Chirp Complex Cepstrum-based Decomposition for Asynchronous Glottal Analysis

Thomas Drugman, Thierry Dutoit

TCTS Lab, University of Mons, Belgium

Abstract

It was recently shown that complex cepstrum can be effectively used for glottal flow estimation by separating the causal and anticausal components of speech. In order to guarantee a correct estimation, some constraints on the window have been derived. Among these, the window has to be synchronized on a Glottal Closure Instant. This paper proposes an extension of the complex cepstrum-based decomposition by incorporating a chirp analysis. The resulting method is shown to give a reliable estimation of the glottal flow wherever the window is located. This technique is then suited for its integration in usual speech processing systems, which generally operate in an asynchronous way. Besides its potential for automatic voice quality analysis is highlighted.

Index Terms: Glottal Flow, Voice Quality, Mixed-phase Decomposition.

1. Introduction

Complex Cepstrum has recently shown its ability to efficiently estimate the glottal flow [1]. An essential constraint for leading to a correct source-tract separation with this technique is the condition of being synchronized on Glottal Closure Instants (GCIs). On the other hand, the large majority of current speech processing systems operate in an asynchronous way, i.e use a constant frame shift. This paper proposes a modification of the complex cepstrum-based decomposition so as to integrate this technique in such systems.

The paper is structured as follows. Section 2 reviews the principles of the Complex Cepstrum-based Decomposition (CCD). Section 3 extends this formalism to integrate a chirp analysis. Besides an automatic way to find the optimal chirp contour is proposed. In Section 4 the performance of the resulting method is evaluated on a large corpus of emotional speech. It is shown that the chirp CCD technique gives a reliable asynchronous estimation of the glottal flow and can be used for voice quality analysis. Finally Section 5 concludes.

2. Complex Cepstrum for Glottal Source Estimation

The principle of the Complex Cepstrum-based Decomposition (CCD, [1]) relies on the mixed-phase model of speech [2]. According to this model, speech is composed of both minimum-phase (i.e causal) and maximum-phase (i.e anticausal) components. While the vocal tract and the glottal *return phase* can be considered as minimum-phase systems, it has been shown [3] that the glottal *open phase* is a maximum-phase signal. The key idea of the mixed-phase decomposition is then to separate both minimum and maximum-phase components of speech, where the latter is only due to the glottal contribution. In previous works, we proposed two algorithms achieving the mixed-phase decomposition: the Zeros of the Z-Transform (ZZT) algorithm [4], and the Complex Cepstrum-based Decomposition (CCD, [1]). Although both techniques are functionnaly equivalent, CCD was shown [1] to be much faster than ZZT. For this reason, this paper only focuses on the use of the Complex Cepstrum.

The Complex Cepstrum (CC) $\hat{x}(n)$ of a discrete signal $x(n)$ is defined by the following equations [5]:

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

$$\log[X(\omega)] = \log(|X(\omega)|) + j\angle X(\omega) \tag{2}$$

$$\hat{x}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log[X(\omega)] e^{j\omega n} d\omega \tag{3}$$

where Equations 1, 2, 3 are respectively the Discrete-Time Fourier Transform (DTFT), the complex logarithm and the inverse DTFT (IDTFT). Our decomposition arises from the fact that the complex cepstrum $\hat{x}(n)$ of an anticausal (causal) signal is zero for all n positive (negative). Retaining only the negative indexes of the CC should then estimate the glottal open phase contribution. CCD consequently achieves the mixed-phase decomposition using the quefrency origine as a discriminant.

Nonetheless it was shown in [1] that windowing is crucial and dramatically conditions the efficiency of the method. It is indeed essential that the window applied to the voiced segment of speech $x(n)$ respects some constraints in order to exhibit correct mixed-phase properties. Among these constraints, the window should [1]:

- be synchronized on a Glottal Closure Instant (GCI),
- have an appropriate shape and length (proportional to the pitch period).

Although some works aim at estimating the GCI positions directly from the speech signal [6], or use ElectroGlottoGraphs (EGGs, [7]), the great majority of current speech processing systems do not have this information available and consequently operate in an asynchronous way. In the following Section, we extend the Complex Cepstrum formalism so as to remove the GCI-synchronization constraint and propose an automatic way to achieve this.

