Neurocomputing 132 (2014) 121–125

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Speech polarity determination: A comparative evaluation

Thomas Drugman *, Thierry Dutoit

TCTS Lab, University of Mons, Belgium

ARTICLE INFO

Article history:
Received 20 February 2012
Received in revised form
25 April 2012
Accepted 6 June 2012
Available online 18 October 2013

Keywords:
Speech processing
Speech analysis
Speech polarity
Glottal source
Phase information

ABSTRACT

The performance of various speech processing applications may be dramatically affected by an inversion of the speech polarity, which depends upon the recording setup. As a consequence, automatically detecting the speech polarity is a necessary preliminary step to guarantee a correct behaviour of such methods. The goal of this paper is two-fold. First a new approach for polarity determination based on the calculation of higher-order statistical moments is introduced. These moments oscillate at the local fundamental frequency with a phase shift which is dependent on the speech polarity. Secondly, a thorough comparative evaluation between the proposed method and three other state-of-the-art techniques is carried out. Experiments are led on a large amount of data with 10 speech corpora. In addition to an analysis in clean conditions, the robustness of these methods to both an additive noise and to reverberation is also investigated.

1. Introduction

The polarity of speech may affect the performance of several speech processing applications. This polarity arises from the asymmetric glottal waveform exciting the vocal tract resonances. Indeed, the source excitation signal produced by the vocal folds generally presents, during the production of voiced sounds, a clear discontinuity occurring at the Glottal Closure Instant (GCI, [1]). This discontinuity is reflected in the glottal flow derivative by a peak delimitating the boundary between the glottal open phase and return phase. Polarity is said positive if this peak at the GCI is negative, like in the usual representation of the glottal flow derivative, such as in the Liljencrants–Fant (LF) model [2]. In the opposite case, polarity is negative.

When speech is recorded by a microphone, an inversion of the electrical connections causes the inversion of the speech polarity. Human ear is known to be insensitive to such a polarity change [3]. However, this may have a dramatic detrimental effect on the performance of various techniques of speech processing. In unit selection based speech synthesis [4], speech is generated by the concatenation of segments selected from a large corpus. This corpus may have been built through various sessions, possibly using different devices, and may therefore be made up of speech segments with different polarities. The concatenation of two speech units with different polarities results in a phase discontinuity, which may significantly degrade the perceptual quality when taking place in voiced segments of sufficient energy [3]. There are also several synthesis techniques using a pitch-synchronous overlap-add (PSOLA) which suffer from the same polarity sensitivity. This is the case of the well-known Time-Domain PSOLA (TDPSOLA, [5]) method for pitch modification purpose.

Besides, efficient techniques of glottal analysis require to process pitch synchronous speech frames. For example, the three best approaches considered in [1] for the automatic detection of GCI locations are dependent upon the speech polarity. An error on its determination results in a severe impact on the reliability and accuracy performance. There are also some methods of glottal flow estimation and for its parameterization in the time domain which assumes a positive speech polarity [6].

This paper proposes a new approach for the automatic detection of speech polarity which is based on the phase shift between two oscillating signals derived from the speech waveform. Two ways are suggested to obtain these two oscillating statistical moments. One uses non-linearity, and the other exploits higher-order statistics. In both cases, one oscillating signal is computed with an *odd* non-linearity or statistics order (and is *dependent* on the polarity), while the second oscillating signal is calculated for an *even* non-linearity or statistics order (and is *independent* of the polarity). These two signals are shown to evolve at the local fundamental frequency and have consequently a phase shift which depends on the speech polarity.

This paper is structured as follows. Section 2 gives a brief review on the existing techniques for speech polarity detection. The proposed approach is detailed in Section 3. A comprehensive evaluation of these methods is given in Section 4. Methods are compared on several large speech corpora in clean conditions, and their robustness to an additive noise and to reverberation is studied as well. Finally Section 5 concludes the paper.

* Corresponding author.
E-mail address: thomas.drugman@umons.ac.be (T. Drugman).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.06.063

2. State-of-the-art methods

Very few studies have addressed the problem of speech polarity detection. We here briefly present three state-of-the-art techniques achieving this purpose.

2.1. Gradient of the spurious glottal waveforms (GSGWs)

The GSGW method [7] focuses on the analysis of the glottal waveform estimated via a framework derived from the Iterative Adaptive Inverse Filtering (IAIF, [8]) technique. This latter signal should present a discontinuity at the GCI whose sign depends on the speech polarity. GSGW therefore uses a criterion based on a sharp gradient of the spurious glottal waveform near the GCI [7]. Relying on this criterion, a decision is taken for each glottal cycle and the final polarity for the speech file is taken via majority decision.

2.2. Phase cut (PC)

The idea of the PC technique [9] is to search for the position where the two first harmonics are in phase. Since the slopes are related by a factor 2, the intersected phase value ϕ_{cut} is

$$\phi_{cut} = 2 \cdot \phi_1 - \phi_2, \tag{1}$$

where ϕ_1 and ϕ_2 denote the phase for the first and second harmonics at the considered analysis time. Assuming a minimal effect of the vocal tract on the phase response at such frequencies, ϕ_{cut} closer to 0 (respectively π) implies a positive (respectively negative) peak in the excitation [9]. PC then takes a single decision via a majority strategy over all its voiced frames.

