Text-Independent Phoneme Segmentation Combining EGG and Speech Data

Lijiang Chen, *Member, IEEE*, Xia Mao, *Member, IEEE*, and Hong Yan, *Fellow, IEEE*

***Abstract*—A new approach for text-independent phoneme segmentation at sampling point level is proposed in this paper. The algorithm consists of two phases: First, the voiced sections in speech data are detected using the information of vocal folds vibration contained in electroglottograph (EGG). A Hilbert envelope feature is adopted to achieve sampling point level detection accuracy. Second, the voiced sections and other sections are treated separately. Each voiced section is divided into several candidate phonemes using the Viterbi algorithm. Then adjacent candidate phonemes are merged based on a Hotellings T-square test method. For other sections, the unvoiced consonants are detected from silence based on a singularity exponent feature. Comparison experiments show that the proposed method has better performance than the existing ones for a variety of tolerances, and is more robust to noise.**

***Index Terms*—Electroglottograph, Hilbert Envelope, Viterbi algorithm, Hotellings T-square test, Singularity Exponent.**

I. Introduction

THE identification of the start and end boundaries of phonemes in continuous speech is an important problem in many areas of speech processing [1]. Although manual labeling produces the most accurate results, it may not be able to maintain the consistency of labeling. Furthermore, it is also costly. As a result, automatic segmentation is often required.

In the last twenty years, many methods have been proposed for automatic phoneme segmentation. A widely used approach is forced alignment which requires two inputs: recorded audio and phone transcriptions [2]. In this approach, the segmentation phase is reduced to the problem of finding the best path in a graph, which can be effectively solved using many tools such as dynamic time warping (DTW) and Hidden Markov Models (HMMs) [3], [4]. Normally, the utterances to be segmented should be pronounced correctly so as to match the given transcriptions exactly. An overview of machine learning techniques exploited for phone segmentation was done by Adell [6]. However, these algorithms are all based on text information included in the waveform [7], [8]. These text-dependent algorithms are useful in some fields, such as automatic labeling of a big database, to prepare the training or test data sets for further processing. However, the text information included in the waveform is not available in many other cases, including real-time phoneme-based speech recognition, accent conversion system, real-time translation system, and computer-aided language learning system [9]. In these cases, text-independent phoneme segmentation algorithms are needed.

Text-independent phoneme segmentation algorithms can be typically categorized into decoder-guided, model-based, and metric-based approaches [10]. In the decoder-guided method, the phone transcriptions are obtained by a speech recognition system, which decodes the spoken audio stream first. In the model-based and metric-based approach, acoustic changes between two adjacent phonemes are used to find the existing boundaries. Aversano introduced a method for text-independent speech segmentation in which the preprocessing is based on critical-band perceptual analysis. It results in 74% segmentation accuracy while over-segmentation is limited to a minimum value [11]. Qiao formulated the optimal segmentation into a probabilistic framework, then developed three and two optimal objective functions in 2008 and 2013 respectively [12], [13]. Scharenborg investigated fundamental problems for unsupervised segmentation algorithms, and suggested that one-stage bottom-up segmentation methods should be expanded into two-stage top-down segmentation methods [14]. Khanagha presented the Microcanonical Multiscale Formalism (MMF) for phonetic segmentation. The results showed higher accuracy than other methods [15], [16]. Zhao focused on the post-processing method, which refines the preliminary boundaries generated by an existed system [9].

The conventional methods have the following limitations:

Firstly, despite using different algorithms, most text-independent automatic phoneme segmentation methods achieve imperfect performance in terms of boundary detection. Especially under the condition of low signal to noise ratios (SNR), the performance and robustness of text-independent phoneme segmentation degrades significantly.

Secondly, most of these algorithms are based on frame segmentation, which decomposes the speech signal into a sequence of overlapping frames [1]. The frame-overlap (normally between 10 ms to 20 ms) is the smallest tolerance for phoneme segmentation. It is not sufficient for some short unvoiced consonants, which are important for many further applications, such as emotion recognition.

This paper introduces a robust text-independent phoneme segmentation algorithm, with high accuracy in the sampling

Manuscript received June 11, 2015; revised November 06, 2015 and January 19, 2016; accepted February 19, 2016. Date of publication February 23, 2016; date of current version April 29, 2016. This work was supported by the National Research Foundation for the Doctoral Program of Higher Education of China (no. 20121102130001), in part by the Fundamental Research Funds for the Central Universities (no. YWF-14-DZXY-015), and in part by the City University of Hong Kong (Project 9610308). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. James Glass.

L. Chen and X. Mao are with the Department of Electronic and Information Engineering, Beihang University, Beijing 100191, China (e-mail: moukyoucn@aliyun.com).

H. Yan is with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China.

Digital Object Identifier 10.1109/TASLP.2016.2533865

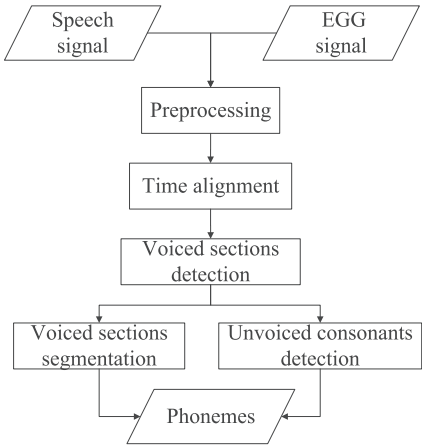


Fig. 1. Overview of our phoneme segmentation method.

point level. The proposed method combines the information from EGG (Electroglottograph) and speech signals. EGG gives the information of vocal folds vibration by measuring the electrical resistance between two electrodes placed around the throat. The EGG directly reveals the source of phonation without influence of vocal tract and aerodynamic noise. Therefore, EGG can provide useful information about speech, and has been used to extract pitch frequency and to detect laryngeal lesions in the field of medicine and rehabilitation [17], [18]. In this paper, EGG is used for the detection of voiced sections.

