Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (San Francisco, CA), pp. I-81-I-84, March 1992

Time-Scale Modification of Speech Using an Incremental Time-Frequency Approach with Waveform Structure Compensation

Benoit Sylvestre [1] and Peter Kabal [1,2]

[1] Department of Electrical Engineering
McGill University
Montreal, Quebec H3A 2A7

[2] INRS-Télécommunications
Université du Québec
Verdun, Quebec H3E 1H6

Abstract

This paper first tries to identify the primary sources of distortion in a non-recursive *time-scale modification* (TSM) algorithm which is based on the short-time Fourier transform (STFT) (Portnoff, [1]). A simpler version of this TSM algorithm is then proposed for processing speech, where *incremental* estimators eliminate the need for explicit linear time-scaling operations. Also featured in the design is a *waveform structure compensation* stage to prevent excessive deterioration of the rate-changed output. A *polar* (i.e. magnitude-phase) synthesis equation is used for increased efficiency. The new TSM method is capable of generating high-quality rate-changed speech at a reasonable computational cost.

1. Introduction

Time-scale modification (TSM) is a process whereby signals are compressed or expanded in time in a manner which preserves (within practical limits) their original frequency characteristics. For example, a listener perceives changes in the apparent rate of articulation of *rate-changed* speech, but not in the speaker dependent features such as pitch and timbre. Potential applications include audio monitoring, reading machines for the blind, audio data compression/expansion and film-to-soundtrack synchronization.

At least two approaches to the problem of generating rate-changed speech have received considerable attention over the last decade:

- **Least-squares error estimation (LSEE)** [2]. Recursive technique whereby a rate-changed signal is estimated by minimizing, in the mean-square sense, the euclidean distance between the short-time Fourier transform (STFT) of the original and rate-changed signals.
- **Time-frequency models** [1, 3]. Usually based on the STFT, these models allow time and frequency to be manipulated independently to achieve a speech modification goal. The speech signal is normally represented as a linear combination of sinusoids [4] or complex exponentials [5].

LSEE algorithms, by the very fact that they are recursive and transparent to the characteristics of the speech signal, are more likely to generate a structurally sound output than current "single pass" time-frequency algorithms. The fragility of the source models used in the latter case and the non-linear transformations required to achieve TSM are to blame. However, the performance impact of these factors is currently not well understood. Further investigation in this area should prove useful, given the speech modification potential of one particular time-frequency scheme, *sinusoidal speech modeling* (SSM) [3].

We first highlight the causes for structural distortion in rate-changed speech. A key paper in the field of time-frequency TSM (Portnoff, [1]) will serve as the basis for our discussion. A novel method [6] which restricts the distortion within reasonable bounds is proposed along with several important simplifications to Portnoff's original design. The revised algorithm, which combines a new *incremental* parameter modification scheme [6] with the advantages of polar synthesis [4], is more robust than its predecessors, while affording high-quality rate-changed speech at a computational cost comparable to, if not less than, that of the SSM-based version.

2. Review and Definitions

The reader is assumed to be familiar with the material presented in [1] and [5]. Similar notation will be adopted.

The signal or sequence $x(n)$, where n is integer valued, represents the samples of a continuous-time, bandlimited waveform $x(t)$ with the sampling interval normalized to unity. The STFT of $x(n)$ is defined as

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} h(n-m)x(m)e^{-j\omega m}, \tag{1}$$

where ω is continuous and $h(n)$ is an analysis window of length M. The STFT $X(n,\omega)$ is periodic in ω with period 2π and may be regarded as a sequence in n with ω treated as a parameter. The STFT synthesis equation is

$$x(n) = \frac{1}{2\pi h(0)} \int_{-\pi}^{\pi} X(n,\omega)e^{j\omega n}d\omega. \tag{2}$$

The time-scaling factor of a sequence is represented by β, a rational number. The range $\beta > 1$ corresponds to time-scale *compression*, and the range $0 < \beta < 1$, to time-scale *expansion*. The rate-changed version of a sequence $f(n,\ldots)$ is written as $f^{\beta}(n,\ldots)$, whereas its *linearly* time-scaled version, as $f(\beta n,\ldots)$. In general, a rate-changed sequence is obtained through non-linear means.

