

A NEW APPROACH TO SEGMENT AND DETECT SYLLABLES FROM HIGH-SPEED SPEECH

D.W. Ying^{1,2}, W. Gao^{1,2}, W.Q. Wang¹

¹(Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P. R. China, 100080)

²(Graduate School of Chinese Academy of Sciences, Beijing, P. R. China, 100039)

{dwyang, wgao, wqwang}@jdl.ac.cn

ABSTRACT

In this paper, we present a novel method to detect sound onsets and offsets, and apply it to detect and segment syllables from high-speed speech according to the Mandarin characteristic. Our system detects onsets and offsets in 8 frequency bands by a two-layer integrate-and-fire neural network. The continuous speech is segmented based on the timing of onsets and offsets. And the energy is used as another cue to locate the segmentation point. In order to improve the accuracy of segmenting, we introduce three time constraints by defining three refractory periods of neurons, which make syllable length no less than the minimum. Although the boundaries between syllables in high-speed speech are not salient, our system can still segment individual syllables from speech robustly and accurately.

1. INTRODUCTION

As the development of information technology, the more powerful tools are needed to manage and storage speech data. The traditional methods take sentence or paragraph as the meta-data, which can't meet the future requirements. We need a novel method to process speech data based on syllable meta-data. For example, in large transcribed news speech database, we need to align the syllable with the text and be able to process or query speech data in finer level.

Recently, a number of studies dealing with syllable detection and segmentation in different languages have appeared. Shastri [1] applied a temporal flow neural network and modulation-filtered spectral features to segment English syllables. The onset of syllabic constituents was identified by a two-level dynamic thresholding method. She evaluated her algorithm based on the examination to several hundred individuals and reported an accuracy of 84%. Shire [2] estimated the locations of syllable onsets based on the log-RASTA features and the spectral features. A MLP neural network is utilized to produce a measure of syllable onset probability. Smith [3] presented an approach of sound

segmentation based on detecting onsets and offsets. According to the timing of the onsets and offsets, a multi-level tree is built to segment sound. The paper gives no exact results of experiment. According to the characteristic of Czech, Kopecek [4] performs syllable segmentation based on optimization of the coincidence of segment boundary positions and segment boundary attributes. Meinedo [5] derives a method from Wu [6] to detect syllable onsets in Portuguese language. Both of them analyze energy trajectories in critical bands to estimate the location of syllable onsets. Zhang Hong [9] also used half-wave differential spectra to segment Mandarin speech into syllables. She detects offsets and onsets by one-order difference in intensity envelope.

Compared with other languages, Mandarin is a single-syllable language. Every word has only one syllable, consisting of a consonant and no more than two adjacent vowels. The onsets and offsets of vowels are much more salient than that of consonant. They provide cues to segment speech.

However, in high-speed speech, the onsets of some syllables are not salient enough to be detected. Moreover, there exist lots of liaisons, and the boundaries of adjacent syllables within liaisons are so fuzzy that segmenting liaisons into individual syllables is very difficult. We present a novel method to segment Mandarin syllables. Our system is a two-layer neural network architecture incorporating the Integrate-and-Fire Model. Compared with Smith's scheme, it can calculate the onset (or offset) time directly, since a two-layer neural network is applied. The neural network is simple, no feedback between neurons, and the differential intensity envelope with a 4ms window is calculated, which greatly accelerates the calculation efficiency. Three time constraints are exerted in the system to improve the segmentation accuracy and computing speed. The rest of this paper is organized as follows. Section 2 describes the proposed algorithm in detail. Then experimental evaluation of the system is given in Section 3. Section 4 concludes the paper.