The paper is structured as follows: section 2 gives a detailed description of the proposed phoneme segmentation method, including preprocessing, time alignment, voiced sections detection, voiced section extraction and unvoiced section detection; section 3 introduces the experiments and the database used, as well as the experiment results; section 4 draws our conclusion.

II. Phoneme Segmentation Method

A. Overview

The phoneme segmentation method includes four steps as depicted in Fig. 1. Firstly, the preprocessing module resamples the speech and EGG signals to 8 kHz and normalizes the amplitude level by the maximum absolute value. Since the speech information can be reliably conveyed through band-limited telephone speech, the sampling rate of 8 kHz is suitable. EGG requires a much narrower band than speech signal. Secondly, the time alignment module makes the speech and EGG aligned in time.

After the above pre-processing steps, EGG and speech signals are used for precise phoneme segmentation in two steps. Firstly, the voiced sections in speech data are detected with the information of vocal folds vibration contained in Electroglottograph (EGG). A Hilbert Envelope feature is adopted to achieve an accuracy of sampling point level detection. Secondly, the voiced sections and other sections are treated separately. Each voiced section is divided into several candidate phonemes using the Viterbi algorithm. Then adjacent candidate phonemes are merged based on a Hotellings T-square test method. For other sections, the unvoiced consonants are detected from silence based on a Singularity Exponent feature.

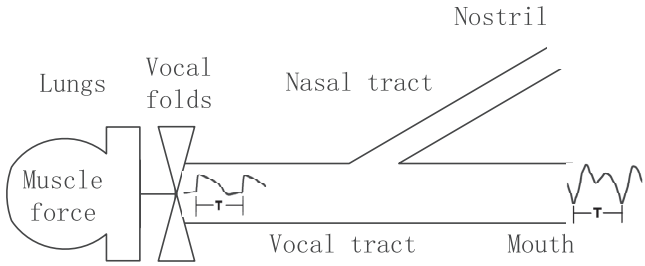


Fig. 2. Schematic model of the speech production system.

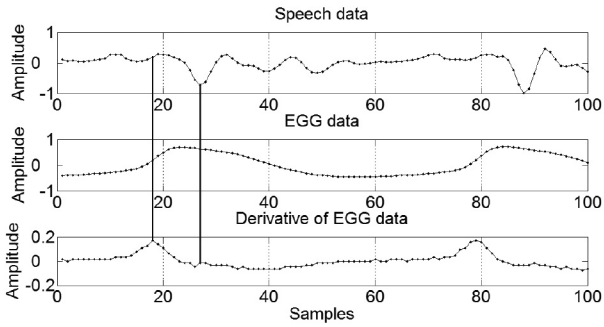


Fig. 3. EGG and speech waveform in one pitch-cycle.

B. Time Alignment

Since the speed of sound is about 340 m/s, EGG is obtained earlier than speech. Assuming that the distance between microphone and speaker's mouth is 0.5 m, the time difference is approximately 2 ms, and is a significant fraction of a pitch period. It should not be ignored for the fine phoneme segmentation using EGG and speech together.

A schematic diagram of the speech production mechanism is given in Fig. 2. It shows the vocal tract and the nasal tract as a Y-type tube that is bounded at one end by the vocal folds and at the others by the mouth and nostril openings. Voiced sounds are produced when the vocal tract tube is excited by pulses of air pressure resulted from vibration of vocal folds.

Numerical techniques can be used to create a complete physical simulation of sound generation and transmission in the vocal tract [19]. In this study, we only consider the relationship between EGG and speech waveform which are recorded near vocal folds and mouth respectively, as shown in Fig. 2. EGG is considered to have a direct relationship with the glottal opening area that causes the transglottal airflow [20], [21].

The detailed structures of EGG and speech waveform in one pitch-cycle are shown in Fig. 3. The horizontal axis represents the sampling point whose sampling frequency is 8 kHz. The vertical axis represents normalized amplitude. The derivative of EGG data has an apparent peak which represents that the glottal opening area decreases rapidly (resistance decreases rapidly). This peak is called the glottal closure time which is labeled by a vertical line in Fig. 3. According to the aerodynamics, the intraglottal air pressure is smaller at the glottal closure time than that at other times [22]. After a period of time Δt, the low pressure

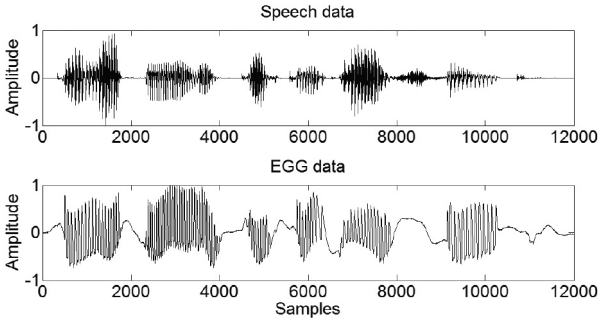


Fig. 4. Speech and EGG waveform of an utterance.

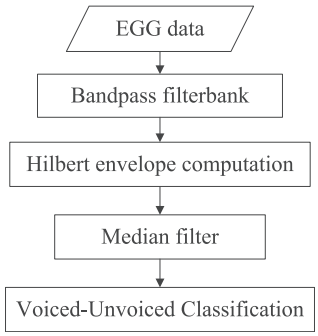


Fig. 5. Process of voiced speech detection.

