Voiced/Nonvoiced Detection Based on Robustness of Voiced Epochs

N. Dhananjaya and B. Yegnanarayana, *Senior Member, IEEE*

***Abstract*—In this paper, a new method for voiced/nonvoiced detection based on epoch extraction is proposed. Zero-frequency filtered speech signal is used to extract the instants of significant excitation (or epochs). The robustness of the method to extract epochs in the voiced regions, even with small amount of additive white noise, is used to distinguish voiced epochs from random instants detected in nonvoiced regions. The main feature of the proposed method is that it uses the strength of glottal activity as against using the periodicity of the signal. Performance of the proposed algorithm is studied on TIMIT and CMU ARCTIC databases, for two different noise types, white and vehicle noise from the NOISEX database, at different signal-to-noise ratios (SNRs). The proposed method performs similar or better than the popular normalized crosscorrelation based voiced/nonvoiced detection used in the open source utility *wavesurfer*, especially at lower SNRs.**

***Index Terms*—Excitation source, glottal activity detection, glottal closure instant, voiced/nonvoiced detection, zero-frequency filtering.**

I. Introduction

VOICED/NONVOICED (V/NV) detection involves identifying the regions of speech when there is significant glottal activity (i.e., the vibration of vocal folds). Such regions of speech are generally referred to as *voiced speech*. The nonvoiced regions of speech include both silence (or background noise) as well as unvoiced speech (such as voiceless fricatives and stops). Note that here the term voiced regions is used to refer to those regions where the vibration of the vocal folds is strong, and it is not necessary that the vibrations be regular (i.e., periodic) always, as in the case of strong aspiration or creaky voices. Any method to detect such regions should not depend critically on the property of periodicity of waveform in successive glottal cycles. The novelty of the method proposed in this paper lies in exploring the strength of glottal activity for detecting the voiced regions.

Approaches for glottal activity detection fall into three broad categories, namely, time-domain, frequency-domain and statistical approaches. The time-domain and frequency-domain approaches measure one or more acoustic features which reflect the production characteristics of the voiced sounds such as energy, periodicity and short-term correlation. Some parameters used are zero crossing rate, autocorrelation coefficient at the first lag, the first coefficient of a p^{th}-order ($p = 12$) linear prediction (LP) analysis, long-term normalized autocorrelation peak strength (in the range 2–15 ms), normalized LP error, normalized low-frequency energy, cepstral peak strength, harmonic measure from the instantaneous frequency amplitude spectrum [1]–[3]. Voiced/nonvoiced decisions are taken by setting thresholds on individual parameter values (chosen empirically), and the decisions are combined in a hierarchical manner. The main problem with these methods is in setting thresholds which are critical in determining the performance of V/NV detection. Also, most of these measures of voicing are susceptible to noise, and the performance deteriorates with decreasing signal-to-noise ratio (SNR). Statistical models such as neural network models, Gaussian mixture models (GMM) or hidden Markov models (HMM) are also used for combining evidence from multiple features [1], [4]. These methods do not depend critically on threshold setting, but require training data for different types of background noises. Statistical approaches are more popular in voice activity detection (VAD) algorithms used in speech coding applications [5], [6]. They assume different models of random process for speech and background noise, and estimate the parameters of the underlying distributions. Performance of these approaches depends on the choice of the probability distributions, and the ability to estimate the parameters of the noise distribution. Generally these methods do not make use of the knowledge of speech production mechanism in any significant way. Also, most of these methods do not evaluate separately the performance of detecting voiced and unvoiced regions of speech.

In this paper, we propose a new approach for detecting the regions of glottal activity in continuous speech based on the presence of impulse-like excitation (epochs) around the instant of glottal closure (GCI). Zero-frequency (ZF) resonator output of the speech signal is used to extract epochs, which was shown to be robust against different types of degradations even at very low SNRs [7]. The paper is organized as follows. Section II describes the method for ZF filtering of speech signal and computation of the instants of significant excitation. The key idea for V/NV decision or glottal activity detection is presented in Section III. Some issues on the robustness of the proposed method for varying levels of noise is discussed in Section IV. Performance of the proposed method for varying SNRs is given in Section V. Section VI gives a summary of the paper and discusses some issues that need to be addressed.

