DETERMINING POLARITY OF SPEECH SIGNALS BASED ON GRADIENT OF SPURIOUS GLOTTAL WAVEFORMS

Wen Ding and Nick Campbell

ATR Interpreting Telecommunications Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
E-mail : ding@itl.atr.co.jp

ABSTRACT

Speech polarity is crucial in many speech processing fields. We present a novel method to determine polarity of speech signals from gradient of spurious glottal waveforms. We use the iterative adaptive LPC inverse filtering to cancel effect of vocal tract transfer function while maintaining the most properties of source excitation. Then we take the first-derivative (gradient component) of spurious glottal waveforms to capture the sharp gradient near the glottal closure instant. By using the gradient components of the spurious glottal waveforms, we detect speech polarity, *i.e.*, the polarity of glottal waveforms, by finding whether the glottal closure instants are located above or below the zero-line. Furthermore, a frame-based decision technique is applied to get robust results. Experimental results with a wide variety of speech utterances reveal a high performance and the computation complexity is much more less than a previously proposed method.

1. INTRODUCTION

Polarity of speech signals may be disturbed (up-down inversed) by the recording equipment or in transition channels. In the research areas of speech processing such us epoch detection based on the instants of glottal closure [1], joint estimation of voice source and vocal tract parameters [2], two-channel speech (voice and EGG signals) analysis [3], pitch mark detection of PSOLA synthesis technique [4], etc., polarity detection of speech waveforms should be considered. It is obvious that for epoch detection using glottal closure instants of LPC inverse-filter signals, unreliable results can be obtained if speech signals with negative polarity (up-down inversed) are analyzed. In our phoneme-based concatenative speech synthesis system [5], the speech databases may be collected from a wide range of sources. Special attentions should be paid to speech polarity, since pitch marks extracted from speech signals with positive polarity are different to those of negative polarity. The connection of speech units with antithetic polarity may cause the boundary discontinuity in the synthesized speech.

The definition of polarity stems from the glottal volume velocity waveform generated from the vocal cords, which has an asymmetric structure, i.e., a round curve from glottal opening instant following a sharp return near *glottal closure instant* (*GCI*) as shown in Fig. 1. The derivative glottal waveform is used to approximate the lip radiation characteristics when we assume the speech production process to be a linear combination of voice source, vocal tract and lip radiation. Positive polarity means that *GCI*s are in their normal positions, i.e., *GCI*s are located below zero-line as the negative peaks of the derivative glottal waveforms.

This asymmetric attribute of the glottal waveforms can also be observed in the speech signals. In most cases, instead of auto-detection, the polarity of speech signals are usually judged by visually inspecting the speech waveform. But this may be a tedious processing when a large amount of speech data present and introduce unstable subjective factors. A method has been proposed to estimate pitch epochs and polarity from speech signals with dynamic programming technique, but it suffers from a time-consuming search [6].

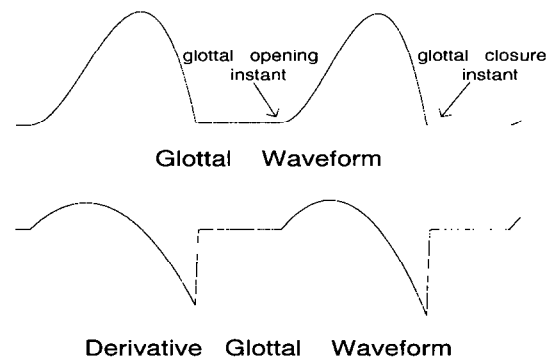

Figure 1. Description of typical glottal volume velocity waveform.

We aim to realize a reliable and fast speech polarity detection based on the asymmetric attribute of spurious glottal waveforms. In order to get the glottal waveform, the IAIF (Iterative Adaptive Inverse Filtering) method has been proposed [7]. We modified and implemented it as one part of our proposed method. The output of inverse filtering provides a noisy derivative glottal waveform, and a low-pass filter is used to cut the noise component. Then we take gradient of the spurious glottal waveform and capture the position of the glottal closure instant. Finally we make decision by checking whether the glottal closure instant is located in its normal direction or not.

