SPEECH ANALYSIS WITH THE FAST CHIRP TRANSFORM

Luis Weruaga and *Marián Képesi*

Austrian Academy of Sciences, Donau-City Strasse 1, A-1220 Vienna, Austria.
luis.weruaga@oeaw.ac.at

*Vienna Telecommunications Research Center, Donau-City Strasse 1, A-1220 Vienna, Austria.
kepesi@ftw.at

ABSTRACT

The Chirp transform is a powerful analysis tool for variable frequency signals such as speech. The computational load represents the main limitation of its original formulation, discouraging its use in real-time applications. This paper analyzes a fast implementation, based on performing time-warping on the signal under analysis, combined with the Fast Fourier Transform. The performance of the Fast Chirp transform depends on the one hand on the estimation of the time-warping operation based on the signal characteristics, and, on the other hand on the interpolation technique used for the warping. Observations from the analysis of speech signals support the method and the further lines.

1. INTRODUCTION

A fine, precise representation of non-stationary signals in the time-frequency domain is of great importance in many fields [1]. In the last few years there has been an interest in the use of time-varying signals as the analysis basis: the so-called adaptive Chirplet transform [2], warped wavelets [3], a piecewise Fourier-based fast Chirp transform [4], etc. The adaptation of this basis according to the time-frequency characteristics of speech signals is a research avenue that has not yet been fully explored.

In [5] an adaptive transform for analyzing speech signals was proposed. The so-called Short-time Chirp transform (STChT) achieves higher time-frequency representation than the Fourier analysis; the signal can be synthesized from its STChT, and thus it is suitable for analysis and filtering. The performance of this transform lies on an cost-efficient automatic estimation of the chirp rate, the only parameter required to construct the analysis basis. However, the prohibitive computational load does not encourage its use in real-time applications.

In this paper several variants of a fast implementation of the Chirp transform are analyzed. This method is based on combining time-warping with the Fourier transform, a technique that has already been proposed for a general framework in [3]. Nevertheless, this last method has not been very popular, and time-frequency methods such as Wigner-Ville distribution [6] and the Wavelet transform [7] have attracted more attention. The analysis proposed here, with several interpolation techniques along with the proposed automatic chirp rate prediction technique from [5] represents a competitive techniques for high performance time-frequency analysis of speech in real-time.

The paper is structured as follows: section 2 presents the Chirp transform and its fast implementation based on time-warping, section 3 deals with the practical aspects of the implementation, and in section 4 the different variants are compared on speech signals.

2. THE CHIRP TRANSFORM

2.1 Preliminary

The analysis and synthesis equations of the Chirp transform of a signal $x(t)$ defined in the interval $[0,\, T]$ are

$$X_\phi(f) = \frac{1}{T}\int_0^T \phi'(t)\,x(t)\,e^{-j2\pi f\phi(t)}\,dt \tag{1}$$

$$x(t) = \sum_{k=-\infty}^{\infty} X_\phi\left(\tfrac{k}{T}\right)\,e^{j2\pi\frac{k}{T}\phi(t)}, \tag{2}$$

where $\phi(t)$ is a warping-time function defined in the interval $[0,\, T]$. This function $\phi(t)$ fulfills the following conditions:

- it is biunivocal and increases monotonically in $[0,\, T]$,
- its derivative, $\phi'(t)$ is continuous in $[0,\, T]$,
- $\phi(0) = 0$ and $\phi(T) = T$.

In spite of the quasi-orthogonality of the chirp basis, perfect reconstruction from the Chirp transform is possible [8], and, thus (1) (2) describe the direct and inverse transform.

If the warping function is defined as

$$\phi(t) = (1 + \alpha\,(t - T))\,t, \tag{3}$$

the resulting analysis/synthesis basis consists of quadratic chirps or linearly frequency-modulated sinusoids, where α is the chirp rate. In order for the quadratic chirps to fulfill the requirements exposed previously, the range of α is to be confined in the following interval

$$-\frac{1}{T} < \alpha < \frac{1}{T}. \tag{4}$$

If $\phi(t) = t$ then (1) (2) represents the trivial case of the analysis and synthesis equations of the Fourier series.

2.2 Fourier-based Implementation

Evaluation of (1) or (2) in a discrete domain requires $O(N^2)$ operations, N being the number of samples of the signal in $[0,\, T]$ This prohibitive computational load does not encourage the use of that transform in real-time applications.

