Syllable Boundaries based Speech Segmentation in Demi-Syllable Level for Mandarin with HTK

Tao Jianhua[1] H.Udo Hain[2]

[1] Department of Computer Science and Technology,
Tsinghua University, Beijing, China, 100084
[2] CT IC 5, Siemens AG, D-81730, Munich, Germany

e-mail : [1] jhtao@ tsinghua.edu.cn [2] Horst-Udo.Hain@mchp.siemens.de

1. Abstract

The paper offers an optimized method in speech segmentation of Mandarin speech database by using Hidden Markov Model (HMM). The method takes the syllable boundaries into account. Testing shows that the accuracy of results is improved to 95% from 88% compared to the normal method. Especially, most of the boundaries between two vowels can also be well detected with the new method. The paper also analyzes the influence of the amount of HMM states and the amount of the training corpus.

1 Introduction

Segmentation and labeling of speech material according to phonetic or similar linguistic rules is a fundamental task in speech processing. Currently, most of the Mandarin Text to Speech systems are based on syllable level. Part of reason is that the articulation of the synthesis result can not be ensured in sub-syllable level. But now, more and more effort has been done to use the demi-syllables as the basic acoustic unit for Mandarin TTS and managed to achieve great success, such as Siemens TTS system PAPAGENO. Traditionally, the segmentation and labeling of speech data in demi-syllable level was done manually by a well trained phonetician using the listening and visual cues. It's rather tedious, time consuming, boring and error prone. The automatic speech segmentation is rather important, especially while the TTS system is used in PDA where it is impossible to use large database. The automatic speech segmentation algorithms found in the literature can be divided into two broad categories: Hierarchical and Non-hierarchical. The Hierarchical speech segmentation procedures involve a multi-level, fine-to-coarse, segmentation description; sometimes displayed in a tree-like fashion called dendogram. The best segmentation is the multi-level segmentation search space [1]. The Non-hierarchical speech segmentation algorithms attempt to score metric of acoustic models [5][6]. As we know, Mandarin is a tonal language which includes four lexical tones and one neutral tone. The sole-syllable has a relatively constant pitch contour. And the typical phonetic structure of the syllable is rather simple, such as CV or V. It is easy to acquire the syllable boundaries in much high accuracy with the help of pitch contours, energy contours, zero-across rate contours. The paper offers the method to segment the Mandarin speech database in demi-syllable level based on multi-level best path searching, facilitated by syllable boundaries, and acquires a very high quality of the segmentation. In section 2, the paper describes how to detect the demi-syllable boundaries facilitated by syllable boundaries. In section 3, all of demi-syllables (initials and finals) are defined and modeled into HMM with appropriate state number. The short pause and silence model are also generated to better fit the long sentence. Triphone models are made and tree-based clustering method was used to make the set of triphone models smaller and more flexible to database. And then, all of the results are analyzed in section 4. Results show that the segmentation results are improved by 7% from normal method. The influence of the HMM state number and the amount of training corpus are also discussed in

this section. In section 5, the paper makes a further analysis of the influence of the distribution of demi-syllables for the accuracy of segmentation.

2 HOW TO USE SYLLABLE BOUNDARIES FOR SPEECH SEGMENTATION

Let's suppose vector $\vec{T}$ as the transcription of the sentence in the corpus.

$\vec{T} = (t_1, t_2, \cdots, t_n \cdots, t_N)$

t_n denotes the n'th phoneme or demi-syllable in the transcription. N means the amount of all of the units.

The speech is also represented into vector $\vec{S}$

$\vec{S} = (s_1, s_2, \cdots, s_m \cdots, s_M)$

s_m is the m'th frame of the speech signal.

The speech segmentation is to dispatch all frames of speech into N elements in appropriate time sequence. Thus, we can get K possibilities. A typical align result is shown as following,

$(\vec{G}_{k1}, \vec{G}_{k2}, \cdots, \vec{G}_{kn}, \cdots, \vec{G}_{kN})$, $k = (1 \sim K)$

Here, $\vec{G}_{kn}$ contains several frames of speech which fits n'th element, $\vec{G}_{kn} = (s_p, s_{p+1}, \cdots, s_{p+q})$, p and q are positive integer and satisfy the condition $p + q < M$.

