Speech Analysis with the Short-Time Chirp Transform

Luis Weruaga¹, Marián Képesi²

¹Department of Information and Communication Technologies
Cartagena University of Technology, Spain.
weruaga@ieee.org

²Forschungszentrum Telekommunikation Wien,
Vienna, Austria.
kepesi@ftw.at

Abstract

The most popular time-frequency analysis tool, the Short-Time Fourier Transform, suffers from blurry harmonic representation when voiced speech undergoes changes in pitch. These relatively fast variations lead to inconsistent bins in frequency domain and cannot be accurately described by the Fourier analysis with high resolution both in time and frequency. In this paper a new analysis tool, called Short-Time Chirp Transform is presented, offering more precise time-frequency representation of speech signals. The base of this adaptive transform is composed of quadratic chirps that follow the pitch tendency segment-by-segment. Comparative results between the proposed STCT and popular time-frequency techniques reveal an improvement in time-frequency localization and finer spectral representation. Since the signal can be resynthesized from its STCT, the proposed method is also suitable for filtering purposes.

1. Introduction

A precise multidimensional representation of non-stationary signals is of great importance in many fields of digital signal processing. Signal analysis in the time-frequency domain is still a field of active research, especially in case of speech processing. Challenging problems such as speech enhancement, de-noising, blind separation of speakers, and separation of music instruments require not only sophisticated data-processing algorithms but also fine multidimensional signal representation. Several time-frequency techniques with different properties have been applied on speech and other acoustic signals. Let us briefly review the state of the art in time-frequency analysis, focusing on works mainly used in the field of speech and acoustic signal processing. The three basic, and also most important time-frequency analysis techniques are: the Short-Time Fourier Transform (STFT) [1]-[3], which has a linear nature, the Wigner-Ville Distribution (WVD) [4]-[7] with a quadratic nature, and the Wavelet transform [8]-[11], which uses an analysis base of signals confined in both time and frequency whose time-frequency properties resemble those of the analyzed signal. Even though these techniques are well developed and widely used, we meet serious limitations when applying on speech signals. The STFT provides a poor time-frequency localization of harmonics from speech with a quick varying fundamental frequency (which is the situation in normal conversation). The increase in time resolution can be achieved only by decreasing the frequency resolution, and viceversa. Therefore, if the frequency of the analyzed signal changes during the analysis window, and the output of STFT is a smeared representation, decrease of the analysis window length would lead to lower frequency resolution. Application of the Wavelet transform on speech signals is difficult since there is no optimal fixed base that could be commonly used in speech to achieve good resolution (different bases should be used simultaneously depending on the sound analyzed). Even though the WVD shows good time-frequency localization properties, important details of speech stay hidden, the transform is computationally expensive, and it has no easy inverse transform. In the last few years there has been an interest in the use of time-varying signals as the analysis base. The so-called Chirp transform and Chirplet banks [12]-[17] have been defined. The selection of the appropriate bank according to the time-frequency characteristics of the analyzed signal is a research topic that has not yet been fully explored.

In this paper an analysis technique suitable for pitch varying speech signals based on tracking the trajectory of the mean frequency component is proposed. This technique supports the use of a quadratic-chirp base, which is designed adaptively to match the expected basic components of the speech signal under analysis. The paper is structured in the following way: Section 2 introduces the Short-Time Chirp Transform (STCT), which is defined as a more general formulation of STFT; Section 3 describes its application on speech signals with a pitch-varying formulation of speech; Section 4 presents results comparing the proposed technique against classical techniques and finally conclusions will be presented and close the paper.

2. The Short-Time Chirp Transform

The idea of the STCT is derived from the formulation of the STFT and aims to get better time-frequency representation of signals with linearly changing frequency components. The concept of STCT is presented along with a synthetic example.

2.1. STCT definition

The short-time Fourier transform (STFT) is defined as

\[ S_x(n, k) = \sum_{m=0}^{N-1} x[n + mM] w[m] \exp(-j 2\pi mk/n/N) \]  

where \( x[n] \) and \( w[n] \) are samples of the analyzed discrete signal, resp. the sliding analysis window with a length of \( N \) samples, \( n \) is discrete time, and \( k \) represents discrete frequency.