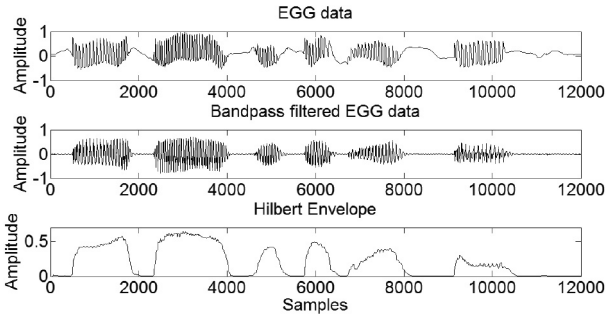


Fig. 6. An EGG utterance and Hilbert envelope.

state is transferred from glottis to microphone, which is labeled by another vertical line in Fig. 3. In this instance, the distance between glottis and microphone is calculated as:

$$D_{\Delta} = V_s * \Delta t = V_s * \Delta n / f_s \tag{1}$$

where D_{Δ} is the distance, V_s is the velocity of sound in the air taking 340 m/s, Δn is the number of samples between the two vertical lines in Fig. 3 taking 9, fs is 8 kHz, and the distance is calculated to be about 0.383 meters.

In fact, the shape of the speech waveform varies in different phonemes, the distance may also change because of the head movement. We extract the Δn in equation (1) in each voiced section (discussed later), then the mean of Δn is used to align the EGG and speech waveform in the time domain.

C. Voiced-Unvoiced Classfication

In this study, voiced speech is defined as a waveform generated when the vocal folds vibration occurs. It includes all the vowels and part of the consonants. The speech and EGG waveform of an utterance are shown in Fig. 4.

Fig. 4 shows that, phonemes have a distinctive appearance in the speech waveform. The voiced sounds are highly structured and quasi-periodic, while the unvoiced look like random noises. These differences result from the distinctively different ways in which these sounds are produced.

Because EGG can easily indicate the voiced segments in speech, we have proposed a voiced speech segmentation method based on multiple threshold [23]. In this study, the method is enhanced to the accuracy of sampling point level. The process of proposed voiced speech detection is shown in Fig. 5.

It is well known that, the vibration of vocal folds is quasi-periodic, whose fundamental frequency fluctuates within a small range due to different speakers, contents and emotions. Therefore, EGG can be approximated as a narrow-band signal.

A butterworth band-pass filter is designed with a passband frequency from 100 Hz to 1000 Hz. EGG data is filtered by the band-pass filter to remove low-frequency jitter and high-frequency noise caused by the muscle movement and equipment thermal noise. The filtered EGG can be seen as a frequency and amplitude modulated signal shown as:

$$x(n) = A_n \cos((\omega_0 + \omega_n)n + \phi), n \in \{1, 2, \ldots, N\} \tag{2}$$

where A_n and ω_n are time-varying amplitude and angular frequency. N is the number of samples.

The temporal envelope (or the Hilbert envelope) is calculated from $x(n)$ as the magnitude of the complex analytic signal $x(n) + j\mathrm{H}\{x(n)\}$, where $\mathrm{H}\{\}$ denotes the Hilbert transform. Hence,

$$\mathrm{E}(n) = |\hat{x}(n)| = \sqrt{x^2(n) + \mathrm{H}^2\{x(n)\}}, n \in \{1, 2, \ldots, N\} \tag{3}$$

Since EGG is not a strictly narrowband signal, the Hilbert envelope is only approximately correct [24]. It needs to be smoothed by a median filter.

Fig. 6 shows an example of EGG utterance and its Hilbert envelope. The upper diagram in Fig. 6 is the original EGG data. The middle diagram in Fig. 6 is the filtered EGG data. The lower diagram in Fig. 6 is the Hilbert envelope after median smoothing.

An appropriate threshold can be selected to divide the samples into two categories according to the Hilbert envelope value. However, the determination of threshold is very difficult and database related. A clustering method based on the assumption of Normal Distribution is adopted to find the voiced speech. The method is shown as follows:

- The input $\mathrm{E}(n), n \in \{1, 2, \ldots, N\}$ is divided into two categories $c_0(n), n \in \{1, 2, \ldots, N_0\}$ and $c_1(n), n \in \{1, 2, \ldots, N_1\}$, according to the median value.
- Calculate the mean $\{m_0, m_1\}$ and standard deviation $\{\sigma_0, \sigma_1\}$ for each class. Since the Hilbert envelope is one-dimensional, the covariance matrix is a scalar for each class.

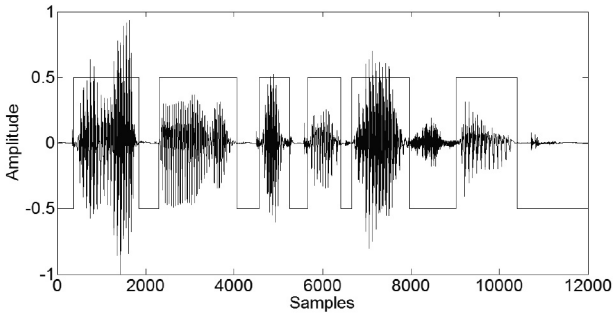


Fig. 7. Result of voiced speech detection.

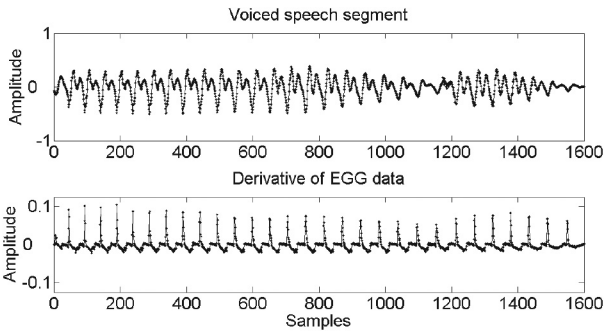


Fig. 8. Speech and derivative of EGG in a voiced section.

