一種用於語音信號線性預測殘留以提取 聲門關閉瞬間的過濾法

胡懷祖

國立宜蘭技術學院電子工程系

摘要

本文提出一種以過濾方法來告示潛藏於低通過濾後之殘留信號中的重大變動訊息。由於聲門關閉瞬間是以負波峰的型態出現於處理過的殘留信號，我們試圖增強和搜尋各式語音信號的負波峰，納入實驗驗證的案例則以一些過去引發問題的單音節為主，這包括有 /u/、 /m/、 /n/、 /z/、 /α^{w}/、 /L/ 與 /h/。而這種方法於白雜訊場合的穩健性亦是我們的探討重點，其效能是以一個低音階的母音 /u/ 於 10 dB 信噪比的情況作為示範，若與其它三種方法比較，我們所提出的方法最能在聲門關閉瞬間呈現明顯對比。此外，集合 24 句語料庫所得到的實驗結果亦指出此法在一般的信噪比環境下的工作情形令人滿意。

關鍵字：聲門關閉瞬間、線性預測殘留、過濾。

A filtering method for extracting glottal closure instants from linear prediction residual of speech signal

Hwai-Tsu Hu

Department of Electronic Engineering, National I-Lan Institute of Technology

Abstract

A filtering method is proposed to signify the epochal feature residing in the lowpass filtered residual. While the glottal closure instant (GCI) manifests itself as a negative peak in the processed residual, our method aims at the enhancement and retrieval of these peaks for a large variety of speech signals. The voiced sounds considered in our experiments include /u/, /m/, /n/, /z/, /α^{w}/, /Λ/, and /η/, all of them were reported to be problematic. The robustness in the presence of white Gaussian noise is also under our investigation. Its performance is demonstrated by trying out a low-pitched vowel /u/ with SNR = 10 dB. Compared with three other methods, the proposed method produces an evident contrast at the GCI's. Furthermore, the results based on a database consisting of 24 sentences indicates that this method works well in a moderate SNR environment. The accurate determination of GCI's not only helps with the

extraction of acoustic features of speech signals, but facilitates the application of a pitch synchronous approach to speech processing.

Key words: glottal closure instant, linear prediction residual, filtering.

I. Introduction

Linear predictability of speech signals has long been considered an informative source to find the glottal closure instants (GCI). Such an idea emerges from the observation that the main excitation for the vocal tract system comes from the abrupt closure of glottis. Since an autoregressive process adequately models the vocal tract system, speech samples except for those at GCI's are describable by a set of linear equations. As a result, the prediction residual obtained by inverse filtering the speech signal often exists sharp epoch pulses coinciding with the GCI's [1,2]. Methods such as the Frobenius norm [3] and determinant of the autocovariance matrix [4,5] originated from the same understanding. Attempts were also directed to detect abrupt changes from a statistic viewpoint [6].

The above introduction seems to make the prediction residual very attractive in identifying the GCI's. However, the epoch pulse may be obscured by other spurious subpulses due to various reasons, such as noise contamination, vocal turbulence (e.g., voiced fricative or sounds with incomplete glottal closure), improper modeling (e.g., nasal-tract coupling), and the lack of discriminative predictive errors (e.g., sinusoidal-like waveform). The confusion of the epoch excitations with other subpulses makes the unambiguous identification of epochs a very difficult task. Thus, restrictions of certain vowels are often imposed on the application of the residual signal. We believe that the restriction is removable through a proper interpretation of the residual signal. To overcome such a drawback, we consider the use of filtering operations to signify the most important features of the residual signal before extracting GCI information. Moreover, the proposed epoch determination algorithm should cover a large variety of speech signals.

II. Usefulness of lowpass filtered residual

In recent years many speech coders accept the lowpass filtered residual or its

variants as a prevailing tool to acquire pitch periods [7-9]. This is because the lowpass filtered residual conveys the information of glottal activities while eliminating formant resonances to allow clear pitch estimation. To determine the useful features of the lowpass filtered residual, it is instructive to verify the relationship between the residual and the glottal flow function. According to the source-filter theory [10], the production of voiced speech is the result of passing the glottal flow through a filter retaining a transfer function of the vocal tract system and lip radiation, i.e.,

$$S_v(z) = G(z)V(z)R(z), \qquad (1)$$

where $S_v(z)$, $G(z)$, $V(z)$, and $R(z)$ are the z-transforms of the voiced speech signal, glottal pulse, vocal tract system, and lip radiation, respectively. As the effect of lip radiation is generally simulated by a differentiator, an equivalent representation for the speech production becomes

$$S_v(z) = G^{(1)}(z)V(z), \qquad (2)$$

where $G^{(1)}(z)$ denotes the z-transform of the differentiated glottal pulse.

