

Oscillating Statistical Moments for Speech Polarity Detection

Thomas Drugman, Thierry Dutoit

TCTS Lab, University of Mons, Belgium

Abstract. An inversion of the speech polarity may have a dramatic detrimental effect on the performance of various techniques of speech processing. An automatic method for determining the speech polarity (which is dependent upon the recording setup) is thus required as a preliminary step for ensuring the well-behaviour of such techniques. This paper proposes a new approach of polarity detection relying on oscillating statistical moments. These moments have the property to oscillate at the local fundamental frequency and to exhibit a phase shift which depends on the speech polarity. This dependency stems from the introduction of non-linearity or higher-order statistics in the moment calculation. The resulting method is shown on 10 speech corpora to provide a substantial improvement compared to state-of-the-art techniques.

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 Liljencrant-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 of speech segments with different polarities. The concatenation of two speech units with different polarity 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-know 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 assume 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* on 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, providing an objective comparison on several large databases. Finally Section 5 concludes the paper.

2 Existing 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 (GSGW)

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:

$$\phi_{cut} = 2 \cdot \phi_1 - \phi_2, \quad (1)$$

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 from 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 (RPS's), denoted $\theta(k)$ and defined as:

$$\theta(k) = \phi_k - k \cdot \phi_1, \quad (2)$$

where ϕ_k is the instantaneous phase of the k^{th} 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 3kHz, 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 to oscillate 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:

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

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

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

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

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

$$w_t(n) = w(n - t) \quad (5)$$

As in [1], the window length is recommended to be proportionnal 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.

Figure 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 sinusoids evolving at the local fundamental frequency and whose relative phase shift depends upon the order p_1 . Note that a similar conclusion can be drawn when inspecting the effect of p_2 . The principle of the proposed method is that $y_{p_1,p_2}(t)$ is polarity-dependent if $p_1 \cdot p_2$ is odd (i.e the oscillating moment is inverted with a polarity change), and is polarity-independent if $p_1 \cdot p_2$ is even.

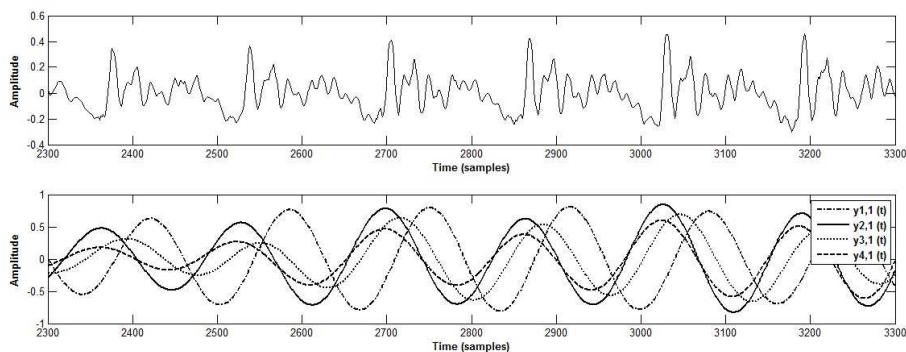


Fig. 1. Illustration of the oscillating moments. *Top plot:* the speech signal. *Bottom plot:* the resulting oscillating moments with various values of p_1 and for $p_2 = 1$.

In the following tests, for the sake of simplicity, only the oscillating moments $y_{1,1}(t)$ and $y_{1,2}(t)$ (or $y_{2,1}(t)$) are considered. Figure 2 shows, for the several speakers that will be analyzed in Section 4, how the distribution of the phase shift between $y_{1,1}(t)$ and $y_{1,2}(t)$ is affected by an inversion of polarity. Note that these histograms were obtained at the frame level and that phase shifts are expressed as a function of the local T_0 . Figure 2 suggests that fixing a threshold around -0.12 could lead to an efficient determination of the speech polarity.

Our proposed method, called Oscillating Moment-based Polarity Detection (OMPD), works as follows:

- Estimate roughly with an appropriate technique the *mean* pitch value $T_{0,mean}$ (required for determining the window length) and the voicing boundaries.

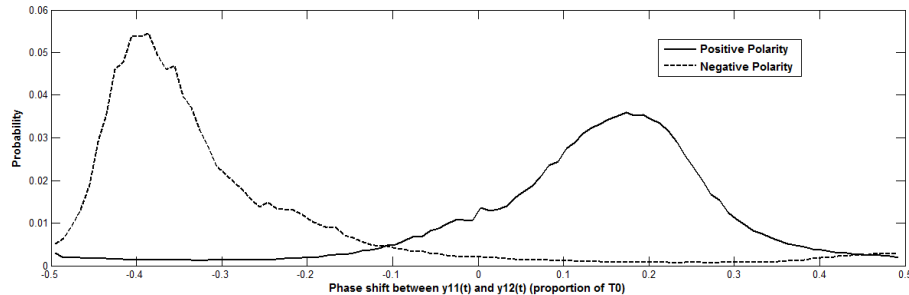


Fig. 2. Distribution of the phase shift (in local pitch period) between $y_{1,1}(t)$ and $y_{1,2}(t)$ for a negative and positive polarity.

- Compute from the speech signal $s(n)$ the oscillating moments $y_{1,1}(t)$ and $y_{1,2}(t)$, as indicated by Equations 3 to 5.
- For each voiced frame, estimate the local pitch period T_0 from $y_{1,1}(t)$ (or equivalently from $y_{1,2}(t)$) and compute the local phase shift between these two signals.
- Apply a majority decision over the voiced frames, a frame being with a positive polarity if its phase shift is comprised between -0.12 and 0.38.