2. ANALYSIS FRAMEWORK

The overall framework of detecting speech polarity is illustrated in Fig. 2. The strategy is to find the polarity of the spurious glottal waveform which is the same of the corresponding speech polarity. In a brief summary of the process, the recorded polarity-unknown speech signal $s(n)$ is analyzed by LPC and the coefficients of the vocal tract transfer function is obtained. After inverse filtering, the

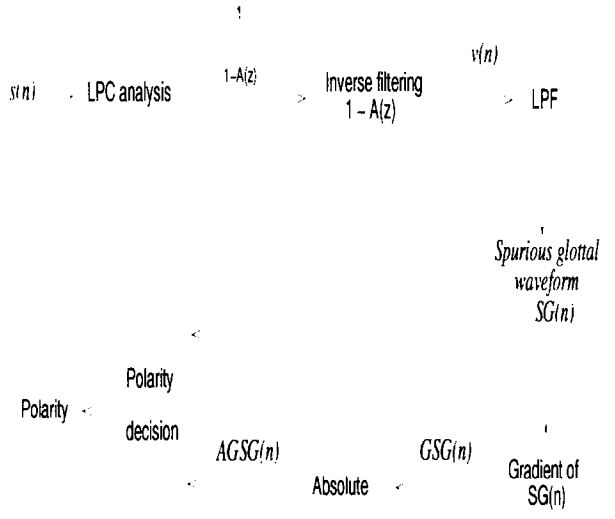

Figure 2. Framework of polarity detection approach.

noisy output $v(n)$ is filtered by a low-pass filter to cancel its noise component and the spurious glottal waveform $SG(n)$ is obtained. Then we take the first derivative (gradient component) of $SG(n)$ and get the absolute value of the gradient spurious glottal waveform $AGSG(n)$. Finally, the polarity of $SG(n)$ is determined using both $SG(n)$ and $AGSG(n)$. All these procedures will be described detailed in the following sections.

2.1. SPURIOUS GLOTTAL WAVEFORM

To do LPC analysis, speech production process is assumed to be an auto-regressive (AR) model, also known as LPC model:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n), \qquad (1)$$

where $s(n)$ is a speech sample at time n, $u(n)$ is a normalized excitation and G is the gain of the excitation. The coefficients $a_1, a_2, \cdots, a_p$ are assumed constant over a predefined analysis frame. By taking the z-transform of (2) we get the transform function

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{1 - A(z)}. \qquad (2)$$

The digital filter provides approximation of the vocal tract acoustic properties over the analysis period.

Inverse filter analysis is employed to devise a filter operating on the acoustic speech signal which reverses the transform performed by the vocal tract, revealing glottal airflow signal [8].

$$v(n) = s(n) + a_1 v(n-1) + a_2 v(n-2) + \cdots + a_p v(n-p). \quad (3)$$

The IAIF (Iterative Adaptive Inverse Filtering) method [7] has been proposed to get the inverse-filtered signals from speech. I In order to speed it up, the method is modified by just taking the preemphasis of speech signals instead of 1st-order LPC and inverse-filtering. The whole procedure is illustrated in Fig. 3. It proceeds in a pipe-in and pipe-out way and the first step is replaced by only using the

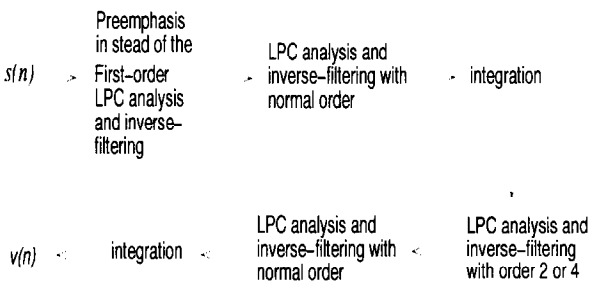

Figure 3. Block diagram of the modified IAIF approach.

preemphasis of speech signals. This modification has been shown to produce a reliable estimation while reducing the computation compared to the original one.

But $v(n)$ is usually noisy because the uncorrect inverse filter settings and the noises embedded in the speech signals. To remove the noise component of $v(n)$, a low-pass filter (LPF) with cutoff frequency 1000 Hz is designed to cut off the noise frequency. At this moment, the filtered signal is called *spurious glottal waveform* $SG(n)$ since it contains the main waveform properties of the glottal waveform. Figure 4 (b) and (c), respectively, shows an example of the inverse filtered waveform and the spurious glottal waveform after eliminating the noise component of $v(n)$. We can see that $SG(n)$ is stable in its low frequency domain and has distinguishing waveform feature (sharp gradient) near the glottal closure instant. The next step is to characterize the sharp gradient and to determine the polarity that is whether the position of the glottal closure instant is above or below the zero-line.