Onset	Phoneme_1	Phoneme_2	···	Phoneme_n	Offset

Fig. 9. Typical structure of voiced section.

- Calculate probability B_{ij} for each class as equation (4):

$$B_{ij} = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(E(j)-m_i)^2}{2\sigma_i^2}} \tag{4}$$

  where $i \in \{0, 1\}$, $j \in \{1, 2, \ldots, N\}$.
- Re-categorize each sample based on the principle of maximizing the probability. Repeat the second and third step, until the new centers do not change anymore, or the maximum number of cycles is achieved.

The result of voiced speech detection is shown in Fig. 7.

Fig. 7 shows that the utterance is divided into six voiced sections and seven unvoiced sections. Each section will be treated separately.

D. Voiced Sections Segmentation

In a voiced section, the vocal folds vibrate continuously. In most languages, one voiced section may contain more than one vowel, semivowel or consonant, each of which should be separated from the whole voiced section. Fig. 8 shows the speech and derivative of EGG in a voiced section.

In Fig. 8, the speech data has been aligned to the derivative of EGG. It shows that the shape of speech data in each pitch-cycle changes slowly due to the muscle movement. Similar pitch-cycles belong to the same phoneme. Generally, a voiced speech section has the structure shown in Fig. 9.

The Onset and Offset are transitional regions to the adjacent unvoiced sections. There are a few immature pitch-cycles in Onset or Offset region. In Onset or Offset time, the shape of vocal tract is transformed from unvoiced consonants to voiced phonemes or reverse. Between Onset and Offset, there exist several phonemes. The number of pitch-cycles in each phoneme changes widely, from a few to dozens.

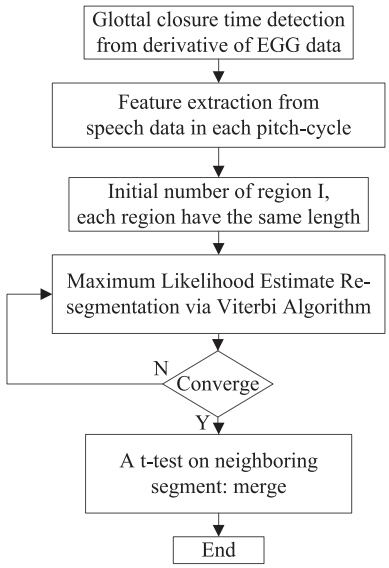


Fig. 10. Adaptive Voiced Sections Segmentation.

For the voiced section in Fig. 8, five regions can be divided such as [Onset P1 P2 P3 Offset], and the number of pitch-cycles allocated to each region is [1 19 3 5 2]. Based on the above analysis, we propose a method to segment a voiced section into phonemes based on a **divide-merge** strategy. The method contains four main steps shown in Fig. 10.

First, the glottal closure times are detected from derivative of EGG data. Then features are extracted from the speech data in each pitch-cycle. These pitch-cycles are clustered into several regions through the Viterbi algorithm, called **divide** process. The number of divided regions should not be less than the real number of phonemes plus Onset and Offset. Finally, a Hotellings T-square test is used to **merge** the adjacent two successive segments.

1) Feature Extraction: The vocal folds create phonation in the glottis and then the phonation is modified by the vocal tract into different vowels and consonants. Generally, the first three formants are enough to disambiguate the vowel. However, there are not only vowels in one voiced section but also some semivowels or consonants. Mel frequency Cepstral Coefficient (MFCC) is a popular feature which has been proved useful in many speech processing fields[25], [26].

The calculation of the MFCC includes the following steps:

- The speech data in each pitch-cycle is padded with trailing zeros to a fixed length n (128 is selected).
- The hamming window is used to avoid the spectrum leakage.
- The magnitude squared discrete Fourier transform (DFT) turns the windowed speech data into the frequency domain so that the short-term power spectrum P(f) is obtained.
- The spectrum P(f) is then filtered by a group of triangular band-pass filters along the Mel-frequency axis. The output is a set of sub-band energies E(d), d = 1,2,. . .D.

- The MFCC is calculated by taking a type II DCT on the logarithm of E(d) as equation (5):

$$C(i) = \sqrt{\frac{2}{D}} \sum_{d=1}^{D} \left[log(E(d)) \cdot \cos \frac{(2d-1)i\pi}{2D} \right] \quad (5)$$

where i = 1,. . .D.

The number of MFCC coefficients D is selected by experiments.

2) Divide by the Viterbi Algorithm: The Viterbi algorithm aims at finding the optimum segmental state sequence to achieve the re-assignment of a sequence to several clusters [27]. To implement the Viterbi algorithm, the following parameters are indispensable:

- $O = (o_1 o_2 o_3 \dots o_N)$: the observation sequence which represents a feature sequence of a voiced section;
- o_j: the observation vector consists of acoustic features;
- I: the voiced section may be divided into I successive segments;
- Segment centroid is assumed to have the *p.d.f* of multivariate Gaussian $p(o_j|\Phi_i)$:

$$p(o_j\,|\Phi_i\,) = \frac{1}{2\pi\sqrt{|\Sigma_i|}} e^{-\frac{1}{2}(o_j-\mu_i)^t \Sigma_i^{-1}(o_j-\mu_i)} \quad (6)$$

The parameter Φ_i includes the mean-vector μ_i and covariance matrix $\sum_i$, which can be obtained from the n_i observation points of the *i*th segment by an unbiased estimator. It is used for deciding which segment the point o_j belongs to.

$$\mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} o_{ij} \quad (7)$$

$$\Sigma_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (o_{ij} - \mu_i)(o_{ij} - \mu_i)^t \quad (8)$$

- The transition probability matrix A, whose (i, j) th element is a_{ij}:

$$A = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & \dots & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & \dots & 0 \\ & \dots & \dots & & \\ 0 & 0 & \dots & & 1 \end{pmatrix}_{I \times I} \quad (9)$$

- $T(j,m) = max_i(T(i,m-1)a_{ij}p(o_m|\Phi_i)), i \in (1,2,..,I), j \in (1,2,\dots,I), \quad m \in (1,\dots,n)$ that finds the state i which most probably precedes state j at time m.