It is worth to mention that an important advantage of OMPD, with regard to the techniques described in Section 2, is that it just requires a rough estimate of the mean pitch period, and not an accurate determination of the complete pitch contour. This gives the method also an advantage for performing in adverse conditions.

4 Experiments

The experimental evaluation is carried out on 10 speech corpora. Several voices are taken from the CMU ARCTIC database [10], which was designed for speech synthesis purpose: AWB (Scottish male), BDL (US male), CLB (US female), JMK (Canadian male), KSP (Indian male), RMS (US male) and SLT (US female). About 50 min of speech is available for each of these speakers. The Berlin database [11] is made of emotional speech (7 emotions) from 10 speakers (5F - 5M) and consists of 535 sentences altogether. The two speakers RL (Scottish male) and SB (Scottish female) from the CSTR database [12], with around 5 minutes per speaker, are also used for the evaluation.

Results of polarity detection using the four techniques described in the previous sections are reported in Table 1. It can be noticed that GSGW gives in general a lower performance, except for speaker SB where it outperforms other approaches. PC generally achieves high detection rates, except for speakers SB and SLT. Although RPS leads to a perfect polarity determination in 7 out of the 10 corpora, it may for some voices (KSP and SB) be clearly outperformed by

other techniques. As for the proposed OMPD method, it works perfectly for 8 of the 10 databases and gives an acceptable performance for the two remaining datasets. In average, over the 10 speech corpora, it turns out that OMPD clearly carries out the best results with a total error rate of 0.15%, against 0.64% for PC, 0.98% for RPS and 3.59% for GSGW.

Speaker	GSGW			PC			RPS			OMPD		
	OK	KO	Acc. (%)	OK	KO	Acc. (%)	OK	KO	Acc. (%)	OK	KO	Acc. (%)
AWB	1134	4	99.64	1138	0	100	1138	0	100	1138	0	100
BDL	1112	19	98.32	1131	0	100	1131	0	100	1131	0	100
Berlin	356	179	66.54	528	7	98.69	535	0	100	525	10	98.13
CLB	1131	1	99.91	1132	0	100	1132	0	100	1132	0	100
JMK	1096	18	98.38	1109	5	99.55	1114	0	100	1114	0	100
KSP	1103	29	97.43	1132	0	100	1059	73	93.55	1132	0	100
RL	50	0	100	50	0	100	50	0	100	50	0	100
RMS	1082	50	95.58	1132	0	100	1129	3	99.73	1132	0	100
SB	49	1	98	37	13	74	42	8	84	47	3	94
SLT	1125	6	99.38	1101	30	97.35	1131	0	100	1131	0	100
TOTAL	8238	307	96.41	8490	55	99.36	8461	84	99.02	8532	13	99.85

Table 1. Results of polarity detection for 10 speech corpora using the four techniques. The number of sentences whose polarity is correctly (OK) or incorrectly (KO) determined are indicated, as well as the detection accuracy (in %).

5 Conclusion

This paper investigated the use of oscillating moments for the automatic detection of speech polarity. The proposed technique is based on the observation that these moments oscillate at the local fundamental frequency and have a phase shift which is dependent upon the speech polarity. The introduction of this polarity-dependency is made by considering non-linearity or higher-order statistics. The resulting method is shown through an objective evaluation on several large corpora to outperform existing approaches for polarity detection. On these databases, it reaches an average error rate of 0.15% against 0.64% for the best state-of-the-art technique. Besides the proposed method only requires a rough estimate of the *mean* pitch period for the considered voice.

Acknowledgments

Thomas Drugman is supported by the “Fonds National de la Recherche Scientifique” (FNRS).

References

1. T. Drugman, M. Thomas, J. Gudnason, P. Naylor, T. Dutoit: *Detection of Glottal Closure Instants from Speech Signals: a Quantitative Review*, IEEE Trans. on Audio, Speech and Language Processing, *To appear*.
2. G. Fant, J. Liljencrants, Q. Lin: *A four parameter model of glottal flow*, STL-QPSR4, pp. 1-13, 1985.
3. S. Sakaguchi, T. Arai, Y. Murahara: *The Effect of Polarity Inversion of Speech on Human Perception and Data Hiding as Application*, ICASSP, vol. 2, pp. 917–920, 2000.
4. A. Hunt, A. Black: *Unit selection in a concatenative speech synthesis system using a large speech database*, ICASSP, pp. 373-376, 1996.
5. E. Moulines, J. Laroche: *Non-parametric techniques for pitch-scale and time-scale modification of speech*, Speech Communication, vol. 16, pp. 175-205, 1995.
6. T. Drugman, B. Bozkurt, T. Dutoit: *A comparative study of glottal source estimation techniques*, Computer Speech and Language, vol. 26, pp. 20-34, 2012.
7. W. Ding, N. Campbell: *Determining Polarity of Speech Signals Based on Gradient of Spurious Glottal Waveforms*, ICASSP, pp. 857–860, 1998.
8. P. Alku, J. Svec, E. Vilkmán, F. Sram: *Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering*, Speech Communication, vol. 11, issue 2-3, pp. 109–118, 1992.
9. I. Saratxaga, D. Erro, I. Hernez, I. Sainz, E. Navas: *Use of harmonic phase information for polarity detection in speech signals*, Interspeech, pp. 1075–1078, 2009.
10. J. Kominek, A. Black: *The CMU Arctic Speech Databases*, SSW5, pp. 223-224, 2004.
11. F. Burkhardt, A. Paseschke, M. Rolfes, W. Sendlmeier, B. Weiss: *A Database of German Emotional Speech*, Interspeech, pp. 1517-1520, 2005.
12. P. Bagshaw, S. Hiller, M. Jack: *Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching*, Eurospeech, pp. 1003–1006, 1993.