1pSCb6. Investigation of calculation methods of effective duration on autocorrelation-based Chinese speech recognition

Kun Peng Huang* and Yoichi Ando

*Corresponding author's address: GuoGuang Electric Company Ltd., Guangzhou, 510800, Guangdong, China, ohmycar@163.com

The factors extracted from running autocorrelation function (ACF) have phonetic and syllabic meanings, and they can be used for Chinese speech recognition. One of the factors is the effective duration of ACF envelope $\tau_e$, which reveals the temporal repetitive continuity of the speech signals. The values of $\tau_e$ vary when different calculation methods are adopted. By adopting different calculation methods to several sets of Chinese vowels and characters, which were recorded from a group of speakers, different values of $\tau_e$ were obtained. Computed results were analyzed by combining with other ACF-based factors.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

Previous studies show that temporal factors extracted from autocorrelation function (ACF) are associated with auditory percepts (Ando, 1998), and they can be used for speech recognition (Ando et al., 2009). The factors of \( W_{\phi(0)} \), \( \tau_1 \) and \( \phi_1 \) are defined as shown in Figure 1(a). Timbre (spectral tilt) is described by \( W_{\phi(0)} \). Pitch period and pitch strength are described respectively by \( \tau_1 \) and \( \phi_1 \) (Ando, 1998; Inoue et al., 2001), while loudness is described by listening level (LL) and the effective duration \( \tau_e \) (Ando et al., 2000, Sato et al., 2002).

The effective duration \( \tau_e \) is the delay time when the normalized running ACF (r-ACF) decreases to 10% of zero-lag ACF(0), and it is extrapolated from the initial envelope of r-ACF in decibels as shown in Figure 1(b). \( \tau_e \) describes signal repetitiveness. The low value of \( \tau_e \) indicate signals are less repetitive or changing more abruptly. Hence \( \tau_e \), especially the local minimal \( (\tau_e)_{\text{min}} \) is important for speech recognition in finding signals transition points, e.g. the points where consonants transit to vowels.

The envelope is usually formed by the “peak detection method”, and how the peaks are selected or the computation methods of \( \tau_e \) have been investigated by some studies (D’Orazio et al., 2011, Kato et al., 2007, Sato et al., 2011). These studies are mainly on music signals or on a general scope. D’Orazio (2011) pointed out that in some cases like instrumental attacks or vocal glitches, peak detection above the -5dB range of normalized r-ACF, that is with a threshold of -5dB may produce meaningless results. This study will investigate the computation method of \( \tau_e \) on speech signals, especially on Chinese speech signals.

COMPUTATION METHODS

Chinese Speech Signals Recording and Analysis Methods

Introduction to Chinese language syllables

Mandarin language is the most spoken language among Chinese population, and each sentence or phrase is composed by Chinese characters. Each character has one spoken syllable, and each syllable has an initial consonant and a last vowel (CV). The combination of consonant and vowel is spoken in a specific tone. Usually there are four kinds of tones in Mandarin syllables. Syllables having the same CV with different tones are spoken by different characters. The phrase of “Good morning” for instance, includes three character or syllables in Chinese, and they are “/zao/ /shang/ /hao/”. The /hao/ means “good” in English, and it is spoken in the third tone. If it is spoken in the fourth tone, it can mean the verb of “love” in “love (doing) something”. According to GB 2312-1980 (National Standards of the P. R. C., 1982), there are 407 syllables in 6763 Chinese characters. If tones are counted, the total syllables are no more than 1500. Hence for speech recognition, to establish a database of such basic syllables is important.
Signals Recording and Analysis

In order to investigate the computation method of \( \tau_e \), some CVs (syllables) in the first tone were recorded by the author and other three speakers, though the set of one speaker is used for study in this paper. The signals were sampled at 16 kHz and 16-bit, and were filtered by an A-weighting network (American National Standards Inst., 2001) to approximate the signal passing through the ear (Ando, 1985, 1998). The short-term r-ACF is computed by

\[
\Phi_p(\tau; t, T) = \frac{1}{2T} \int_{t-T}^{t+T} p(t) p(t+\tau) dt
\]

(1)

Where \( p(t) \) is the A-weighted signal and \( 2T \) is the integration interval, which is selected as 40 ms (Mouri, 2001). For the \( t,e \), an interval of 20 ms is selected to be r-ACF analyzed, i.e., r-ACF is analyzed frame by frame. The normalized r-ACF is computed by

\[
\phi_p(\tau) = \phi_p(\tau; t, T) = \frac{\Phi_p(\tau; t, T)}{\left[\Phi_p(0; t, T)\Phi_p(0; t+\tau, T)\right]^{1/2}}
\]

(2)

When \( \tau=0 \), the zero-lag \( \phi_p(0, t) = 1 \).

