C/V Segmentation on Mandarin Speech Signals via Additional Noise Cascaded with Fourier-Based Speech Enhancement System

Ching-Ta Lu¹,², and Hsiao-Chuan Wang¹

¹Department of Electrical Engineering, National Tsing Hua University, Hsinchu, 300, Taiwan, R.O.C.
² Department of Electronic Engineering, Chin Min Institute of Technology, Miaoli, 351, Taiwan, R.O.C.

Abstract — An efficient and simple approach to segment the consonant and the vowel (C/V) speech signals by slightly corrupting a speech signal with white noise is proposed. This approach is based on the vowels are quasi-periodic, and the consonants can be regarded as broad-spectrum noises. A Fourier-based speech enhancement algorithm performs well in enhancing the corrupted vowel signals, and this algorithm fails in enhancing the corrupted consonant signals. Experimental results show that the novel scheme performs well in detecting the C/V boundaries for a spontaneously spoken utterance.

I. INTRODUCTION

Segmenting the C/V is an important task in Mandarin speech recognition, compression and analysis/synthesis systems. Mandarin speech is a monosyllabic language, its phoneme structure is represented as either vowel or consonant follows by vowel signals. There are 37 vowels and 21 consonants for the constitution of 408 phonologically allowed tone-independent syllables as given in [1]. A transient region exists between consonant and vowel signals. How to segment the C/V is still a challenge in speech signal processing. Many algorithms has been proposed for C/V segmentation, including the method using wavelet transform incorporated with energy profile [1], the method incorporating the cepstrum, the zero crossing rate with energy profile [2] were proposed. The difficulties of using the zero crossing rate and the energy profile for detecting C/V boundaries are in setting an appropriate threshold. Inappropriate selection of the threshold would cause a detection error. A voiced consonant in some sense can be treated as the vowel-like signal with low-energy. However the energy of plosive consonant is usually greater than that of the vowel, e.g. the syllable /pa/ as shown in [1]. Based on the above finding, segmenting the C/V according to energy profile should predefine many complex rules to ensure the performance. This paper proposes a novel scheme to detect C/V boundaries by slightly corrupting speech signals with additive white noise without any predefined threshold. Since the vowels are quasi-periodic, and the consonants can be regarded as broad-spectrum noises. A Fourier-based speech enhancement algorithm performs well in enhancing the corrupted vowel signals, and this algorithm fails to enhance the corrupted consonant signals. Consequently, we propose to use the SNR improvement (SNR_Imp) of a frame to be a major factor in segmenting the C/V. Experimental results show that the proposed approach outperforms the method incorporated with cepstrum, zero-crossing rate and energy profile [2].

The rest of this paper is organized as follows. Section II describes the proposed C/V segmentation algorithm. Section III briefly introduces a Fourier-based speech enhancement system. Section IV demonstrates the experimental results. Conclusions are finally drawn in Section V.

II. C/V SEGMENTATION ALGORITHM

The block diagram of proposed C/V segmentation system is shown in Fig. 1. Speech signal is initially framed by Hanning window, hence the speech signal is slightly corrupted by additive white noise d(n) in the time domain with frame size N. The noisy speech signal y(n) can be expressed as

\[ y(n) = s(n) + d(n), \quad 0 \leq n \leq N - 1 \]  (1)

The corrupted signal y(n) is transformed into frequency domain by fast Fourier transform (FFT). The minimum statistics algorithm incorporated with an optimal smoothing parameter is performed to estimate the power of noise for each frequency bin [3]. This algorithm updates the noise estimate in both speech-activity and speech-pause regions. If the number of noise-dominated subband exceeds a predefined value in a frame, this frame is tagged as silence. Due to the speech-pause regions are detected in a slightly corrupted speech signal, the detected results are reliable. Hence a Fourier-based speech enhancement algorithm incorporated with masking properties of a human ear is implemented to enhance the noisy speech. The proposed method aims at detecting C/V boundaries according to the SNR_Imp which is calculated by

\[ \text{SNR}_{\text{Imp}} = 10 \cdot \log_{10} \left( \frac{\sum_{n=0}^{N-1} |s(n)|^2}{\sum_{n=0}^{N-1} |s(n) - \hat{s}(n)|^2} \right) - \lambda \]  (2)

where s(n) and \( \hat{s}(n) \) denote the clean speech and the enhanced signal, respectively. \( \lambda \) represents the predefined SNR value of noisy speech in dB.