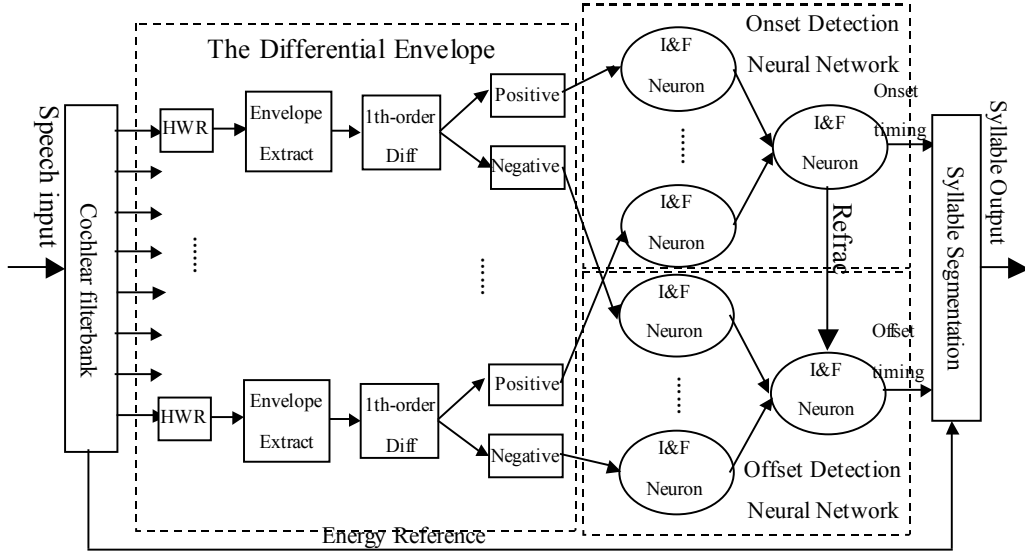


Fig. 1. The framework of the system

2. OUR APPROACH

2.1. Leaky integrate-and-fire neuron model

The simplest leaky integrate-and-fire neuron model [7] consists of a capacitor C in parallel with a resistor R driven by a current $I(t)$. The driving current can be split into two components, $I(t) = I_R + I_C$. I_R can be calculated from Ohm's law as $I_R = u/R$. I_C can be calculated as $I_C = C du/dt$. Thus

$$I(t) = \frac{u}{R} + C \frac{du}{dt} \quad (1)$$

A current $I(t)$ charges the RC circuit. The voltage $u(t)$ across the capacitance (points) is compared to a threshold value. If $u(t)$ reaches the threshold value, an output spike is generated. Immediately the potential is reset to a new value (less than threshold). After the $u(t)$ hold the value for a refractory period, the dynamics is again given by (1) until the next threshold crossing occurs. Let $A(n) = C * u(n)$, $\gamma = \frac{1}{C}$, $R=1$ and we differentiate the (1). Then we get the following equation:

$$A(n+1) = A(n) * (1 - \gamma) + I(n) \quad (2)$$

2.2. Detecting the onsets and offsets

The system tries to detect the timing of those syllable onsets or offsets based on the integrate-and-fire neural network. Then the start and the end of individual syllables are further located. The framework of the whole system is illustrated in Fig.1, which involves a filter-bank, a differential intensity envelope extractor, a two-layer integrate-and-fire neural network, and a syllable segmentation module.

First, the input speech is decomposed into 8 frequency channels using 8th-order gammatone filters [8]. For each band, the output is half-wave rectified. Then the intensity is calculated for each 4ms window to obtain the intensity envelope with 250 samples per second. The intensity envelope is differentiated to capture the energy change trends. To remove impulsive noise, the results are smoothed with the combination of a median filter and a linear filter. At last, the differential envelope is divided into the negative envelope and the positive envelope, and they are respectively input into the onset detection NN and the offset NN.

In onset detection NN, each channel inputs its positive envelope to the corresponding neuron in the first layer. The neuron calculates the integral of the differential envelope. The constant γ in the equation (2) characterizes the rate of leaking. When the rate of inputting exceeds the rate of leaking, the potential $A(n)$ of the neuron increases. If it reaches a threshold, the neuron emits a spike and then the potential resets to an initial value, which means that the frequency channel detects an onset at that time. A time-dependent current of exponential attenuation is used to model the spikes, as the formula (3), where $\alpha_i(n)$ is the output of the i th neuron in the first-layer at the time n . τ is a constant to characterize the attenuation rate, $n^{(i)}$ is the time step at which the i th neuron fires.

$$\alpha_i(n) = \frac{1}{\tau} \exp\left(-\frac{n - n^{(i)}}{\tau}\right) \quad i = 1, 2, \dots, 8 \quad (3)$$

The spikes are transmitted to the corresponding neuron in the second-layer with one-step delay. The input of the neuron in the second layer is specified by the formula (4),

$$I_2(n+1) = \sum_{i=1}^{32} \alpha_i(n) \quad (4)$$

In the same way, when the neuron reaches a threshold, it fires and the system will detect an onset at that time. Fig. 2 shows how the spikes' timing of the first-layer neurons corresponds to the timing of the second-layer neuron.

The similar procedure is applied to detect the offsets. But the corresponding neuron receives the negative input signal. In the post-process step, a different modifying strategy is adopted (as described in 2.3).