II. Epoch Extraction by ZF Filtering of Speech Signal

A ZF resonator which exploits the fact that the effect of an impulse-like excitation is felt throughout the spectrum including the zero frequency, was proposed for accurate estimation of the voiced epochs [7]. A ZF resonator involves a pair of poles on the

Manuscript received October 01, 2009; revised December 03, 2009. First published December 15, 2009; current version published January 20, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Saeid Sanei.

N. Dhananjaya is with the Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India (e-mail: dhanu@cse.iitm.ac.in).

B. Yegnanarayana is with the International Institute of Information Technology, Hyderabad, India (e-mail: yegna@iiit.ac.in).

Digital Object Identifier 10.1109/LSP.2009.2038507

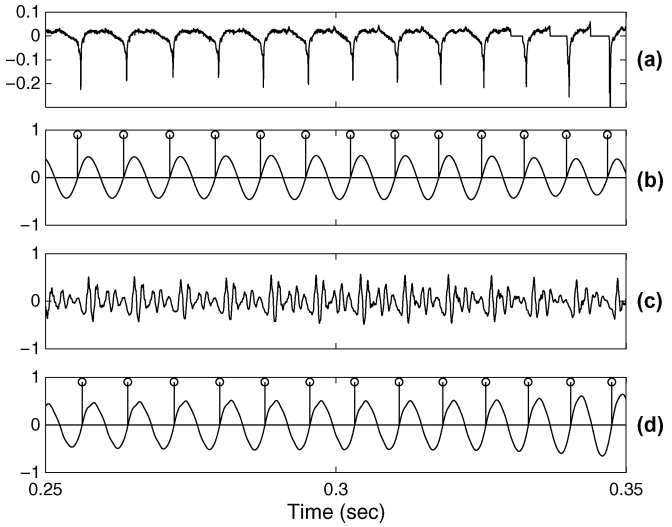

Fig. 1. Epoch extraction using ZF filtered signal. (a) Short segment of dEGG signal. (b) ZF filtered signal derived from the dEGG signal. (c) Speech signal recorded simultaneously with the dEGG signal. (d) ZF filtered signal derived from the speech signal. The hypothesized epochs at the positive zero crossings of the filtered signals are marked in (b) and (d).

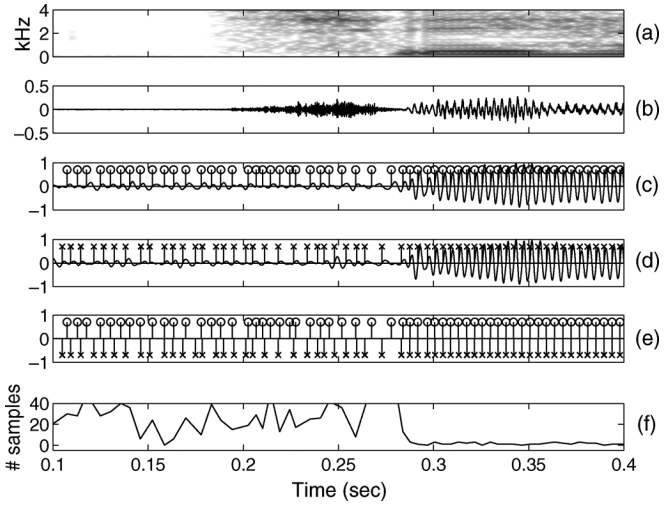

Fig. 2. Epoch extraction using ZF filtered speech signal for two different additive noise sample functions (at 30-dB SNR). (a) Spectrogram. (b) Speech signal. (c) ZF signal for the first noise sample function along with the epochs (E_1). (d) ZF signal for the second noise sample function along with epochs (E_2). (e) E_1 ($+$ve and circles) and E_2 ($-$ve and crosses). (f) Epoch drift measured between E_1 and E_2.

