A NEW FREQUENCY DOMAIN APPROACH TO TIME-SCALE EXPANSION OF AUDIO SIGNALS

Aníbal J. S. Ferreira

Department of Electrical and Computer Engineering
Faculty of Engineering, The University of Porto, Portugal
E-mail: ajf@inescn.pt

ABSTRACT

We present a new algorithm for *time-scale expansion* of audio signals that comprises: time interpolation, *frequency-scale expansion* and modification of a spectral representation of the signal. The algorithm relies on an accurate model of signal analysis and synthesis, and was constrained to a non-iterative modification of the magnitudes and the *wrapped* phases of the relevant sinusoidal components of the signal. The structure of the algorithm is described and its performance is illustrated. A few examples of time-expanded wideband speech can be found on the Internet.

1. INTRODUCTION

Time-scale modification of audio signals is a desired functionality in many applications including for example, non-linear audio editing. Ideally, only the presentation rate of the audio material should be modified without affecting its intelligibility or its quality. Time domain techniques as well as frequency domain techniques have been used to modify the time-scale of audio signals.

Time-domain techniques generate a new signal by concatenating time frames taken from the original sound either in an overlapped fashion (corresponding to time-scale expansion) or in a discontinuous fashion (corresponding to time-scale compression). The concatenation is performed so as to avoid audible artifacts associated to amplitude or phase discontinuities. To prevent these, some small overlap between adjacent frames is usually enforced in which fading and alignment techniques are used. Therefore, under this class of algorithms, methods for time-scale modification differ mainly on the fading criterion, or on the correlation or similarity metrics underlying the alignment criterion.

Frequency domain techniques involve the modification of a spectral representation of the audio signal, which is combined with decimation and/or interpolation operations in order to synthesize the rate modified audio signal. A classic reference is the work developed by Portnoff [1] which comprises phase unwrapping, a model of speech production, and a signal analysis and synthesis scheme based on the Short Time Fourier Transform (STFT). The coefficients of the STFT are related to the parameters of the speech model such that their manipulation leads to the synthesis of rate-changed speech. A large number of other examples could be given such as a technique based on the phase vocoder [2], or a technique developped by Quatieri *et al.* [3], which is based on a 21 band perfect reconstruction filter bank.

Our technique assumes that the audio signal can be represented as a sum of quasi-stationary sinusoids. It relies on an accurate spectral manipulation, which accounts for the specific nature of the filter bank and the shape of the analysis/synthesis time windows, but does not assume a specific source production model. The technique consist in a new approach in the sense that it achieves time-scale expansion through a *spectral expansion*, in the original time-scale, of the spectrum modified in amplitude *and* phase. An independent time-interpolation operation then leads to the time expanded signal which preserves the original pitch and timbre. Other innovative aspects are that the technique accounts explicitly for temporal modulation effects, and uses a new method to estimate the phase and the center frequency of a sinusoidal component beyond the frequency resolution of the analysis/synthesis filter bank [4].

The algorithm was constrained to meet two severe requirements:

1. the algorithm should be integrated within a *perceptual audio coder* (ASC [5]) and therefore should share its analysis/synthesis filter bank,
2. the algorithm should perform all relevant spectral manipulations non-iteratively, on a single frame basis, and without the possibility to look into the past, nor into the future of the signal.

An obvious consequence of these requirement is that they preclude phase unwrapping, which means in practice that only time-scale *expansion* by integer factors is allowed [4].

The paper is organized as follows. In section 2 we present the main principles underlying our approach. We explain in section 3 the relevant properties of the analysis/synthesis filter bank allowing convenient spectral modifications. The main structure of the frequency-scaling and spectral modification algorithm is described in section 4, and its performance is illustrated, with a few examples, in section 5.

2. THE CENTRAL IDEA

The algorithm achieves time-scale expansion through a three step operation as illustrated in Figure 1.

This work was supported by the Portuguese Research Program PRAXIS XXI under research contract 2/2.1/TIT/1644/95.
The author is also affiliated with INESC, Largo Mompilher 22, 4000 PORTO PORTUGAL.

