SPEECH POLARITY DETECTION USING STRENGTH OF IMPULSE-LIKE EXCITATION EXTRACTED FROM SPEECH EPOCHS

Sudarsana Reddy Kadiri[1] *and B. Yegnanarayana*[2]

[1]International Institute of Information Technology, Hyderabad-500032, India.
[2]Birla Institute of Technology & Science Pilani, Hyderabad-500078, India.
sudarsanareddy.kadiri@research.iiit.ac.in;yegna@iiit.ac.in

ABSTRACT

In this paper, we address the issue of speech polarity detection using strength of impulse-like excitation around epoch. The correct detection of speech polarity is a crucial step for many speech processing algorithms to extract suitable information. Occurrence of errors in the detection of speech polarity could have an impact on the performance of speech systems. Automatic detection of speech polarity has become an important preliminary step for many speech processing algorithms. We propose a method based on the knowledge of impulse-like excitation of speech production mechanism. The impulse-like excitation is reflected across all frequencies including the zero frequency (0 Hz). Using the slope around zero crossings of the zero frequency filtered signal, an automatic speech polarity detection method is proposed. Performance of the proposed method is demonstrated on 8 different speech corpora. The proposed method is compared with the three existing techniques such as gradient of the spurious glottal waveforms (GSGW), oscillating moments-based polarity detection (OMPD) and residual excitation skewness (RESKEW). From the experimental results, it is observed that the performance of the proposed method is comparable or better than the existing methods for the experiments considered.

Index Terms— Speech polarity detection, Speech analysis, Zero frequency filtering, Glottal closure instant, Phase.

1. INTRODUCTION

Speech is the result of exciting a time varying vocal tract system with the time varying excitation. The excitation source (glottal flow) produced by the movement of the vocal folds during the production of voiced speech has a clear discontinuity at the epochs or glottal closure instants (GCIs) due to the abrupt closure of the vocal folds. The discontinuity is reflected in the glottal flow derivative signal by a peak between the glottal open phase and return phase. In Liljencrants-Fant (LF) model [1], the speech polarity is defined as positive, if the glottal flow derivative exhibits a strong negative peak at the GCI, otherwise it is said to be of negative polarity. Often, speech signal posses the positive polarity. However, while recording using microphone, the electrical signals may be inversed depending on the electrical polarity connection of the device. This will cause for an inversion of the polarity of the recorded speech. The polarity of the speech stems from the asymmetric shape of the glottal source waveform. Correct detection of speech polarity plays crucial role in several speech processing techniques. For example, in [2–4], epochs are detected from the negative to positive zero crossings (NPZCs) of the zero frequency filtered signal. If the polarity is reversed and no effort is made to correct it beforehand, the location of the epochs may go wrong and thus affect the subsequent analysis. Similar is the case in some other epoch extraction methods, like speech event detection using the residual excitation and a mean-based signal (SEDREAMS) [5] and dynamic plosion index (DPI) [6].

In concatenated speech synthesis, if the concatenated units are of different polarities, it will result in phase discontinuity at the boundary of the concatenated units. The phase discontinuities are perceived by the listener if they occur in the high energy regions of voiced speech [7, 8]. Speech polarity is also an issue in most of the pitch synchronous analysis and synthesis methods such as pitch-synchronous overlap-add (PSOLA) and Time-Domain PSOLA (TD-PSOLA) for pitch modification [9].

Speech modification techniques based on the sinusoidal or harmonic models use phase manipulation procedures that depend on the polarity of the signals [10]. Hence the methods dealing with the phase of the speech signals are polarity dependent in contrast with the techniques which rely on the magnitude spectrum. Further, in recognition systems which use phase based features, polarity detection plays an important role. For example, the speaker recognition systems proposed in [11,12] use the discrete cosine transform (DCT) coefficients of glottal source as features. If training and testing data are from microphone with opposite polarities, it may result in poor performance of the system. This is also the case for speech recognition systems that use the phase-based features such as methods that were proposed in [13, 14].

