INTERSPEECH 2009 BRIGHTON

Use of Harmonic Phase Information for Polarity Detection in Speech Signals

Ibon Saratxaga, Daniel Erro, Inma Hernáez, Iñaki Sainz, Eva Navas

University of the Basque Country, Aholab Signal Processing Laboratory, Bilbao

{ibon, derro, inma, inaki, eva}@aholab.ehu.es

Abstract

Phase information resultant from the harmonic analysis of the speech can be very successfully used to determine the polarity of a voiced speech segment. In this paper we present two algorithms which calculate the signal polarity from this information. One is based on the effect of the glottal signal on the phase of the first harmonics and the other on the relative phase shifts between the harmonics. The detection rates of these two algorithms are compared against others established algorithms.

1. Introduction

The speech signal is asymmetric in amplitude. The polarity of the speech stems from the asymmetric shape of the glottal excitation pulses. When a microphone converts the speech pressure waves into electrical signals, they may end up inversed depending on the electrical polarity connection of the device. Human ear is insensitive to this up-down inversion, but it affects several speech processing techniques, as it is explained next.

In the speech synthesis field, some of the most important state-of-the-art systems are based on selection and concatenation of units taken from a large corpus. If the synthesis corpus is recorded in different sessions or using different recording devices, there may be polarity inconsistencies between different sessions. When two units with different polarity are concatenated a phase discontinuity can appear. Such discontinuities are not perceived by listeners if they occur in unvoiced or low-energy segments, but they are perceptually important if they occur in the middle of vowels, as reported in [1]. Therefore, correct polarity determination would eliminate an important source of synthesis artifacts.

There are also many techniques both for speech synthesis and analysis, which are pitch synchronous. So, they require marking the beginning and end of every pitch period. These pitch marks are used as reference points for segmentation, concatenation and manipulation of speech signals. In order to detect meaningful and consistent pitch epochs, it is usual to search for instants related to the closure of the glottal folds in the larynx of the speaker, which are linked somehow to the positive and negative peaks of the waveform. A usual criterion is to choose either positive or negative local maxima as epochs to mark the pitch period. Polarity inversion obviously affects this criterion, as it converts maxima into minima and vice versa. Hence, it is necessary to ensure common polarity of the signals, to obtain a coherent pitch marking.

Robust polarity detection is also necessary in other areas as data hiding applications [1]. Furthermore, some speech modification techniques based on sinusoidal or harmonic models use phase manipulation procedures that are dependent on the polarity of the signals [2].

In many cases, the polarity of a given speech signal can be visually determined by comparing the sharpness of the positive peaks with that of the negative peaks. However, the waveform distortion introduced by noisy recording conditions or reverberation makes the visual method less reliable, especially for certain voices. If a whole speech database is to be analyzed and it is known that all the recordings have the same polarity, existing automatic polarity detection methods like [3][4] can be applied to each recording separately and a single decision can be taken by counting the number of positive and negative scores. In large databases, this decision is quite reliable regardless of the method. However, in our experience, the same methods can have a higher error rate when applied to automatic polarity determination of separate signals. The need of a robust automatic polarity detector is justified by the importance of increasing the flexibility and portability of speech processing tools such as the above-mentioned ones (synthesizers, analyzers, modifiers…), so that they can be used by anyone with any voice and recording device.

In this paper we propose two new methods for automatic polarity determination based on the phase information of the signals. The reported experiments show that both of them are characterized by a high accuracy and robustness, much better than other existing techniques, being the results consistent for many different voices and recording conditions. Due to their characteristics, the methods are very suitable for speech processing systems based on a harmonic model of speech.

The rest of the paper is structured as follows. In section 2 some theoretical notes about the relation between phase and polarity are presented. Next, sections 3 and 4 explain the basis of each method. Section 5 presents the experiments and results of the evaluation and finally, the main conclusions of this work are summarized in section 6.