Supposed that $S_v(z)$ is corrupted by additive noise $W(z)$ such that the resultant noisy observation $\hat{S}_v(z)$ becomes $S_v(z)+W(z)$, where $W(z)$ may refer to noise excitation, phase diversion, predictive as well as modeling errors. Given that the inverse filter $\hat{A}(z)$ is derived by carrying out LP analysis of $\hat{S}_v(z)$, we decompose the reciprocal of $\hat{A}(z)$ into two components: one relates to the spectral tilt of the differenticated glottal flow, termed $\hat{T}(z)$, and the other corresponds to the

estimated vocal tract system, termed $\hat{V}(z)$. Hence the inverse filtering operation is represented as

$$\hat{A}(z)\hat{S}_v(z) = \frac{G^{(1)}(z)V(z)}{\hat{T}(z)\hat{V}(z)} + \frac{W(z)}{\hat{T}(z)\hat{V}(z)}. \qquad (3)$$

In the above equation, we may assume that the estimated $\hat{V}(z)$ roughly matches its counterpart $V(z)$ and $\hat{T}(z)$ can be approximated by a lowpass filter $I(z)$. Here we choose $I(z)$ as $1/(1-0.95z^{-1})$ to render a spectal roll-off rate of –6 dB/octave. By multiplying $I(z)$ on both sides of eqn. 3, we have

$$I(z)\hat{A}(z)\hat{S}_v(z) \approx G^{(1)}(z) + \frac{W(z)}{\hat{V}(z)}. \qquad (4)$$

It is readily seen that the left-hand-side of eqn. 4 explicitly denotes a lowpass filter residual (LFR). The righ-hand-side of eqn. 4 includes the differentiated glottal pulse apart from a noise term. Since the negative peak of a differentiated glottal pulse is an evident indicator of a GCI [11], the GCI determination turns out to be a procedure of seeking large negative peaks across the LFR. Furthermore, as the energy of the differentiated glottal pulse concentrates in the low frequency region, the waveshape of the LFR would not be seriously disturbed unless the noise accumulates enough energy at low frequencies.

III. Modifications in deriving lowpass filtered residual

Though the foregoing analysis establishes a sensible theoritic basis for GCI determination, our experiments with respect to real speech data reveal that the derived LFR may not show distinct negative peaks. The reason can be attributed to the

assumptions previously made to derive the LFR. For example, the assumption of $V(z)$ to be an all-pole filter is inadequate for analyzing nasal-like sounds. The assumption that separates $\hat{A}(z)$ into $\hat{V}(z)$ and $\hat{T}(z)$ may be invalid for vowels exhibiting strong deceleration of glottal flow during the closing glottis. Moreover, the contaminant noise may also lead to a severe mismatch between $A(z)$ and $\hat{A}(z)$.

Thus, we intend to render a signal more suitable for GCI detection by filtering the LFR. An overlap-and-add approach is adopted to facilitate the filtering operations. The overlap-and-add algorithm starts from segmenting the speech signal into frames of 240 samples. Each frame is weighted by a Hanning window and overlapped 50% with adjacent frames. A series of filtering operations, as will be discussed shortly, is then applied to the LPR to achieve sharper negative peaks. Advantages due to the overlap-and-add approach are multi-fold. As the Hanning window gently declines to zero on frame boundaries, there is no need to concern about filter initialization. Likewise, the contribution due to filter memory can be neglected provided that the filtering process does not leak too much energy into neighboring frames. In other words, no special care is needed to minimize the startup and ending transients. Consequently, the overlap-and-add approach lowers the complexity of the system when we apply filtering operations on a frame-by-frame basis. In addition, the smooth transition inherited in the process of overlap-and-add facilitates the adjustment of the dynamic range of the filtered output. Before we pack these overlapped frames together, the power of the processed signal is always adjusted to be unity.