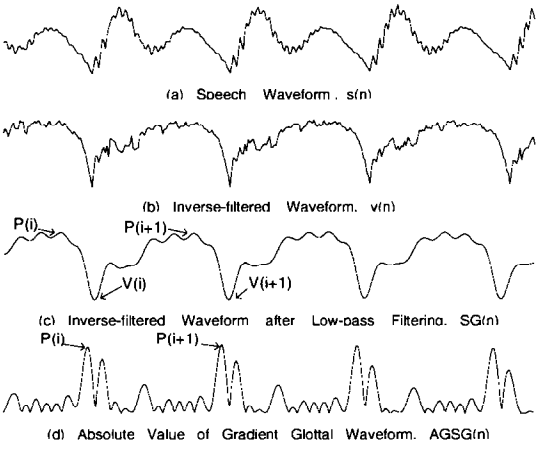

Figure 4. (a) Speech waveform $s(n)$ with positive polarity, (b) inverse-filtered waveform $v(n)$, (c) the spurious glottal waveform $SG(n)$, and (d) the absolute gradient value of $SG(n)$.

2.2. POLARITY DECISION

In order to judge the polarity of the glottal closure instant, we take the absolute value of gradient of the spurious glottal waveform $AGSG(n)$ as follows,

$$AGSG(n) = abs(deriv(SG(n))) \qquad (4)$$

$$= abs(SG(n) - SG(n-1)). \qquad (5)$$

An example of the gradient waveform of the spurious glottal waveform is shown in Fig. 4 (d).

It can be observed that the prominent peaks of $AGSG(n)$, $P(i), P(i+1), \cdots$, occur around the glottal closure instants of SG, $V(i), V(i+1), \cdots$. This is because that the gradient component near the glottal closure instant is much larger than other places in $SG(n)$. We deduce from the above observation that the speech polarity is positive when the peak of $AGSG(n)$ is closer to the negative peak of SG than the positive peak, and vice versa.

The pitch-synchronous polarity decision can be formulated as:

$$\mathrm{P} = \begin{cases} positive & \text{if } dist(P(i)_{AGSG(n)}, V(i)_{SG(n)}) < \\ & \quad dist(P(i)_{AGSG(n)}, P(i)_{SG(n)}) \\ negative & \text{otherwise} \end{cases} \quad (6)$$

where P denotes speech polarity of a pitch period, $dist(P(i)_{AGSG(n)}, V(i)_{SG(n)})$ denotes the time distance between the peaks $P(i)$ of $AGSG(n)$ and the negative peaks $V(i)$ of $SG(n)$. The pitch and voice/unvoice features are extracted with a normal method described in [2]. Now the pitch-synchronous polarity decision can be made based on the above equation. But according to the phonation variation in an utterance, there may exist some segments without the sharp return phase in the source excitation , e.g., the end part of an utterance, the polarity detection of these segments becomes quite difficult. Thus we have to consider an approach to deal with this kind of situation. In the next section, we adopt a frame-based technique to get the robust decision.

2.3. FRAME-BASED DETECTION

In one utterance, the polarity is a constant value everywhere. If we can find a criterion that can be used to search for the most reliable segment in the utterance for polarity decision, then the other segments can be ignored. But unfortunately, making such a criterion seems to be a quite difficult task. As an example, Fig. 5 illustrates a transition segment of a male utterance with negative polarity. In this case, there are two main peaks of $AGSG(n)$ before and after the transition part, and both of these peaks are close to the positive peaks of spurious glottal waveforms $SG(n)$. But in the transition part, the peaks of $AGSG(n)$ become quite ambiguous to make a reliable decision. Due to the phonation type and its dynamic changes in human speaking, this kind of tough segments alway exist.

