Automatic Segmentation of Chinese Mandarin Speech into Syllable-like

Jian Li,Furao Shen
National Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing, China
lj12@software.nju.edu.cn, frshen@nju.edu.cn

***Abstract*—In this paper, we propose a novel approach to automatically segment a continuous Mandarin speech into syllable-like units. The main idea of our algorithm is to merge two round distinct selected boundaries using feature information generated from time and frequency domain. The first round segmentation is mainly based on aggregating the characteristics of different frequency regions and the second round employees zero crossing rate to improve the accuracy of segment results. The experimental results indicate that our hybrid method has high accuracy and coverage with respect to the reference boundaries even in low error tolerance.**

***Keywords*-automatic speech segmentation; Chinese Mandarin Speech; syllable segmentation**

I. Introduction

Speech segmentation plays an important role in speech-related applications such as speech recognition systems and text-to-speech synthesis. Taking speech recognition for example, one of the significant benefits from segmentation is the smaller linguistic search space coming with the computation complexity reduction. And with the help of segmentation, we are able to train acoustic model from end to end without complex alignment before that. However, segmentation is a tedious and time-consuming work, which leads to the research on automatic speech segmentation.

In the last few decades, groups of researchers were focused on extracting valuable features or representations from speech signals and improving the algorithms to process them. As energy is a most typical feature of speech, Jittiwarangkul [1] applied five different versions of energy to segment connected speech based on the local maximum and minimum energy contour and then compared the performance. In fact, energy can be regarded as the same as a loudness function computed by power spectrum. And to some degree, the most influential segmentation algorithm to process this kind of function is Mermelstein's Convex Hull algorithm [2], which the syllable boundaries are detected from the difference between the convex hull of the loudness function and the loudness function itself. In [3], Prasad used minimum phase group delay function to process the short-term energy and get a better representative for syllable boundary detection. Then this method was improved to subband-based version by Murthy [4] through exploiting the additive property of the Fourier transform phase and the deconvolution property of the cepstrum. Despite of these, some hybrid approaches that seem more robust were demonstrated in [5], [6]. Xie described a method to detect syllabic nuclei by employing periodicity and energy. Zhao utilized silence detection, convex hull analysis and spectral variation analysis to segment the Mandarin speech more reliably and illustrated the rules of Zero-Crossing Rate in speech segmentation.

Besides the algorithms that directly get the segmentation results from certain features of speech, the modeling techniques such as HMM/ANN/GMM are widely used to improve the performance. In [7], based on a common syllable model, Nakagawa got the segmentation boundaries by finding the optimal HMM state sequence with Viterbi algorithm. Brugnara [8] also applied HMM on the task of segmenting and labeling of speech. In [9], Shastri delineated the temporal boundaries of syllabic units in continuous speech using a Temporal Flow Model(TFM). Recently, there are also some researchers trying to use deep learning technology to solve this problem.

Although modeling method always seems better, the performance of this kind of methods are subject to the specific language and the train data. Hence, they always need a lot of speeches with segment references and periods of time for training and parameter tuning. Only when the model is well built, can we have the good performance of segmentation. But if there comes some new kind of utterances that far from the previous train data, the model need to be retrained to adapt to it which certainly cost a lot of time. In this paper, we are not going to use model-based method. The main idea of our method is to combine the features from time domain and frequency domain with two round distinct selection, then merge them into final segment result. The algorithm is not only simple and intelligible, but also effective and powerful.

II. Phonology of Chinese Mandarin

Syllable is the basic unit in speech sounds that typically made up of syllable nucleus (most often a vowel) with optional initial and final margins (typically, consonants). For example, the word $hello$ has two syllables: he and llo. However, unlike English, every word in Chinese Mandarin is completely one syllable consisting of consonant and vowel. Hence, segmenting the Mandarin speech into syllable-like also means single-word-like, which will help a lot to the further research such as text-speech exchange systems or speech recognition systems for Chinese.

In this paper, the method proposed below is not going to segment the speech directly to syllable-like at first, because some consonant in Mandarin speech may act as

978-1-4673-9596-0 /15/$31.00 ©2015 IEEE

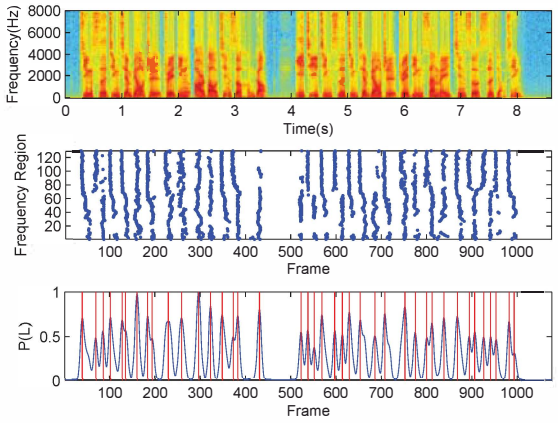

Figure 1. *Spectrogram(top), Time-frequency positions of landmark (middle) and time positions of Landmark (bottom) with respect to P(L).*