Following the signal segmentation and windowing by the overlap-and-add technique, there are several modifications when deriving the LFR. First, a notch filter is employed to remove the d.c. component and low-frequency drift from the speech signal, thus yielding a zero-mean prediction residual. Second, we use a first-order highpass filter, $1-0.925z^{-1}$, to deemphasize the low frequency spectrum of the underlying speech signal. We note that the purpose of preemphasizing the

speech signal is somewhat different from that originally adopted in speech analysis. Our intent is to reduce the strength the glottal flow and, sometimes, the first formant so that the subsequent linear prediction analysis does not take the spectral tilt of the glottal flow into account. This, in turn, helps the separation of glottal spectral tilt, $\hat{T}(z)$, and the vocal tract system, $\hat{V}(z)$.

Third, prior to deriving the residual signal, we damp the coefficients of the inverse filter by a factor of 0.95. This is equivalent to passing the actual residual with a filter $Q(z)$ as

$$Q(z) = \frac{A(z/0.95)}{A(z)}. \qquad (5)$$

This filter especially reinforces the spectrum within the formants of narrower bandwidths. The damping effect alleviates the problem of inverse filtering speech signal with sinusoidal-like waveforms since it partially retains the sinusoidal characteristics.

Fourth, another notch filter with a broader bandwidth is employed to mildly attenuate the low-frequency fluctuation of the differentiated glottal flow. This appears imperative in our experiments. Depending on the phase characteristics of the residual signal, the swinging range of the LFR may extend below the negative peaks. The attenuation of low frequency components can make the negative peaks more distinguishable. Fig. 1 illustrates the effect due to the participation of a broadband notch filter. The speech signal under investigation is a segment of /o/ uttered by a male.

Fifth, while these negative peaks of the LFR is probably smeared by strong comtaminant noise, we ameiorate this problem by using a lowpass filter with a gradual spectral falloff. Though a narrow-band lowpass filter with an abrupt spectral tilt may effectively subdue the noise, we do not recommand this kind of filters since it increases the risk of smoothing out the negative peaks.

As the above steps are accomplished by means of filtering, a composite filter $C(z)$ with all the functions given in step 2 to 5 can represent the entire filtering task. In particular, we would like the phase characteristics to be zero so that the filtering opertions do not affect the locations of the negative peaks. Consequently, we process the residual forward and then backward using the same filter. In our experiments, the composite filter $C(z)$ is chosen as

$$C(z) = \left[\frac{A(z/0.95)}{1-0.95z^{-1}}\right]\left[\frac{\left|1-0.4z^{-1}\right|^2}{\left|1-0.6z^{-1}\right|^2}\right]\left[\frac{1}{\left|1-0.5z^{-1}\right|^2}\right], \qquad (6)$$

where the part given in the first brackets indicates the damped inverse filter associated with lowpass filtering, the part in the second brackets denotes the broadband notch filter, and the third part represents a lowpass filter with a broad bandwidth.

Among the foregoing five steps, the damping effect, in combination of the highpass emphasis, offers a solution to identify GCI's from the sounds with no abrupt discontinuity in the glottal flow. A typical example is given in Fig. 2. The speech signal under investigation is the ending section of a diphthong $/a^{w}/$ uttered by a female. The Fourier spectrum of the speech waveform possesses a predominant sinusoidal frequency with no obvious harmonics. For this type of speech, the performance of linear prediction is simply too good to render any evident indication for GCI detection. However, with the involvement of these two modifications, the resulting LFR still maintains a similar waveform with negative peaks. Again, to maintain the congruity with our former definition of GCI's, we take the periodical minimums of the modified LFR as the GCI's.

IV. Description of the GCI determination algorithm

Our GCI determination algorithm begins with the linear prediction and inverse

filtering of speech signals so as to render a modified LFR. Fig. 3 summarizes all the required filtering operations. As mentioned earlier, searching for the negative peaks of the modified LFR is regarded as an effectual way to detect the GCI's. However, partly due to the fluctuating nature of glottal flow and partly due to the phase distortion of inverse filtering, the modified LFR occasionally exhibits false peaks that confound with the epoch peaks. In order to reduce the chance of picking a wrong GCI, the proposed algorithm is divided into two stages. In the first stage, we take an interval comprising the current frame and extra 80 nearby samples of adjacent frame to perform coarse pitch estimation using the average magnitude difference function [12]. A procedure similar to comb filtering [13] in noise reduction techniques is then adopted to make the largest negative peak more discernible. We take this negative peak as the first identified GCI. The rest GCI's are then obtained by finding the minimum locations approximately one pitch period away from the already identified GCI's. This procedure continues until the searching range is out of the current frame by 20 samples.

