TIME DOMAIN TECHNIQUE FOR PITCH MODIFICATION AND ROBUST VOICE TRANSFORMATION

Rivarol Vergin *Douglas O'Shaughnessy* *Azarshid Farhat*

INRS-Télécommunications
16 Place du Commerce,
Île-des-Sœurs, H3E1H6, Québec, Canada
email: vergin@inrs-telecom.uquebec.ca

ABSTRACT

Modification of speech is a subject of major interest today, with numerous applications including text to speech synthesis. The basic mechanisms behind this process often consist of pitch-scale and time-scale modifications of speech. While giving generally good results, it remains in most of the cases that the same speaker can be associated with the original signal and its modified version, which limits the use of these techniques in some applications where disguising voices is necessary. These paper presents an approach to increase the possibilities of speech modifications while preserving most of the speech quality of the original signal.

1. INTRODUCTION

The underlying model of speech production involving an excitation source and a vocal tract filter is implicit in many analysis methods. The excitation is often analyzed in terms of the periodicity and amplitude of the signal, while variations in the speech spectrum are assumed to derive from vocal tract variations. The observable speech is the result of convolving excitation and vocal tract sample response.

The excitation may be either periodic, resulting in voiced speech, or noisy and aperiodic, resulting in unvoiced speech. The voicing occurs in the larynx, where airflow can be interrupted periodically by the vocal folds. The puffs of air produced by the opening and closing of the vocal folds generate a periodic excitation for the vocal tract. The fundamental frequency of the vocal fold vibration is known as $F0$ and the perceptual feature of speech corresponding to $F0$ is often called "pitch".

Because pitch periods generally associated with male speakers are quite different from those associated with female speakers, modification of speech often results in pitch modification. Indeed, increasing or decreasing the pitch periods of the input speech signal can modify the apparent gender of the current speaker for a male or, respectively, female speaker. This pitch-scale modification affects the apparent rate of the original signal, hence the need for time-scale modifications giving to the modified signal the original apparent rate of articulation.

It remains in most cases that, while the modification is apparent, the same speaker can be associated with the original signal and its modified version, which can be a nuisance in some applications where disguising voices is necessary. This contribution suggests a new technique, based on an interpolation-decimation procedure, to reduce the possibility for the original speaker to be recognized while preserving most of the speech quality of the original signal, when evaluated by some informal listening tests.

2. SPEECH MODIFICATION

Many aspects of speech can be considered in an application where speech modification is a need. Modification of some prosodic parameters (e.g., duration, intensity) can change, in some cases, the meaning of an utterance. However, one of the most common techniques used to modify the aspects of a speech signal involves a modification of its apparent pitch contour [1] [2]. The first step of this procedure is an estimation of the pitch epochs, that is the closing time of the vocal cords corresponding to a high amplitude in the signal.

2.1. Pitch estimation

Because of the nonstationary nature of speech, irregularities in vocal cord vibration, interaction of the vocal tract and the glottal excitation, a perfect evaluation of the pitch periods is not always possible. Many algorithms exist, however, allowing us to have a reasonable estimation of these parameters. Some of them are performed in the frequency domain from a measurement of the harmonic spacing, others are directly performed in the time domain.

Autocorrelation [3] is the method retained in this paper to evaluate the pitch periods. Assuming that $x(n)$ corresponds to a voiced part of the input speech signal, the short-time autocorrelation function is given by:

$$\phi_i(m) = \frac{1}{N'} \sum_{n=0}^{N'-1} x(n+i)x(n+i+m), \qquad (1)$$

where i is the index of the starting sample of the frame, N (corresponding approximately to 30 ms) is the section length being analyzed. N' is generally set to: $N' = N - m$.

The effects of the formant structure are reduced through the use of a low pass filter with a passband of 0 to 900 Hz. The estimated pitch periods, obtained from the autocorrelation function, are used to define the bounds where pitch epochs (here defined by $t_a(i)$) are expected in the input signal $x(n)$. The procedure is completed by a postprocessing algorithm, based on the mean value of the pitch periods, to eliminate irrelevant pitch epochs.

2.2. Pitch-scale modification

Following the evaluation of the pitch epochs, the pitch scale modification is simple; the input signal $x(n)$ is decomposed into a stream of short-segments of signal $x(i,n)$ by multiplying the input by a sequence of pitch-synchronous windows $h(i,n)$, that is:

$$x(i,n) = h(t_a(i) - n)x(n). \qquad (2)$$

The window is centered around the successive pitch epochs, $t_a(i)$, on the voiced part of the signal and at a constant rate on the unvoiced part. $h(n)$ is a Hanning window with a length greater than or equal to two pitch periods. The successive short-term signals then always overlap. Defining α as a scaling factor, generally α takes values between 0.5 and 2; the pitch epochs of the synthesis signal are located at $t_s(i)$, where:

$$t_s(i) = \alpha t_a(i). \qquad (3)$$

The synthesis signal is simply given by:

$$x'(n) = \sum_i \gamma(i)x(i, t_s(i) - n). \qquad (4)$$

$\gamma(i)$ is a time varying normalization factor which compensates for the energy modifications related to the pitch modification procedure.