2. Relationship between phase and polarity

From a signal processing point of view, the speaking process can be described by a source-filter model [5]. The source signal is the airflow crossing the glottis. In voiced sounds, it can be represented as a train of pulses whose instantaneous amplitude is proportional to the opening area of the vocal cords (Figure 1a). The vocal tract is modelled as a filter $V(z)$ that shapes the glottal source signal in frequency according to the position of the physical articulators, and therefore it is characterized by a number of time-varying resonances.

The speech signal can be seen as the result of filtering the glottal source through the vocal tract and radiating it to the open air. The lip radiation effect can be approximated through a derivative filter $R(z)$. Therefore, an equivalent system is obtained if $R(z)$ is suppressed and the derivative of the glottal source is used as excitation of $V(z)$ (Figure 1b). Since the closure of the glottis causes an abrupt variation in the slope of the glottal source, the excitation signal looks like a train of peaky pulses.

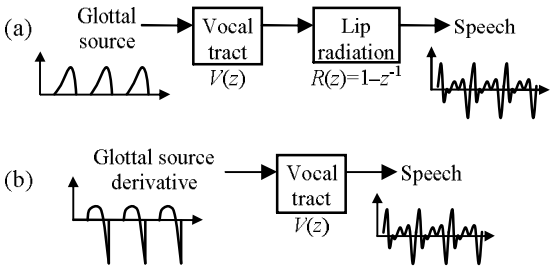

Figure 1: *Diagram of the source-filter model*

According to Fourier's theorem, a steady segment of the excitation signal can be decomposed into a sum of harmonically related sinusoids whose frequencies are multiples of the pitch. If the excitation signal is interpreted as a sum of harmonic sinusoids, the peaks take place at the time instants where the sinusoids are maximally in phase. If the peaks have positive amplitude, the phases of the harmonics are close to zero; if the peaks have negative amplitude, the phases are close to ±π. When the excitation crosses the vocal tract, the phases are slightly modified according to the phase response of $V(z)$. The two polarity detection methods presented in this paper consist of measuring the harmonic parameters directly on the speech signal and inferring whether the underlying excitation (Figure 1b) has positive or negative peaks. They make different assumptions about $V(z)$:

- The Phase Cut (PC) method assumes that the phase contribution of $V(z)$ is negligible at lower frequencies, and searches for the position where the two first harmonics are in phase.
- The Relative Phase Shift (RPS) method assumes that the evolution of the phase response of $V(z)$ along the frequency axis is smooth. It benefits from the fact that when the excitation peaks are positive (the excitation phases are close to zero) the phase increments between harmonics near the peaks are approximately equal to the contribution of $V(z)$.

As the effectiveness of the methods depends on the validity of the assumptions made above, we formulate the hypothesis that they are valid in most of the voiced segments of a given utterance. In this paper, we show that phase-based detection is more robust than other approaches.

3. Phase Cut method (PC)

As it has been mentioned above, the peaks of the excitation signal (the derivative of the glottal source) are the instants where the phase of the sinusoids is maximally close to 0 or π. Assuming that the phases are not drastically modified by the vocal tract, a similar phase structure can be found in the speech waveform. For a small interval centred at a given analysis time instant t_a, the instantaneous phases of the k-th harmonic can be represented as a line passing by the point (t_a, φ_k) with slope equal to $k \cdot 2\pi f_0$, where φ_k is the phase of the harmonic measured at $t=t_a$ and f_0 is the pitch at t_a. We observed that the intersection between the phase lines of the first and second harmonics occurred near 0 or either near π, depending on the polarity of the signal. Figure 4 illustrates this phenomenon.