With the following change in the variable of the integral

$$\tau = \phi(t) \tag{5}$$

(1) can be expressed as a Fourier transform as follows

$$X_\phi(f) = \frac{1}{T}\int_0^T x(\varphi(\tau))\,e^{-j2\pi f\tau}\,d\tau, \tag{6}$$

where $\varphi(t)$ is the inverse function of $\phi(t)$, that is, $\phi(\varphi(t)) = t$, and inherits the same properties of monotonicity and continuity.

The pair $\{\phi(t), \varphi(t)\}$ is the so-called "warping" pair. The signal within the Fourier transform (6) is a warped version of the original $x(t)$. Therefore, the fast implementation of the Chirp transform of signal $x(t)$ is composed of a time-warping process followed by a Fourier transform,

$$x(t) \xrightarrow{\varphi(t)} y(t) = x(\varphi(t)) \tag{7}$$

$$X_\phi(f) = \frac{1}{T}\int_0^T y(t)\, e^{-j2\pi ft}\, dt, \tag{8}$$

and the inverse fast Chirp transform is accordingly

$$y(t) = \sum_{k=-\infty}^{\infty} X_\phi\left(\tfrac{k}{T}\right) e^{j2\pi \frac{k}{T}t}, \tag{9}$$

$$y(t) \xrightarrow{\phi(t)} x(t) = y(\phi(t)) \tag{10}$$

3. DISCRETE-TIME IMPLEMENTATION

Since t is continuous time, the time-warping transformations (7) and (10) do not present the resampling dilemma, and no loss of information occurs (provided that $\phi(t)$ fulfills the conditions listed in Section 2.1). Let us now consider $x[n]$ as a N-sample long segment of $x(t)$ sampled at a Nyquist rate f_s ($N = Tfs$). The translation of the Fourier-based implementation of the Chirp transform to the discrete-time domain is in a first approach

$$x[n] \xrightarrow{\varphi(\cdot)} y[n] = x[f_s\, \varphi(n/f_s)], \tag{11}$$

$$X_\phi[k] = \mathrm{DFT}_N\{y[n]\}, \tag{12}$$

and the translation of the inverse process is

$$y[n] = \mathrm{IDFT}_N\{X_\phi[k]\}, \tag{13}$$

$$y[n] \xrightarrow{\phi(\cdot)} x[n] = y[f_s\, \phi(n/f_s)]. \tag{14}$$

The time-warping operations (11) and (14) present clear drawbacks: firstly the resampling is non-uniform, which means that in some zones the new sample interspace will be larger than the original sampling period, this violating the Nyquist criterion, and secondly, the value $f_s\, \varphi(n/f_s)$ does not have to be integer, and thus (11) and (14) imply an additional computing effort for performing the required inter-sample interpolation.

3.1 Non-uniform sampling

The analysis to determine the time-frequency loss arising from the resampling (11) and (14) is difficult. The resampling becomes lossy once the inter-sample space goes beyond the sampling period. Intuitively if the derivative of the warping pair meets either $\phi'(t) < 1$ or $\varphi'(t) < 1$, a degradation around time t takes place after converting the signal to the discrete domain. But the previous conditions are always fulfilled (except in the trivial case $\phi(t) = t$). One could also argue that the condition $\varphi'(t) < 1$ is the one that determines the loss, because the loss caused by ϕ-warping occurs only in areas of oversampling (produced by φ-warping).

By looking at the initial equations (1) and (2), a better understanding of this lossy procedure arises. Conversion of (1) to a discrete time domain is

$$X_\phi[k] = \frac{1}{T}\sum_{n=0}^{N-1} \phi'(nT_s)\, x[n]\, e^{-j\frac{2\pi}{N}k\phi(nT_s)}, \tag{15}$$

where T_s is the sampling period, and $k \in \{-K, ..., K\}$ is the frequency index that is limited by a certain integer $K \leq \frac{N}{2}$. In (15) no time-warping process takes place, and $x[n]$ is present here at no distortion. Also the windowing term is harmless since its effect can be compensated further by an inverse windowing (operation intrinsic in (2)). The chirp basis is then the major cause of the information loss. The derivative of the exponent of the continuous-time basis gives the instantaneous frequency, which is

$$f(nT_s) = \frac{k}{N} f_s\, \phi'(nT_s). \tag{16}$$

at uniformly-spaced sampling instants. According to Nyquist criterion, $f(nT_s)$ should not go further than $f_s/2$. This sets the range of the frequency index to be

$$|k| < \frac{N}{2} \max\left\{|\phi'(t)|\right\}^{-1} \tag{17}$$

Condition (17) set the highest frequency bin K of the basis so that no spectral overlapping occurs. Since $\phi'(t) = \varphi'(t)^{-1}$, condition (17) is in agreement with the intuitive condition $\varphi'(t) < 1$ to be the source of information loss. Nevertheless, (17) gives the solution to at least mitigate that loss without introducing spectral artifacts in the process.