2.3. Relative phase shift (RPS)

The RPS approach [9] takes advantage of the fact that, for positive peaks in the excitation, phase increments between harmonics are approximately due to the vocal tract contribution. The technique makes use of Relative Phase Shifts (RPSs), denoted $\theta(k)$ and defined as

$$\theta(k) = \phi_k - k \cdot \phi_1, \tag{2}$$

where ϕ_k is the instantaneous phase of the kth harmonic. For a positive peak in the excitation, the evolution of RPS's over the frequency is smooth. Such a smooth structure is shown to be sensitive to a polarity inversion [9]. For this, RPS considers harmonics up to 3 kHz, and the final polarity corresponds to the most represented decisions among all voiced frames.

3. Oscillating moments-based polarity detection (OMPD)

In [1], we proposed a method of Glottal Closure Instant (GCI) determination which relied on a mean-based signal. This latter signal had the property of oscillating at the local fundamental frequency and allowed good performance in terms of reliability (i.e. leading to few misses or false alarms). The key idea of the proposed approach for polarity detection is to use two of such oscillating signals whose phase shift is dependent on the speech polarity. For this, we define the oscillating moment $y_{p_1,p_2}(t)$, depending upon p_1 and p_2 which respectively are the statistical and non-linearity orders, as

$$y_{p_1,p_2}(t) = \mu_{p_1}(x_{p_2,t}) \tag{3}$$

where $\mu_{p_1}(X)$ is the p_1th statistical moment of the random variable X.

The signal $x_{p_2,t}$ is defined as

$$x_{p_2,t}(n) = s^{p_2}(n) \cdot w_t(n) \tag{4}$$

where $s(n)$ is the speech signal and $w_t(n)$ is a Blackman window centered at time t:

$$w_t(n) = w(n-t) \tag{5}$$

As in [1], the window length is recommended to be proportional to the mean period $T_{0,mean}$ of the considered voice, so that $y_{p_1,p_2}(t)$ is almost a sinusoid oscillating at the local fundamental frequency. For $(p_1,p_2)=(1,1)$, the oscillating moment is the mean-based signal used in [1] for which the window length is $1.75 \cdot T_{0,mean}$. For oscillating moments of higher orders, we observed that a larger window is required for a better resolution. In the rest of this paper, we used a window length of $2.5 \cdot T_{0,mean}$ for higher orders (which in our analysis did not exceed 4). Besides, to avoid a low-frequency drift in $y_{p_1,p_2}(t)$, this signal is high-passed with a cut-off frequency of 40 Hz.

Fig. 1 illustrates for a given segment of voiced speech the evolution of four oscillating moments $y_{p_1,p_2}(t)$ respectively for $(p_1,p_2)=\{(1,1);(2,1);(3,1);(4,1)\}$. It can be noticed that all oscillating moments are quasi-sinusoids evolving at the local fundamental frequency and whose relative phase shift depends upon

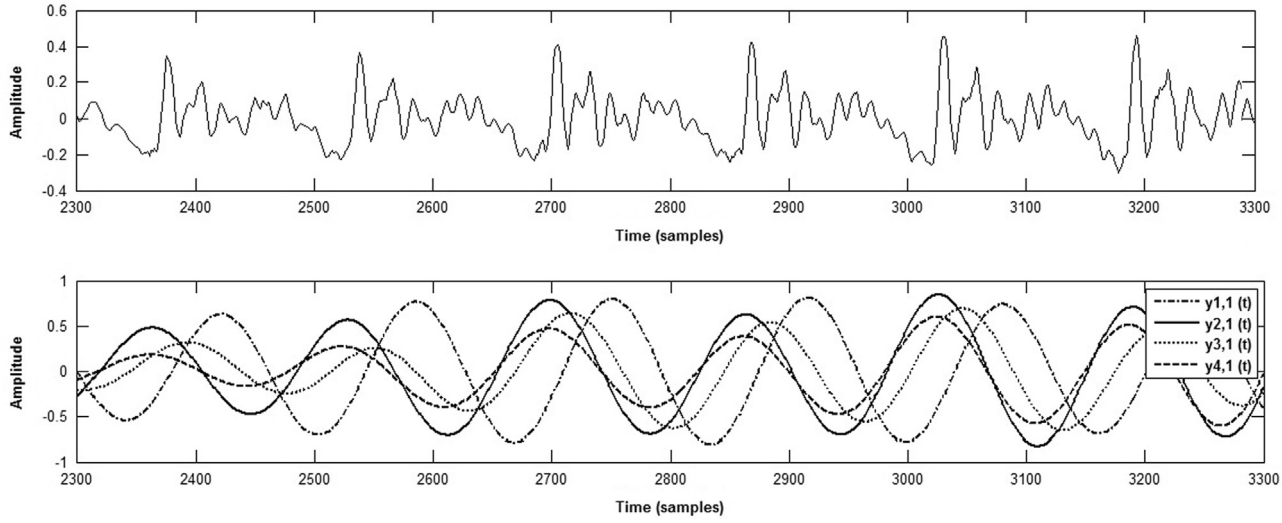

Fig. 1. Illustration of the oscillating moments (sampling rate = 16 kHz). *Top plot*: the speech signal. *Bottom plot*: the resulting oscillating moments with various values of p_1 and for $p_2 = 1$.