unit circle at zero Hertz, which can be implemented in terms of simple cumulative sum operations. To highlight the small fluctuations in the output of the resonator, a trend removal operation is used by subtracting the local mean computed over a short window size. The size of the window is in the range of one to two pitch cycles. The ZF filtered signal exhibits high energy in the voiced regions due to significant contribution from the impulse-like excitation as compared to the nonvoiced regions of speech. Also, the filtered signal has the property that its positive zero crossings (negative to positive) are synchronized with the instants of glottal closure, called *epochs*. To illustrate this, a segment of speech along with the simultaneously recorded electroglottogram (EGG) signal from the CMU ARCTIC database is used [8]. Fig. 1(b) shows the ZF filtered signal derived from the differenced electroglottogram (dEGG) signal shown in Fig. 1(a). It can be seen that the positive zero crossings of the filtered signal are synchronized with the large negative peaks in the dEGG signal which correspond to the instants of glottal closure. Fig. 1(c) and (d) show that the information about the instants of glottal closure can be derived directly from the speech signal. Another useful property of the ZF filtered signal is that the slope or the rate of zero crossing (negative to positive) is proportional to the strength of excitation [9].

III. Epoch-Based Voiced/Nonvoiced Detection

The key idea exploited in this paper is that addition of a small amount of noise to the speech signal does not affect the zero crossings of the ZF filtered signal in the voiced region, whereas it leads to zero crossings at random locations in the nonvoiced region. The glottal closure during the production of voiced sounds impart the most significant impulse-like excitation to the vocal tract system. These high SNR epochs are robust to noise. The ZF filtered signal can be used to locate these instants with a high degree of precision and accuracy even in the presence of severe degradation [7]. Lack of any significant excitation in the nonvoiced regions result in zero crossings located at random instants, and these locations can easily get affected by the addition of even a small amount noise.

A small amount of white Gaussian noise is added to the speech signal (effective SNR of about 30 dB). The ZF filtered signal and the epochs are computed. Another sample function of white Gaussian noise is added to the speech signal, and the epochs are computed again. Fig. 2(c) and (d) show the two ZF filtered signals and the corresponding epochs obtained for two different sample functions of noise. It can be seen from Fig. 2(e) that the two epoch sequences are in coherence within the voiced region, and are located at random instants in the nonvoiced region. The precision of the epochs for different noise sample functions is measured in terms of the drift in the epoch locations from one noise sample function to the other. For every epoch from the first noise sample function, the drift is measured as the distance in number of samples to the nearest epoch from the second noise sample function. The drift in epochs for two different sample functions of noise is shown in Fig. 2(f). Only those epochs which drift by not more than 1 ms are hypothesized as voiced epochs.

The spurious epochs that could still be present in the silence or unvoiced region are eliminated using the instantaneous pitch period and jitter measured at each epoch. The instantaneous pitch period at each epoch (in terms of number of samples, N_0) is computed as the minimum of the distances with the epochs on either side. Similarly, at every epoch the change in pitch period (ΔN_0) is computed over the next two epochs on either side, and the minimum is chosen as the instantaneous jitter. Only those epochs which have a pitch period less than 15 ms and a jitter within 1 ms are retained as the voiced epochs. These voiced epochs are further validated based on the strength of excitation to eliminate any spurious epochs. Any epoch with an excitation strength less than 1% of the maximum strength of excitation is marked as nonvoiced. Note that while the proposed algorithm requires some thresholds or limits to be set on the epoch drift, pitch period, jitter and excitation strength, none of these are critical for the performance of the method. The final voiced epochs