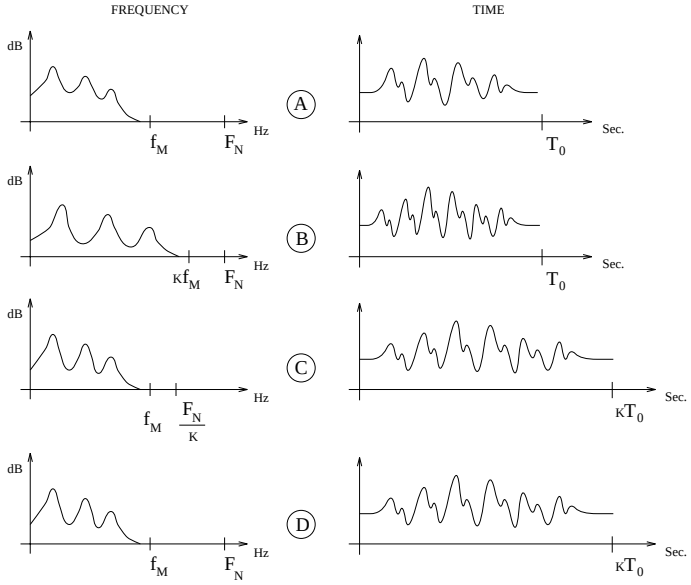

Figure 1: A frequency domain approach to time-scale expansion: (A) original signal, (B) spectral modification, (C) same as B with sampling frequency $\frac{2}{K}F_N$ instead of $2F_N$, (D) same as C after interpolation by K.

The signal represented in (A) is assumed to have its significant spectral content limited to f_M below the Nyquist frequency F_N. The first step consists in a spectral modification (in amplitude and phase) followed by expansion along the frequency axis (not along the time axis as in [1] and [3]) by the desired time expansion factor, K, as illustrated in step (B). This spectral expansion is a delicate operation because the frequency and phase relationships among the most relevant signal components must be preserved in order to keep the fine temporal structure of the signal.

If, as represented in (C), the sampling frequency of the signal is changed to $\frac{2}{K}F_N$ (resampling), the pitch of the original signal is maintained and the time-scale is expanded by K. In order to recover the original sampling frequency, the signal must be interpolated by K, as represented in (D).

As results clear from Figure 1, the interpolation should be performed before the frequency manipulation in order to avoid reducing the signal bandwidth by the time-scale expansion factor.

3. PROPERTIES OF THE ODD-DFT BASED ANALYSIS/SYNTHESIS SYSTEM

The spectral expansion process uses properties of the analysis/synthesis scheme of ASC. This scheme can be reduced to a 50% overlap signal analysis according to a time window $h(n)$ of length N and a N point Odd-DFT [6] direct transformation (1), a N point Odd-DFT inverse transformation (2), and a 50% overlap-and-add synthesis (3) using the same time window $h(n)$. The block index m identifies the position of successive transformations.

$$X(k,m) = \sum_{n=0}^{N-1} h(n)x(n - m\frac{N}{2})e^{-j\frac{2\pi}{N}(k+\frac{1}{2})n} \qquad (1)$$

$$\hat{x}(n,m) = \frac{1}{N}\sum_{k=0}^{N-1} X(k,m)e^{j\frac{2\pi}{N}(k+\frac{1}{2})n} \qquad (2)$$

$$\tilde{x}(n) = \sum_{m=-\infty}^{\infty} h(n - m\frac{N}{2})\hat{x}(n,m) \qquad (3)$$

The window $h(n)$ should provide perfect reconstruction in the absence of spectral modification [7], *i.e.* $\tilde{x}(n) = x(n-d)$, where d is the system delay. It shoud also provide good spectral selectivity and good attenuation of time aliasing effects due to spectral sub-sampling and filtering. A simple and yet satisfactory solution given the above requirements can be derived from the Hanning window [4]:

$$h(n) = \sin\frac{\pi}{N}(n+\frac{1}{2}) \ , \quad 0 \le n \le N-1 \ . \qquad (4)$$

This window is adequate to implement the spectral analysis and modification processes since, in addition, it is analytically tractable. For example, the main lobe of its frequency response can be approximated by:

$$|\widehat{H(\omega)}|_{dB} = G\log\left[\cos\frac{N}{6}\omega\right] \ , \quad |\omega| < \frac{3\pi}{N} \ , \qquad (5)$$

where G is a gain parameter and, except for a constant additive term, the normalized magnitude in dB of the envelope of the frequency response, for $|\omega| > \frac{3\pi}{N}$, is given by:

$$|\widehat{H(\omega)}|_{dB} = 20\log\left|\frac{1}{\sin\frac{1}{2}\left(\frac{\pi}{N}-\omega\right)} + \frac{1}{\sin\frac{1}{2}\left(\frac{\pi}{N}+\omega\right)}\right| \ . \qquad (6)$$

