Glottal Source Estimation Using an Automatic Chirp Decomposition

Thomas Drugman\textsuperscript{1}, Baris Bozkurt\textsuperscript{2}, and Thierry Dutoit\textsuperscript{1}

\textsuperscript{1} TCTS Lab, Faculté Polytechnique de Mons, Belgium
\textsuperscript{2} Department of Electrical & Electronics Engineering, Izmir Institute of Technology, Turkey

Abstract. In a previous work, we showed that the glottal source can be estimated from speech signals by computing the Zeros of the Z-Transform (ZZT). Decomposition was achieved by separating the roots inside (causal contribution) and outside (anticausal contribution) the unit circle. In order to guarantee a correct deconvolution, time alignment on the Glottal Closure Instants (GCIs) was shown to be essential. This paper extends the formalism of ZZT by evaluating the Z-transform on a contour possibly different from the unit circle. A method is proposed for determining automatically this contour by inspecting the root distribution. The derived Zeros of the Chirp Z-Transform (ZCZT)-based technique turns out to be much more robust to GCI location errors.

1 Introduction

The deconvolution of speech into its vocal tract and glottis contributions is an important topic in speech processing. Explicitly isolating both components allows to model them independently. While techniques for modeling the vocal tract are rather well-established, it is not the case for the glottal source representation. However the characterization of this latter has been shown to be advantageous in speaker recognition \textsuperscript{1}, speech disorder analysis \textsuperscript{2}, speech recognition \textsuperscript{3} or speech synthesis \textsuperscript{4}. These reasons justify the need of developing algorithms able to robustly and reliably estimate and parametrize the glottal signal.

Some works addressed the estimation of the glottal contribution directly from speech waveforms. Most approaches rely on a first parametric modeling of the vocal tract and then remove it by inverse filtering so as to obtain the glottal signal estimation. In \textsuperscript{5}, the use of the Discrete All-Pole (DAP) model is proposed. The Iterative Adaptive Inverse Filtering technique (IAIF) described in \textsuperscript{6} isolates the source signal by iteratively estimating both vocal tract and source parts. In \textsuperscript{7}, the vocal tract is estimated by Linear Prediction (LP) analysis on the closed phase. As an extension, the Multicycle closed-phase LPC (MCLPC) method \textsuperscript{8} refines its estimation on several larynx cycles. In a fundamentally different point of view, we proposed in \textsuperscript{9} a non-parametric technique based on the Zeros of the Z-Transform (ZZT). ZZT basis relies on the observation that speech is a mixed-phase signal \textsuperscript{10} where the anticausal component corresponds to the vocal folds open phase, and where the causal component comprises both the glottis
closure and the vocal tract contributions. Basically ZZT isolates the glottal open phase contribution from the speech signal, by separating its causal and anticausal components. In [11], a comparative evaluation between LPC and ZZT-based decompositions is led, giving a significant advantage for the second technique.

This paper proposes an extension to the traditional ZZT-based decomposition technique. The new method aims at separating both causal and anticausal contributions by computing the Zeros of a Chirp Z-Transform (ZCZT). More precisely, the Z-transform is here evaluated on a contour possibly different from the unit circle. As a result, we will see that the estimation is much less sensitive to the Glottal Closure Instant (GCI) detection errors. In addition, a way to automatically determine an optimal contour is also proposed.

The paper is structured as follows. Section 2 reminds the principle of the ZZT-based decomposition of speech. Its extension making use of a chirp analysis is proposed and discussed in Section 3. In Section 4, a comparative evaluation of both approaches is led on both synthetic and real speech signals. Finally we conclude in Section 5.

2 ZZT-Based Decomposition of Speech

For a series of $N$ samples $(x(0), x(1), ..., x(N−1))$ taken from a discrete signal $x(n)$, the ZZT representation is defined as the set of roots (zeros) $(Z_1, Z_2, ..., Z_{N−1})$ of the corresponding Z-Transform $X(z)$:

$$X(z) = \sum_{n=0}^{N−1} x(n)z^{−n} = x(0)z^{−N+1} \prod_{m=1}^{N−1} (z - Z_m)$$

(1)

The spectrum of the glottal source open phase is then computed from zeros outside the unit circle (anticausal component) while zeros inside it give the vocal tract transmittance modulated by the source return phase spectrum (causal component). To obtain such a separation, the effects of the windowing are known to play a crucial role [12]. In particular, we have shown that a Blackman window centered on the Glottal Closure Instant (GCI) and whose length is twice the pitch period is appropriate in order to achieve a good decomposition.

3 Chirp Decomposition of Speech

The Chirp Z-Transform (CZT), as introduced by Rabiner et al. [13] 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. Nowadays CZT reaches several fields of Signal Processing such as time interpolation, homomorphic filtering, pole enhancement, narrow-band analysis, ...

As previously mentioned, the ZZT-based decomposition is strongly dependent on the applied windowing. This sensitivity may be explained by the fact that ZZT implicitly conveys phase information, for which time alignment is known to be crucial [14]. In that article, it is observed that the window shape and onset may lead to zeros whose topology can be detrimental for accurate pulse