The model used in [1, 5] views a speech signal as the response of a linear time-varying filter to an excitation source which is either a "quasi-periodic" unit-sample train in the case of voiced speech, or white noise in the case of unvoiced speech. Major speech parameters such as pitch and vocal tract filter response are expressed as a function of time and are assumed to be "nearly fixed" for the duration of the vocal tract filter response.

The rate-changed version of the speech model is obtained by linearly time-scaling the speech parameters. For voiced speech,

however, there is one important exception: the values of the instantaneous phase of the excitation source must be scaled by $1/\beta$ to preserve the original pitch (and bandwidth) of the signal [1].

The following polar STFT expression [5] is used to estimate the necessary speech parameters for implementing the rate-change modification:

$$X(n,\omega) = M(n,\omega)\exp[j\theta(n,\omega)] \quad (3)$$
$$= M(n,\omega)\exp\Big[j\Big(\alpha(n,\omega)+\vartheta(n,\omega)\Big)\Big] \quad (4)$$

where

$$M(n,\omega) = |X(n,\omega)| \quad (5)$$
$$\theta(n,\omega) = \arg X(n,\omega) + 2\pi I(n,\omega) \quad (6)$$
$$\alpha(0,\omega) = \theta(0,\omega) \quad (7)$$
$$\vartheta(0,\omega) = 0. \quad (8)$$

The magnitude of the frequency response of the vocal tract filter is proportional to $M(n,\omega)$. The integer $I(n,\omega)$ guarantees the uniqueness of the *unwrapped* STFT phase $\theta(n,\omega)$. The quantity $\theta(n,\omega)$ is expressed as the sum of two unwrapped phase terms, $\alpha(n,\omega)$ and $\vartheta(n,\omega)$, which are both unknown. The phase modulation term $\alpha(n,\omega)$ represents the phase induced by the slow variations in the source pitch and the vocal tract filter response. The frequency modulation (FM) term $\vartheta(n,\omega)$, by comparison, varies much more quickly in n because it is proportional to the instantaneous phase of the excitation source. Portnoff developed an algorithm for estimating both phase quantities from $\theta(n,\omega)$.

The accuracy of the STFT-based magnitude and phase estimators depend to some extent on the design of the analysis window. The duration of $h(n)$ should be sufficiently short (no greater than 20ms) so that the speech parameters appear nearly fixed over the analysis interval. In spectral terms, the bandwidth of the analysis window should be wide enough to pass the speech parameters with negligible distortion. However, this bandwidth should be less than one half the source pitch to allow proper resolution of voiced speech spectra [5]. If $h(n)$ is a 20ms Hamming window, an "accurate" bandpass STFT representation of a voiced speech signal can be obtained only if its pitch exceeds 200Hz [1]. Portnoff argued that the same analysis filter bandwidth is adequate for processing both voiced and unvoiced speech [5].

The rate-changed version of $X(n,\omega)$ was postulated as

$$X^{\beta}(n,\omega) = X(\beta n,\omega)\exp\left[j\left(\frac{1}{\beta}-1\right)\vartheta(\beta n,\omega)\right] \quad (9)$$
$$= M(\beta n,\omega)\exp\Big[j\Big(\alpha(\beta n,\omega)+\vartheta(\beta n,\omega)/\beta\Big)\Big]. \quad (10)$$

Rate-change modifications are achieved by linearly time-scaling the original STFT and applying a non-linear phase modification to preserve the original bandwidth. It was suggested that the result is suitable for both voiced and unvoiced speech [1]. The rate-changed signal $x^{\beta}(n)$ is obtained by substituting (9) into the STFT synthesis equation.