In order to reduce the influence of such unreliable segments on polarity detection, a robust approach considered in this paper is to separate the utterance into many frames and make the frame-based decision using (6). For a frame with several pitch periods, the polarity decision of the frame is made by the majority of decisions of pitch-synchronous positive and negative polarity. Then each frame has a value of polarity. For the whole sentence, the polarity is made using the same philosophy, the majority of the frame numbers with either positive or negative polarity. In such a way, one unreliable segment with many pitch periods may produce only one bad decision and has little influence on the final decision.

$$\mathrm{Polarity} = \begin{cases} positive & \text{if } count(Positive_{(frame)}) < \\ & \quad count(Negative_{(frame)}) \\ negative & \text{otherwise} \end{cases} \quad (7)$$

where $Polarity$ means the final judgement of the whole sentence. $Positive_{(frame)}$ means the frame detected to be positive polarity (positive frame). $Negative_{(frame)}$ means the frame with the negative polarity (negative frame). Then we *count* the number of the positive and negative frames, and the polarity of the sentence is decided based on the majority. The frame length is $N_{(frm)}$ and the shift length is $N_{(shft)}$ as shown in Fig. 5.

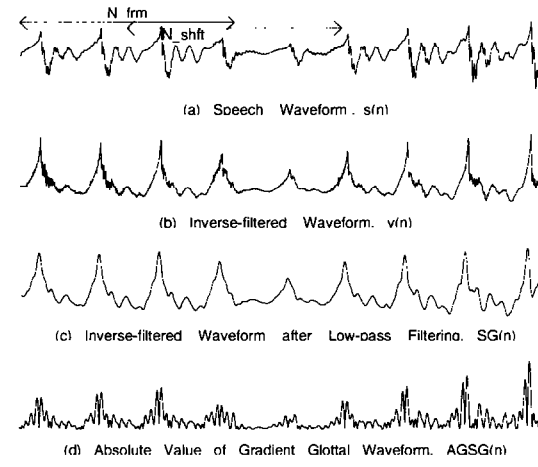

Figure 5. (a) Speech waveform $s(n)$ of a male voice with negative polarity, (b) inverse-filtered waveform $v(n)$, (c) the spurious glottal waveform $SG(n)$, and (d) the absolute gradient value of $SG(n)$.

3. RESULTS AND DISCUSSIONS

Since the behavior of source excitation of a female voice (short return phase) are usually quite different from that of a male voice (relatively long return phase), it is meaningful to inspect an example. Figure 6 shows a segment of a female voice with positive polarity. It can be seen that the spurious glottal waveform has a round curve (short return phase) near the GCI. Since the gradient around the GCI is still very strong than the glottal open phase, the speech polarity still can be detected as positive.

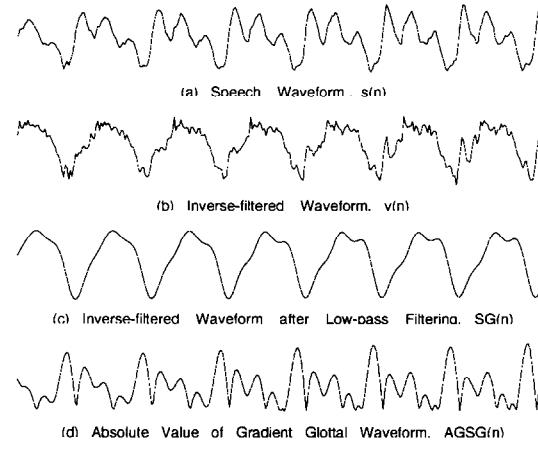

Figure 6. (a) Speech waveform $s(n)$ of a female voice with positive polarity, (b) inverse-filtered waveform $v(n)$, (c) the spurious glottal waveform $SG(n)$, and (d) the absolute gradient value of $SG(n)$.

In order to evaluate the proposed method, language/speaker variations and recording environments are

considered in the experiments. We use several ATR speech databases of English/Japanese spoken by male/female speakers, which were collected from varied sources. Part of the speech signals were sampled at 12 kHz, and the others were sampled at 16 kHz. The polarity of all the sentences has been checked by human inspection beforehand and used in the evaluation. We set $N_{frm} = 80ms$ and $N_{shft} = 40ms$, used in (7).

Table 1 gives a fragment of frame-based polarity decision (positive v.s. negative), from which we can investigate the stages of how the polarity decision being made. From the table, it can be observed that the final decision is achieved based on the majority of the frame numbers in positive or negative polarity. This indicates that even if some segments are difficult/failed to be detected or even with a wrong pitch estimation, the final result can be correct using (7).

Table 1. A FRAGMENT OF FRAME-BASED POLARITY DECISION OF SPEAKER *sally*. FROM LEFT TO RIGHT : FILE NAME, NUMBER OF FRAMES JUDGED TO POSITIVE POLARITY, NUMBER OF FRAMES JUDGED TO NEGATIVE POLARITY, THE FINAL RESULT OF THE SENTENCE.