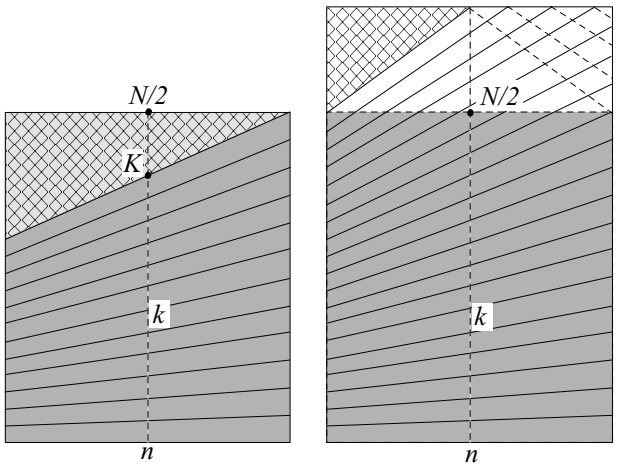
3.2 Lossless variant

After the time-resampling (11), the DFT computation in (12) presents artifacts above a certain frequency index K. There are ways to implement the fast chirp transform without spectral aliasing and even lossless:

1. The simplest solution to avoid that spectral aliasing is not to count with the frequency indices beyond index K. This idea is shown graphically in Figure 1.a): the meshed area corresponds to the time-frequency domain that must be disregarded.
2. Another possibility is to carry out the fast chirp implementation on an oversampled signal and consider all the frequency bins that cover the useful time-frequency space. This allows a lossless implementation as long as the spectral aliasing takes place only on blank time-frequency regions (above $f_s/2$). Figure 1.b) shows this case graphically (the white area represents the blank part of the time-frequency space, result of the oversampling). This option increases the computing load.

4. RESULTS

In order to validate the ideas of the previous sections in a practical framework, we performed the analysis and synthesis on speech signals (telephony quality) with different warping techniques combined with oversampling. The methods compared are the Short-time Fourier (STFT) and the Short-time Chirp Transforms [5] (STChT). The STChT is based on a similar procedure as the STFT, with the exception that the



a) b)

Figure 1: Time-frequency coding by a quadratic chirp basis: a) lossy solution. components that do not introduce spectral aliasing are considered only; b) oversampled signal in order to code the relevant time-frequency space.

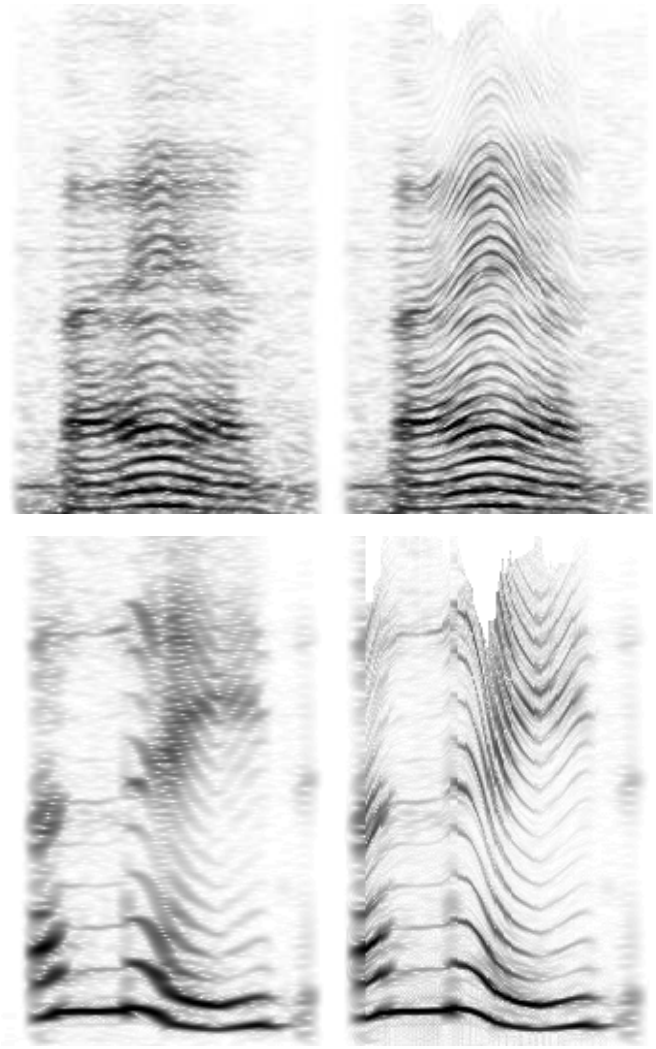
analysis / synthesis technique is either the Chirp or the Fast Chirp transform (instead of the Fourier one).