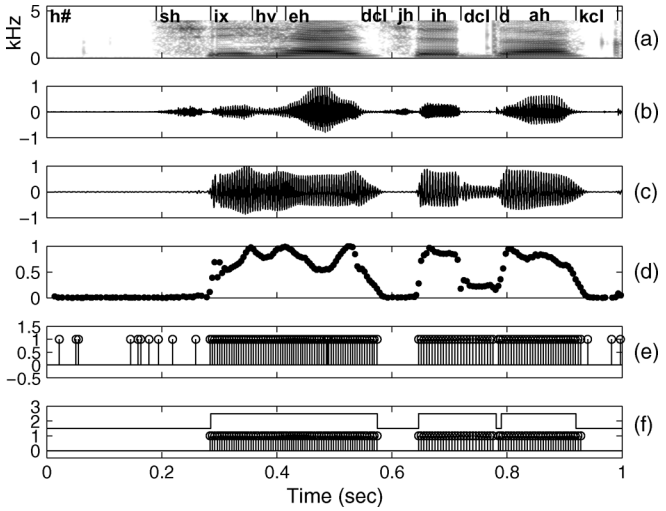

Fig. 3. Detection of voced epochs using noise sample functions. (a) Spectrogram. (b) Speech signal. (c) ZF filtered speech signal. (d) Excitation strength at the epochs. (e) Voiced epochs hypothesized based on epoch drift. (f) Final voiced epochs obtained after validations based on pitch period, jitter and excitation strength. The reference or ground truth for voiced/nonvoiced detection is plotted above the epochs.

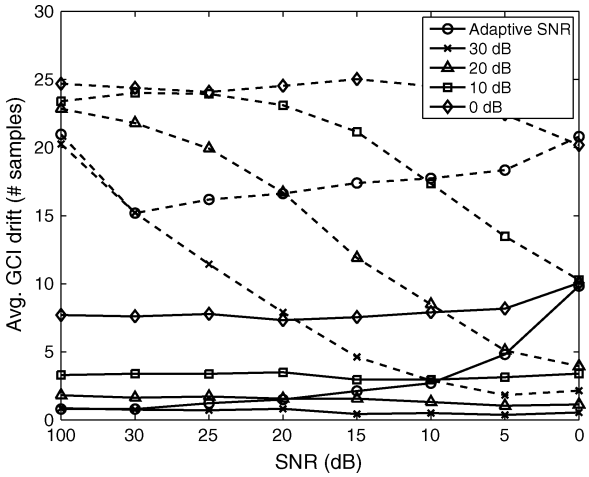

Fig. 4. Epoch drift in voiced (solid lines) and nonvoiced (dashed lines) regions for varying input signal SNR. The legend at the top right corner shows the amount of noise used for epoch detection. Adaptive SNR is the case when the amount of noise chosen for epoch detection is same as the input signal SNR.

obtained are shown in Fig. 3(f), along with the manually marked ground truth for reference. The epochs hypothesized as voiced based on the drift in epochs are shown in Fig. 3(e), and the excitation strength used for validating these epochs is shown in Fig. 3(d). It can be seen that the excitation strength provides good evidence for V/NV decision. But relying only on the excitation strength or the filtered signal energy makes the setting of threshold a difficult task. It can be seen that even the weak voice bar regions (corresponding to the regions marked as /dcl/ between time instants 0.5 to 0.6 s and 0.7 to 0.8 s) are detected. Also, the region with weak voicing towards the tail of the vowel /ah/ at around 0.9 s is also detected by the proposed method, while it is ignored during manual marking.

IV. Analysis of Drift in Epochs Induced by Noise

In this section, a discussion on the drift the epochs undergo in voiced and nonvoiced regions due to addition of noise is given. Also, a discussion on the suitable amount of noise that can be added to the speech signal at different SNRs is given. Fig. 4 shows the epoch drift for voiced (solid lines) and nonvoiced regions (dashed lines) for varying SNRs of the input speech signal, and for different amounts of noise (30, 20, 10, and 0 dB) added for the detection of voiced epochs. Note that noise is added to the clean signal to generate a degraded signal for a specified SNR. Then different sample functions of noise are added at different levels to determine the voiced epochs. It is seen that the average drift in the voiced region is small even when the added noise is 0 dB, indicating the robustness of the epoch extraction method. But, as can be seen from the dashed lines for nonvoiced regions the drift in epochs is not significant to be discriminated from the voiced epochs, when the SNR of the input signal is greater than the amount of noise added for the detection of voiced epochs. The epoch drifts plotted for the case of "adaptive SNR", where the amount of noise added is equal to the signal SNR, show that the best results may be obtained if an estimate of the signal SNR is available. At the same time, looking at the plots for 10-dB noise (marked by squares), one can infer that it can give equally good (in terms of low drift for voiced and large drift for nonvoiced epochs) results up to 10-dB SNR of the input signal. A constant 10-dB additive white Gaussian noise is used for the experiments reported in this paper. Also, it can be seen that setting of threshold on the epoch drift for separating voiced epochs from nonvoiced is not very difficult. A threshold of 1 ms (16 samples at 16 kHz) is chosen for the experiments described in this paper.