V. Performance evaluation

The efficiency of the proposed method is illustrated by showing the experimental results with respect to speech signals that usually cause problems by other methods. Cases considered here includes the phoneme transition between /Λ/ and /ŋ/, two nasal consonants /m/ and /n/, a voiced fricative /z/, and, and a vowel /u/ which contains no abrupt glottal closure. All the utterances were sampled at 8 KHz with 16 bit resolution. The linear prediction order is set to 10. During the phoneme transition from /Λ/ to /ŋ/, the resulting LFR exhibits narrow peaks at GCI's (see Fig. 4), thus leading to an easy and unambiguous detection of the GCI's. Also note that the LFR holds greater similarity than the corresponding speech signal for different

phonemes. In this case the effectiveness due to the use of the modified LFR is observed.

For the nasal consonant /m/ presented in Fig. 5, the modified LFR seems not affected by nasal-tract coupling since distinct negative peaks has been shown at the GCI's. We note that the importance of using the residual instead of the speech signal can be justified based on the result presented in Fig. 5. One may argue that the GCI's can be found by directly seeking the negative peaks just in front of large positive peaks of the speech signal using simple logical relations. This argument is quite persuasive for the vowels that exhibit abrupt discontinuity in the glottal flow, but it is invalid for this type of speech. Fig. 5 demonstrates such a discrepancy very well. On the other hand, the performance for the nasal /n/ is presented in Fig. 6. In this case, it is almost impossible to identify the epochs directly form the residual signal. The modified LFR, in contrast, shows clear negative peaks around GCI's. For the voiced fricative /z/, the excitation incorporates both voiced and unvoiced sources due to incomplete closure of glottis. As shown in Fig. 7, though the widths of the resulting negative peaks in the modified LFR are not so spiky as those observed in other voices, the GCI's can still be identified.

The last voiced sound considered in the illustration is a high-pitched vowel /u/. The resulting modified LFR for the vowels /u/ is noted in Fig. 8. As pointed by other researchers, the lack of abrupt discontinuity in the glottal flow has been known a great obstacle, but the modified LFR does not seem to encounter such a problem.

In addition to the problematic cases discussed above, we also concern about the noise sensitivity of the modified LFR. A low-pitched vowel /u/ serves as a representative example since it provides sufficient evidences regarding the advantages of the modified LFR. In this example, we contaminate the vowel /u/ by adding white Gaussian noise with signal-to-noise ratio (SNR) set to 10 dB. For the sake of comparison, we also present the results obtained from three other methods, namely, the epoch filtering of linear prediction residual (EFLPR) [2], the maximum likelihood epoch determination (MLED) [14], and the Frobenius norm (FN) [3]. The

autocorrelation method of order 10 are adopted to carry out the LP analysis. Settings required by the other three methods follow those given in the original papers. Furthermore, in order not to mislead the reader's judgement, procedures that were previously employed to produce more pulse-like signals are disallowed.

Fig. 9 shows the results of GCI determination for the noise-contaminated vowel /u/. It is seen that the EFLPR simply fails to offer any indication. The FN responses with pulse-like signals near the GCI's, but a certain degree of ambiguity exists between epoch pulses and other subpulses. On the other hand, the MLED provides rather useful information by showing hunches around the GCI's. The performance of the modified LFR is also satisfying, although it exhibits less clear negative peaks around GCI's.