3. Distortion in Rate-Changed Speech

The structural difference between $X(n,\omega)$ and $X^{\beta}(n,\omega)$ is the major cause of distortion in rate-changed speech. Since $X(n,\omega)$ is a function of a finite number of samples, a "regular" STFT can be said to have finite memory. Though it may be argued that $X(\beta n,\omega)$ has finite memory also, $X^{\beta}(n,\omega)$ clearly does not—the $\vartheta(\beta n,\omega)/\beta$ term has *infinite* memory (i.e. each phase value depends on the location of the time origin, as integer multiples of 2π are not invisible for $\beta \neq 1$). Therefore, FM component estimation errors accumulate in $X^{\beta}(n,\omega)$ indefinitely.

Even if the quantity $\vartheta(\beta n,\omega)$ were known exactly, some degree of phase error in $X^{\beta}(n,\omega)$ would still be unavoidable. Since the FM component will often deviate from the local phase linearity assumption of the speech model, the $1/\beta$ scaling factor in (10) merely ensures that the *average* frequency of $X^{\beta}(n,\omega)$ matches that of $X(n,\omega)$. The phase deviations may occur within the scope of an STFT frame due to waveform transients or other features which violate the source model.

The accumulation of phase disturbances eventually destroys the original phase relationship of the frequency components, thereby impairing the *waveform structure* of $x^{\beta}(n)$. The next section proposes a method which restricts the memory in $X^{\beta}(n,\omega)$ to improve long-term performance.

4. Waveform Structure Compensation

Suppose that the speech signal $x(n)$ is segmented into a string of contiguous sub-waveforms or *segments* $x_i(m)$, each of length L (in samples). The sample index m spans the segment length and is linearly related to n, i.e. $n = m + iL$. The STFT of the i-th segment, $X_i(m,\omega)$, is just $X(n,\omega)$ expressed as a function of i and m.

The goal is to construct the rate-changed signal $x^{\beta}(n)$ by concatenating its individually rate-changed segments $x_i^{\beta}(m)$. Each one of these is synthesized from the rate-changed version of $X_i(m,\omega)$, namely

$$X_i^{\beta}(m,\omega) = X_i(\beta m,\omega)\exp\left[j\left(\frac{1}{\beta}-1\right)\vartheta_i(\beta m,\omega)\right], \quad (11)$$

where $\vartheta_i(0,\omega) = 0$. The initial condition on the FM component indicates that the STFT phase unwrapping process is reset at intervals of L samples on the original time-scale.

As Figure 1 illustrates, the $x_i^{\beta}(m)$ are not necessarily phase-continuous at their edges because the average frequency of each rate-changed component ω is preserved while the segment lengths are altered by a factor of $1/\beta$. The initial phases (at $m = 0$) of the frequency components in each segment are not affected by a rate-change modification, as (11) confirms. Consequently, the quantities $\arg X_i(0,\omega)$ serve as a convenient reference for correcting the phase relationship of the frequency components.

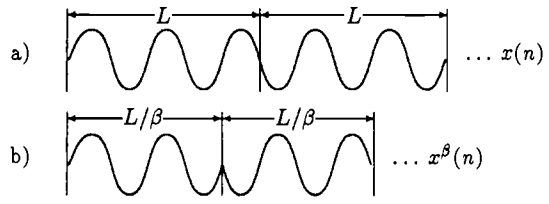

Fig. 1 a) Original signal with segment boundaries. b) Concatenation of individually rate-changed segments.

In order to prevent the phase error of one segment from propagating to future segments, it is proposed to force adjacent rate-changed segments to be phase-continuous. The amount of phase correction required at the i-th segment boundary per frequency component is

$$\Theta_i(\omega) \approx d_{\pi}\Big[\arg X_{i+1}(0,\omega) - \epsilon_{i+1} - \arg X_i^{\beta}(L^{\beta},\omega)\Big], \quad (12)$$

where $L^\beta = \lfloor L/\beta \rfloor$. The floor operator $\lfloor \cdot \rfloor$ truncates the fractional part of its argument. The $d_\pi[\cdot]$ operator adds or removes multiples of 2π to its argument until the results lies in the $-\pi$ to π range. The quantity $X_i^\beta(L^\beta, \omega)$ corresponds roughly to the same sample index along $x^\beta(n)$ as $X_{i+1}(0, \omega)$. The constant ϵ_{i+1} offsets the starting phase of $X_{i+1}(m, \omega)$ to reduce the amount of phase correction for *perceptually* important frequency components. Since ϵ_{i+1} is in effect a complex plane rotation factor, the waveform structure of the $(i+1)$-th segment remains intact. The phase offset for the i-th segment is given by

$$\epsilon_i = \int_{-\pi}^{\pi} W_{i-1}(\omega) d_{2\pi}\Big[\arg X_i(0,\omega) - \arg X_{i-1}^\beta(L^\beta, \omega)\Big] d\omega. \quad (13)$$