The PC method consists on determining the phase where the phase lines of the two first harmonics intersect, φ_{cut}. Since the slopes are related by a factor 2, it is immediate to prove that the phase value at the intersection is given by:

$$\varphi \quad \hat{}\varphi \quad \varphi_{:} \qquad (1)$$

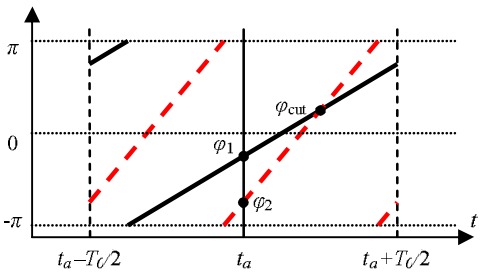

Figure 2: *Instantaneous phase evolution of the first (solid) and second (dashed) harmonics.*

If φ_{cut} is closer to 0, the excitation is considered to have positive peaks; if φ_{cut} is closer to π, the peaks are negative. For a given utterance, a single decision can be taken by processing all its frames separately (at 10ms frame rate) and comparing the number of positive and negative answers. Note that the two lowest harmonics are the best choice for several reasons: they have only one intersection point per cycle, they are less influenced by the vocal tract formants than others, and their high amplitude makes their phase accurately measurable.

4. Relative Phase Shift method

The second method derives from a novel representation of the phase information of the harmonic analysis, described in [6]. Harmonic analysis models each frame of a signal by means of a sum of sinusoids harmonically related to the pitch or fundamental frequency.

$$h\ t\ = \sum_{k=1}^{N} A \quad \acute{\varphi} \quad \grave{} \quad \varphi \quad \acute{} \quad \grave{} \quad \pi \quad \theta_{\Gamma} \qquad (2)$$

where N is the number of bands, A_k are the amplitudes, $\varphi_k(t)$ is the instantaneous phase, f_0 the pitch or fundamental frequency and θ_k is the initial phase shift of the k-th sinusoid.

Usually, the methods used to calculate the parameters of the model give the whole instantaneous phase of every sinusoid, $\varphi_k(t)$, instead of the initial phase shift θ_k. This instantaneous phase changes depending on the analysis instant as well as on the frequency of the harmonic, due to the linear phase term $2\pi k f_o$. On the contrary, the initial phase shift (θ_k) is constant while the waveform shape is stable under the assumption of local stationarity, regardless of the time instant chosen for the analysis.

The initial phase shift determines the waveform shape of the signal. For a given set of harmonic sinusoids the resulting waveform shape depends only on the differences between the initial phase shifts (θ_k) of the components, which we call Relative Phase Shifts (RPS's). These RPS's are also constant as long as the initial phase shifts are so. Thus, they can be calculated at any analysis point wherever local stationarity conditions can be assumed, avoiding the necessity of determining any special point for the analysis. Being relative, the RPS's are computed using a common reference. The fundamental frequency, F0, being the basic harmonic component, constitutes the natural one.

We have developed an expression to obtain the relative differences of the initial phase shifts from the measured instantaneous phases. Let us consider two sinusoids:

$$x_{\_}\ t\ = \quad \acute{} \ \pi_{\searrow\_} \quad \theta \grave{} \qquad \acute{} \ \pi_{\searrow\ldots} \quad \theta_{\kappa}) \qquad (3)$$

where $x_1(t)$ will be the reference sinusoid with frequency f_1 and $x_k(t)$ another sinusoid with frequency $f_k > f_1$. θ_k is the initial

phase shift and t stands for time. For the sake of simplicity we will consider θ_1=0, which implies setting the time origin at the point where $x_1(t)$ has instantaneous phase 0. For any arbitrary analysis point (t_a) the instantaneous phases are:

$$\varphi_1(t_a) = \pi \qquad \varphi_k(t_a) = \pi + \theta_k \qquad (4)$$

In the case of harmonic analysis, f_1 will be the fundamental frequency (f_o) and the frequencies of the two sinusoids will be harmonically related, so $f_k=kf_1$. Applying this condition, we get the relative phase shift (RPS):

$$\theta_k = \varphi_k(t_a) - k\varphi_1(t_a) \qquad (5)$$

Finally the RPS is wrapped to values in the [-π, π] interval.