3. Extension to the Chirp Analysis

The Chirp Z-Transform (CZT), as introduced by Rabiner et al [8] in 1969, allows the evaluation of the z-transform on a spiral contour in the z-plane. Its first application aimed at separating too close formants by reducing their bandwidth. In a previous work [9], we showed for the Zeros of the Z-Transform (ZZT) algorithm that considering a contour possibly different from the

unit circle makes the method more robust to GCI location errors. In its original version [4], the ZZT technique achieves the mixed-phase decomposition by separating the roots of the polynomial $X(z)$ located inside and outside the unit circle in the z-plane (respectively corresponding to the minimum and maximum-phase components of speech). The key idea of the work described in [9] was then to evaluate the CZT on a circle of radius R (with possibly $R \neq 1$) such that the root distribution is split into two well-separated groups. More precisely, it was observed that the significant impulse present in the excitation at the GCI results in a gap in the root distribution. When analysis is exactly GCI-synchronous, the unit circle perfectly separates causal and anticausal roots. On the opposite, when the window moves off from the GCI, the root distribution is transformed. Such a decomposition is then not guaranteed for the unit circle and another boundary is generally required.

Figure 1 gives an example of root distribution for a natural voiced speech frame for which a timing error is made on the actual GCI position. It is clearly seen that using the traditional z-transform ($R = 1$) for this frame will lead to an erroneous decomposition. In contrast, it is possible to find an optimal radius leading to a correct separation, as indicated with a solid line in Figure 1. In [9] it was also demonstrated that, for a two pitch period-long Blackman window (satisfying the second constraint of Section 2), the optimal radius is comprised within the bounds $exp(\pm\frac{50\pi}{17L})$ where L denotes the frame length in samples (these bounds are indicated in Figure 1 in dotted lines).

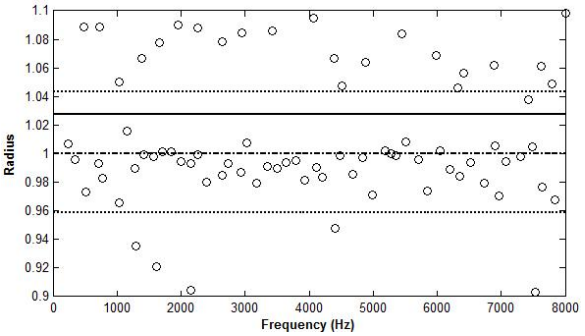

Figure 1: Representation of the Zeros of the Z-Transform in polar coordinates. The optimal chirp circle (solid line) gives the best separation of the root distribution within the bounds $exp(\pm\frac{50\pi}{17L})$ (dotted lines) and leads to a correct source-filter separation. It is clearly seen that, in this case, the unit circle (dashdotted line) will lead to an erroneous decomposition.

Relying on these conclusions, it is here proposed to integrate the chirp analysis whitin the Complex Cepstrum-based Decomposition (CCD) and to automatically find the optimal circle *without requiring the computation of the root distribution* (so as to still benefit from the speed of the CCD algorithm).

Achieving a chirp ZZT-based decomposition is straightforward since it is only necessary to modify the radius used to isolate the maximum-phase component. In order to integrate the chirp analysis for the CCD technique, let us consider the signal $x(n)$. Its CZT evaluated on a circle of radius R can be written as [8]:

$$X(Rz) = \sum_{n=0}^{L-1} x(n)(Rz)^{-n} = \sum_{n=0}^{L-1} (x(n)R^{-n})z^{-n} \qquad (4)$$

and is consequently equivalent to evaluating the z-transform of a signal $x'_R(n) = x(n)R^{-n}$ on the unit circle. The chirp CCD computed on a circle of radius R can therefore be achieved by applying the traditional CCD framework described in Section 2 to $x'_R(n)$ instead of $x(n)$.

In order to automatically estimate the radius giving an optimal separation between minimum and maximum-phase contributions, the unwrapped phase $\phi'_R(\omega)$ of $x'_R(n)$ is inspected. More precisely, the radius axis is uniformly discretized in N values ($N = 60$ in our experiments) between the bounds $exp(\pm\frac{50\pi}{17L})$. For each radius value R, $\phi'_R(\omega)$ is computed and the linear phase component is characterized by $\phi'_R(\pi)$ (with $\phi'_R(0) = 0$ by definition). From this, we define the variable $n_d(R)$:

$$n_d(R) = \frac{\phi'_R(\pi)}{\pi} \qquad (5)$$

as the number of samples of circular delay, i.e the number of samples that $x'_R(n)$ shoud be circularly shifted so as to remove its linear phase.