The sliding window with stepsize \( M \) emphasizes the local frequency properties of the signal and the analysis base is defined by complex exponentials.
When assuming that the instantaneous frequency of the component to represent in detail changes linearly during the analysis interval, a more precise time-frequency representation can be accomplished by using linearly varying complex exponentials, or "chirps" as analysis base in time interval \([0, N - 1]\):

\[
\xi(m, k, \alpha) = \exp(j2\pi km(1 + \alpha (m - N))/N) \tag{2}
\]

where \(\alpha\) is the normalized frequency variation rate.

This new base replaced in (1) gives rise to the Short-Time Chirp Transform (STCT), which could be defined in discrete form as

\[
C_x(n, k, \alpha_n) = \sum_{m=0}^{N-1} x[m + nM] \overline{w[m]} \xi(m, k, \alpha_n)^* \tag{3}
\]

where superscript * denotes complex conjugate and \(\alpha_n\) is the discretized time-variant chirp-rate for the \(n\)-th segment.

The time-frequency resolution of STCT relies on the correct estimation of chirp rate \(\alpha_n\). This parameter is derived from the averaged frequency variation of the main signal components in the \(n\)-th segment, normalized by the mean frequency value and the segment-length.

\[
\alpha_n = \frac{\Delta f_0}{f_0 M} = \frac{2(f_n - f_{n-1})}{M(f_n + f_{n-1})} \tag{4}
\]

Here \(n\) is discrete time-index, \(f_n\) and \(f_{n-1}\) are values of mean frequency \(f_0\) calculated from the actual and previous segments. (4) holds at \(n=0\).

Since linearly varying chirpy signals are often related to quadratic chirps, this transform can be considered the quadratic version of the STCT. Quadratic chirps are nearly orthogonal in the interval of analysis

\[
\frac{1}{T} \sum_{n=0}^{N-1} \xi(n, k_1, \alpha) \overline{\xi(n, k_2, \alpha)}^* =
\frac{1}{T} \sum_{n=0}^{N-1} \xi(n, k_1 - k_2, \alpha) \approx \delta(k_1 - k_2) \tag{5}
\]

where \(\delta(\cdot)\) is a Kronecker delta, so the STCT becomes invertible. The inverse transform is computed by applying the conjugate of the analysis base as synthesis base, followed by either an overlap-add or overlap-save method, analogously to the STFT.

The STCT resembles the standard STFT. Additionally, when applying this transform one should be aware of cross-terms that could appear at highest frequencies due to the aliasing raising from the sweeping base-waves at high chirp-rates \(\alpha\).

### 2.2. A synthetic example

In this section the time frequency representation of the STFT and the proposed STCT of a signal composed of two tones whose frequencies change smoothly and independently on each other will be analyzed. The frequency rate of each tone follows different artificial rules. The upper side of Figure 1 represents the time-frequency representation achieved by the STFT over that signal. As can be seen, the frequency resolution of a particular component is different in each time instant with the level of its frequency variation. The higher the frequency variation, the stronger the smearing of the component is. None of the two tones is represented in detail.

Figure 1: (a) STFT and (b)-(c) STCT time-frequency representations of a composition of two chirpy components, both changing from 0 to 1 Hz in a 1 sec interval.

By applying (3) on the previous signal whereas \(\alpha_n\) is chosen as the variation of the frequency of one of the two tones, a more precise representation is achieved, shown on the two lower images of Figure 1. As can be seen, only one of the components is represented in detail, the one whose frequency variation rule matches at each time the value of the analysis chirp rate, \(a\). The other component is represented even blurrier than in the STFT since its frequency variation rate is more distant to \(\alpha\) than to 0.