Among other interesting properties of the RPS's (detailed in [6]) a major feature is that it reveals a structured pattern in the phase information of the voiced segments. This can be noticed in Figure 3 which shows a "RPS phasegram" which, as its magnitude counterpart the spectrogram, shows the evolution along time of the RPS's for each harmonic. Figure 3 shows a phasegram of the voiced speech segment of five sustained vowels |*aeiou*|, where the stable pattern of every vowel can be clearly distinguished. Smooth evolution of the RPS's along frequency agrees with the assumptions of smooth vocal tract frequency response and in-phase glottal excitation.

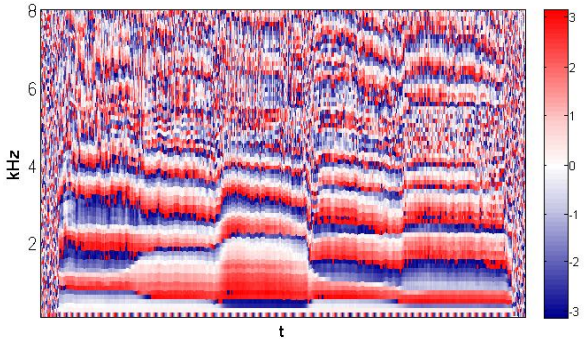

Figure 3: *RPS phasegram of a voiced speech signal /aeiou/.*

This phase structure is sensitive to the polarity inversion as it is shown in Figure 4, where the RPS phasegram of the up-down inverted signal is depicted.

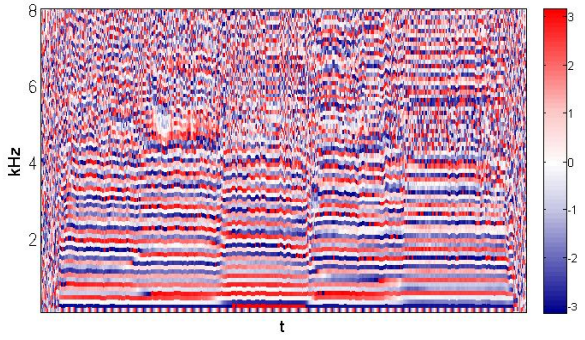

Figure 4: *RPS phasegram of an inverted voiced speech segment /aeiou/.*

The smoothness of the response has disappeared producing an interleaved pattern. If the original signal is inverted, then its instantaneous phases ($\varphi(t)$'s) are shifted π radians. Thus the expression for the phase difference presented in (5), the relative initial phase shift, becomes

$$\theta_k = \varphi_k^i - k\varphi_1^i = \varphi_k - k\varphi_1 - k\pi = \theta_k - k\pi \qquad (6)$$

where φ_k^i and φ_o^i (time dependency of $\varphi(t)$ is omitted for concision) stand for the inverted instantaneous phases. This equation shows that the RPS's are shifted by π but only for the even harmonics while the odd ones remain unchanged. This alternation explains the interleaved pattern in the phasegram, and allows distinguishing between inverted and non inverted speech signals.

4.1. The RPS algorithm

This algorithm exploits the above phenomenon calculating the "ripple" of the RPS's along frequency axis for each analysis frame. The analysis is performed typically every 10ms on the voiced frames only. A Hann window of three pitch period length is used to extract the frame. It uses a previously calculated pitch estimation using the CDP pitch detection algorithm [7].

The RPS algorithm calculates first the RPS's of the frame. To do this we only need to obtain the instantaneous phases of the model. It is not necessary to calculate all the parameters of the harmonic model, which would require solving a quite complex system of equations. Instantaneous phases can be extracted in a straightforward way from the complex spectrum calculated by a Fourier transform. Then the pitch value is used to compute the RPS's of the harmonic components using equation (5), taking the instantaneous phases φ directly from the value of the complex phase of the spectrum at the given frequencies. This way, we get a number of RPS's corresponding to each harmonic component that lies in the frequency spam of the signal. In fact, the analysis is limited to the frequencies below 3 kHz, as higher components are more severely affected by noise.