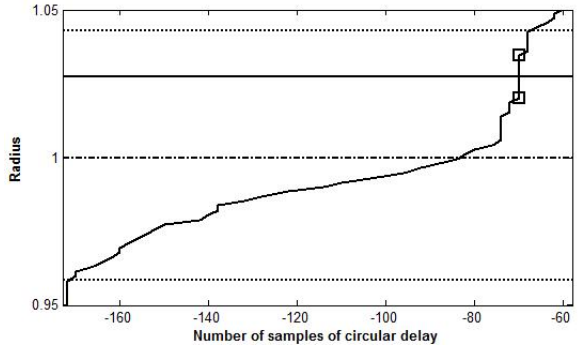

Figure 2: Evolution of $n_d(R)$ for the same signal as in Figure 1. The optimal radius (solid line) is defined as the middle of the largest interval (indicated by squares) for which $n_d(R)$ stays constant, within the bounds $exp(\pm\frac{50\pi}{17L})$ (dotted lines). The unit circle used in the traditional CCD is represented in dashdotted line.

Figure 2 shows the evolution of $n_d(R)$ for the same signal as used in Figure 1. $n_d(R)$ is actually a step function where gaps are due to the passage of some roots from the inside to the outside of the considered chirp circle. These phase discontinuities are illustrated in Figure 3. Indeed consider a zero which, initially located inside the circle of radius R_1 used for the evaluation of the CZT, is now passed outside of the circle of radius R_2 (with $R_1 > R_2$). When the CZT is evaluated on a point close to this zero in the z-plane, this results in a phase jump of $-\pi$ (see angles α_1 and α_2 in Fig. 3) which is then reflected in $\phi'_R(\pi)$. The difference $n_d(R_1) - n_d(R_2)$ is consequently interpreted as the number of zeros which, initially inside the circle of radius R_1, have passed the boundary to be now located outside the circle of radius R_2. In other words, inspecting the variable $n_d(R)$ allows us to detect the discontinuities in the root distribution (without requiring its whole computation). Similarly to [9], the optimal radius used for the chirp CCD is then defined as the middle of the largest interval for which $n_d(R)$ stays constant, within the bounds $exp(\pm\frac{50\pi}{17L})$.

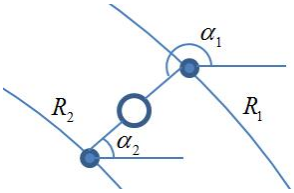

Figure 3: Illustration of a phase jump of $-\pi$ due to the passage of a zero from the inside of the circle of radius R_1 to the outside of the circle of radius R_2.

4. Experiments

Experiments are carried out on the De7 corpus. This database was designed by Marc Schroeder as one of the first attempts of creating diphone databases for expressive speech synthesis [10]. The database contains three voice qualities (modal, soft and loud) uttered by a German female speaker, with about 50 minutes of speech available for each voice quality. Besides, GCI positions are estimated by the algorithm described in [6].

The goal of this Section is to compare the traditional and the proposed chirp CCD techniques by studying their efficiency for glottal source estimation. Experiments are divided into two parts. In the first one, the sensitivity of both methods to GCI location errors is investigated. In the second part, the whole expressive speech database is analyzed by the two techniques and it is shown that chirp CCD leads to results similar as with the traditional CCD, but without the requirement of operating in a GCI-synchronous way.

4.1. Robustness to GCI location errors

When performing the mixed-phase separation, it may appear for some frames that the decomposition is erroneous, leading to an irrelevant high-frequency noise in the estimated glottal source [4]. As a criterion deciding whether a frame is considered as correctly decomposed or not, the spectral center of gravity is inspected. The distribution of this feature is displayed in Figure 4 for the loud voice using the traditional CCD. A principal mode at around 2kHz clearly emerges and corresponds to the majority of frames for which a correct decomposition is carried out. A second minor mode at higher frequencies is also observed. It is related to the frames where the mixed-phase decomposition fails, leading to a maximum-phase signal containing an irrelevant high-frequency noise. It can be noticed from this histogram (and it was confirmed by a manual verification of numerous frames) that fixing a threshold at around 2.7kHz makes a good distinction between frames that are correctly and incorrectly decomposed.

Given this criterion, the sensitivity of both traditional and chirp CCD techniques to a GCI location error is displayed in Figure 5 for the loud dataset of the De7 corpus. The constraint of being GCI-synchronous for the traditional CCD is clearly confirmed on this graph. It is indeed seen that the performance dramatically degrades for this technique as the window center moves off from the GCI. On the contrary, the chirp CCD method gives a high rate of correctly decomposed frames (however slightly below the performance of the GCI-centered traditional CCD) wherever the window is located.