The ability of signifying the primary excitation pulses in moderate SNR's demonstrates the robustness of the proposed method. For most cases encountered in a noise-free or moderate SNR environment, the modified LFR is quite efficient for detecting the GCI's. However, the derived LFR still suffers more or less purturbation in the presence of strong noise. Cheng and O'shaughnessy suggested to take the Hilbert envelope of the speech signal to construct a selection signal [14]. This selection signal in conjunction with the MLED was reported to consititute a robust algorithm for GCI determination. In fact, the EFLPR has took advantage of the Hilbert transform while processing the residual signal. In this paper, we do not incline to exploit this feature based on two reasons: 1) the Hilbert transform is computational expensive as compared to filtering operations, and 2) an empirical adapation in time-shifting is required. The Hilbert envelope with respect to the lowpass filtered speech signal presented in Fig. 9 is only for the purpose of better illustration.

Our attention in the following discussion, instead, turns to the noise influence to the modified LFR based on measured pertubation. A data base consisting of four Mandarin sentences spoken by six speakers (three males and three females) is under our study. A total of 9392 GCI's are identified from clean speech signals by using

the proposed algorithm and verified by visual inspection. We use these GCI's as reference points for judging. Fig. 10 shows the mean absolute differences (MAD) between the reference GCI's and the GCI's extracted from noisy speech with the SNR's in the range of 0-30 dB. On the other hand, Fig. 11 presents the proportion of the deviation within a scope between –2 and 3 samples to the entire number of GCI's. The deviation scope considered here is asymmetric around zero since the distribution of the differences of the GCI positions tends to have a positive mean. While the MAD reflects the consequence of noise perturbation, the percentage within the designated deviation range implies the reliability of the modified LFR in the presence of various levels of noise. The degrading accruacy in low SNR's reflects the limitation of the modified LFR in GCI extraction. In accordance with the results shown in Figs 10 and 11, it is appropriate for us to come up with a conclusion that the modified LFR is capable of carrying out GCI detection with moderate SNR's. A detailed anlysis reveals that the degraded performance in low SNR's can also be ascribable to the incompetence in obtaining an accurate inverse filter. While errors due to the inverse filtering affect an entire frame of the modified LFR, such errors certainly lead to misjudgement of GCI in a domino effect. Hence further adjustments and refinements of the modified LFR seems indispensible in case the speech signals are severely corrupted by noise. However, we have to point out that the miss of GCI locations does not means the loss of pitch tracking. As illustrated in Fig. 11, the modified LFR exhibits rather clear periodicity even though there appear several false negative peaks.

VI. Conclusions

In this paper we revisit the topic of glottal closure detection by taking advantage of the lowpass filtered residual. The proposed method takes various aspects of the residual into consideration for signifying the epochal features of the residual signal.

The use of an overlap-and-add approach smoothes the transition of the filtered output across frames, and allows the adjustment of signal power to achieve steady amplitude. The processes that constitute a modified LFR for better GCI identification include the damped inverse filtering, broadband notch filtering, and broadband lowpass filtering. We illustrate the effectiveness of the modified LFR using several problematic sounds, which comprise /Λ/, /η/, /m/, /n/, /z/, /a^w/, and /u/. The robustness of the proposed method is demonstrate by examining a low-pitched vowel /u/ in the presence of white Gaussian noise with SNR = 10 dB. The proposed method performs satisfactorily well while compared with other methods. Our experiments based on a data set consisting of 24 sentences also indicate that the proposed method is capable of achieving reliable GCI detection under a noise-free or moderate SNR condition. The satisfying performance in GCI identification thus provides a powerful tool to support the pitch-synchronous analysis, text-to-speech synthesis, and prosodic modifications by means of the PSOLA technique [15]. Furthermore, as the essential computations for deriving the modified LFR involve only simple filtering operations, the affordable computational cost allows the proposed method to be easily incorporated into a real-time speech coder that exploits the glottal features [16,17].

Acknowledgment

This research was supported by the National Science Council, Taiwan, ROC, under Grant NSC87-2218-E-197-001.