V. Performance Evaluation

The performance of the proposed method for voiced/nonvoiced detection is evaluated on the TIMIT database [10]. A subset of the TIMIT database, consisting of 38 speakers (24 male and 14 female) uttering ten short (3 to 5 s) sentences each, is used for these evaluations. The performance is measured in terms of the number of epochs missed in the voiced regions and the number of spurious or false epochs hypothesized in the nonvoiced region. Epochs derived from the clean speech using a ZF resonator [7], and the V/NV decision derived from the manual markings, are used to obtain the reference epochs in the voiced regions. An epoch in the voiced region (reference epoch) is said to be missed if there is no epoch hypothesized within 1-ms duration on either side of the reference epoch. Any epoch hypothesized in the nonvoiced region of the V/NV decision obtained from the manual markings is a false detection. Performance of the proposed method is evaluated for two different noise types (white and vehicle) from the NOISEX-92 database and for different SNRs of the input signal. The percentage of voiced speech samples in each of the utterances is maintained at 40% by appending requisite duration of silence before the addition of noise samples [6]. The results are given in Table I. As a comparison the performance of the V/NV decisions given by *wavesurfer*, an open source utility which relies on normalized crosscorrelation based pitch tracking refined by dynamic programming, is given [11]. The proposed method performs similar or better (at higher noise levels) than the decisions given by *wavesurfer* in terms of percentage classification accuracy, which is computed as $P_c = 1 - (0.4 \times P_m + 0.6 \times P_f)$. Here, P_m denotes the percentage of epochs missed in the voiced regions, and P_f denotes the percentage of epochs in the nonvoiced regions falsely

TABLE I
PERFORMANCE OF VOICED/NONVOICED DETECTION

		SNR	*Proposed*		*Wavesurfer*	
			P_c	(P_m, P_f)	P_c	(P_m, P_f)
TIMIT	White	Clean	**94.4**	(3.1,7.2)	**94.4**	(10.6,2.3)
		30 dB	**94.2**	(3.2,7.6)	**94.2**	(12.0,1.7)
		20 dB	**92.8**	(5.8,8.1)	**93.0**	(15.7,1.2)
		10 dB	**91.6**	(7.1,9.2)	**89.7**	(24.9,0.6)
		5 dB	**89.8**	(10.4,10.1)	**85.3**	(36.4,0.3)
		0 dB	**85.7**	(16.9,12.5)	**77.7**	(55.7,0.1)
	Vehicle	Clean	**94.4**	(3.2,7.2)	**94.4**	(10.6,2.3)
		30 dB	**93.3**	(4.4, 8.3)	**94.2**	(11.5,2.0)
		20 dB	**92.7**	(5.0,8.8)	**93.5**	(13.8,1.5)
		10 dB	**91.8**	(6.9,9.1)	**91.5**	(19.8,1.0)
		5 dB	**89.7**	(9.9,10.5)	**89.1**	(26.1,0.7)
		0 dB	**86.4**	(16.1,11.9)	**84.9**	(37.1,0.4)
ARCTIC	White	Clean	**96.0**	(2.1, 5.2)	**95.3**	(8.7,2.1)
		30 dB	**95.9**	(2.6, 5.0)	**95.0**	(9.8,1.8)
		20 dB	**95.8**	(3.2,4.9)	**94.1**	(12.9,1.3)
		10 dB	**94.6**	(4.8,5.8)	**91.8**	(19.8,0.4)
		5 dB	**92.7**	(7.4,7.1)	**87.7**	(30.2,0.3)
		0 dB	**89.1**	(13.5,9.2)	**83.0**	(42.3,0.1)
	Vehicle	Clean	**96.0**	(2.1,5.2)	**95.3**	(8.7,2.1)
		30 dB	**95.5**	(3.1,5.5)	**94.3**	(10.7,2.4)
		20 dB	**95.2**	(3.3,5.8)	**93.9**	(12.1,2.1)
		10 dB	**94.7**	(4.0,6.1)	**91.9**	(17.9,1.5)
		5 dB	**92.2**	(7.1,8.2)	**88.6**	(27.2,0.9)
		0 dB	**88.3**	(14.1,10.1)	**85.7**	(35.1,0.4)