When (4) is combined with the Odd-DFT filter bank, a number of interesting results are obtained:

- any DC component of the input signal, is compacted on the first spectral line (*i.e.* at $k=0$) [7],
- if $x(n)$ is a pure tonal signal whose frequency is $\frac{2\pi}{N}(\ell+\Delta\ell)$, where ℓ is an integer number in the range $1 \le \ell \le \frac{N}{2}-1$ and $\Delta\ell$ is a real number in the range $0.0 \le \Delta\ell < 1.0$, then the frequency line at $k=\ell$ of the Odd-DFT corresponds to the strongest spectral line, and any existing fixed phase ϕ of the tonal signal can be recovered from the phase at $k=\ell-1$ after adding $\frac{\pi}{2N}$ and an offset approximated by $\Delta\ell\pi$ [4]; moreover, when $\Delta\ell = 0$, the difference between the phase at $k=\ell$ and the phase at $k=\ell-1$ is exactly $\pi(\frac{1}{N}-1)$, regardless of ϕ,
- in the case of time modulated tones, the temporal envelope is approximately maintained even if the tones are displaced in frequency, by keeping the relative phase relationships among the neighboring spectral coefficients, and provided that the temporal modulation is relatively smooth when compared to the duration of the analysis/synthesis window.

A detailed illustration of these properties is provided in [4].

The spectral expansion implied in step (B) of Figure 1 is accompanied by a modification of both the frequency and the fixed initial phase of each relevant tonal component of the quasi-stationary input signal, according to the time expansion factor. Given that the discrete transform gives access only to a finite number of sampled frequencies, it is necessary to rely on an accurate estimation of both the integer

frequency, $\frac{2\pi}{N}\ell$, and the fractional frequency $\frac{2\pi}{N}\Delta\ell$ of each relevant tonal component. Although several methods of frequency estimation could be used such as the "quadratic fit" or the "phase vocoder" [8], we developed an alternative method that is matched to the peculiar combination of the chosen analysis/synthesis window with the Odd-DFT [4].

4. SPECTRAL EXPANSION OF NARROW AND WIDEBAND SIGNAL COMPONENTS

Both narrowband and wideband components of an audio signal keep their bandwidth when subjected to magnitude and phase changes during the frequency expansion process, in order to preserve the original pitch and the fine temporal structure. The major algorithmic steps are summarized below (the block index m is omitted for clarity).

1. Following the time-frequency transformation, all spectral peaks are found using a peak detector similar to the one used in the Psychoacoustic Model 1 of the MPEG-Audio Standard. These peaks are further decimated in order to avoid low level variations and in order to retain the most *relevant* signal components in a *psychoacoustic* sense [9]. The location of each relevant spectral peak (local maximum) as well as the location of the center of each spectral valley (local minimum) are stored.
2. The center frequency of each local maximum is obtained by estimating accurately the ℓ and $\Delta\ell$ of all relevant signal components.
3. As the time expansion factor, K, is specified and the original frequency position $\ell_a + \Delta\ell_a$ of each relevant signal component is known, the new frequency position $\ell_b + \Delta\ell_b$ is easily obtained.
4. The bandwidth of each spectral component is found using (7) if $\Delta\ell > 0.5$ or (8), otherwise.

$$W = \max\left\{6.0, \frac{\pi(2\Delta\ell+1)}{\arccos\left(10^{\frac{X(\ell-1)_{dB}-X(\ell)_{dB}}{30}}\right)}\right\} \quad (7)$$

$$W = \max\left\{6.0, \frac{\pi(3-2\Delta\ell)}{\arccos\left(10^{\frac{X(\ell+1)_{dB}-X(\ell)_{dB}}{30}}\right)}\right\} \quad (8)$$

5. The Power Spectral Density at the center frequency $(\ell_a+\Delta\ell_a)$ of each signal component is estimated using its bandwidth, at the closest discrete frequency:

$$P_{MAX} = X(\ell_a)_{dB} - 30\log\left[\cos\frac{\pi}{W}(2\Delta\ell_a - 1)\right] . \quad (9)$$

6. The fixed phase, ϕ, for each signal component is estimated by removing the constant phase contributions:

$$\phi_a = \phi(\ell_a - 1) + \pi(\frac{2}{N} - \Delta\ell_a) . \quad (10)$$

This estimation assumes that any time modulation is sufficiently smooth so that, as a consequence, the phase at $k = \ell_a - 1$ is not modified substantially.