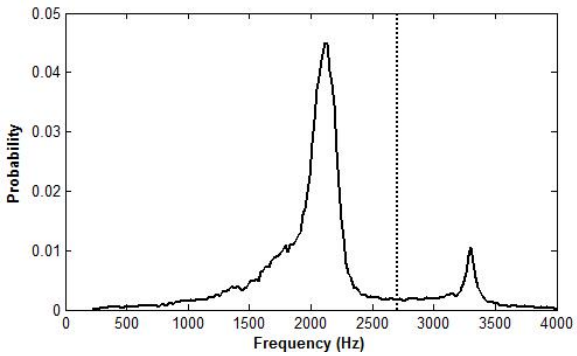

Figure 4: Distribution of the spectral center of gravity of the maximum-phase component, computed for the whole dataset of loud utterances. Fixing a threshold around 2.7kHz makes a good separation between correctly and incorrectly decomposed frames.

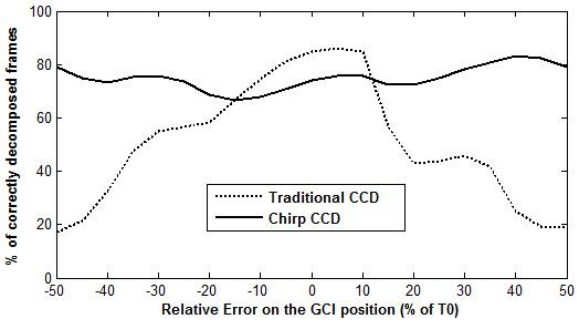

Figure 5: Robustness of both traditional and chirp CCD methods to a timing error on the GCI location.

4.2. Asynchronous glottal analysis of emotional speech

In this Section, we confirm the potential of the chirp CCD technique for asynchronously estimating the glottal flow on a large speech corpus. For this, the whole De7 database with its 3 voice qualities is analyzed. The glottal flow is estimated by 2 techniques:

- the traditional CCD: analysis is GCI-synchronous,
- the chirp CCD: analysis is asynchronous. A constant frame shift of $10ms$ is considered, as widely used in many speech processing systems. Note however that a two pitch period-long Blackman window is applied, as this is essential for achieving a correct mixed-phase decomposition (see Section 2).

In a first time, we evaluate the proportion of frames that are correctly decomposed by these two techniques using the spectral criterion of Section 4.1. Overall results for the three voice qualities are summarized in Table 1. The traditional CCD gives relatively high rates of correct decomposition with around 85% for the three datasets. It can also be observed that the chirp CCD method makes double the erroneous decompositions than the traditional approach. Nevertheless a correct estimation of the glottal source is carried out by the chirp CCD for around 70% of speech frames, which is rather high for real connected speech.

In a second time, frames of the glottal flow that were correctly estimated are characterized by the three following fea-

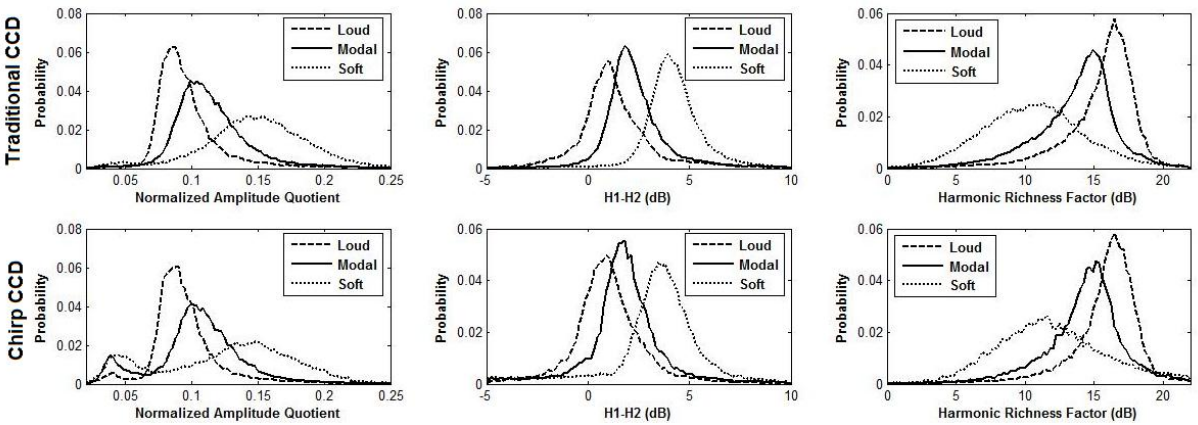

Figure 6: Distributions of glottal parameters estimated by (from top to bottom) the traditional and chirp CCD techniques, for three voice qualities. The considered glottal features are (from left to right): the Normalized Amplitude Quotient (NAQ), the H1-H2 ratio and the Harmonic Richness Factor (HRF).