The $d_{2\pi}[\cdot]$ operator is identical to $d_\pi[\cdot]$ operator, except that the final result lies in the 0 to 2π range. The weight function $W_{i-1}(\omega)$ favors the perceptually important frequency components of the $(i-1)$-th segment. The weights may be chosen on the basis of spectral energy and absolute frequency. The human ear will be more sensitive to distortion at the lower end of the auditory spectrum.

We define a phase-modulated version of $x_i^\beta(m)$ as

$$y_i^\beta(m) = \frac{1}{2\pi h(0)} \int_{-\pi}^{\pi} X_i^\beta(m,\omega) \times \exp\Big[j\Big(\mu_i(m,\omega) + \epsilon_i\Big)\Big] e^{j\omega m} d\omega, \quad (14)$$

where

$$\mu_i(m,\omega) = \frac{\Theta_i(\omega)}{L^\beta} m \qquad \text{for } 0 \le m < L^\beta. \quad (15)$$

The amount of phase correction is uniformly distributed so as to minimize the amount of distortion per sample. Hence, the average frequency of $X_i^\beta(m, \omega)$ is shifted by a fixed quantity which does not exceed π/L^β. To summarize (14), $\mu_i(m, \omega)$ ensures that $y_i^\beta(m)$ is phase-continuous with $y_{i+1}^\beta(m)$, while ϵ_i reduces the perceptual impact of phase modulation in $y_{i-1}^\beta(m)$.

5. Synthesis and Parameter Modification

In practice, the *discrete* version of the STFT synthesis equation (and analysis equation) will be used, i.e.

$$x_i^\beta(m) = \frac{1}{Nh(0)} \sum_{k=0}^{N-1} X_i^\beta(m,\omega_k) e^{j\omega_k m}, \quad (16)$$

where $\omega_k = 2\pi k/N$. The constant N denotes the number of frequency samples. If (16) were computed using an N-point Fast Fourier Transform (FFT) algorithm, a polar to rectangular coordinate conversion would be required before (11) could be substituted into (16). Since the unwrapped STFT phase $\theta_i^\beta(m, \omega_k)$ is readily available (because it serves in estimating the exponential phase term of (11)), it is more efficient to synthesize $x_i^\beta(m)$ from the polar rather than rectangular representation of $X_i^\beta(m, \omega_k)$. Furthermore, $x_i^\beta(m)$ is real, and so (16) can be reduced to

$$x_i^\beta(m) = \frac{2}{Nh(0)} \sum_{k=0}^{N/2-1} M_i^\beta(m,\omega_k) \cos\Big(\theta_i^\beta(m,\omega_k) + \omega_k m\Big). \quad (17)$$

Unlike the SSM synthesis equation [3, 4], (17) uses the *entire* spectrum as well as fixed rather than time-varying base frequencies. The advantages of (17) become more apparent when $X_i(m, \omega)$ is downsampled in time: the interpolation procedure for polar parameters is usually simple for achieving high-quality synthesis due to their relative smoothness [4]. This property may be further exploited for evaluating $X_i^\beta(m, \omega_k)$. The following *incremental* estimators are proposed:

$$\hat{M}_i^\beta(m,\omega_k) = M_i(\lfloor \beta m \rfloor, \omega_k) \quad (18)$$

$$\hat{\theta}_i^\beta(m,\omega_k) = \theta_i(0,\omega_k) + \sum_{r=1}^{m} \nabla_{\lfloor \beta r \rfloor}^b \theta_i(\lfloor \beta r \rfloor, \omega_k), \quad (19)$$