The voiced section can be divided into *I* segments after the Viterbi search. The number of *I* should not be less than the real number of phonemes plus Onset and Offset.

3) Merge by Hotellings T-square Test: A Hotellings T-square test can be used to determine whether the two multivariate data sets are significantly different from each other.

The Hotellings T-square test to determine whether the two adjacent segments should be merged can be carried out as follows:

- The *I* segments obtained using the Viterbi algorithm are merged in a single left to right pass. The length of the window is two and frame shift is one.
- Hypothesis H0: $\mu_1 = \mu_2$, the two segments in the window have no significant difference and should be merged. Select a significance level α which is very important and will be selected by experiment.
- Calculate test statistic value as equation (10) and (11):

$$\mathrm{T}^2 = (\mu_1 - \mu_2)^T \left[\Sigma(\frac{1}{n_1} + \frac{1}{n_2}) \right]^{-1} (\mu_1 - \mu_2) \quad (10)$$

where Σ is the pooled sample covariance matrix namely:

$$\Sigma = \frac{(n_1 - 1)\Sigma_1 + (n_2 - 1)\Sigma_2}{n_1 + n_2 - 2} \quad (11)$$

where Σ_1 and Σ_2 are the covariance matrix of the two segments. μ_1 and μ_2 are the mean of the samples. n_1 and n_2 are the number of elements in each segment.

- For n_1 and n_2 sufficiently large, T^2 follows the chi-square distribution. However, for small n_1 and n_2 as in the case of this article, a better estimate is achieved as follows:

$$F = \frac{n - D}{D(n-1)} T^2 \sim F(D, n - D) \quad (12)$$

where D is the length of feature vector, $n = n_1 + n_2 - 1$.

- Check the F_{crit} using the inverse of the F cumulative distribution function $F_{crit} = finv(1-\alpha)$, if $F < F_{crit}$, accept the hypothesis and merge the two segments. Otherwise, reject the hypothesis.
- Finally, if the Onset or Offset exists (edge regions including only one to two pitch-cycles), they will be classified into the neighboring unvoiced section.

E. Unvoiced Sections Segmentation

Unvoiced sections contain two components: unvoiced consonants and silence. Unvoiced consonants are produced by creating a constriction somewhere in the vocal tract tube and forcing air through that constriction, thereby creating non-linear and turbulent air flow. Multifractal theory is a powerful tool to describe the non-linear and turbulent systems [28].

The Singularity Exponent (SE) h(n), for a signal s(n), describes the local degree of singularity around the point 'n'. Khanagha adopted SE to execute phonetic segmentation of the speech signal [16]. SE can be estimated by evaluation of the power-law scaling behavior of a multi-scale functional Γ_r over a set of fine scales r as equation (17):

$$\Gamma_r(s(n)) = \alpha(n) r^{h(n)} + o(r^{d+h(n)}), \quad r \to 0 \quad (13)$$

where $\Gamma_r(\cdot)$ can be any multi-scale functional complying with the power-law. $\alpha(n)$ is independent of the scale and can be separated from h(n). $o(\cdot)$ means negligible for an adequately small scale.

There exist several methods to estimate SE [29]. For a practical application, we adopt the method proposed by khanagha [16] shown as follows:

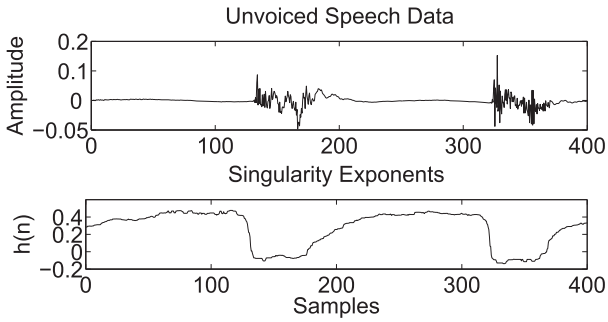


Fig. 11. Singularity Exponents of two adjacent unvoiced sections.

$$\Gamma_{r0}(s(n)) \leftarrow |s(n) - s(n-1)| \tag{14}$$

$$\Gamma_{r1}(s(n)) \leftarrow |s(n) - s(n-1)| + |s(n+1) - s(n)| \tag{15}$$

$$\Gamma_r^K(s(n)) = \sqrt{\Gamma_{r1}(s(n))\Gamma_{r0}(s(n))} \tag{16}$$

$$h(n) = \log \frac{\Gamma_r^K(s(n))}{\langle \Gamma_{r_0}(s(n)) \rangle} / \log r_0 \tag{17}$$

where $r_0 = \frac{1}{fs}$, $\langle \cdot \rangle$ denotes the average value over the whole utterance.

Finally, the curve $h(n)$ is smoothed by a median filter (5 taps). Fig. 11 shows the Singularity Exponents of two adjacent unvoiced sections.

A clustering method based on the Normal Distribution is adopted to distinguish consonants from unvoiced sections, just like the method used to detect voiced speech (see subsection II-C).

III. Experiments

A. Database Description

In order to facilitate other researchers to compare with their own method, some well-known open database should be chosen, such as the TIMIT database [2], [9], [16]. However, the promoted method needs dual-mode data of EGG and voice. The existing open databases consistent with this condition are very rare.