identified as voiced. Note that here a fixed level of noise (10-dB SNR) is used for the extraction of voiced epochs irrespective of the SNR of the input signal. Since decisions are made at several levels using different parameters, it is not straightforward to use a single parameter to control the tradeoff between P_m and P_f in both the proposed method as well as in the method used by wavesurfer. Hence, the percentage classification accuracy is used as a measure to evaluate and compare the performance of both these methods.

The main source of error in the case of TIMIT dataset is the manual marking. There are two kinds of errors introduced by manual labeling. 1) The boundaries may not be very precise, and a few milliseconds of error is inevitable. Some weak voiced regions towards the vowel ending are typically overlooked. Also, the aspiration produced during some stop consonants tends to extend into the following vowel making the boundary fuzzy. 2) The other type of manual errors are due to mismatch between speaker articulation and listener anticipation. Some sounds or regions that are susceptible to such errors are stop consonants (the lack or presence of voicing during the closure period) and voiced fricatives.

The performance of the proposed method is also evaluated on the CMU ARCTIC database [8], which has simultaneous recordings of speech and EGG signals. A subset of the database with three different speakers each uttering 100 short sentences (4 to 5 s) is used. The EGG signal is used for deriving the ground truth so as to minimize human error in labeling. Zero-frequency filtered EGG signal is used to detect the epochs and the excitation strength. A simple threshold on the excitation strength is used to detect the reference voiced epochs which are later verified manually. The performance of the proposed method for different noise conditions is given in Table I. The performance is better than the TIMIT dataset owing to lack of any manual errors.

VI. SUMMARY AND CONCLUSIONS

A new method for voiced/nonvoiced detection was proposed based on the ability of the ZF filtered signal to detect the voiced epochs with high precision, and on the accuracy of detecting the epochs even in the presence of degradation. One of the main features of the proposed method is that it depends entirely on the excitation source information, as the vocal tract spectral information is more prone to noise. Moreover, it uses the strength of glottal activity as against using the periodicity in the signal. Another feature of the method is the injection of a small amount of noise to detect the high SNR instants of glottal closure, and hence the voiced regions. Also, threshold setting is not very critical in the proposed method. One of the limitations of the proposed method is that a fixed amount of noise is added irrespective of the input SNR. Since the method uses zero frequency filtering, it may not work well when the signal is bandlimited by removing the low-frequency component as in the telephone speech.

REFERENCES

[1] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process,*, vol. ASSP-24, no. 3, pp. 201–212, Jun. 1976.

[2] D. Arifianto, "Dual parameters for voiced-unvoiced speech signal determination," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, Honolulu, HI, May 2007, pp. IV-749–IV-752.

[3] C. Shahnaz, W. P. Zhu, and M. O. Ahmad, "A multifeature voiced/nonvoiced decision algorithm for noisy speech," in *Proc. Int. Symp. Circuits and Systems*, Kos, Greece, May 2006, pp. 2525–2528.

[4] A. P. Lobo and P. C. Loizou, "Voiced/unvoiced speech discrimination in noise using Gabor atomic decomposition," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, Hong Kong, Apr. 2003, pp. I-820–I-823.

[5] R. Tahmasbi and S. Rezaei, "Change point detection in GARCH models for voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 1038–1046, Jul. 2008.

[6] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 412–424, Mar. 2006.

[7] K. Sri Rama Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[8] J. Kominek and A. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, 2004, pp. 223–224 [Online]. Available: http://festvox.org/cmu_arctic/index.html

[9] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Process. Lett.*, accepted for publication.

[10] J. S. Garofolo *et al.*, TIMIT Acoustic-Phonetic Continuous Speech Corpus Linguistic Data Consortium. Philadelphia, PA, 1993.

[11] K. Sjolander and J. Beskow, "Wavesurfer—An open source speech tool," in *Proc. Int. Conf. Spoken Language Processing*, Beijing, China, Oct. 2000, pp. 464–467.