### 3. STCT analysis of speech signals

In contrast to the composition analyzed in the previous section, the harmonic structure of speech signal is related to a fundamental frequency, \(f_0\). Variations in the fundamental frequency affect all the harmonic components, the more, the higher the frequency of that harmonic is. This fact produces a harmful effect in the Fourier time-frequency analysis, resulting in poor spectral representation. The upper side of Figure 2 represents the Fourier analysis of a voiced segment of speech cut from a normal conversation, whose fundamental frequency undergoes
the chirp-rate of analyzed segments derived from pitch trajectory. The parameter that builds the STCT is the pitch-rate, which is the first derivative of the pitch trajectory. In case of clean speech this parameter could be easily predicted from the pitch-values of the previous segments. Besides the standard pitch-tracking algorithms (zero-crossing, autocorrelation, spectrum and cepstrum-based methods) new approaches of pitch estimation and pitch tracking have been presented. One of them is the correlogram-based ‘perceptual pitch detector’ [18] based on auditory models. Once known the fundamental frequency for the analyzed segments, the pitch variation rate can be easily computed as the pitch-change among subsequent segments normalized by itself. If the speech is unvoiced, or the pitch value is hard to estimate, the analyzed segments are considered as unvoiced, the frequency variation rate \( \alpha_n \) is set to zero and the STCT resembles the standard STFT.

In the examples presented in this paper none of the automatic pitch-tracking algorithms was used. Instead we developed an interactive graphical interface to ‘manually’ track the pitch by inspecting the spectrogram. The reason for using this tool was the need for controlled errorless pitch-trajectory estimation. This tool created in Matlab uses several manually set samples of pitch trajectory by clicking on a standard spectrogram as an input. Its output is the samples of the pitch trajectory calculated by using spline-based interpolation over the selected samples.

4. Results

The analysis tool proposed in this paper was tested on real speech recordings. The sampling rate of 8 kHz was chosen, and analysis windows with a constant step-size \( (M = 64) \) with a combination of different segment-lengths were applied. Figure 3 represents clean continuous speech signal from a female speaker analyzed by using the following techniques: Short-time Fourier transform with a segment-length of 32ms and 64 ms \( (N = 256, \text{ resp. } N = 512) \), Smoothed-Pseudo-WignerVille Distribution \((L_s = 255; L_h = 511)\) and finally the proposed STCT driven by a manual estimation of pitch. No pre-emphasis was applied and the Hamming window was used for all methods. The effect of using different window-lengths in case of STFT is well known: a short window provides better time resolution, and fast time-frequency changes (pitch variation) results in fair harmonic representation. On the contrary, a larger window provides more accurate harmonic localization if the pitch remains stationary, and blurry representation when it varies. The SPWVD provides good time-frequency localization of the harmonics, but several details of speech, both in time and frequency are lost, due to the time-frequency smoothing - especially at higher harmonics that are fast time-variant -, that this method requires. The STCT proposed in this paper also provides sharp time-frequency harmonic representation, when pitch fluctuates. The time-frequency resolution of the STCT is more detailed than that of the STFT, mainly for voiced sounds, even for larger window lengths, and also finer than the time-frequency resolution of the SPWVD. Since reconstructing the signal from its STCT image is feasible, the proposed technique is also suitable for filtering applications.

5. Conclusions

This paper introduces the Short-Time Chirp Transform as a powerful tool for time-frequency analysis of speech signals. The new technique uses quadratic base-functions designed adaptively, based on the pitch variations of analyzed segments. Since the base function of STCT can exactly match the harmonic structure of voiced speech, this technique gives a much more detailed representation of speech than the commonly used STFT, and also outperforms SPWVD by preserving all the formant information even at low-level time-frequency regions. Implementation of STCT is fast and simple and the only additional parameter required is the chirp-rate of analyzed segments derived from pitch trajectory.
6. References


Figure 3: Comparison of different time-frequency representations of female speech: (a) STFT (segment-length $N = 256$, step-size $M = 64$); (b) STFT ($N = 512, M = 64$); (c) Smoothed Pseudo Wigner-Ville Distribution ($L_g = 255; L_h = 511$); (d) proposed STCT ($N = 512; M = 64$), $\alpha_n$ derived from manual pitch tracking.