Normally polarity of the speech signal is attributed to the glottal source signal. Most of the speech polarity detection methods rely on the glottal source signal derived from the speech after removing the predictable (second order correlations) portion. The excitation source or glottal source signal is usually derived by performing linear prediction (LP) analysis [15] of the speech signal. The prominent peak in the estimated derivative of the glottal source signal is seen as the epoch, and if it is in the negative going half, the polarity is said to be positive, and negative otherwise [1]. Based on this, in [16], gradient of the spurious glottal waveforms (GSGW) method was proposed that use the fact that the peaks in the glottal flow derivative (estimated from iterative adaptive inverse filtering (IAIF) using LP analysis [17]) should be close to the epoch locations. In [18], based on the observation that the first two harmonics are in phase near the epochs, phase cut (PC) method was proposed. An extension of PC method that uses higher harmonics for speech polarity detection was proposed in [18] and was named as relative phase shift (RPS) method. In [19], the oscillating moments-based polarity detection (OMPD) method was proposed that uses phase shifts in the even and odd ordered statistical moments oscillating at the local pitch period. Based on the observation that the distribution of

sample values of the estimated glottal source signal have a negative skew and that of the traditional LP residual a positive skew, residual excitation skewness (RESKEW) method was proposed in [20]. In this, three decision rules have been presented, using skewness of only the approximately voice source (RESKEW-glot), the traditional LP residual signal (RESKEW-res), and the difference of the skewness of the approximated glottal source and traditional LP residual signal, named as RESKEW method. In order to reduce the computational complexity, long-term weighted skew (LT-WSKEW) was proposed in [12]. The GSGW, PC, RPS and OMPD methods requires fundamental frequency (F_0) and voicing decisions. From the recent studies in [12,20], it was observed that RESKEW method was found to be more reliable.

In this paper, we propose a method to automatically detect the speech polarity based on the strength of the impulse-like excitation around epoch computed from the slope around the zero crossings of the zero frequency filtered (ZFF) signal. Since the excitation source at the glottal closure is reflected as impulse-like discontinuity, it is expected that slope around negative to positive zero crossings (NPZCs) is high if the signal polarity is positive.

The paper is organized as follows: Section 2 describes the basis for the present study. In Section 3, we present a method to detect the polarity of the speech signal. In Section 4, performance of the proposed method is compared with the three existing methods for speech polarity detection. Finally, Section 5 gives a summary of the study.

2. BASIS FOR THE PRESENT STUDY

The present study is motivated from the studies made in [2], for epoch extraction using zero frequency filtering (ZFF) method. In [2], the authors showed that rapid changes occur around the negative to positive zero crossings (NPZCs) of the ZFF signal, and hence NPZCs can be considered as instants of glottal closure or epochs (GCIs). Detecting the correct speech polarity is a necessary step to ensure whether the NPZCs or positive to negative zero crossings (PNZCs) to be considered for GCIs. In recent studies [21,22], lower accuracies were observed in the epoch extraction using ZFF method for some recorded databases. One reason for this may be that the recorded signals have variations in speech polarity.

From this observation and also from the fact that rapid change at epochs occur around the zero crossings of the ZFF signal, a new speech polarity detection method is proposed. The method is based on the slope around the zero crossings of the ZFF signal. For illustration, a sequence of randomly spaced impulses with arbitrary strengths are shown in the Figs. 1(a) and 2(a), and the corresponding zero frequency filtered signals are shown in the Figs. 1(b) and 2(b), respectively. It can be observed from Fig. 1(b) that the NPZCs of the ZFF signal have rapid change i.e., the slope is high when compared to PNZCs and vice versa can be seen from Fig. 2(b). In the next section, we also show that this observation is valid even for speech signals.

3. PROPOSED SPEECH POLARITY DETECTION METHOD

For the detection of speech polarity, we are using the ZFF method proposed in [2] for epoch extraction. This method exploits the knowledge of the speech production mechanism. The motivation behind that study was the effect of impulse-like excitation is reflected across all frequencies including zero frequency (0-Hz). The

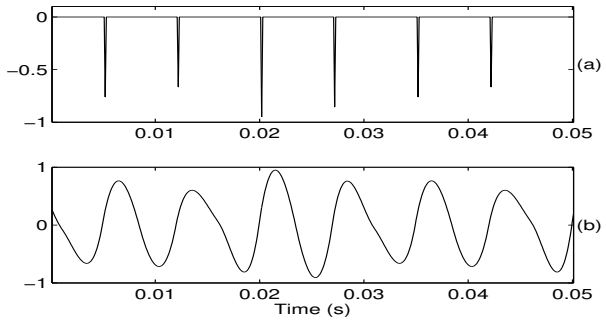

Fig. 1. Illustration of speech polarity for a sequence of negative polarity impulses. (a) Aperiodic sequence of impulses with varying amplitudes and nonuniform intervals. (b) Zero-frequency filtered signal.