The Database of German Emotional Speech (Berlin EmoDB) is adopted [30]. The Berlin EmoDB has ten actors (5 males and 5 females), 10 German utterances (5 short and 5 longer sentences) which could be used in everyday communication. Besides, there are seven emotions in the Berlin EmoDB, notably (German terms in brackets) neutral (neutral), anger (Wut), fear (Angst), joy (Freude), sadness (Trauer), disgust (Ekel) and boredom (Langeweile).The emotional utterances make the phoneme segmentation more difficult. The database includes 816 utterances.

B. Performance Measures

The segmentation quality can be evaluated and analyzed according to the reference transcription of database. Suppose NT is the total number of detected boundaries (correct and incorrect), NH is the number of correctly detected boundaries and NR is the total number of boundaries in the reference transcription. The relationship of these parameters is shown in Fig. 12.

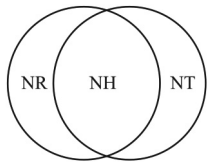


Fig. 12. Relationship of segmentation parameters.

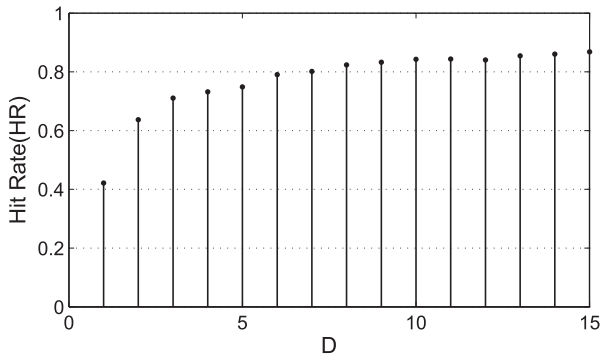


Fig. 13. HR value changes with feature dimension D.

There are two scores: the Hit Rate (HR) and the False Alarm Rate (FA), which can be calculated as equation (18) and (19):

$$HR = \frac{NH}{NR} \tag{18}$$

$$FA = \frac{NT - NH}{NT} \tag{19}$$

In order to assess the overall quality of a segmentation method, a global measure which simultaneously takes these scores into account is required. A well known measure is the F_1 value as equation (20):

$$F_1 = \frac{2 \times (1 - FA) \times HR}{(1 - FA) + HR} = \frac{2 \times NH}{NR + NT} \tag{20}$$

C. Parameters Selection

There are three parameters which may affect the quality of segmentation, including the feature dimension *D*, the number of Viterbi algorithm states *I*, and the Hotellings T-square test significance level α.

The feature dimension D is the number of MFCC coefficients. We fix other parameters as $I = 6$, $\alpha = 0.05$, and change *D* from 1 to 15. The HR value with 30 ms tolerance is shown in Fig. 13.

It shows that the value of HR rises with increasing *D* at first. However, the upward trend becomes unobvious when *D* comes greater than 10, which is selected as the value of *D* in subsequent experiments.

The parameters *I* and α are used in the **divide-merge** procedure for detecting phonemes within a voiced section. Since the real number of phonemes in a voiced section is unknown, the value of *I* should not be less than the real number of phonemes plus Onset and Offset. However, if the value of *I* is too big, the FA will increase correspondingly even. The Hotellings T-square

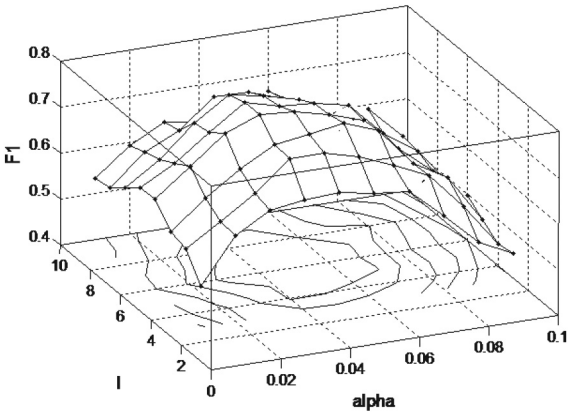


Fig. 14. F_1 values based on different *I* and α.

TABLE I
SEGMENTATION RESULTS FOR DIFFERENT TOLERANCES

Results	HR			FA			F_1		
Tolerances	M0	M1	M2	M0	M1	M2	M0	M1	M2
5 ms	0.25	0.35	0.43	0.57	0.39	0.34	0.31	0.45	0.52
10 ms	0.29	0.43	0.61	0.55	0.39	0.38	0.36	0.50	0.61
15 ms	0.48	0.56	0.70	0.50	0.36	0.33	0.49	0.60	0.68
20 ms	0.51	0.61	0.82	0.42	0.26	0.20	0.54	0.67	0.81
25 ms	0.62	0.69	0.88	0.33	0.20	0.11	0.64	0.74	0.89
30 ms	0.81	0.81	0.98	0.26	0.14	0.09	0.77	0.83	0.94

test significance level α is used to provide a standard for similarity discriminant. The F_1 values based on different *I* and α with 30 ms tolerance are shown in Fig. 14.

The contour lines in Fig. 14 indicate that the F_1 is highest when *I* and α fall in a specific range. Probable reasons are surmised that the range of *I* is close to the real number of phonemes, while the range of α reflects the real differences of phonemes statistical features. The *I* and α are selected as 6 and 0.05 in subsequent experiments.

D. Comparison Test

The proposed methods are compared with the method presented by Khanagha [16]. Khanagha's method is selected for two reasons: first of all, the segmentation result of that method is significantly more accurate than the state-of-the-art ones; Moreover, it is also a sample-based text-independent segmentation method as ours, which does not use any priori information such as known phone sequence or word sequence.

In addition, a text-dependent conventional GMM-HMM method is used by performing a forced alignment against the transcripts. The used features are the same as the proposed methods. The number of Gaussians per state are selected by experiments.

Table I shows the segmentation results of the Khanagha's method (M0), our proposed method (M1) and text-dependent method (M2) for different tolerances.