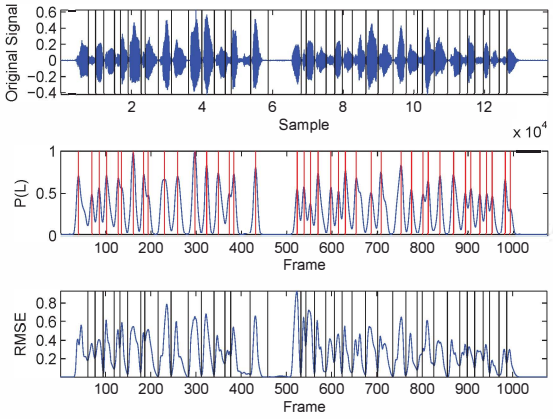

Figure 2. *Landmark generated by P(L) (middle) and boundary* (bottom) generated through RMSE of the original speech signal (top).*

syllabic one such as $z\ c\ s\ zh\ ch\ sh$ and lead to the wrong segment boundary that separate the consonant and vowel in one integral syllable. In order to make accurate segmentation, the method proposed should be able to correct this kind of mistakes which is essential to the Mandarin segmentation performance.

III. Proposed Method

Based on the phonology of Chinese Mandarin, our algorithm is made up of these steps:
1) Extract estimated power from frequency domain to detect the landmark of the speech.
2) Greedily, make first round selection of boundary candidates based on the energy and the landmark.
3) Cautiously, make second round selection of boundary candidates based on zero-crossing rate.
4) Drop the redundant boundaries and merge the previous two round candidates.

A. Landmark Selection through Frequency Domain

In this first step, we are going to select frames as landmarks which power concentrates on to represent the center of syllables. In order to get more accurate landmarks, we estimate the power on frequency domain after getting the spectrogram by Fast Fourier Transform.

$$\hat{P}_i(f) = \left| \sum_{i=0}^{W-1} x(n) e^{-j\omega n} \right|^2 \tag{1}$$

where $\hat{P}_i(f)$ denotes to the power of i^{th} windowed frame (length is W) on different frequency region f and $x(n)$ is the original speech signal.

Combining the proposed representation of power $\hat{P}(f)$ with normalization form $\hat{P}_{norm}(f)$ and the Nicolas Obin's idea[10], the landmarks can be detected over time-frequency domain as:

$$landmark^{(k)} = localmax(\hat{P}_{norm}(f_k)) \tag{2}$$

where k denotes to the k^{th} frequency region and each region will have its own landmark vector $landmark^{(k)}$

Then, through exploiting the above landmarks over time-frequency domain, the landmarks over time domain can be generated: The more frequent is observed a time position of a candidate over frequencies, the more likely is the presence of a landmark. Naturally, It works well to calculate the frequent by integrating landmark vectors into a single probability density function P(L). Meanwhile, P(L) need to be smoothed by using a moving average window (usually 3 frame, around 20ms) for several times to get higher level vision so that it will not be too sensitive.

Finally, the selection of landmark candidates over time domain can be determined as $Landmark$ (Figure 1) by detecting the local maximum of $P(L)$.

However, some syllables may have two landmarks: one for syllabic consonant and the other for the vowel because both two sometimes have their own power centers. Since it is supposed to be only one landmark (center) for a single syllable (word), this step can be regarded as a greedy selection and will be corrected later in this paper.

B. First Round Boundary Candidates

The Landmark generated above make it easier and more accurate to find the boundary between two syllables. But instead of using the estimated power $\hat{P}(f)$ again, root means square energy (RMSE) is chosen to detect the boundary because of its better performance on representing the pause part of speech.

$$E_n = [\frac{1}{W} \sum_{i=0}^{W} x(i)_n^2]^{\frac{1}{2}} \tag{3}$$

Where W is the length of window and $x(i)_n$ denotes the i^{th} windowed speech sample in frame number n

Then, the first round boundary candidates will be generated by detecting the minimum frame of the energy E_n between landmarks:

$$boundary_i^* = \arg\min_n E_n \tag{4}$$

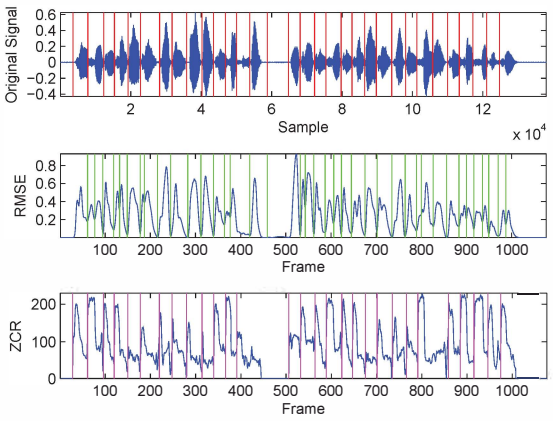

Figure 3. *Greedily selection result* $boundary^*$ *generated by RMSE (middle), strict selection result* $boundary^{**}$ *generated by ZCR (bottom), and merge version of selection result* $Boundary$ *(top) for the* $00087.wav$ *utterance in CASIA-863 speech database.*

Where i denotes the i^{th} boundary and n is the frames between two landmarks: $Landmark_i \leq n \leq Landmark_{i+1}$.