The warping function $\phi(t)$ from the Chirp transform used in the experiments is quadratic (3), and thus the chirp basis is composed of linearly frequency-modulated sinusoids (as also depicted in the time-frequency representation in Fig. 1). In order to compute the chirp rate α for each segment the automatic method described in appendix A was used (the chirp rate is computed based on a the temporal tracking of the pitch of voiced sounds). In all methods the Hamming window is used, and the window length is set to $N = 512$.

A visual comparison between the STFT and the STChT (Fast) is shown in Figure 2. The upper images come from a male speaker pronouncing the word /bad/, and the lower images from a female saying /a lathe/. In both cases the intonation of the speakers is natural (the male pitch fluctuates 20%, and the female 85%). The left column contains the classical spectrogram, and the right one the chirp-based spectrogram. In case of pitch variation the chirp analysis clearly outperforms the Fourier-based approach. Noticeable is that the option chosen in both cases was not to consider frequency bins higher than K (the blank areas at high frequencies) when the pitch variation is dominant.

In order to evaluate the losses of the warping operation in the fast Chirp transform, several variants were considered. The details of the methods are summarized in Table 1. Two output parameters define the comparative study: the signal-to-error ratio (SER) after resynthesis, and the computational load of each method versus the STFT.

Comparison between the STFT (method M.1) and the STChT (M.2 and M.3) reveals the STChT as a precise analysis and synthesis technique. Comparison between methods M.2 and M.6 suggests the use of its fast implementation. Comparison between M.2 and M.3 proves that oversampling the signal preserves the time-frequency space at high frequencies (comparison between M.7 and M.5-6 gives also the same conclusion). Comparison between the STChT and its fast versions, shows the expected dramatic decrease



a) b)

Figure 2: Fourier and Chirp transform comparative: upper images - male voice; lower images - female voice; a) STFT; b) STChT.

in computational load, but reveals degradation after resynthesis, even when high frequencies are preserved (M.3 vs. M.7). This fact is explainable by the interpolation techniques considered (linear and cubic), leading to additional loss distributed throughout the frequency domain (as can be deduced by analyzing the spectrogram of the error). Thus, our current research efforts are directed towards designing more accurate time-warping techniques.

5. CONCLUSIONS

This paper analyzes the implementation of the Fast Chirp transform as a combination of time-warping and the Fast Fourier transform on speech signals. The study of the original and the fast formulation together allows one to address the problems of its practical implementation. The time-warping operation may introduce two types of signal degradation: the first one from the spectral aliasing of the higher components of the chirp basis, and the second one from the type of interpolation used in the resampling. The results ob-

Method	I	H	R	SER	Cost
1. STFT	- -	$N/2$	1	77	1.0
2. STChT	- -	$N/2$	1	51	711
3. STChT	- -	$> N/2$	2	61	- -
4. STChT Fast	1	K	1	22	13
5. STChT Fast	3	K	1	28	14
6. STChT Fast	3	$N/2$	1	29	14
7. STChT Fast	3	$> N/2$	2	34	35

Table 1: Comparison of STCT and STChT variants on the female speech. The column labels represent: I - order of the interpolation method (1-linear, 3-cubic Hermite); H - highest frequency bin index; R - oversampling ratio; SER - ratio in dBs between the signal energy and that of the residual error; Cost - ratio of the computing time between the method vs. STFT (the results are based on a Matlab implementation including the chirp rate estimation algorithm).

tained on pitch-varying speech signals support the use of the fast Chirp transform for time-frequency speech analysis and resynthesis. In order to achieve a more accurate implementation, better interpolation techniques are required.