References

1. ANANTHAPADMANABHA, T. V., and YEGNANARAYANA, B.: ‘Epoch extraction of voiced speech ’, *IEEE* Trans. Acoust., Speech, Signal Processing, 1975, 23, (6), 562-570
2. ANANTHAPADMANABHA, T. V., and YEGNANARAYANA, B.: ‘Epoch extraction from linear prediction residual for identification of closed glottis interval’, *IEEE* Trans. Acoust., Speech, Signal Processing, 1979, 27, (4), 309-319
3. Ma, C, KAMP, Y., and WILLEMS, L. F.: ‘A Frobenius norm approach to glottal closure detection from the speech signal', *IEEE* Trans. Speech, Audio Processing, 1994, 2, (2), 258-265
4. STRUBE, H. W.: ‘Determination of the instant of glottal closure from the speech wave’, J. Acoust. Soc. Am., 1974, 56, (5), 1625-1629
5. WONG, D. Y., MARKEL, J. D., and GRAY, A. H.: ‘Least squares glottal inverse filtering from the acoustic speech waveform’, *IEEE* Trans. Acoust., Speech, Signal Processing, 1979, 27, (4), 353-362
6. MOULINES, E., and DI FRANCESCO, R.: ‘Detection of the glottal closure by jumps in the statistical properties of the speech signal’, Speech Commun., 1990, 9, 401-418

7. McCREE, A. V., and BARNWELL III, T. P.: ‘Improving the performance of a mixed excitation LPC vocoder in acoustic noise’, Proceedings of ICASSP, 1992, (II), 137-140

8. LAURENT, P. A., and DE LA NOUE, P.: ‘A robust 2400 bps subband LPC vocoder’, Proceedings of ICASSP, 1995, 500-503

9. LAFLAMME, C., SALAMI, R., MATMTI, R., and ADOUL, J.-P.: ‘Harmonic-stochastic excitation (HSX) speech coding below 4 Kbit/s’, Proceedings of ICASSP, 1996, 204-207

10. FANT, G.: ‘Acoustic Theory of speech production’, (Mouton, Paris, 1960)

11. CHILDERS, D. G., and LEE, C. K.: ‘Vocal quality factors: analysis, synthesis, and perception’, J. Acoust. Soc. Am., 1991, 90, (5), 2394-2410

12. ROSS, M. J., SHAFFER, H. L., COHEN, A., FREUDBERG, R., and MANLEY, H. J.: ‘Average magnitude difference function pitch extractor’, *IEEE* Trans. Acoust., Speech, Signal Processing, 1974, 22, (5), 353-362

13. HU, H. T.: ‘Comb filtering of noisy speech using overlap-and-add approach’, Electronic Lett., 1998, 34, (1), 16-18

14. CHENG, Y. M., and O'SHAUGHNESSY, D. : ‘Automatic and reliable estimation of glottal closure instant and period’, *IEEE* Trans. Acoust., Speech, Signal Processing, 1989, 37, (12), 1805-1815

15. VALBRET, H., MOULINES, E., TUBACH, J. P.: ‘Voice transformation using PSOLA technique’, Speech Commun., 1992, 11, 175-187

16. HEDELIN, P.: ‘High-quality glottal LPC vocoding’, Proceedings of ICASSP, 1986, 465-468

17. CHILDERS, D. G., and HU, H. T.: ‘Speech synthesis by glottal excited linear prediction’, J. Acoust. Soc. Am., 1994, 96, (4), 2026-2036