with the initial conditions

$$\hat{M}_i^\beta(0,\omega_k) = M_i(0,\omega_k) \quad (20)$$

$$\hat{\theta}_i^\beta(0,\omega_k) = \theta_i(0,\omega_k). \quad (21)$$

The $\nabla_q^b[\cdot]$ operator is the first backward difference with respect to q. Equations (18) and (19) define a TSM system where no explicit linear time-scaling nor multiplications by $1/\beta$ are required. Time-scale compression ($\beta > 1$) is in effect achieved by periodically *deleting* sample "intervals" from $x_i(m)$, whereas time-scale expansion ($0 < \beta < 1$) consists of periodically *repeating* sample "intervals" in $x_i(m)$.

Due to its incremental structure, the rate-changed phase sequence $\hat{\theta}_i^\beta(m, \omega_k)$ retains essentially the same smoothness properties as the original. However, $\hat{M}_i^\beta(m, \omega_k)$ is in general discontinuous, which is of no great concern because $M_i(m, \omega_k)$ is relatively smooth. Since the above scheme disregards the source model entirely, the waveform deterioration process is expected to be more rapid.

Incorporating the phase modulation terms and the incremental estimators into (17) yields

$$y_i^\beta(m) = \frac{2}{Nh(0)} \sum_{k=0}^{N/2-1} \hat{M}_i^\beta(m,\omega_k) \cos\Big(\hat{\theta}_i^\beta(m,\omega_k) + \mu_i(m,\omega_k) + \epsilon_i + \omega_k m\Big). \quad (22)$$

The rate-changed signal $x^\beta(n)$ is constructed by concatenating the $y_i^\beta(m)$.

6. Simulation

A software simulation of the TSM system previously described was performed using a 20ms Hamming analysis window ($= h(n)$), a 512-point short-time Fourier analysis stage and a sampling rate of 16kHz ($= f_s$). Each (discrete) STFT frame was computed at sample intervals equal to $M/4$, where M is the length of $h(n)$. Both the original STFT magnitude and unwrapped phase sequences were upsampled via linear interpolation by a factor of $M/4$ so that the parameter modification procedure given by (18) and (19) could be applied.

In the absence of any parameter modification ($\beta = 1$), the original speech signal (Figure 2) and its reconstructed version (Figure 3) were virtually indistinguishable. The result suggests that the distortion caused by polar parameter interpolation is negligible.

TSM tests were conducted over the range $0.5 \le \beta \le 2.0$ using speech signals in the male and female speaker categories. Examples are shown in Figure 4 and Figure 5. In general, rate-changed signals whose voiced portions satisfy the minimum pitch bound which ensures proper resolution of voiced speech spectra can be rated as "very good" to "excellent"—the output is free from artifacts and its subjective quality is similar to that of the original signal. Some distortion due to poor resolution of voiced speech

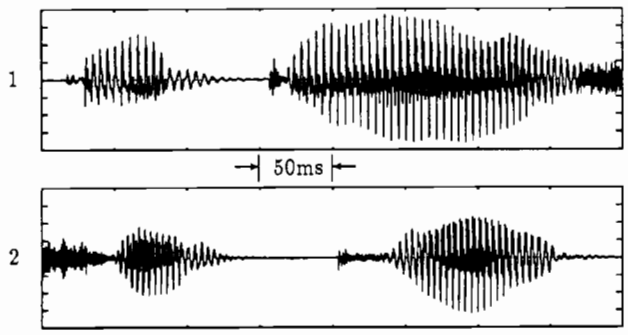

Fig. 2 Original signal represented as two consecutive parts.

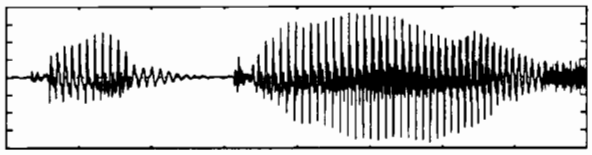

Fig. 3 Reconstructed version ($\beta = 1$) of Figure 2.1.