The ripple of the RPS is then calculated. To do this, the RPS's are first unwrapped. The absolute differences between every harmonic and its predecessor are accumulated. The same calculation is performed with the inverted signal. This is actually done just by adding π (before the wrapping) to the even RPS's.

Finally, both sums are compared. As explained before, the smoother the RPS's are the smaller the sum should be. If the non-inverted frame results smoother than the inverted one, the frame is marked as non-inverted, and vice versa.

This decision is made for every frame of the signal and a single final decision is taken comparing the number of frames in each category.

5. Experiments

5.1. Evaluation databases

Nine databases with known polarity have been used to test the proposed algorithms. Due to the dual decision of the polarity, the probability of correct answers by chance is very high; thereby databases have to be large in order to produce significant results. Moreover, we have selected databases with different features, languages and speakers allowing testing the algorithms under very different conditions. All databases are sampled at 16 kHz. These databases are:

- Karolina & Pello [8]: Female and male voices acted emotional speech database in Basque language. 702 identical sentences in six emotions (happiness, anger, fear, surprise, disgust and sadness) plus neutral for each voice.
- TC-Star Laura DB [9]: Female UK English speaker. Studio quality, neutral style. 5558 sentences.

- CMU_ARCTIC_SLT & CMU_ARCTIC_BDL [10]: Female and male voices in US English with 1132 sentences for each voice, recorded by a female and a male experienced voice talent.
- Berlin DB of Emotional Speech [11]: Acted emotional speech in German language. 10 speakers in six emotions (happy, angry, anxious, fearful, bored and disgusted) plus neutral. 535 sentences altogether.
- Subset of SPEECON Spanish DB [12]: 30 male and 30 female speakers recorded simultaneously by three channels with different SNR, using different microphones and in different locations like cars, offices, public places, etc. Channel C0, was recorded with a close-talk microphone (SNR around 30 dB). C1 was recorded with a Lavalier microphone and C2 with a directional microphone 1 metre away from the speaker (SNR around 15 dB). 1020 sentences per channel.

5.2. Compared polarity detection algorithms

As well as evaluating the two proposed algorithms, we have tested other renowned algorithms so that we get comparative data. These algorithms are:

- GSGW [3]: Implementation of the algorithm for speech polarity determination based on the gradient of the spurious glottal waveforms.
- RAPT [13]: An implementation of the robust algorithm for pitch tracking. This algorithm uses peak detection and dynamic programming to calculate the pitch, and determines signal polarity in the process.

5.3. Results

Experiments have shown (Table 1) that the phase based algorithms (RPS and PC) have very good results in almost every database, outperforming the other methods. They perform equally well both with male and female voices, languages and phonation styles. For these algorithms, in contrast to the RAPT and GSGW, we have not found any voice for which detection performance decays noticeably.

Table 1. *Results of the experiments.*

	RPS			**PC**		
	OK	NOK	Acc. (%)	OK	NOK	Acc. (%)
Karolina	4911	3	**99,94**	4901	13	99,74
Pello	4912	2	**99,96**	4907	7	99,86
Laura	5558	0	**100**	5558	0	**100**
SLT	1132	0	**100**	1108	24	97,88
BDL	1132	0	**100**	1132	0	**100**
Berlin	534	1	**99,81**	524	11	97,94
C0	1003	17	**98,33**	993	27	97,35
C1	1020	0	**100**	939	81	92,06
C2	587	433	57,55	845	175	82,84

	RAPT			**GSGW**		
	OK	NOK	Acc. (%)	OK	NOK	Acc. (%)
Karolina	4810	104	97,88	4590	324	93,41
Pello	1850	3064	37,65	4886	28	99,43
Laura	5557	1	99,98	3454	2104	62,14
SLT	1074	58	94,88	1132	0	**100**
BDL	1111	21	98,14	1119	13	98,85
Berlin	533	2	99,63	274	261	51,21
C0	876	144	85,88	996	24	97,65
C1	760	260	74,51	976	44	95,69
C2	443	577	43,43	927	93	**90,88**

For the clean databases (i.e. excluding C2), the average accuracy is 99,89% for the RPS and 99,19% for the PC. The other methods give lower averages (81,93% for the RAPT and 86,17% for the GSGW).