The result of first round selection is shown on the Figure 2. Up to now, what we have done is segmenting the speech greedily into "syllables" while the "syllable" here is not always the true syllable but sometimes a consonant and a vowel.

C. Second Round Boundary Candidates

Another feature is always used to segment speech especially for Chinese Mandarin: Zero-Crossing Rate (ZCR).

$$ZCR_n = \sum_{i=0}^{W} |sgn[x_n(i)] - sgn[x_n(i+1)]| \quad (5)$$

Where $sgn[(x(i)_n]$ denotes the sign of i^{th} windowed speech sample in frame number n

According to Zhang[11], the vowel in Chinese usually has low zero-crossing rate while the initial consonant usually has high zero-crossing rate. Hence, ZCR will increase emergently at the beginning frame of a new syllable and decrease fast at the frame between consonant and vowel. In order to capture this kind of irregular frames, ZCR slope (ZCRS) is introduced to show the changes of ZCR by calculating the differences between neighbor frames. Then, the second round boundary candidates $boundary^{**}$ can be generated by detecting the local maximum of $ZCRS$ with a minimum threshold: $minzcrs$ (usually 20). Only when $ZCRS$ is larger than $minzcrs$ which means ZCR going up sharply, will the frame be evaluated as irregular enough to be the boundary of syllables. In fact, this second round selection is quite cautious and strict because there will not be a segment until the frame is irregular enough (Figure 3).

D. Merge Two Rounds Candidates

Thus far, there are two round boundary candidates generated above. However, $boundary^*$ and $boundary^{**}$ are not the same because the former round selection is greedy while the latter is quite strict. In Figure 3, it is shown that $boundary^*$ has many boundary candidates including several wrong boundaries that segment syllable into consonant part and vowel part, by contrast, $boundary^{**}$ has a few strict candidates that are definitely the correct segment boundaries. So, merging them and make a final round selection will be a most important work.

Before giving the merging step, it is necessary to illustrate two parameters: $slength$ and $minzcrs$. The first one denotes to the length of a full syllable. The second parameter $minzcrs$ is used again to detect the irregular frame. But in this step, it is used for searching the redundant boundary between consonant and vowel where ZCR decreases rapidly. In this case, the function $MisSeg$ will return true.

In short, the general idea of this step can be concluded as a $boundary^{**} - centric$ strategy: drop the boundaries in $boundary^*$ that is very near to the boundaries in $boundary^{**}$ or the boundaries that seems to redundantly segment the consonant and vowel in one syllable. Specifically, the pseudo code is demonstrated in Algorithm 1.

Algorithm 1 Merge Algorithm

```
for u ∈ boundary** and v ∈ boundary* do
    if (−slength/3 < v − u < slength/2) then
        boundary* ← boundary* − v
    end if
    if (0 < v − u < slength) and MisSeg(v) then
        boundary* ← boundary* − v
    end if
end for
Boundary ← boundary* + boundary**
```

In Figure 3, we can find the merging step fine-tune the previous two round results so that $Boundary$ finally performs well on the syllable segmentation task.

IV. Experiment and Evaluation

A. Experiment Method

In the experiments, we used CASIA-863 speech database as our test data. The speech signals ranges from 3-15s, sampled at 16000Hz and every speech has its reference file containing the specific word and its pinyin in Chinese. However, the database does not provide the syllable(word) alignment, the reference alignment is manually produced. The test data for evaluation contains around 3700 syllables and the average duration of syllable is around 250ms.

The method we used for evaluation is F-measure method:

$$F = \frac{2 * Pricse * Recall}{Pricse + Recall} \quad (6)$$

Where $Pricse$ is the number of correct detected results divided by the number of all boundaries given by certain method, and $Recall$ is the number of correct detected results divided by the number of boundaries in reference.

B. Result Discussion

We compared our method with the same kind of segmentation methods respectively proposed in [2][3][5][6] using F-value . In particular, we applied simple pause detection method to capture the silent intervals so as to meet the requirement of Convex Hull algorithm. Besides, we add a simple energy-based step for detecting the syllable boundaries to Xie's system after its getting the landmarks. The performance of these methods cames as below when the tolerance is 50ms.