:
sc140.wav 67 : 4 Positive
sc141.wav 67 : 5 Positive
sc142.wav 60 : 4 Positive
sc143.wav 57 : 5 Positive
sc144.wav 42 : 2 Positive
sc145.wav 47 : 8 Positive
sc146.wav 55 : 8 Positive
sc147.wav 55 : 5 Positive
sc148.wav 57 : 9 Positive
sc149.wav 52 : 2 Positive
sc150.wav 48 : 11 Positive
sc151.wav 32 : 8 Positive
:

The results of all the databases are shown in Table 2. Based on the above frame-based results, every speaker has a score in sentence number of correct and wrong decision. The accuracy of one speaker is computed as the percentage of the correct decisions divided by the number of the utterances. The results show a high performance obtained by the proposed method. Through the investigation of those wrong-decision sentences, it is shown that the cause may come from the soft-like voices and the incorrect inverse-filtering settings.

The average running time of the algorithm in sentence level is 0.6 real-time, that is nearly 10 times faster than the method in [6], on a SUN sparc 20 workstation.

4. CONCLUSION

We have proposed a fast and robust method to deal with the detection of speech polarity using gradient of LPC inverse-filter signals, which takes the advantage that there is usually a sharp gradient of the spurious glottal waveform near GCI. The modified adaptive inverse-filter approach was shown to possess the ability of providing a stable estimation of the spurious glottal waveform. The sharp gradient was captured by using $AGSG(n)$. Robust estimation was realized by using frame-based decision. From our experiments, we used multi-language, multi-speaker speech databases and

Table 2. POLARITY DETECTION RESULTS OF MULTI-LINGUAL, MULTI-SPEAKER SPEECH DATABASES.

LANGUAGE	SPEAKER	UTTERANCE	ERROR*	ACCURACY
English	gsw	200	3	98.5
	sally	200	0	100
Japanese	MHN	503	0	100
	MHT	503	0	100
	MTH	378	14	99
	FKT	503	14	99
	FMP	503	0	100
	FKS	503	0	100

ERROR* means the utterances with the wrong decision.

the results showed a reliable detection of speech polarity with the frame-based decision. The results also indicated the fact that since the glottal waveforms change a lot according to the voice phonation or speaking style, the accuracy of detecting speech polarity based on inverse filtering may be affected by these factors. Voices of normal phonation and stress or pressed phonation type usually possess the typical glottal waveform with a sharp return phase near GCIs, while the breathy and harsh voice usually has a round curve around GCIs. Therefore, the proposed method is expected to yield a robust and accurate estimation of the polarity of normal or pressed voices.

ACKNOWLEDGMENTS

The authors would like to thank Christophe d'Alessandro for his helpful discussion and providing the IAIF approach.

REFERENCES

[1] C. X. Ma, Y. Kamp, and L. F. Willems, "A Frobenius norm approach to glottal closure detection from the speech signal", *IEEE Trans. SPEECH & AUDIO PROCESSING*, Vol.2, pp.258-264, 1994.

[2] W. Ding, H. Kasuya and S. Adachi, "Simultaneous estimation of vocal tract and voice source parameters based on an ARX model", *IEICE Trans. Inf. & Syst.*, Vol. E78-D, No. 6, pp. 738-743, 1995.

[3] A. K. Krishnamurthy and D. G. Childers, "Two-channel speech analysis", *IEEE Trans. ASSP*, Vol. 24, pp. 730-740, 1986.

[4] C. Hamon, E. Moulines, F. Charpentier, "A diphone synthesis system based on time-domain modification of speech", *Proc. ICASSP*, Glasgow, 1989.

[5] W. N. Campbell & A. W. Black, "Prosody and the selection of units for speech synthesis", *Progress in Speech Synthesis*, eds Santen et al, , Olive, Hirschberg & Sproat, Springer New York, pp. 279-292, 1996

[6] T. David, etc., Entropic Research Laboratory, Inc., 1993.

[7] P. Alku, "Glottal wave analysis with pitch-synchronous iterative adaptive inverse filtering", *Speech Communication*, Vol.11, pp. 109-118, 1992.

[8] P. Milenkovic, "Glottal inverse filtering by joint estimation of an AR system with a linear input model", *IEEE Trans. ASSP*, Vol. 34, pp. 28-42, 1986.