We learn from Table I that the results of text-dependent method are much better than that of text-independent ones. In the case of text-independent methods, the proposed method has better performance than Khanagha's method based on Berlin EmoDB database. In particular, the smaller the tolerance, the more obvious the relative improvement is. The better performance can be attributed to the combination of EGG and speech.

TABLE II
SEGMENTATION RESULTS FOR DIFFERENT SNR

Results	HR			FA			F_1		
SNR(dB)	M0	M1	M2	M0	M1	M2	M0	M1	M2
clean	0.81	0.82	0.99	0.27	0.14	0.07	0.77	0.84	0.96
20	0.66	0.73	0.89	0.38	0.20	0.10	0.64	0.76	0.89
10	0.55	0.67	0.83	0.51	0.35	0.28	0.52	0.66	0.77
5	0.44	0.49	0.62	0.67	0.49	0.42	0.37	0.50	0.60

Speech database recorded in the lab can achieve a high Signal to Noise Ratio (SNR). However, in practice, the effect of noise cannot be ignored. Since EGG is obtained through electrodes, the interference of noises in the air can be avoided in EGG data. Therefore, most of the noises in speech data can be restrained by combining EGG and speech. Table II shows the segmentation results of the three methods for different SNR with 30 ms tolerance. Different intensity of white noises are added to the speech signal.

From Table II, we can find that the segmentation results of all methods become worse with the decreasing of the SNR. However, the downward trend of the proposed method is more slowly than that of the Khanagha's method and text-dependent method. EGG has made major contributions to this result.

E. Applications

The key practical utility of the proposed method is to be used as the first step in processing data from an unknown source such as real-time voice recognition and Human-Computer Interaction. There are still some significant challenges in Human-Computer Interaction through voice, especially the noisy environments, emotional speakers, and unknown sources. The typical application scenarios include driving automatic control system, remote online education, gaming equipment, etc.

Speech emotion recognition is a major challenge in the field of speech processing. A speech emotion recognition experiment is implemented based on a method proposed by us previously, which combined acoustic information and 'emotional point' information for robust speech emotion recognition [31]. 'Emotional point' was defined as speech clips which reflect rich emotional information. This definition is consistent across different languages and cultures. Phoneme Segmentation is a very important step for 'emotional point' extraction. The proposed phoneme segmentation method is used for speech emotion recognition based on $EmoDB$. In addition, a method without EGG information is tested. The Hilbert envelope of the bandpass filtered audio signal is obtained instead of finding the Hilbert envelope of the EGG signal. Tables III and IV show the results of speech emotion recognition.

It can be seen from Tables III and IV that EGG can significantly improve the results of speech emotion recognition. The average recognition rate increases 6.2 percent. Especially for sadness, fear and boredom, the correct rates increase 13.0, 10.3 and 7.7 percent respectively.

TABLE III
SPEECH EMOTION RECOGNITION RESULTS WITHOUT EGG

Correct	On average: **53.6**					
rate (%)	sad.	ang.	bor.	fea.	joy	dis.
sad.	**54.5**	0	0	24.2	0	21.4
ang.	2.8	**74.4**	8.9	0	13.9	0
bor.	2.7	6.1	**46.4**	11.3	21.5	12.0
fea.	19.8	0	6.8	**43.3**	4.3	25.8
joy	0	0	36.3	0	**54.1**	9.6
dis.	11.3	0	4.4	10.9	24.6	**48.7**

TABLE IV
SPEECH EMOTION RECOGNITION RESULTS WITH EGG

Correct	On average: **59.8**					
rate (%)	sad.	ang.	bor.	fea.	joy	dis.
sad.	**67.5**	0	0	23.3	0	9.2
ang.	4.8	**72.8**	10.9	0.6	9.1	1.9
bor.	3.3	9.7	**54.1**	1.9	16.0	15.0
fea.	21.3	1.4	0	**53.6**	0.6	23.2
joy	1.1	4.5	30.7	1.7	**57.5**	4.6
dis.	6.7	0	6.0	9.8	23.9	**53.6**

IV. CONCLUSIONS

In this paper, EGG is adopted to design a precise and robust text-independent phoneme segmentation method. Unlike the traditional methods, phonemes are firstly classified into two categories named voiced (including vowels, semivowels and some consonants) and unvoiced (other consonants). This classification is performed with sampling point level accuracy by using the Hilbert Envelope feature obtained from EGG. Secondly, the voiced sections and other sections are treated based on a divide-merge strategy and a Singularity Exponent feature separately. Comparison experiments show that the proposed method has excellent performance for a variety of tolerances and SNRs.

REFERENCES

[1] G. Almpanidis and C. Kotropoulos, "Phonemic segmentation using the generalised Gamma distribution and small sample Bayesian information criterion," *Speech Commun.*, vol. 50, no. 1, pp. 38–55, 2008.

[2] J. Yuan *et al.*, "Automatic phonetic segmentation using boundary models," in *Proc. Interspeech*, 2013, pp. 2306–2310.

[3] H. Romsdorfer and B. Pfister, "Phonetic labeling and segmentation of mixed-lingual prosody databases," in *Proc. Interspeech*, 2005, pp. 3281–3284.

[4] S. Hoffmann and B. Pfister, "Fully automatic segmentation for prosodic speech corpora," in *Proc. Interspeech*, 2010, pp. 1389–1392.

[5] J. Amit and E. W. Carol, "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines," in *Proc. Int. Joint Conf. Neural Netw.*, 2003, vol. 1, pp. 675–679.

[6] J. Adell and A. Bonafonte, "Towards phone segmentation for concatenative speech synthesis," in *Proc. 5th ISCA Workshop Speech Synth.*, 2004, pp. 139–144.

[7] D. T. Toledano, L. A. H. Gomez, and L. V. Grande, "Automatic phonetic segmentation," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 617–625, Nov. 2003.

