. 632-637, 1982.

[9] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," JASA, vol. 87, pp. 820-857, FEB 1990.

[10] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," *Proc. IEEE Int'l Conf. on Computer Vision*, pp. 839-846, 1998.