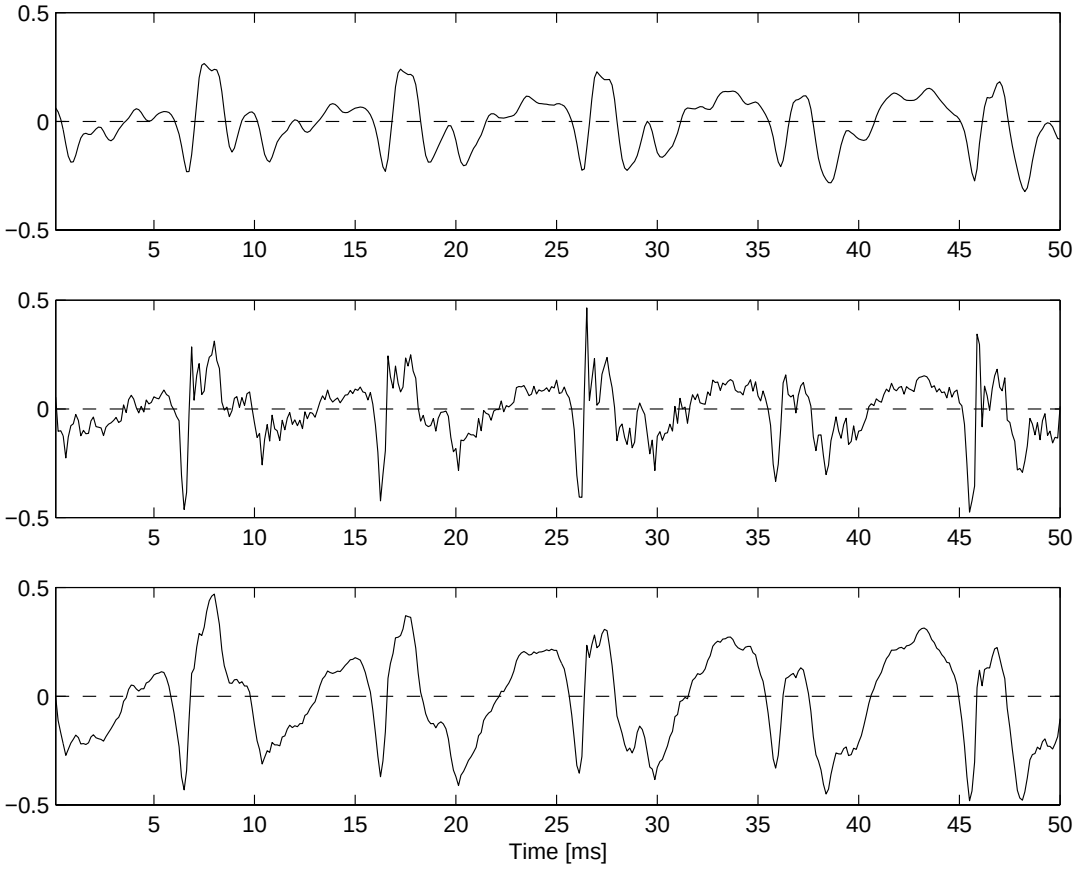

Fig. 1. Effect due to the broadband notch filter; from top to bottom: the speech signal /o/, lowpass filtered residual with and without notch filtering.