7. New magnitudes are synthesized at $k_L = \ell_b - 1$, $k_0 = \ell_b$ and at $k_R = \ell_b + 1$ as given respectively by:

$$\hat{X}(k_L)_{dB} = P_{MAX} - 30\log\left[\cos\frac{\pi}{W}(2\Delta\ell_b + 1)\right] \quad (11)$$

$$\hat{X}(k_0)_{dB} = P_{MAX} - 30\log\left[\cos\frac{\pi}{W}(2\Delta\ell_b - 1)\right] \quad (12)$$

$$\hat{X}(k_R)_{dB} = P_{MAX} - 30\log\left[\cos\frac{\pi}{W}(2\Delta\ell_b - 3)\right] \quad (13)$$

The new magnitudes from $k_{Rb} = \ell_b + 2$ to the next local minimum on the right, and from $k_{Lb} = \ell_b - 2$ to the next local minimum on the left, are synthesized by keeping the original differential magnitudes in dB:

$$\hat{X}(k_{Rb}) = X(k_{Ra}) + \hat{X}(\ell_b + 1) - X(\ell_a + 1) ,$$

$$\hat{X}(k_{Lb}) = X(k_{La}) + \hat{X}(\ell_b - 1) - X(\ell_a - 1) .$$

The magnitude of each local minimum is further modified using (6) and the spectral distance to one of the surrounding maxima, whichever produces a predominant effect.

8. The phase of each component expanded in frequency is synthesized by reconstructing first the phase at $k = \ell_b - 1$ as follows:

$$\hat{\phi}(\ell_b - 1) = K\phi_a - \pi(\frac{2}{N} - \Delta\ell_b) . \quad (14)$$

The phases from $k_b = \ell_b$ to the next local minimum on the right and from $k_b = \ell_b - 2$ to the next local minimum on the left, are synthesized by maintaining the original differential phases:

$$\hat{\phi}(k_b) = \phi(k_a) + \hat{\phi}(\ell_b - 1) - \phi(\ell_a - 1) .$$

9. The above spectral expansion and modification process generates spectral holes which are avoided by interpolating the magnitudes and phases of all coefficients of the spectral hole. The interpolation of phases was preferred to a simple randomization in order to avoid critical phase transitions such as $-\pi$.

5. A FEW RESULTS

We can make an objective assessment of the quality of the whole spectral modification by comparing the output of the algorithm at step (B), with the ideal output signal, since the final stages of resampling and time interpolation do *not* affect the quality of the signal. The ideal output signal can be easily generated by synthesizing the desired spectral modification subjected to the time envelope of the original signal.

Figure 2 illustrates the response of the algorithm (bottom) to a stationary and harmonic input signal (top) when the scale expansion factor is 2, the sampling frequency is 48 KHz, and $N = 1024$. The signal in the middle of Figure 2 is the signal ideally expanded in frequency by 2. The input signal comprises four tones, harmonically related, whose fundamental frequency corresponds to $(\ell + \Delta\ell) = 12.3$. In

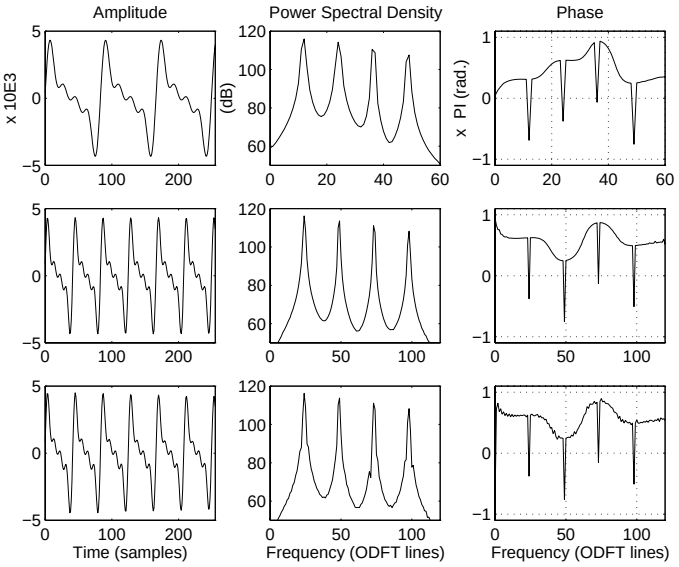

Figure 2: Frequency expansion of a four tone harmonic signal by a factor of two. Top: original signal. Middle: synthesized signal. Bottom: processed signal.

all cases, the spectrogram corresponds to a 1024-point time-frequency Odd-DFT transformation based on the sine window (4). For clarity only one fourth of the time segment is displayed. The fixed initial phases of all four tones of the input signal are zero.

