Speech Polarity Determination: A Comparative Evaluation

Thomas Drugman, Thierry Dutoit
TCTS Lab, University of Mons, Belgium

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.

Keywords: Speech Processing, Speech Analysis, Speech Polarity, Glottal Source, Phase Information.

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 occuring 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 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. 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.

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 (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:

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

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 (RPS's), 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 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 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_1^{th} statistical moment of the random variable X. The signal $x_{p_2,t}$ is defined as:

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

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.

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 quasi-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.

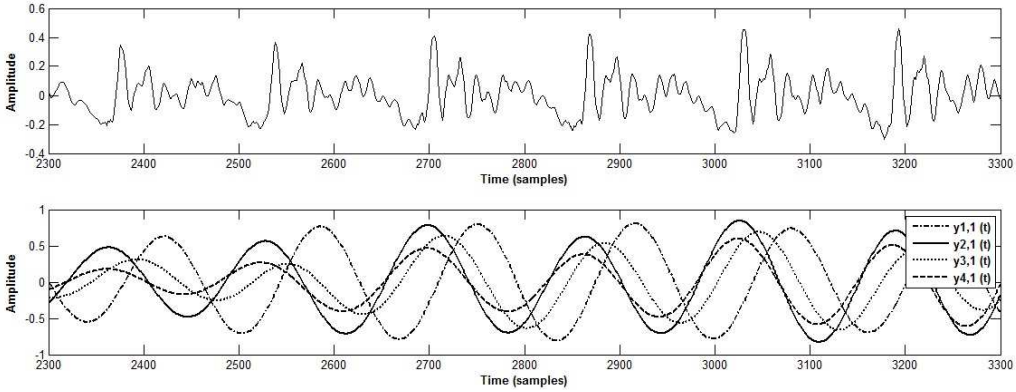

Figure 1: Illustration of the oscillating moments (sampling rate = 16 kHz). *Top plot*: the speech signal. *Bottom plot*: the resulting oscillating moments with various values of p_1 and for $p_2 = 1$.

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:

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

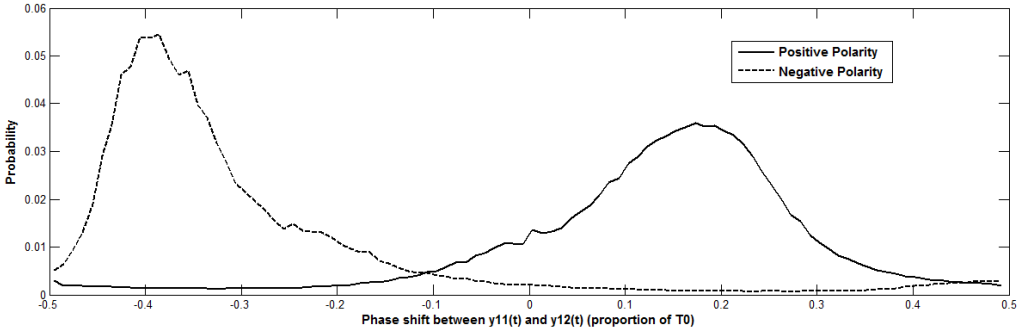

Figure 2: Distribution of the phase shift (in local pitch period) between $y_{1,1}(t)$ and $y_{1,2}(t)$ for a negative and positive polarity.

- 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.

The choice of the threshold -0.12 arises from the inspection of Figure 2 which was obtained on a large amount of data. It is however important to note that we observed the phase shift distribution to be only poorly sensitive to the considered speaker or to recording conditions. Besides, the fact that the second threshold 0.38, which was fixed for symmetry purpose, does not exactly correspond to the second intersection point of positive and negative polarity distributions in Figure 2 lets us think that the proposed method is insensitive to the threshold setting in a certain extent around the suggested value.

It is worth mentioning 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 of performing in adverse conditions.

4. Experiments

In some speech processing applications, such as speech synthesis, utterances are recorded in well controlled conditions. For such high-quality speech

signals, the performance of speech polarity detection techniques is studied in Section 4.1. For many other types of speech processing systems however, there is no other choice than capturing the speech signal in a *real world environment*, where noise and/or reverberation may dramatically degrade its quality. The goal of Sections 4.2 and 4.3 is to evaluate how speech polarity detection methods are affected respectively by an additive noise and by reverberation.

Experiments are 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.

For experiments of robustness, a quarter from each of the 10 speech corpora was used per noisy/reverberant configuration (except for the CSTR database which contains less data, and where the whole dataset was used). This way of proceeding still ensures an important amount of data per noisy condition, so that it does not affect the conclusions that will be drawn in the following.