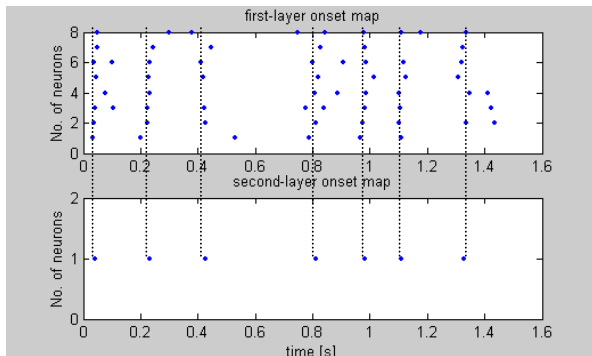


Fig.2. Onset maps from the speech “jian|jiao|hou|fa|zhong|liang|guo is the first-layer onset map of 8 neurons, the below panel is the second-layer onset map of one neuron.

To improve robustness and accuracy, the system exploits three time constraints to limit the minimum of syllable length. The constraints are implemented by three classes of refractory periods. The onset refractory period constrains the minimum interval between the beginning points of adjacent syllables. The offset refractory period constrains that between the ends of adjacent syllables. The onset-offset refractory period (the broken line in fig.1) constrains the minimum interval between the syllable start and its end. All the intervals must be no less than the corresponding minimum period. The values of the refractory periods depend on the speed of continuous speech.

2.3 Syllable segmentation and detection

In fact, it is difficult to detect the onsets or offsets of all syllables. Many syllables miss either onsets or offsets. Onsets have to combine with offsets to provide sufficient cues. And the timing of onsets or offsets can't be regarded as the actual syllable starts or ends. Usually, the energy at syllable starts and ends is the local minimum. We need to infer the segmental points from the timing of onsets and offsets, according to the energy. Since there is always a consonant before the vowel in the syllable for Mandarin, syllable onsets lag behind syllable starts. On the other hand, because the decay process of signal energy will not end immediately, syllable offsets produce ahead of end. All the timing of onsets and offsets are placed on a time axis to segment speech in a complementary manner. The corresponding strategies of syllable segmentation are described as follows:

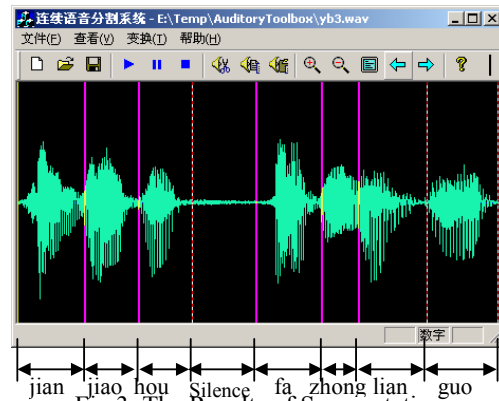


Fig.3. The Results of Segmentation

1. In the range of 0~80ms prior to the timing of onsets, the point, which has a local minimum energy, is taken as the syllable start. In the range of 0~80ms behind the offset timing, the point with a local minimum energy is taken as the syllable end.
2. If the interval of two adjacent points doesn't exceed 30ms, the energy of them is compared and the point with smaller energy is deleted.
3. If the interval of two adjacent points doesn't exceed 50ms, the part between the points is an invalid syllable.
4. If the syllable energy is less than 30% of the average, the syllables are invalid. In most cases, they are either silence or noise.

The fig.3 is the automatic segmentation result of the seven syllables ‘jian|jiao|hou|fa|zhong|liang|guo’. The solid lines represent the starts of the syllables and the broken lines represent the ends.

3. EXPERIMENTAL EVALUATIONS

The experimental data comes from the news of China Central TV station, which is spoken in the standard Mandarin at the speed of 300~320 words/min (ordinary speed: 120~250 words/min). We select 100 sentences as the test data, which contains 3502 words (or syllables) and lasts 61s totally. The sampling rate of data is 16kHz. To check the correction of the automatic segmental points, we label the segmental points by hands. If the interval between the automatic segmental point and the labeled point doesn't exceed 30ms, the automatic points are regarded as a correct point; otherwise, it is regarded as an added point. If there are no automatic points nearby the labeled point, we take the automatic point as a leaked point.

To evaluate the performance of the system, the following three measures are used. The add rate is defined as $P_A = \frac{N_A}{N_T} \times 100\%$, where N_A is the number of added points

and N_T is the total number of segmental points labeled by hand. The Leak rate is defined as $P_A = \frac{N_L}{N_T} \times 100\%$, where N_L

is the number of leaked segmental points. The error rate is the sum of P_A and P_L , defined as $P_E = P_A + P_L$.