For the noisy C2 database, the methods which use reduced bandwidth for the analysis (PC and GSGW) limit the noise energy and perform notably better than the methods which use a wider bandwidth. The waveform distortion produced by noise, affects phase information and impacts negatively in the results of RPS and, to a lesser extent, in those of PC.

6. Conclusions

We have presented two methods to detect the polarity of the speech signal using phase information. Experiments prove that this phase information is a reliable indicator of the speech polarity regardless the type of voice, language and speaking style, decreasing the error rate of other methods in 1 or 2 orders of magnitude.

7. Acknowledgements

The authors want to thank the members of the ECESS (http://www.ecess.eu) Consortium for granting the use of the subset of the SPEECON Spanish and TC-Star Laura database. We also want to acknowledge the free use of CMU ARCTIC and Berlin Emo-DB databases.

This work was partially supported by the Avivavoz project, MEC (TEC2006-13694-C03-02/TCM) and the ANHITZ program of the Basque Government (IE06/185).

8. References

[1] Sakaguchi, S., Arai, T. and Murahara, Y., "The Effect of Polarity Inversion of Speech on Human Perception and Data Hiding as Application" Procs. ICASSP00, vol. 2, 917-920, 2000

[2] Erro, D., Moreno, A., Bonafonte, A., "Flexible Harmonic/ Stochastic Speech Synthesis", 6th ISCA Workshop on Speech Synthesis, 2007.

[3] Ding, W. and Campbell, N., "Determining Polarity of Speech Signals Based on Gradient of Spurious Glottal Waveforms", Procs. ICASSP 98, 857-860, 1998.

[4] Legát, M., Tihelka, D., and Matoušek, J. "Pitch Marks at Peaks or Valleys?", LNCS 4629 Text, Speech and Dialogue. 502-507, Springer Berlin/Heilderberg, 2007.

[5] Fant, G., "Acoustic Theory of Speech Production", Mouton, 1960.

[6] Saratxaga, I., Hernaez, I., Erro, D., Navas, E. and Sanchez, J., "Simple representation of signal phase for harmonic speech models", Electronics Letters, vol. 45, Issue 7:381 – 383, 2009

[7] Luengo, I., Saratxaga, I., Navas, E., Hernáez, I., Sanchez, J., and Sainz, I. "Evaluation Of Pitch Detection Algorithms Under Real Conditions", Procs. ICASSP 07, 1057-1060, 2007

[8] Saratxaga, I., Navas, E., Hernaez, I., and Luengo, I., "Designing and Recording an Emotional Speech Database for Corpus Based Synthesis in Basque", Proc. LREC 2006, pp. 2126—2129, 2006

[9] Höge, H., Kacic, Z., Kotnik, B., Rojc, M., Moreau, N. and Hain, H.-U., "Evaluation of Modules and Tools for Speech Synthesis. The ECESS Framework", Procs. LREC 2008, 91-95, 2008.

[10] Kominek, J. and Black, A.W., "The CMU Arctic Speech Databases", SSW5-2004, 223-224, 2004.

[11] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. and Weiss, B., "A Database of German Emotional Speech", Proc. Interspeech 2005, 1517-1520, 2005.

[12] Kotnik B., Höge H., and Kacic Z., "Evaluation of Pitch Detection Algorithms in Adverse Conditions". Proc. 3rd International Conference on Speech Prosody, 149-152, 2006.

[13] Talkin, D., "A robust algorithm for pitch tracking (RAPT)", Speech Coding and Synthesis, Elsevier Science, 495-518, 1995.