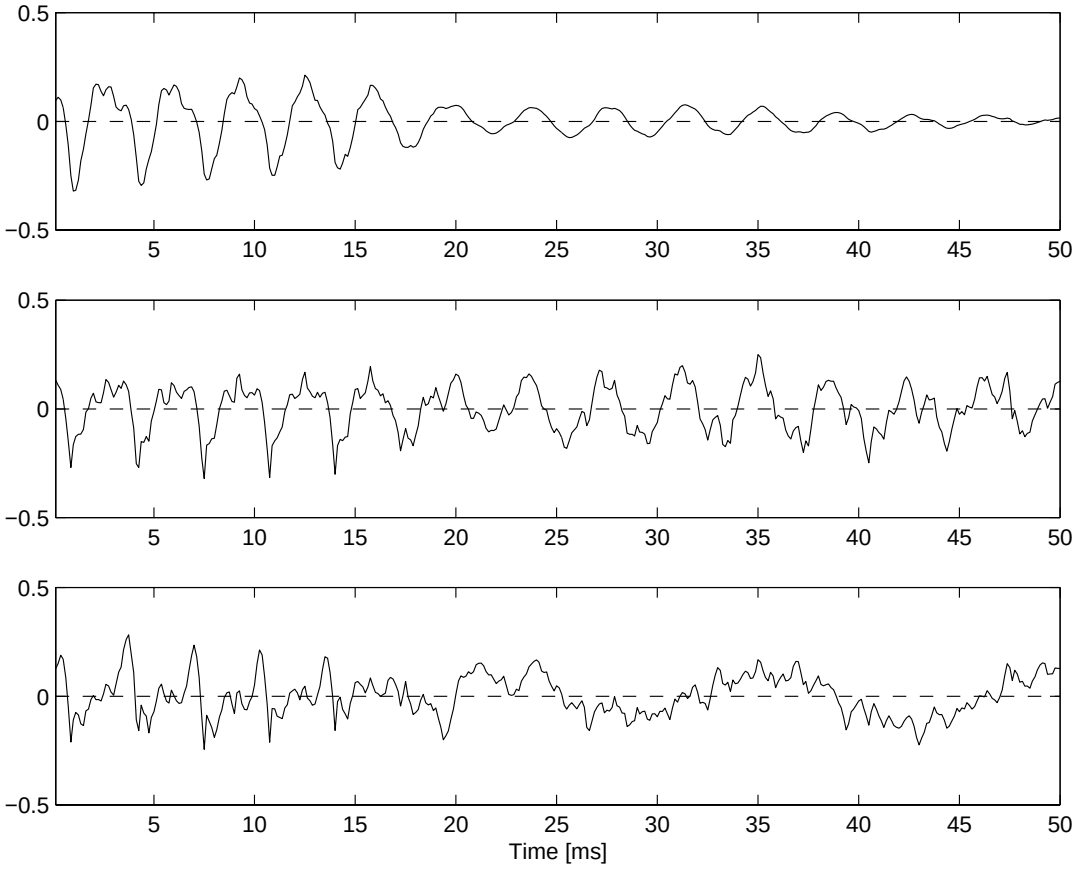

Fig. 2. Effect due to the highpass emphasis and damped inverse filtering; from top to bottom: the speech signal /a^w/, lowpass filtered residual with and without highpass preemphasis and damped inverse filtering.

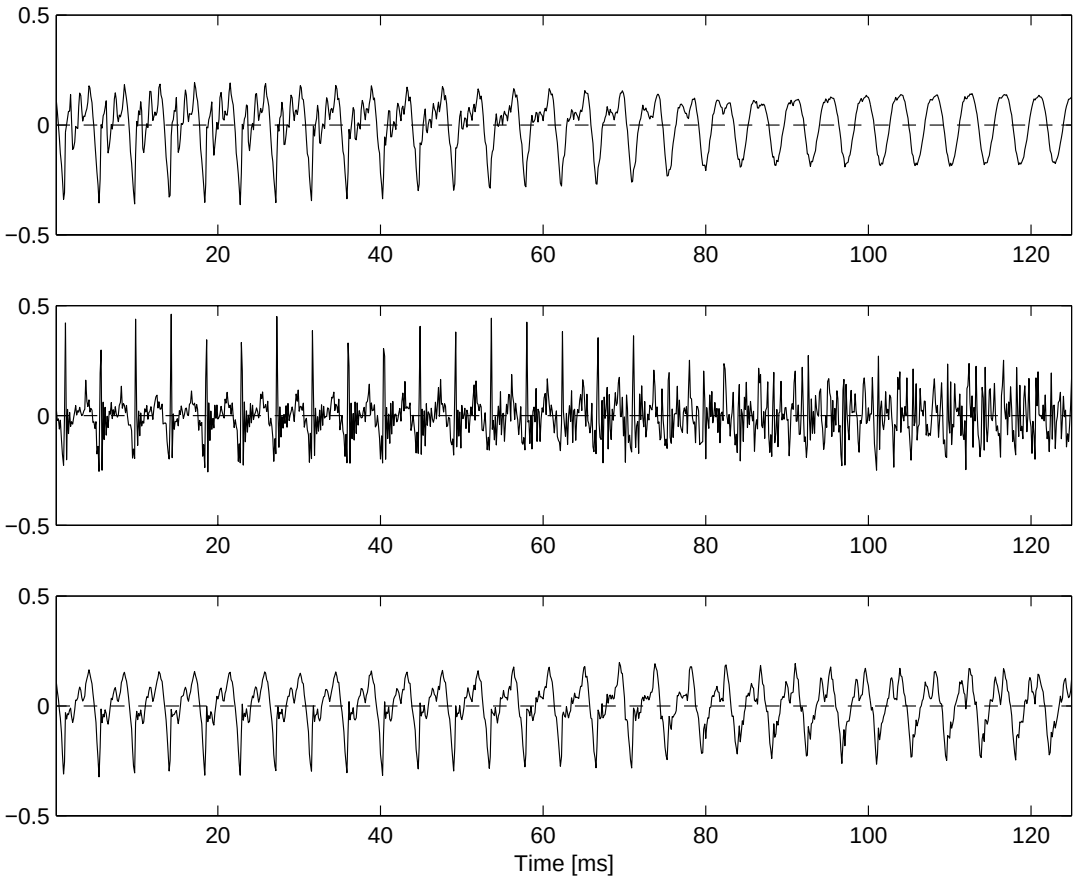

Fig. 4. GCI detection for phoneme transition between /Λ/ and /η/; from top to bottom: the speech signal, the residual signal, and the modified LFR.