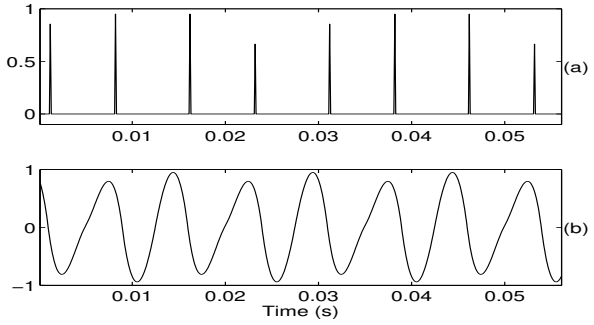

Fig. 2. Illustration of speech polarity for a sequence of positive polarity impulses. (a) Aperiodic sequence of impulses with varying amplitudes and nonuniform intervals. (b) Zero-frequency filtered signal.

advantage of choosing the zero-frequency resonator is that the characteristics of the time varying vocal tract system will not affect the characteristics of the discontinuities in the output of the resonator.

The steps involved in the proposed speech polarity detection method are as follows:

1. The speech signal ($s[n]$) is differenced to remove any unwanted very low frequency components. The differenced signal is given by,

$$x[n] = s[n] - s[n-1]. \tag{1}$$

2. The differenced signal ($x[n]$) is passed through a cascade of two zero-frequency resonators, given by the following equation [2]:

$$y_o[n] = \sum_{k=1}^{4} a_k y_o[n-k] + x[n], \tag{2}$$

where $a_1 = +4$, $a_2 = -6$, $a_3 = +4$, $a_4 = -1$. The signal $y_o[n]$ is equivalent to integration of speech signal four times, hence it approximately grows/decays as a polynomial function of time.

3. The trend in $y_o[n]$ is removed by subtracting the local mean computed over the average pitch period at each sample. The

resulting signal ($y[n]$) is called zero frequency filtered (ZFF) signal. That is

$$y[n] = y_o[n] - \frac{1}{2N+1} \sum_{i=-N}^{N} y_o[n+i], \quad (3)$$

where $2N + 1$ corresponds to the number of samples in the window used for trend removal.

4. The slopes around the NPZCs and PNZCs of the ZFF signal ($y[n]$) are computed by the following equations:

$$C_{npzc}[n] = y_{npzc}[n+1] - y_{npzc}[n-1], \quad (4)$$
$$C_{pnzc}[n] = y_{pnzc}[n+1] - y_{pnzc}[n-1], \quad (5)$$

where $C_{npzc}[n]$ and $C_{pnzc}[n]$ correspond to the slopes around the NPZCs and PNZCs of $y[n]$, respectively.

5. Decide the polarity of the speech signal as,

$$Polarity = \begin{cases} +, & \text{if } \Sigma(C_{npzc}[n]) > \Sigma(C_{pnzc}[n]) \\ -, & \text{otherwise,} \end{cases} \quad (6)$$

where *polarity* refers to the final judgment of the whole utterance.

This method is abbreviated as ZFF-SPD (zero-frequency filtering based speech polarity detection) in this paper. For illustration, the positive polarity signal and its corresponding ZFF signal are shown in Figs. 3(a) and 3(b), the negative polarity signal and its corresponding ZFF signal are shown in Figs. 4(a) and 4(b), respectively.

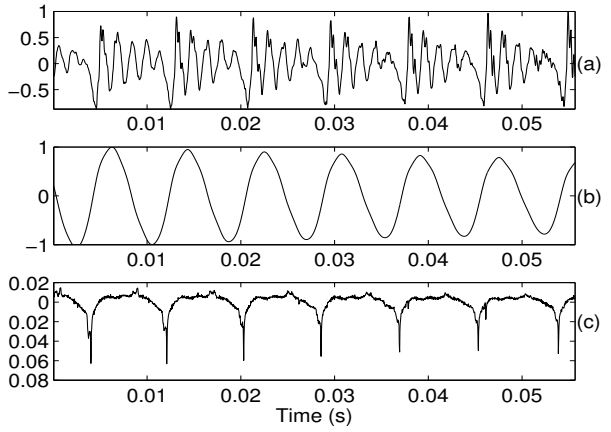

Fig. 3. Illustration of speech polarity for a positive polarity speech signal. (a) Segment of a speech signal, (b) Zero-frequency filtered signal, and (c) Differenced EGG (dEGG) signal.