The performance of a silence frame should be ignored, so that the SNR_Imp of a speech-pause frame is set to...
zero. If the SNR_Imp of a frame exceeds zero, this frame is decided to vowel. Conversely, a frame is decided to consonant if the SNR_Imp is smaller than zero. An energy variation between previous and present frames $T(m)$ is used to ensure whether a frame is consonant or not. $T(m)$ is defined as

$$T(m) = E(m) - E(m-1)$$

where $E(m)$ and $E(m-1)$ represent the energy values of present and previous frames, respectively.

In case of energy increasing, i.e. $T(m) > 0$, or being a turning point of energy profile, i.e. $T(m) < 0$ & $T(m+1) > 0$, a frame may be a consonant, enabling the candidate flag of consonant, i.e. $C_{\text{flag}} = 1$. In case of the candidate flag of consonant being enabled ($C_{\text{flag}} = 1$), and the SNR_Imp being less than zero, a frame is decided to be a consonant.

$$|\hat{S}(\omega)|^2 = g(\omega) |Y(\omega)|^2$$

where $g(\omega)$ is the gain factor of a frequency bin. The gain factor of a frequency bin can be optimized by minimizing the short-time power associated with the speech distortion $P[E_s(\omega)]$, subject to a constraint on the short-time power related to residual noise $P[D(\omega)]$, below the square value of noise masking threshold (NMT) $T^2(\omega)$:

$$\min_{g(\omega)} \{P[E_s(\omega)]\} \leq P[D(\omega)]$$

subject to the constraint $P[D(\omega)] \leq T^2(\omega)$

The detailed process of estimating the NMT can be found in [4]. If the level of residual noise is below the NMT, then the human ear cannot perceive the residual noise. Accordingly, the spectra of noisy speech are retained to reduce the speech distortion. The spectra of noisy speech should not be changed to preserve the speech quality and to maintain the speech distortion in an acceptable level. In contrast, the spectra of noisy speech should be suppressed when the noise level exceeds the NMT. Hence the perceptual gain factor $g(\omega)$ can be derived according to the same procedure as [5][6], the gain factor is given as

$$g(\omega) = \left(1 + \max\left(\frac{|D(\omega)|^2}{T(\omega)} - 1, 0\right)\right)$$

If the level of corrupting noise exceeds the NMT of a frame, the ratio between the short-time power spectrum of the noise $|D(\omega)|^2$ and the NMT $T(\omega)$ exceeds unity, this gain factor in Eq. (8) becomes $T(\omega)/|D(\omega)|^2$. Conversely, when the level of corrupting noise is below the NMT of a frame, the human ear cannot perceive this

![Block diagram of the proposed C/V segmentation system.](image)
noise. Maintaining the speech quality is obtained by keeping the spectra of noisy speech unchanged. The ratio between the short-time power spectrum of the noise $|D(\omega)|^2$ and the NMT $T(\omega)$ is belows unity. The corresponding gain factor in Eq. (8) is unity. Accordingly, the perceptual gain factor is finally summarized as

$$g(\omega) = \begin{cases} \frac{T(\omega)}{|D(\omega)|^2}, & \text{if } |D(\omega)|^2 \geq T(\omega) \\ 1, & \text{o.w} \end{cases}$$ (9)

Obviously, the gain factor in Eq. (9) is very compact. Adequately estimating the short-time power spectrum of the noise $|D(\omega)|^2$ and the NMT $T(\omega)$ of a frame is sufficient to determine this gain factor without any empirical parameter.

IV. EXPERIMENTAL RESULTS

In the experiments, speech signals are Mandarin Chinese spoken by five female and five male speakers. Noisy speech signals are obtained by slightly corrupting the speech signals with white noise (average segmental SNR = 10 dB). The minimum statistics algorithm incorporated with an optimal smoothing parameter is performed to estimate the power of noise for each frequency bin [3]. The following parameters are used in the experiments: (1) sampling frequency is 8 kHz; (2) the frame size $N$ is 256 with 50% overlap; (3) Hanning window is utilized; (4) total number of critical bands is 18.