[8] K. Prahallad, A. W. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2006, pp. I1–CI4.

[9] S. Zhao *et al.*, "A hybrid refinement scheme for intra-and cross-corpora phonetic segmentation," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 81–97, 2015.

[10] S. Chen *et al.*, "Automatic transcription of broadcast news," *Speech Commun.*, vol. 37, no. 1, pp. 69–87, 2002.

[11] G. Aversano *et al.*, "A new text-independent method for phoneme segmentation," in *Proc. 44th IEEE Midwest Symp. Circuits Syst. (MWSCAS)*, 2001, pp. 516–519.

[12] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons," in *Proc. Acoust. Speech Signal Process. (ICASSP'08)*, 2008, pp. 3989–3992.

[13] Y. Qiao, D. Luo, and N. Minematsu, "Unsupervised optimal phoneme segmentation: Theory and experimental evaluation," *IET Signal Process.*, vol. 7, no. 7, pp. 577–586, Sep. 2013.

[14] O. Scharenborg, V. Wan, and M. Ernestus, "Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries," *J. Acoust. Soc. Amer.*, vol. 127, no. 2, pp. 1084–1095, 2010.

[15] V. Khanagha *et al.*, "Improving text-independent phonetic segmentation based on the microcanonical multiscale formalism," in *Proc. Acoust. Speech Signal Process. (ICASSP)*, 2011, pp. 4484–4487.

[16] V. Khanagha *et al.*, "Phonetic segmentation of speech signal using local singularity analysis," *Digital Signal Process.*, vol. 35, pp. 86–94, 2014.

[17] R. E. Kania *et al.*, "Fundamental frequency histograms measured by electroglottography during speech: A pilot study for standardization," *J. Voice*, vol. 20, no. 1, pp. 18–24, 2006.

[18] S. M. Prasanna and D. Govind, "Analysis of excitation source information in emotional speech," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010, vol. 1, pp. 781–784.

[19] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA, USA: MIT Press, 2000.

[20] C. M. Sapienza, E. T. Stathopoulos, and C. Dromey, "Approximations of open quotient and speed quotient from glottal airflow and EGG waveforms: Effects of measurement criteria and sound pressure level," *J. Voice*, vol. 12, no. 1, pp. 31–43, 1998.

[21] S. Granqvist *et al.*, "Simultaneous analysis of vocal fold vibration and transglottal airflow: Exploring a new experimental setup," *J. Voice*, vol. 17, no. 3, pp. 319–330, 2003.

[22] J. Neubauer *et al.*, "Coherent structures of the near field flow in a self-oscillating physical model of the vocal folds," *J. Acoust. Soc. Amer.*, vol. 121, no. 2, pp. 1102–1118, 2007.

[23] L. Chen *et al.*, "Speech emotional features extraction based on electroglottograph," *Neural Comput.*, vol. 25, no. 12, pp. 3294–3317, 2013.

[24] N. E. Huang *et al.*, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," in *Proc. Roy. Soc.*, 1998, pp. 903–995.

[25] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *J. Comput. Sci. Technol.*, vol. 16, no. 6, pp. 582–589, 2001.

[26] S. Nakagawa, K. Asakawa, and L. Wang, "Speaker recognition by combining MFCC and phase information," *Spectrum*, vol. 60, p. 76.4, 2007.

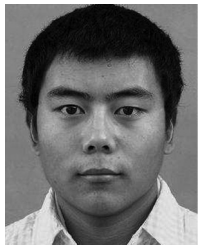
[27] G. D. Forney Jr., "The Viterbi algorithm," *Proc. IEEE*, vol. 61, no. 3, pp. 268–278, Mar. 1973.

[28] D. Harte, *Multifractals: Theory and Applications*. Boca Raton, FL, USA: CRC Press, 2001.

[29] A. Turiel, C. J. Perez-Vicente, and J. Grazzini, "Numerical methods for the estimation of multifractal singularity spectra on sampled data: A comparative study," *J. Comput. Phys.*, vol. 216, no. 1, pp. 362–390, 2006.

[30] F. Burkhardt *et al.*, "A database of German emotional speech," in *Proc. Interspeech*, 2005, pp. 1517–1520.

[31] L. J. Chen *et al.*, "Mandarin emotion recognition combining acoustic and emotional point information," *Appl. Intell.*, vol. 37, no. 4, pp. 602–612, 2012.



Lijiang Chen (M'12) received the B.S. and Ph.D. degrees in electronic and information engineering from Beihang University, Beijing, China, in 2007 and 2012, respectively. He is currently a Lecturer with the School of Electronic and Information Engineering, Beihang University. His research interests include speech signal processing, pattern recognition and speech emotion recognition.

Xia Mao (M'12) received the M.S. and Ph.D. degrees from Saga University, Saga, Japan, in 1993 and 1996, respectively. She is currently a Professor with the School of Electronic and Information Engineering, Beihang University, Beijing, China. Her research interests include affective computing, artificial intelligence, pattern recognition and human-computer interaction. So far, she has authored over 140 pieces of papers both domestically and overseas, many of them have been cited by the SCI, EI, ISTP, etc. She is leading several projects supported by the National High-tech Research and Development Program (863 Program), the National Natural Science Foundation and Beijing Natural Science Foundation.

Hong Yan (F'06) received the Ph.D. degree from Yale University, New Haven, CT, USA. He was a Professor of Imaging Science with the University of Sydney, Sydney, N.S.W., USA, and currently is a Professor of Computer Engineering with City University of Hong Kong, Kowloon Tong, Hong Kong. His research interests include image processing, pattern recognition and bioinformatics. He has authored or co-authored over 300 journal and conference papers in these areas. He was elected an IAPR Fellow for contributions to document image analysis.