From Fig. 3(b), it is clear that the ZFF signal has rapid changes around the NPZCs, hence the signal is said to be of positive polarity. Similarly from Fig. 4(b), it can be observed that the ZFF signal has rapid changes around the PNZCs, hence the signal is said to be of negative polarity. This is also evident from the differenced Electroglottograph (dEGG) signal shown in Figs. 3(c) and 4(c). Hence, the instants of NPZCs of ZFF signal correspond to the epoch locations if the signal polarity is positive otherwise the PNZCs correspond to the epoch locations.

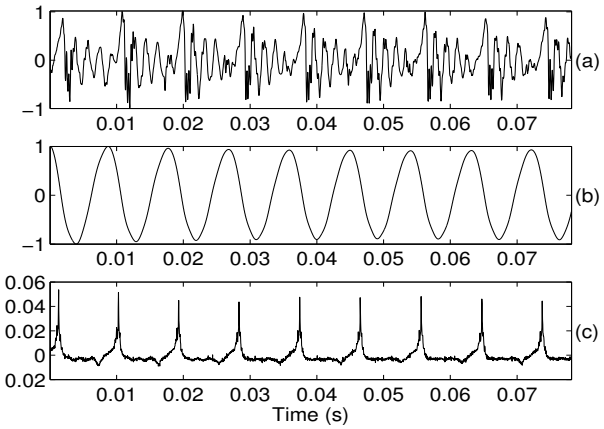

Fig. 4. Illustration of speech polarity for a negative polarity speech signal. (a) Segment of a speech signal, (b) Zero-frequency filtered signal, and (c) Differenced EGG (dEGG) signal.

4. COMPARISON OF PROPOSED ZFF-SPD WITH OTHER METHODS

In this section the proposed method of speech polarity detection is compared with three existing methods. The three methods chosen for the comparison are: gradient of the spurious glottal waveforms (GSGW) [16], oscillating moments-based polarity detection (OMPD) [19] and residual excitation skewness (RESKEW) [20].

4.1. Methods for comparison

A brief discussion on the implementation details of the three chosen methods for comparison are given below:

Gradient of the spurious glottal waveforms (GSGW): This method uses the glottal waveform estimated from inverse filtering technique (modified Iterative Adaptive Inverse Filtering (IAIF [17])). Since the glottal waveform signal should present a discontinuity at the GCI and whose sign depends on the speech polarity, this method uses a criterion based on a sharp gradient of the spurious glottal waveform near the GCI [1]. From the gradient of the spurious glottal waveform, speech polarity is determined by finding whether the GCIs are located above or below the zero line [16]. Based on this criterion, decisions are made for pitch synchronous wise (i.e., for each glottal cycle) and each frame wise, and the robust polarity for the speech signal is taken over majority decision over all the frames of the speech signal.

Table 1. *Description of the databases used for the evaluation (Here M refers to male and F refers to Female).*

Database	Type of speaker(s)	Amount of data (min)
AWB	Scottish male	83
BDL	US male	56
EMO-DB	German (5-M, 5-F)	25
CLB	US female	64
JMK	Canadian male	58
KSP	Indian male	37
RMS	US male	66
SLT	US female	56

Table 2. *Results of polarity detection for 8 different speech corpora using the four techniques (Here cor. refers to correct, incor. refers to incorrect and acc. refers to accuracy).*