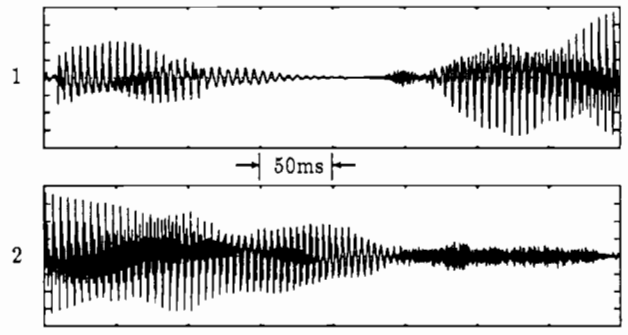

Fig. 4 Expanded version ($\beta = 0.5$) of Figure 2.1.

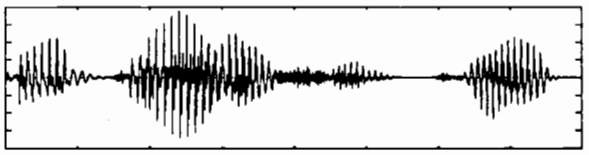

Fig. 5 Compressed version ($\beta = 2.0$) of Figure 2.

spectra occurs for lower-pitched signals, particularly for those in the male speaker category. Time-scale compression tends to mask this distortion because the rate of articulation is accelerated. For time-scale expansion, however, quavering in the voiced portions (especially near unvoiced-to-voiced speech boundaries) and smearing in the unvoiced portions become more evident as β is decreased. Expanded speech signals in the male speaker category can be rated as "good" because they are free from artifacts and are still quite intelligible.

The speech segment lengths $L = \lfloor 100 \times 10^{-3} f_s \rfloor$ for $\beta > 1$ and $L = \lfloor 100\beta \times 10^{-3} f_s \rfloor$ for $0 < \beta < 1$ gave good results. The sampling frequency f_s affects the quality of rate-changed speech in at least three ways. As f_s is increased,

- the granularity of the incremental estimators decreases.
- the effect of the waveform structure compensation section becomes less noticeable because the amount of phase correction is distributed over a larger number of samples.
- less frequency aliasing and phase unwrapping errors occur for perceptually important frequency components because they are shifted away from the Nyquist frequency.

Injecting background noise (eg. speech or music) into the source signal resulted in no significant loss of quality in the rate-changed output.

7. Conclusion

Current time-frequency TSM algorithms suffer from at least two weaknesses: the irremediable time-frequency resolution compromise of the STFT analysis filter and the cumulative structural distortion caused by the non-linear transformations involved. In general, the quality of rate-changed speech signals is better for higher pitched sources (>200Hz). Waveform structure compensation does not address the fundamental problem of estimating the unwrapped phase of the unknown rate-changed STFT. Ideally, $\theta^{\beta}(n,\omega)$ should be a non-linear function of past and future rate-changed samples in the vicinity of n on the *new* time-scale. It is not likely, therefore, that $\theta^{\beta}(n,\omega)$ can be estimated from phase data calculated on the original time-scale alone, without some form of recursion or analysis-by-synthesis. However, the simple TSM system which has been described may be more than adequate for certain applications. Moreover, variable TSM can be easily implemented by letting β vary as a function of time in (18) and (19).

References

[1] M. R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Transactions on Acoust., Speech, Signal Processing*, vol. 29, pp. 374–390, June 1981.

[2] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoust., Speech, Signal Processing*, vol. 32, pp. 236–242, Feb. 1984.

[3] T. F. Quatieri and R. J. McAulay, "Speech transformations based on a sinusoidal representation," *IEEE Transactions on Acoust., Speech, Signal Processing*, vol. 34, pp. 1449–1464, Dec. 1986.

[4] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoust., Speech, Signal Processing*, vol. 34, pp. 744–754, Aug. 1986.

[5] M. R. Portnoff, "Short-time Fourier analysis of sampled speech," *IEEE Transactions on Acoust., Speech, Signal Processing*, vol. 29, pp. 364–373, June 1981.

[6] B. M. Sylvestre, *Time-Scale Modification: a Time-Frequency Approach*, M. Eng. Thesis, McGill University, April 1991.