A. CHIRP-RATE ESTIMATION

The objective of this part is to track the pitch evolution of voiced segments, and thereupon to estimate the chirp rate α_m from the previous result. The first step is to compute the pitch of the $(m-1)$-th segment. Let $\rho_{m-1}[k]$ be the $(m-1)$-th spectral energy frame, that is, $\rho_{m-1}[k] = |X_{m-1}[k]|^2$, and let $f_o[m-2]$ be the estimated pitch of the previous frame. The method to estimate $f_o[m-1]$ is based on the mean-shift algorithm[1] with additional constraints: by adding successively the center of masses of higher speech harmonics an accurate final pitch value is obtained. The method is described in Algorithm 1.

Algorithm 1 Pitch estimation of the $(m-1)$-th frame

$f = f_o[m-2]$
for $h = 1$ to *number_of_harmonics* **do**
  for $i = 1$ to h **do**
    $\mu_i = \sum_{|k-if|<\frac{f}{2}} \rho_{m-1}[k]$
    $f_i = \mu_i^{-1} \sum_{|k-if|<\frac{f}{2}} k\,\rho_{m-1}[k]$
    $N_i = (f-f_w)^{-1} \sum_{\frac{f_w}{2}<|k-f_i|<\frac{f}{2}} \rho_{m-1}[k]$
    $S_i = \max(\rho_{m-1}[f_i] - N_i, 0)$
    $w_i = \frac{Si}{N_i} \rightarrow \hat{w}_i = i\,w_i$
  end for
  $f = \sum_j (\hat{w}_j f_j / j) / \sum_j \hat{w}_j$
end for
$f_o[m-1] = f$
$\lambda[m-1] = 10 \log_{10} \left(\sum_j (w_j S_j) / \sum_j w_j \right)$

[1] Compute the mean of the data distribution within the window and translate the center of the window to this mean point; repeat the process till this translation becomes depreciable.

The internal parameters μ_i, f_i, N_i, S_i, and w_i, $\hat{w}_i$ of the algorithm are respectively the spectral mass, the estimated frequency, the noise energy and the harmonic energy of the i-th harmonic component, and two additional SNR-based parameters that weight the importance of this harmonic when computing the output parameters λ and f_o. The estimated pitch is computed as the weighted linear combination of the frequency of each harmonic component f_i divided by its index i. The estimated noise N_i is computed as the sum of the spectral mass $f_w/2$ bins farther to the central harmonic position (f_w is the spectral resolution of the analysis window). The net energy S_i is computed with a spectral substraction approach. Output parameter λ is the confidence value containing the net energy that is actually present in the harmonic: this parameter serves as a reliable indication of whether the analyzed speech segment is voiced or unvoiced.

The frequency variation rate is obtained from the pitch values of the last L segments, $f_o[m-L], ..., f_o[m-1]$, by fitting a straight line on them (let the slope of the resulting straight line be η), and applying the following operation

$$\alpha_m = \begin{cases} \dfrac{\eta/M}{f_o[m-1]+\eta} & \text{if } \lambda[m-1] > \lambda_s, \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

Parameter λ_s is the voiced/unvoiced decision threshold and M is the time step between consecutive spectral frames.

In the computation of the frequency variation rate (18) the numerator contains the slope per sample, and ideally the denominator should contain the pitch of the m-th segment; since this segment is not available yet, a linearly-predicted estimation is used instead. When computing η, only reliable pitch values are used, that is, only those that correspond to a voiced segment ($\lambda > \lambda_s$).

REFERENCES

[1] P. Flandrin, *Time–Frequency / Time–Scale Analysis*, Academic Press, San Diego, 1999.

[2] S. Mann and S. Haykin, "Adaptive 'Chirplet' Transform: an adaptive generalization of the wavelet transform," *Optical Engineering*, 31(6), pp. 1243-1256, 1992.

[3] R.G. Baraniuk and D.L. Jones, "Warped wavelet bases: unitary equivalence and signal processing," *Proc. IEEE ICASSP*, pp. 320-323, Minneapolis, April 1993.

[4] F.A. Jenet and T.A. Prince, "Detection of variable frequency signals using a fast chirp transform," *Phys. Rev. D.* 62(12), p.122001, 2000.

[5] M. Képesi and L. Weruaga, "Adaptive chirp-based time-frequency analysis of speech signals," *IEEE Trans. Audio and Speech Proc.*, accepted in July. 2003.

[6] *The Wigner distribution - theory and applications in signal processing*, W. Mecklenbräuker, F. Hlawatsch, Eds. Amsterdam, The Netherlands: Elsevier, 1997.

[7] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, Prentice Hall, Englewood Cliffs, NJ, 1995.

[8] L. Weruaga and M. Képesi, "Speech analysis with the short-time chirp transform," *Eurospeech*, Geneve, Sept. 2003.