Method	Loud	Modal	Soft
traditional CCD	87.09	84.41	83.68
chirp CCD	76.43	68.07	67.48

Table 1: Proportion (%) of correctly decomposed frames using the traditional and the chirp CCD techniques for the three voice qualities of the De7 database.

tures: the Normalized Amplitude Quotient (NAQ), the H1-H2 ratio between the two first harmonics and the Harmonic Richness Factor (HRF). These glottal parameters were shown in [11] and [12] to lead to a good separation between different types of phonation. The histograms of these parameters estimated by both traditional and chirp CCD methods are displayed in Figure 6 for the three voice qualities. Two main conclusions can be drawn from this Figure. First, it turns out that the distributions obtained by both techniques are strongly similar. A minor difference can however be noticed for NAQ histograms, where the distributions obtained by the chirp method contain a weak irrelevant peak at low NAQ values. The second important conclusion is that both techniques can be efficiently used for glottal-based voice quality analysis, leading to a clear discrimination between various phonation types.

5. Conclusion

This paper proposed an extension of the traditional Complex Cepstrum-based Decomposition (CCD). For this, the z-transform was evaluated on a contour in the z-plane possibly different from the unit circle. Circular contours were considered and an automatic way to find the optimal radius leading to well-separated groups of zeros was proposed. The resulting method was shown to be much more robust to GCI location errors than the traditional CCD approach. More particularly a reliable estimation of the glottal flow was obtained in an asynchronous way on real connected speech. Besides this technique showed its potential to be used for automatic voice quality analysis.

6. Acknowledgments

Thomas Drugman is supported by the "Fonds National de la Recherche Scientifique" (FNRS). Authors are also thankful to M. Schroeder for providing the De7 database.

7. References

[1] T. Drugman, B. Bozkurt, T. Dutoit, *Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation*, Interspeech Conference, 2009.

[2] B. Bozkurt, T. Dutoit, *Mixed-phase speech modeling and formant estimation, using differential phase spectrums*, VOQUAL'03, pp. 21-24, 2003.

[3] B. Doval, C. d'Alessandro, N. Henrich, *The voice source as a causal/anticausal linear filter*, Proceedings ISCA ITRW VOQUAL03, pp. 15-19, 2003.

[4] B. Bozkurt, B. Doval, C. d'Alessandro, T. Dutoit, *Zeros of Z-Transform Representation With Application to Source-Filter Separation in Speech* IEEE Signal Processing Letters, vol. 12, no. 4, 2005.

[5] A. Oppenheim, R. Schafer, *Discrete-time signal processing*, Prentice-Hall, chap. 12, 1989.

[6] T. Drugman, T. Dutoit, *Glottal Closure and Opening Instant Detection from Speech Signals*, Interspeech Conference, 2009.

[7] N. Henrich, C. d'Alessandro, B. Doval, M. Castellengo, *On the use of the derivative of electroglottographic signals for characterization of non-pathological phonation*, J. Acoust. Soc. Am., vol. 115, pp. 1321-1332, 2004.

[8] L. Rabiner, R. Schafer, C. Rader, *The Chirp-Z transform Algorithm and Its Application*, Bell System Technical Journal, vol. 48, no.5, pp. 1249-1292, 1969.

[9] T. Drugman, B. Bozkurt, T. Dutoit, *Chirp Decomposition of Speech Signals for Glottal Source Estimation*, ISCA Workshop on Non-Linear Speech Processing, 2009.

[10] M. Schroeder, M. Grice, *Expressing vocal effort in concatenative synthesis*, Proc. 15th Int. Conference of Phonetic Sciences, pp. 2589-2592, 2003.

[11] P. Alku, T. Backstrom, E. Vilkman, *Normalized amplitude quotient for parametrization of the glottal flow*, the Journal of the Acoustical Society of America, vol. 112, pp. 701-710, 2002.

[12] P. Alku, C. Magi, S. Yrttiaho, T. Backstrom, B. Story, *Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering*, the Journal of the Acoustical Society of America, vol. 125, no. 5, pp. 3289-3305, 2009.