For all experiments, the Summation of Residual Harmonics (SRH, [14]) algorithm was used for both estimating the fundamental frequency contour and detecting the voiced-unvoiced segment boundaries. It is important to note that, for experiments in noisy/reverberant conditions, pitch was extracted from degraded recordings. This therefore justifies our choice of using the SRH pitch tracker since it was shown in [14] to clearly outperform the state-of-the-art in adverse conditions, providing better robustness results in various types of noise. Nonetheless, since the pitch tracking performance degrades with a decreasing SNR, it can be expected that this will affect the existing techniques of polarity determination (which require the whole pitch contour) in a larger extent, since the OMPD only needs a rough $F_{0,mean}$ estimate which can be obtained rather easily.

4.1. Results in Clean Conditions

Results of polarity detection in clean conditions 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. On 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.

		GSGW		PC		RPS		OMPD	
Speaker	# files	KO	Acc.	KO	Acc.	KO	Acc.	KO	Acc.
AWB	1138	4	99.64	0	**100**	0	**100**	0	**100**
BDL	1131	19	98.32	0	**100**	0	**100**	0	**100**
Berlin	535	179	66.54	7	98.69	0	**100**	10	98.13
CLB	1132	1	99.91	0	**100**	0	**100**	0	**100**
JMK	1114	18	98.38	5	99.55	0	**100**	0	**100**
KSP	1132	29	97.43	0	**100**	73	93.55	0	**100**
RL	50	0	**100**	0	**100**	0	**100**	0	**100**
RMS	1132	50	95.58	0	**100**	3	99.73	0	**100**
SB	50	1	**98**	13	74	8	84	3	94
SLT	1131	6	99.38	30	97.35	0	**100**	0	**100**
TOTAL	8545	307	96.41	55	99.36	84	99.02	13	**99.85**

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

4.2. Robustness to an Additive Noise

In this experiment, noise was added to the original speech waveform at various Signal-to-Noise Ratios (SNRs). Both a White Gaussian Noise (WGN) and a babble noise (also known as cocktail party noise) were considered. The noise signals were taken from the Noisex-92 database [13], and were added so as to control the overall SNR without silence removal.

The influence of the noise level and type on the polarity error rate is displayed in Figure 3. In the presence of a White Gaussian Noise (WGN), it

can be observed that OMPD remains the best technique at any SNR value. It can be indeed understood that WGN is, at the scale of the OMPD window, an almost zero-mean signal which, after addition to the speech signal, will only slightly affect the evolution of the oscillating moments. The OMPD technique is therefore highly robust to WGN. With the increase of the noise level, the performance of RPS stays almost unchanged while PC has a slight degradation. The most affected technique with a WGN is GSGW, with an absolute increase of its error rate of 2% at 10dB of SNR (compared to clean conditions).

In a babble noise, this degradation is even much stronger. This is especially true for GSGW whose error rate reaches 41% in the noisiest conditions. It is indeed known [6] that the performance of glottal flow estimation techniques rapidly degrades in such environments. Although the proposed OMPD method remains the best approach up to 20dB of SNR, it is clearly outperformed in more severe environments. One reason for this is that babble noise, relatively to WGN, contains stronger lower frequencies which perturb the oscillating signals. In such conditions, these latter signals are not sinusoids anymore (as it was the case in clean recordings) and their phase shift becomes less reliable. In the most severe condition, the best techniques turn out to be PC and RPS whose results are interestingly almost insensitive to an additive noise.

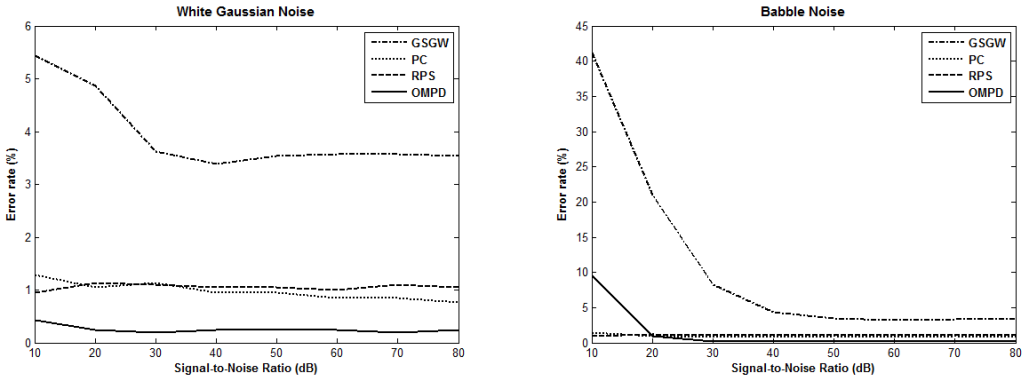

Figure 3: Robustness of polarity detection techniques to an additive noise. *Left panel*: using a white Gaussian Noise, *Right panel*: using a babble noise.