We observe the onset detection is much more reliable than the offset detection. The refractory period of onset detection NN in all of the parameters plays an important role on the system performance. We keep the offset period (0.08s) and the onset-offset period (0.07s) invariant, and change the onset period. The table.1 tabulates the experimental results under different onset periods. The statistics in the table.1 shows the refractory period has different influence on the added rate and the leak rate. The add rate declines as the period increases, while the leak rate increases. The phenomenon can be understood and explained from two aspects. First, syllables in the liaison usually are shorter than ordinary syllables, so a shorter refractory period is desirable to segment liaisons. Second, some syllables with double adjacent vowels have two onsets in a syllable. If the refractory period is too short, the syllables will be oversegmented, which makes the added rate increase. Thus a larger refractory period, which exceeds the interval between two adjacent onsets within a syllable, is advantageous. But the liaison will be unsegmented in the case, and the liaison rate will increase. Therefore it is necessary to balance between the add rate and the leak rate. The experimental results in the table.1, suggest onset period between 0.04~0.08s seems to be an appropriate range, and error rate is relatively smaller.

Period	P_A	P_L	P_E
0.02	7.25%	4.03%	11.28%
0.04	3.97%	4.58%	8.55%
0.06	3.47%	5.20%	8.67%
0.08	2.85%	6.01%	8.86%
0.10	3.03%	7.50%	10.53%
0.12	3.04%	10.16%	13.20%
0.14	2.85%	12.08%	14.93%

Table.1 Performance Under Different Onset Refractory Period

If the refractory period is short enough, almost all the liaison syllables excluding the elision can be segmented. In fact, most of the leak rate is produced by elision.

The time constraints not only lower the error rate, but also accelerate the computing speed. The refractory periods restrain lots of added segmental points, which make the system waste little time to process the added points. The computing speed is improved when the periods are extended. So the computing speed is another factor to be considered. If the error rates are close enough, we are inclining to choose the longer period. By experiment, when the onset period is about 0.06 sec, the

system performance is most excellent. It can run in real-time mode (CPU1.4GHz, 256M Memory) with the error rate 8.67%.

Comparing with [9], we present a more robust method to detect sound onsets and offsets. And offset is used as a complementary cue, which declines the leak rate. The time constraints not only accelerate the computing speed, but also lower the add rate. Although the speech speed of our corpus is twice as much as that of ordinary speech, our segmentation accuracy is close to [9] (error rate: 6.79%).

4. CONCLUSIONS AND FUTURE WORK

The experiments show that the algorithm has a good performance to detect syllables from high-speed speech. Compared with other algorithms, our algorithm runs more accurately and robustly in real-time manner. However the energy feature isn't sufficient enough to validate outputted syllables. The more advanced feature such as pitch or cepstrum should be introduced into the algorithm.

Our algorithm can be applied to not only Mandarin, but also English. Segmentation is not an end in itself: the effectiveness of any technique will depend on the eventual application.

5. REFERENCES

- [1] L. Shastri, et al, Syllable detection and segmentation using temporal flow neural networks, *JCPHS*, 1999.
- [2] Shire, M. L. Syllable onset detection from acoustics, *Master's Thesis*, EECS Dept., University of California, Berkeley, 1997.
- [3] L.S Smith. Onset-based sound segmentation. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing System 8*, MIT Press, 1996.
- [4] I. Kopecek, Automatic Segmentation into Syllable Segments, *Proceedings of First Int. Conference on Language Resources and Evaluation*, 1998, pp. 1275-1279.
- [5] H. Meinedo, et al, *The use of syllable segmentation information in continuous speech recognition hybrid systems applied to the Portuguese language*, In Proceedings ICSLP, 2000.
- [6] Wu, S.-L., Kingsbury, B., Morgan, N. and Greenberg, S. "Incorporating information from syllable-length time scales into automatic speech recognition, *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998.
- [7] W. Gerstner, *Spiking Neuron Models: Signal Neurons, Populations, Plasticity*. Cambridge University Press, 2002.
- [8] Slaney, M. An efficient implementation of the Patterson-Holdsworth auditory filter bank. *Apple Computer Technical Report #35*, 1993.
- [9] Zhang Hong, et al, Segmentation of speech signal based on the half-wave differential spectrum, *China Journal of Acoustic*, 2000.