	GSGW			OMPD			RESKEW			ZFF-SPD		
	cor.	incor.	acc. (%)	cor.	incor.	acc. (%)	cor.	incor.	acc. (%)	cor.	incor.	acc. (%)
AWB	1134	4	99.65	1138	0	100	1138	0	100	1138	0	100
BDL	1112	19	98.32	1131	0	100	1131	0	100	1131	0	100
CLB	1131	1	99.91	1132	0	100	1132	0	100	1132	0	100
JMK	1096	18	98.38	1114	0	100	1114	0	100	1114	0	100
KSP	1103	29	97.44	1132	0	100	1132	0	100	1132	0	100
RMS	1082	50	95.58	1132	0	100	1132	0	100	1132	0	100
SLT	1125	6	99.47	1131	0	100	1131	0	100	1131	0	100
EMO-DB	356	179	66.54	518	17	96.82	525	10	98.13	530	5	99.07
Total	8139	306	94.41	8428	17	99.60	8435	10	99.77	8440	5	99.88

Oscillating moments-based polarity detection (OMPD): The OMPD method [19] calculates on a sample-by-sample basis statistical moments oscillating at the local fundamental frequency. The key idea of OMPD is to compute two oscillating moments with an odd order and an even order, such that their phase shift allows determination of the correct polarity. Local decisions are taken for each voiced frame. The final polarity is decided using majority voting.

Residual excitation skewness (RESKEW): This method is based on the observation that excitation signal contains relevant information about speech polarity, as the behavior reflects in the asymmetry of glottal production. In this method, two excitation signals are considered for speech polarity detection. One is the traditional residual signal and other is a rough approximation of the glottal flow derivative. The residual signal exhibits positive peaks at GCI locations, if the polarity of the signal is positive. It is observed that estimated glottal flow derivative shows negative peaks around GCIs. As these two excitation signals exhibit asymmetry at GCIs, the skewness of the distribution of their samples is computed. The polarity detection based on the skewness measured on the first excitation signal is called RESKEW-res and the secondary excitation signal as RESKEW-glot. As both signals have skewness with opposite sign, this method (RESKEW) uses the sign of differences in skewness of the first excitation signal and the secondary excitation signal [20].

4.2. Databases

Experiments were carried out with 8 speech corpora. Out of which seven voices are taken from the CMU ARCTIC database [23], which was designed for the purpose of speech synthesis: AWB (Scottish male), BDL (US male), CLB (US female), JMK (Canadian male), KSP (Indian male), RMS (US male) and SLT (US female). These databases are available on the Festvox webpage [24]. The eighth database is the German emotional speech database (EMO-DB) [25], which consists of seven types of emotions (7 emotions: happy, angry, anxious, fearful, bored, disgusted and neutral) from 10 speakers (5 female and 5 male) and consists of 535 sentences. The details of the databases used for the evaluation are given in Table 1. The total number of speech files over the 8 speech corpora is 8445. EGG signals are used as the ground truth for polarity detection, which are available with the corresponding speech signals. If the dEGG signal shows the large negative peaks, then the signal is said to have positive polarity and vice versa for negative polarity.

4.3. Results and discussion

Results of speech polarity detection using the four methods described above are given in Table 2. It can be noticed that GSGW method gives the lowest performance. Although OMPD gives a good polarity detection performance in seven databases, in one database its performance is lower. RESKEW method gave very high accurate speech polarity detection rates for seven databases and its performance is lower for emotional speech database. The proposed ZFF-SPD method works perfectly for seven of the eight databases and gives the best performance for the emotional speech database. On average, over the eight speech corpora, it turns out that ZFF-SPD clearly carries out the best results with a total error rate of 0.12% against 0.23% for RESKEW, against 0.40% for OMPD and against 5.59% for GSGW. The second best method is RESKEW with a total error rate of 0.23%. Finally, with an averaged detection error rate of only 0.12% (5 erroneous speech files out of the 8445), the proposed ZFF-SPD method clearly outperforms all other methods. Note that these 5 errors are spread across four emotion datasets. The emotion database contains breathy voices making the polarity detection more difficult. It is also observed that, for some utterances epochs or GCIs are less evident, in the sense that the discontinuities in the excitation around the epoch is much less pronounced. Further investigation is required for speech polarity detection for speech with different voice qualities and for expressive data to provide further insights into the methods.

5. CONCLUSION

In this paper, a new approach for a speech polarity detection is proposed based on the strength of the impulse-like excitation around epoch. This is based on the fact that significant excitation takes place due to abrupt closure of the vocal folds at the glottis. For extraction of strength of the impulse-like excitation around an epoch, the zero frequency filtering method was used. The proposed method was compared to three state-of-art approaches on 8 large speech corpora. The proposed technique was shown to have comparable or better performance with the state-of-art methods. The proposed speech polarity detection method gives an averaged error rate of 0.12% compared to 0.23% for the best existing method. In future, we want to investigate the performance of the proposed method for conversational speech, degraded conditions (such as additive noises, channel degradations such as telephone speech etc.), various voice qualities and expressive voices etc.