Method	Precise(%)	Recall(%)	F-value(%)
Convex Hull	66.3	71.8	68.8
Xie	79.8	78.8	79.2
Group Delay	85.9	78.2	81.4
Hybrid Zhao	88.1	84.6	86.2
Power	74.3	84.3	78.9
Power+RMSE	77.6	88.1	82.3
ZCR	**97.2**	67.9	79.7
Proposed Method	96.7	**91.5**	**94.0**

Table I
EVALUATION RESULT.

As listed in the table, Mermelstein's convex hull algorithm performs as a baseline since it is a simplest and most popular method. Xie improved segmentation performance using two round of this algorithm on periodicity and energy. And benefiting from the ZCR rules, Zhao's hybrid method outperforms other three algorithms including the group delay function. We also test the results of two round distinct selection respectively, the first round's performance get improved by adding the RMSE counter, and the boundaries generated from ZCR always has high accuracy which helps a lot on fine-tuning the result. Our method merge these two rounds of selection and we get the better performance.

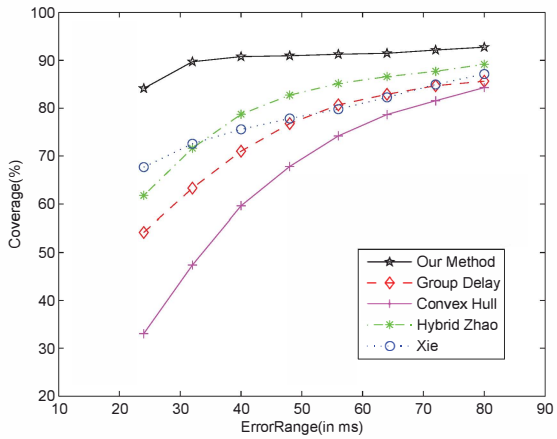

Figure 4. *the coverage of the mental reference syllable boundaries by different methods with different error range .*

Figure 4 shows the coverage of reference boundaries observed on different error range from different methods. The black line with pentagram shows that our method cover almost 83% of the true syllable boundaries even in the low error range.

V. CONCLUSION AND FUTURE WORK

In our system, combined with both time and frequency perspective, greedy round and cautious round of segmentation are merged together to get accurate syllable boundaries. This algorithm is tested on CASIA-863 speech database and the result outperforms other conventional methods both in precise rate and recall rate. Even comparing with model based methods, our system is also competitive while considering the cost of time and other resources. Clearly, the presented segmentation strategy is meaningful and valuable.

REFERENCES

[1] N. Jittiwarangkul, S. Jitapunkul, and S. Luksaneeyanavin, et al. "Thai syllable segmentation for connected speech based on energy," *Circuits and Systems, 1998. IEEE APCCAS 1998. The 1998 IEEE Asia-Pacific Conference on*, 169 - 172.

[2] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *The Journal of the Acoustical Society of America*, 58, 4, 880-883.

[3] V. K. Prasad, T. Nagaraja, H A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions[J]," *Speech Communication*, 2004, 42:429C446.

[4] H. A. Murthy, T. Nagaraja, "Subband-Based Group Delay Segmentation of Spontaneous Speech into Syllable-Like Units[J]," *EURASIP JOURNAL ON APPLIED SIGNAL PROCESSING*, 2004,2004(17):2614-2625.

[5] Z. Xie, P. Niyogi, "Robust Acoustic-Based Syllable Detection," *International Conference on Spoken Language Processing, Pittsburgh, USA*, 2006, pp. 1571C1574.

[6] X. Zhao, D. O'Shaughnessy, " A new hybrid approach for automatic speech signal segmentation using silence signal detection, energy convex hull, and spectral variation[J]," *Electrical and Computer Engineering, 2008. CCECE 2008. Canadian Conference on*, 2008:000145 - 000148.

[7] S. Nakagawa, Y. Hashimoto, "A method for continuous speech segmentation using hmm," *in Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1988, pp. 960C962.

[8] F. Brugnara, D. Falavigna, M. Omologo, "Automatic segmentation and labeling of speech based on hidden Markov models," *Speech Communication,* vol. 12, 1993.

[9] L. Shastri, S. Chang, S. Greenberg, "Syllable Detection And Segmentation Using Temporal Flow Neural Networks[J]," *Proceedings of the Fourteenth International Congress of Phonetic Sciences*, 1999:1721–1724.

[10] N. Obin, F. Lamare, and A. Roebel, "Syll-O-Matic: An adaptive time-frequency representation for the automatic segmentation of speech into syllables," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE*, 2013, 6699-6703.

[11] J Y. Zhang, F. Zheng, and S. Du, et al. "The Merging-Based Syllable Detection Automaton in Continuous Chinese Speech Recognition[J]," *JOURNAL OF SOFTWARE*, 1999.