It can be concluded that as the spectral separation between the tones of the expanded version is larger than in the case of the input signal, the spectral valleys reach lower PSD values. Secondly, as the $\Delta\ell$ of the second and fourth harmonics of the input signal are equal to the $\Delta\ell$ of the first and second harmonics of the ideally expanded version, respectively, their phases are also equal as results from (14). For the same reason, the magnitude shape of these spectral peaks is also similar as suggested by (11), (12) and (13).

It can also be seen that the fine temporal structure of the original signal is maintained. This example assumes signal stationarity and absence of temporal modulation.

Other more realistic examples were also tried. Table 1 presents the Signal-to-Noise Ratios (SNR) expressing the error in the time domain between the ideal output signal and the actual output signal for four different test signals. The first signal corresponds to the signal illustrated in Figure 2. The second signal consists in the fundamental of the previous signal modulated linearly at rise and fall during 50 ms. The third signal corresponds to a similar modulation that lasts only (rise and fall) 5 ms (*tonal pulse*). The fourth sinal is a sine sweep whose frequency increases from 100 Hz to 1 KHz in 40 ms. In all cases the duration of the signal corresponds to four overlap-and-add analysis/synthesis operations. The SNR figures in Table 1 should be interpreted qualitatively rather than quantitatively since the ideal and output signals sound indistinguishable except for the sine sweep.

Not surprisingly, the best score is obtained by the harmonic complex tone because it is a stationary signal and

Table 1: Qualitative evaluation of the time-scaling algorithm for four increasingly difficult signals.

item	harmonic	mod. tone	pulse	sweep
SNR (dB)	28.4	15.4	14.6	1.6

this condition minimizes the errors introduced by the algorithm. In all other cases, the deviations from stationarity get more and more aggressive, particularly because the duration of the strong changes is comparable or less than the size of the transform (N). The low SNR in the case of the sine sweep just reflects the fact that due to the nature of the analysis/synthesis filter bank, the algorithm has an inherent difficulty in dealing with frequency modulated signals.

In the case of *very* wideband signals such as impulsive noise, and sound attacks, the algorithm fragments the wideband character of the signals during the process of spectral expansion, which makes transient smearing audible.

In the case of wideband speech, the time-expanded signals sound free from obvious artifacts and preserve the speaker dependent features, even when the expansion factor is as large as 4. In fact, the naturalness of the unvoiced regions as well as of the original pitch of the speech are absolutely maintained, despite a few subtle distortions arising when the filter bank does not have enough frequency resolution to resolve the reciprocal of the pitch period. A few examples of time-scaled wideband speech are available on the Internet (http://www.inescn.pt/~ajf/timescale.html).

6. REFERENCES

[1] Michael R. Portnoff, "Time-Scale Modification of Speech based on Short Time Fourier Analysis", *IEEE Trans. on Acoust., Speech and Signal Proc.*, vol. 29, no. 3, pp. 374–390, June 1981.

[2] Mark Dolson, "The Phase Vocoder: A Tutorial", *Journal of Computer Music*, vol. 84, no. 4, pp. 14–27, 1986.

[3] T. F. Quatieri, R. B. Dunn, and T. E. Hanna, "A Subband Approach to Time-Scale Expansion of Complex Acoustic Signals", *IEEE Trans. on Speech and Audio Proc.*, vol. 3, no. 6, pp. 515–519, November 1995.

[4] Aníbal J. S. Ferreira, "An Odd-DFT Based Approach to Time-Scale Expansion of Audio Signals", *submitted to* IEEE Trans. on Speech and Audio Proc.

[5] Aníbal J. S. Ferreira, "Audio Spectral Coder", *100th Convention of the Audio Engineering Society*, May 1996, Preprint n. 4201.

[6] Maurice Bellanger, *Digital Processing of Signals*, John Willey & Sons, 1989.

[7] Henrique Malvar, *Signal Processing with Lapped Transforms*, Artech House, Inc., 1992.

[8] Judith C. Brown, "Frequency Ratios of Spectral Components of Musical Sounds", *Journal of the Acous. Soc. of America*, vol. 99, no. 2, pp. 1210–1218, February 1996.

[9] Brian C. J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, 1989.