2.3. Time-scale modification

Because each short-term signal, $x(i,n)$, is only translated to a new centered position $t_s(i)$, the resulting synthesis signal, $x'(n)$, can appear shorter or longer than the original depending if α is less or respectively greater than one. There is hence a need for time-scale modifications [2] to give to the synthesis signal the apparent rate of the original signal. Assuming that:

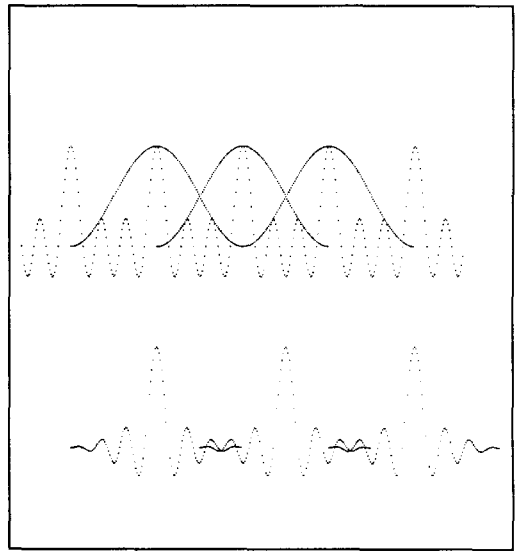

Figure 1. Illustration of the pitch modification procedure: the first curves (at the top) show the original signal and the windows, the second ones show the translation process for $\alpha = 3/2$.

$$t_a(i') \leq t_s(i) \leq t_a(i'+1), \qquad (5)$$

where i and i' are not necessarily equal, the short-term synthesis signal is given by:

$$y(i,n) = (1-\beta)x(i',n) + \beta x(i'+1,n), \qquad (6)$$

with:

$$\beta = \frac{t_s(i) - t_a(i')}{t_a(i'+1) - t_a(i')}. \qquad (7)$$

The new synthesis signal is given by:

$$y(n) = \sum_i \gamma(i)y(i, t_s(i) - n). \qquad (8)$$

This definition permits one to obtain a synthesis signal, $y(n)$, which has the same apparent rate and intensity contour as the original signal.

While giving generally good results, the apparent pitch contour of the synthesis signal is relatively different from the input signal; it remains in most of the cases that the same speaker can be associated with the original signal and its modified version. In some applications however, the need is not only to modify the apparent pitch contour of the speech, but also to reduce the possibility for the original speaker to be recognized. The next section suggests a technique

allowing us to reach this goal partially while preserving most of the speech quality of the original signal.

3. LINEAR SHIFT-VARIANT SYSTEM

The need for linear shift-variant digital filters occurs in many applications where the underlying system model is required to be linear but shift-variant. This is the case of digital-to-digital sampling rate conversion for example. When the sampling rate is decreased from the original sampling rate, the process is called decimation, and when the sampling rate is increased the process is called interpolation. Such systems can be characterized by:

$$y(n) = \sum_{m=-\infty}^{\infty} h(n,m)x(m), \tag{9}$$

where $x(m)$ and $y(n)$ are respectively input and output signals, and $h(n,m)$ is the output at time n of the system due to a unit impulse input applied at time m. For linear shift-invariant systems, $h(n,m)$, is a function of $(n-m)$, and in that case the transfer function is independent of n.

The decimation process consists of retaining only one sample out of each group of M consecutive samples. To avoid some aliaising problems, the original signal $x(m)$ must be sufficiently lowpass, which reduces its initial energy bandwidth by a factor M. The resulting output signal, $y(n)$, is then defined by:

$$y(n) = \sum_{m=-\infty}^{\infty} h(Mn-m)x(m). \tag{10}$$

The reverse process, interpolation, increases the sampling rate of a signal by a factor L, by introducing $L-1$ new samples values between each pair of successive sample values $x(m)$. It can be shown [4] that the resulting output signal, $y(n)$, can be expressed as:

$$y(n) = \sum_{m=-\infty}^{\infty} h(n-iL+mL)x(i-m), \tag{11}$$

where i is the largest integer less than or equal to n/L. Clearly equations 10 and 11 are linear but shift variant. Generally sampling rate conversion by a rational number, M/L, is obtained by first increasing the sampling rate by L, and then decreasing it by M. This concept, consisting of changing the sampling rate by a rational number, is applied in the next section to modify the current aspect of a speech signal through the use of a new interpolation procedure.