In Fig. 2, an utterance is spontaneously spoken by a female speaker with syllables /chi shi ta shii huan he cha dan ta being u i jii sher fan ing i jiian/. Fig. 3 demonstrates an example of utterance fluently spoken by a male speaker with syllables /ta men shuo gen jiui mei gu shiian fa i ou gu hui iou jiuan li shiuan ian/. It can be found that despite an utterance is spoken by a female or a male speaker, the proposed feature SNR improvement (SI) is the most discriminative.

Figs. 2 and 3 compare three types of feature used in C/V segmentation. Obviously, the proposed feature SNR improvement (SI) is more discriminative than cepstrum (Cep) and energy (Egy) profile, so the SNR improvement is proposed to be a major feature for C/V segmentation.

In Fig. 4, an utterance is spontaneously spoken by a female speaker with syllables /chi shi ta shii huan he cha dan ta being u i jii sher fan ing i jiian/. Fig. 3 demonstrates an example of utterance fluently spoken by a male speaker with syllables /ta men shuo gen jiui mei gu shiian fa i ou gu hui iou jiuan li shiuan ian/. It can be found that despite an utterance is spoken by a female or a male speaker, the proposed feature SNR improvement (SI) is the most discriminative.
comparison [2]. This algorithm uses a multi-feature classification scheme as well as signal-dependent initial-thresholds, and a different cepstral weighting function, which improves the detection of low-frequency pitch. Fig. 4 demonstrates an example of C/V segmentation. The C+Z+E method is slightly better able to detect consonant regions. In contrast, the proposed scheme significantly outperforms the C+Z+E method in detecting vowel regions.

A C/V segmentation system has two possible types of error. The first one is described by precision (PRC) rate which specifies an error if a detected C/V segmentation does not correspond to a true C/V segmentation in the reference (manually labeled version). The second error is represented by recall (RCL) rate which specifies an error when detected C/V segmentation does not correspond to a true C/V segmentation in the reference, the PRC and the RCL rates are given as [7]

\[
\text{PRC} = \frac{\text{number of correctly found C/V}}{\text{total number of C/V found}}
\]

and

\[
\text{RCL} = \frac{\text{number of correctly found C/V}}{\text{total number of correct C/V found}}
\]

In order to compare the performance of different C/V segmentation system, the F-measure is used and is defined as [7]

\[
F = \frac{2 \times \text{PRC} \times \text{RCL}}{\text{PRC} + \text{RCL}}
\]

The F-measure varies from 0 to 1, with a higher F-measure indicating better performance. Table 1 compares the C/V segmentation results with references. A correct condition occurs when the segmented C/V frame is correctly classified as reference. The C+Z+E algorithm slightly performs better than the proposed method in terms of PRC rate, however the proposed approach significantly outperforms the C+Z+E method in terms of RCL rate. It enables the performance of our approach to be better than that of the C+Z+E method in terms of F-measure, which fact reveals to that the proposed method is suitable to be applied in C/V segmentation.

Table 1. Performance comparison in terms of precision (PRC) rate, recall (RCL) rate and F-measure for female and male speakers

<table>
<thead>
<tr>
<th>measure</th>
<th>method</th>
<th>consonant</th>
<th>vowel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td>PRC</td>
<td>C+Z+E</td>
<td>0.73</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>proposed</td>
<td>0.64</td>
<td>0.75</td>
</tr>
<tr>
<td>RCL</td>
<td>C+Z+E</td>
<td>0.41</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>proposed</td>
<td>0.52</td>
<td>0.94</td>
</tr>
<tr>
<td>F-measure</td>
<td>C+Z+E</td>
<td>0.45</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>proposed</td>
<td>0.49</td>
<td>0.83</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS

The Fourier-based speech enhancement system fails to enhance the noisy speech slightly corrupted by white noise (with average segmental SNR = 10 dB) for a consonant, and performs well for a vowel. Segmenting C/V is obtained via additional noise cascaded with a Fourier-based speech enhancement system. Experimental results show that the proposed approach outperforms a method incorporated with cepstrum, zero-crossing rate, and energy profile.

ACKNOWLEDGEMENT

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC-93-2213-E-007-094.

REFERENCES