6. ACKNOWLEDGMENTS

The authors would like to thank Tata Consultancy Services (TCS) for supporting first author PhD.

7. REFERENCES

[1] G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 26:1–13, 1985.

[2] K. Sri Rama Murty and B. Yegnanarayana. Epoch extraction from speech signals. *IEEE Trans. Audio, Speech, Lang. Process.*, 16(8):1602–1613, Nov. 2008.

[3] Sudarsana Reddy Kadiri and B. Yegnanarayana. Analysis of singing voice for epoch extraction using zero frequency filtering method. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4260–4264, April 2015.

[4] Sudarsana Reddy Kadiri and B. Yegnanarayana. Epoch extraction from emotional speech using single frequency filtering approach. *Speech Communication*, 86:52 – 63, 2017.

[5] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. *Interspeech*, pages 2891–2894, 2009.

[6] A.P. Prathosh, T.V. Ananthapadmanabha, and A.G. Ramakrishnan. Epoch extraction based on integrated linear prediction residual using plosion index. *IEEE Trans. on Audio, Speech, and Language Processing*, 21(12):2471–2480, Dec 2013.

[7] S. Sakaguchi, T. Arai, and Y. Murahara. The effect of polarity inversion of speech on human perception and data hiding as an application. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 917–920, 2000.

[8] AJ. Hunt and AW. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 373–376, May 1996.

[9] Eric Moulines and Jean Laroche. Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16(2):175–205, 1995.

[10] Doh-Suk Kim. On the perceptually irrelevant phase information in sinusoidal representation of speech. *IEEE Transactions on Speech and Audio Processing,*, 9(8):900–905, Nov 2001.

[11] A. G. Ramakrishnan, B. Abhiram, and S. R. Mahadeva Prasanna. Voice source characterization using pitch synchronous discrete cosine transform for speaker identification. *The Journal of the Acoustical Society of America*, 137(6):469–475, 2015.

[12] B. Abhiram, A.P. Prathosh, and A.G. Ramakrishnan. A fast algorithm for speech polarity detection using long-term linear prediction. In *International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5, July 2014.

[13] M.S.E. Langarani, H. Veisi, and H. Sameti. The effect of phase information in speech enhancement and speech recognition. In *International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pages 1446–1447, July 2012.

[14] R. Schluter and H. Ney. Using phase spectrum information for improved speech recognition performance. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 133–136, 2001.

[15] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, April 1975.

[16] Wen Ding and N. Campbell. Determining polarity of speech signals based on gradient of spurious glottal waveforms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 857–860, May 1998.

[17] P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11:109–118, 1992.

[18] Saratxaga I, Erro D, Hernez I, Sainz I, and Navas E. Use of harmonic phase information for polarity detection in speech signals. In *Interspeech*, pages 1075–1078, 2009.

[19] Thomas Drugman and Thierry Dutoit. Detecting speech polarity with high-order statistics. *Cognitive Computation*, 5(4):442–447, 2013.

[20] T. Drugman. Residual excitation skewness for automatic speech polarity detection. *IEEE Signal Processing Letters*, 20(4):387–390, April 2013.

[21] John Kane and Christer Gobl. Evaluation of glottal closure instant detection in a range of voice qualities. *Speech Communication*, 55(2):295–314, 2013.

[22] Onur Babacan, Thomas Drugman, Nicolas D'Alessandro, Nathalie Henrich, and Thierry Dutoit. A quantitative comparison of glottal closure instant estimation algorithms on a large variety of singing sounds. In *INTERSPEECH*, pages 1702–1706, 2013.

[23] Kominek J and Black A. The cmu arctic speech databases. In *SSW5*, pages 223–224, 2004.

[24] The Festvox Website. Source: `http://festvox.org/cmu_arctic/index.html`.

[25] Burkhardt F, Paseschke A, Rolfes M, Sendlmeier W, and Weiss B. A database of german emotional speech. In *Interspeech*, pages 1517–1520, 2005.