3.1. Interpolation and decimation process

The most common method used to raise the sampling rate by some factor, say L, via interpolation, involves an introduction of $L-1$ zero-valued samples between each pair of successive sample values, followed by a lowpass filter smoothing out the waveform to have the same shape as the signal before interpolation. If L is relatively high, $L \geq 20$, the lowpass filter must have many coefficients to properly accommodate the smoothing process. High values of L can be necessary if a fine sampling rate conversion by a rational number is desired.

To overcome the filtering process, the same algorithm used for time-scale modification (section 2.3) is applied for interpolation. Defining by $\widetilde{x}(k,m)$, $1 \leq k \leq L-1$, the set of new samples to be included between $x(m)$ and $x(m+1)$, they can be simply taken as equal to:

$$\widetilde{x}(k,m) = (1-\theta)x(m) + \theta x(m+1) \tag{12}$$

with $\theta = \frac{k}{L-1}$. Doing this, we insure that $\widetilde{x}(k,m)$ is more similar to $x(m)$ than $x(m+1)$ if $k \leq L/2$ and $\widetilde{x}(k,m)$ is more similar to $x(m+1)$ than $x(m)$ if $k \geq L/2$. However to take into account variations that can occur between consecutive pairs of sample values, $x(m)$ and $x(m+1)$ in the previous equation have been replaced by $\widehat{x}(m)$ and $\widehat{x}(m+1)$, that is:

$$\widetilde{x}(k,m) = (1-\theta)\widehat{x}(m) + \theta\widehat{x}(m+1), \tag{13}$$

with

$$\widehat{x}(m) = \sum_{i=0}^{I} p(i)x(m-i) \tag{14}$$

and

$$\widehat{x}(m+1) = \sum_{i=0}^{I} p(i)x(m+i). \tag{15}$$

As defined, $\widehat{x}(m)$ is a weighted sum of $x(m)$ and I past sample values and $\widehat{x}(m+1)$ is a weighted sum of $x(m+1)$ and I future sample values. The interpolated signal, $y_{int}(n)$, is obtained by concatenating all short-segment samples $\widetilde{x}(k,m)$, $0 \leq k \leq L-1$. The modification of the sampling rate by a rational number, M/L, is then completed by decimating the interpolated signal $y_{int}(n)$. The output signal, $y_{out}(m)$ is given by:

$$y_{out}(m) = y_{int}(Mn). \tag{16}$$

In our application, the interpolation/decimation procedure described in this section is applied on each short-term synthesis signal, $y(i,n)$, defined by equation 6. The resulting set of short-segments, $y_{out}(i,n)$, are then combined to obtain a new synthesis signal:

$$\widehat{y}(n) = \sum_{i} \gamma y_{out}(i, t_s(i) - n), \tag{17}$$

Depending on the values of M and L, in most cases $\widehat{y}(n)$ appears relatively different from $y(n)$ (equation 8) when evaluated by some informal listening tests. The weighting procedure defined by equations 14 and 15 is implemented using $I = 4$. The weighted coefficients $p(i)$ are taken equal to: $0.4, 0.3, 0.2, 0.1$. Changing the values of the coefficients $p(i)$ is an another method that can also be used to modify the current aspects of the input speech signal.

4. SUMMARY

In this paper we have examined one of the most used methods to modify the current aspect of a speech signal, that is, the one based on a modification of its apparent pitch contour. The first step of this technique is an estimation of the pitch epochs, follwed by a pitch-scale transformation to modify the apparent gender of a current speaker. The apparent rate and intensity contour of the original signal are preserved through the use of a time scale modification. The resulting synthesis signal has an apparent pitch contour relatively different from the input signal.

Because, in most of the cases, the same speaker can be associated with the original signal and its modified version, we have suggested the use of a new interpolation/decimation procedure to overcome this problem. This new technique, when combined with the pitch-scale and time-scale modification procedure, allows us to reach this goal while preserving most of the speech quality of the original signal.

REFERENCES

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

[2] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech", Speech communication, pp. 175-205, 1995.

[3] L. R. Rabiner "On the Use of Autocorrelation Analysis for Pitch detection", IEEE Trans. Acoust., Speech and Signal Processing, vol. ASSP-25 No. 1, February 1977.

[4] N. K. Bose "Digital Filters Theory and Applications". New York: Elsevier Science Publishing, 1985.