4.3. Robustness to Reverberation

In many modern telecommunication applications, speech signals are obtained in enclosed spaces with the talker situated at a distance from the microphone. The received speech signal is distorted by reverberation, caused by reflected signals from walls and hard objects, diminishing intelligibility and perceived speech quality [15, 16].

The observation of reverberant speech at the microphone is:

$$x(n) = h(n) * s(n), \tag{6}$$

where $h(n)$ is the L-tap Room Impulse Response (RIR) of the acoustic channel between the source to the microphone. RIRs are characterised by the value T_{60}, defined as the time for the amplitude of the RIR to decay to -60dB of its initial value. A room measuring 3x4x5 m and T_{60} ranging $\{100, 200, \ldots, 500\}$ ms was simulated using the source-image method [17] and the simulated impulse responses convolved with the clean speech signals (as used in Section 4.1).

Results of robustness in reverberant conditions are shown in Figure 4. It turns out that the best method in such environments is the PC approach, followed by the proposed OMPD technique which yields about 5% more errors than PC for all values of T_{60}. For these two methods, error rate is increased of around 15% from the less to the most reverberant conditions. RPS provides an almost constant performance across all reverberant configurations, with an error rate of approximatively 35%. The main reason for which PC performs much better than RPS is that it only considers the two first harmonics whose phase is only poorly affected by reverberation. On the opposite, RPS makes use of harmonicity up to 3kHz where the effects of reverberation become relatively more important. Finally, as it was the case for additive noise robustness, GSGW is the most sensitive approach in non-clean conditions since it relies on estimates of the glottal flow.

5. Conclusion

The goal of this paper was two-fold. First the use of higher-order statistics for the automatic detection of speech polarity has been investigated. The proposed technique is based on the observation that these statistical moments oscillate at the local fundamental frequency and have a phase shift which is dependent upon the speech polarity. Secondly, an extensive comparative evaluation of existing speech polarity determination methods has

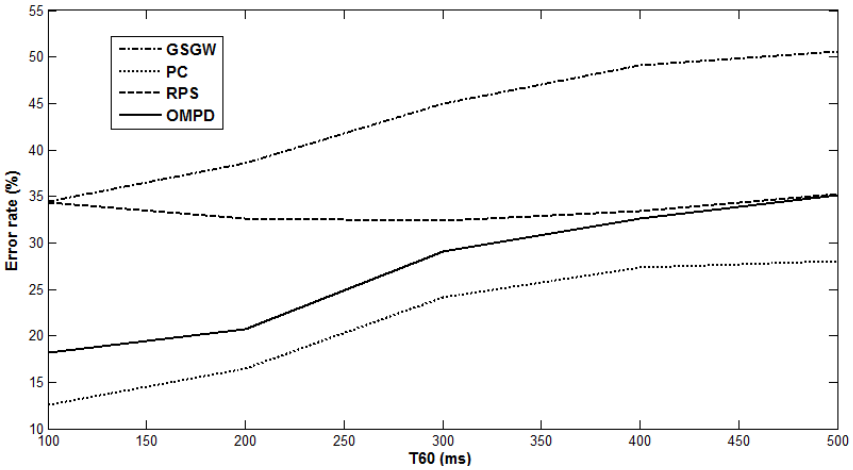

Figure 4: Robustness of polarity detection techniques to reverberation. The higher the T_{60} time constant, the more severe the reverberation.

been carried out on several large speech corpora. On these databases, the proposed approach reached in clean conditions an average error rate of 0.15% against 0.64% for the best state-of-the-art technique. The robustness of these methods to both an additive noise and to reverberation has been studied. In a noisy environment, the proposed approach gave the best results in all conditions, except in the most severe environment with a babble noise at 10dB of SNR. Finally, the Phase Cut approach turned out to be the most suited for being used in reverberant conditions.

Acknowledgments

Authors would like to thank the Walloon Region, Belgium, for its support (grant WIST 3 COMPTOUX # 1017071).

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. Vilkman, 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.

[13] Noisex-92, Online, *http://www.speech.cs.cmu.edu/comp.speech/Sectionl/Data/noisex.html*.

[14] T. Drugman, A. Alwan, *Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics*, Interspeech, pp. 1973–1976, 2011.

[15] R. Bolt, A. MacDonald, *Theory of speech masking by reverberation*, JASA, vol. 21, pp. 577–580, 1949.

[16] H. Kuttruff, *Room Acoustics*, Taylor & France, fourth edition, 2000.

[17] J. Allen, D. Berkley, *Image method for efficiently simulating small-room acoustics*, JASA, vol. 65, no. 4, pp. 943–950, 1979.