For the peak detection in \( \tau_e \) computation, the local maximal peak smaller than the adjacent ones are discarded (D’Orazio et al., 2011), and such procedure is iterated in computation program. The number of iteration times is large enough to discard all peaks smaller than the adjacent ones.

The factor LL is computed and used for the segmentation of each CV, for CVs are recorded continuously with a short pause between two CVs, for instance the sequence composed by 15 syllables or CVs of “/ba/ /bo/ /bai/ … /bian/ /bin/ /bing/”, where the normalized signals above -30dB are accounted for syllables signals. \( W_{\phi(\tau)} \) and \( \tau_1 \) are also computed for helping the analysis of local maxima \( (\tau_e)_{\min} \).

RESULTS AND DISCUSSION

Values of \( \tau_e \) under Different Thresholds

The initial part of r-ACF envelop is used to compute \( \tau_e \), and the threshold settings produce different \( \tau_e \) values. The thresholds from -8dB to -2dB, with interval of 0.2dB are selected to compute \( \tau_e \). Figure 2 is an example of the computed results. As can be shown that different thresholds produce different \( \tau_e \) values, for different initial local maxima are counted for extrapolation. Peaks selected for extrapolation are less when threshold is high, while selected peaks are more when threshold is low. It is hard to say which threshold is the most suitable in the general sense.

FIGURE 2. \( \tau_e \) values computed from different threshold settings of -6dB, -7dB and -8dB. Normalized r-ACF is one frame selected from the CV of /ba/. Each set of line in color represents the threshold and the according linear fitting line obtained by LMS method. The peaks above the threshold are used for linear fitting, and discard procedure is implemented in advance.

The Local Minima Signifying Fluctuations of Pitch Period and Spectral Tilt.

For \( \tau_e \) represent signals changes or fluctuations, the dip points or the local minima \( (\tau_e)_{\min} \) can especially signify where speech signals change abruptly (Ando, 1985, 1998), further more \( (\tau_e)_{\min} \) may also signify the transition point
from the consonant to vowel of each CV, and can help finding the onset points of the syllables. Hence it can be assumed that the changes of pitch period ($\tau_1$) and spectral tilt ($W_\phi(0)$) are closely related to ($\tau_e$)$_{\text{min}}$.

The factors $\tau_1$ and $W_\phi(0)$ are computed for the recorded signals, and combined with the frames where ($\tau_e$)$_{\text{min}}$ are observed, the computation method of $\tau_e$ can be evaluated. Figure 3 is an example of the computed results. For the ($\tau_e$)$_{\text{min}}$ can be quite large like 40ms, though its representation of signal change is not as suitable as that of ($\tau_e$)$_{\text{min}}$ in small value like under 10ms, the ($\tau_e$)$_{\text{min}}$ larger than 40ms are ignored. Figure 3 reveals that

1. Generally, minima appear with the changes of $\tau_1$ or $W_\phi(0)$;
2. The ($\tau_e$)$_{\text{min}}$ as indicated by c1, c3 and c5 can signify the onsets of CVs, which is useful for segmentation of CVs, however there are thresholds under which onsets are not able to be found, either threshold is low (c1) or high (c5).
3. ($\tau_e$)$_{\text{min}}$ as indicated by c2 and c4 show that for a certain frame, minimum is found in one method while not found in other methods.
4. The finding of CV endpoints cannot be counted on ($\tau_e$)$_{\text{min}}$ for minima are rarely found near the time frames of endpoints computed by LL factor. However the value of $\tau_1$ changes abruptly near the endpoint, such phenomenon is quite useful for finding the endpoints of CVs.
5. Different thresholds produced different sets of ($\tau_e$)$_{\text{min}}$ and each set signifies some changes of $\tau_1$ or $W_\phi(0)$.
6. If one threshold has to be chosen for computation, the threshold between -5dB and -6dB is recommended, for the method under such thresholds produce more minima. However threshold-changeable methods could be most suitable, which can be the further investigation.

CONCLUSIONS AND FUTURE DIRECTIONS

Some sets of Chinese CVs or syllables are recorded and analyzed. $\tau_1$, $W_\phi(0)$, and ($\tau_e$)$_{\text{min}}$ are extracted from the signals and analyzed. Different thresholds are set for different computation methods to obtain various $\tau_e$ values. The results reveal that ($\tau_e$)$_{\text{min}}$ are important in finding the onsets of CVs, and signifying the abrupt changes within CVs. Results also show that each set under a different threshold signifies a part of changes of $\tau_1$ or $W_\phi(0)$.

When an auto speech recognition (ASR) system is established, the accuracy rate of the ASR can be very useful in finding a suitable threshold, or a threshold-changeable method should be developed if the fixed-threshold method is not sufficient enough, which will be the future investigations.

ACKNOWLEDGMENTS

The author would like to thank Professor Ando Yoichi, who encouraged and advised the author implemented the ACF method in Chinese speech language, and Kato Kosuke, who gave the author some advice on the computation methods of effective duration times.
REFERENCES