Then, the best speech alignment result can be got from $\arg\max_k[P((\vec{G}_{k1}, \vec{G}_{k2}, \cdots, \vec{G}_{kN}) \mid \vec{T})]$

$$= \arg\max_k[\frac{P(t_1 \mid \vec{G}_{k1})P(t_2 \mid \vec{G}_{k2}) \cdots P(t_N \mid \vec{G}_{kN})}{P(\vec{T})}] \quad (1)$$

Where, the demi-syllables appeared in database are supposed to be equably distributed. Then, $P(\vec{T})$ can thought as a constant. Thus, the best alignment can be got from

$$\arg\max_k[P(t_1 \mid \vec{G}_{k1})P(t_2 \mid \vec{G}_{k2}) \cdots P(t_N \mid \vec{G}_{kN})] \quad (2)$$

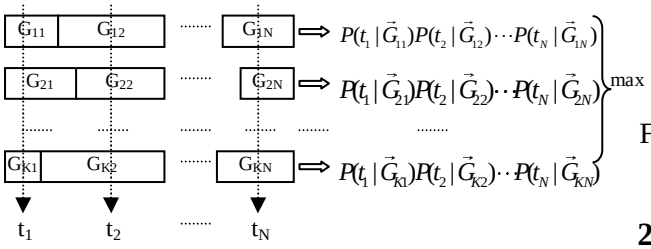

Figure 1, The frame of the speech segmentation

2.1 Speech Segmentation facilitated by syllable boundaries

If the syllable boundaries have been labeled in the corpus, the transcription of the sentence will be,

$\vec{T}' = \{(t_1, t_2, t_3), (t_4, t_5), \cdots, (t_{N-1}, t_N)\}$

And the corresponding speech sequence is,

$\vec{G}' = \{(\vec{G}_{k1}, \vec{G}_{k2}, G_{k3}), (\vec{G}_{k4}, \vec{G}_{k5}), \cdots, (\vec{G}_{N-1}, \vec{G}_N)\}$

The formula (2) will be modified to,

$$\arg\max_k[P(\vec{G}' \mid \vec{T}')]$$

$$= \arg\max_k[P(\vec{G}_{k1} \mid t_1)P(\vec{G}_{k2} \mid t_2)P(\vec{G}_{k3} \mid t_3)]$$

$$\times \arg\max[P(\vec{G}_{k4} \mid t_4)P(\vec{G}_{k5} \mid t_5)]$$

$$\times \cdots \times \arg\max_k[P(\vec{G}_{k(N-1)} \mid t_{(N-1)})P(\vec{G}_{kN} \mid t_N)]$$

$$= \arg\max_k[P(t_1 \mid \vec{G}_{k1})P(t_2 \mid \vec{G}_{k2})P(t_3 \mid \vec{G}_{k3})]$$

$$\times \arg\max_k[P(t_4 \mid \vec{G}_{k4})P(t_5 \mid \vec{G}_{k5})]$$

$$\times \cdots \times \arg\max_k[P(t_{(N-1)} \mid \vec{G}_{k(N-1)})P(t_N \mid \vec{G}_{kN})] \quad (3)$$

Then, the searching space is limited in one syllable. The question of speech segmentation will be replaced by the question of find the best solution of speech sequence within syllables. Not only the accuracy of the segmentation will be improved, but the computing cost will be decreased during the model training.

Figure1 shows method of the speech segmentation. The align searching procedure of optimized method is illustrated in figure 2. The method is extremely helpful in detecting the boundaries between vowel-vowel regions and plosive-vowel regions.

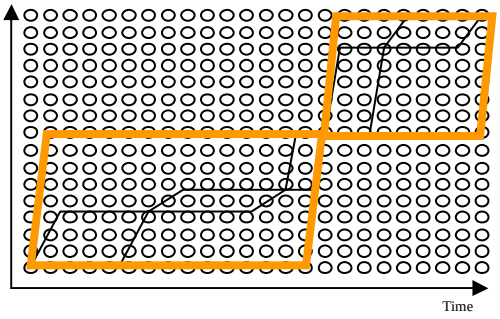

Figure 2, The best path searching within syllable boundaries

2.2 The influence of the distribution of demi-syllables

To get formula (2), a very important supposing is used. That is all elements are supposed to be

well-distributed. Actually, it's very difficult to get that. That means $P(t_n)$ is not real constant. From the experience of natural language processing, $P(t_n)$ can be extended into

$$P(\vec{T}) = P(t_n \mid t_{n-1}t_{n-2}\cdots t_1)P(t_{n-1} \mid t_{n-2}t_{n-3}\cdots t_1)\cdots P(t_1)$$

Thus, the final results of speech segmentation are influenced not only by the acoustic information of the elements, but also by the linguistic information.

3 MODELING AND SEGMENTING

3.1 Defining the models of demi-syllables for Mandarin

The typical HMM of English phonemes contains three states. But, as we know, Mandarin is a tonal language, and the classical minimum speech elements of Mandarin are demi-syllables, which are initials and finals. The initials contain plosives, fricatives and nasals, and the finals contain single-vowels and compound-vowels. The spectrum of compound-vowels are more complicated than the phonemes. While the HMM of initials and single-vowels use three states, the compound-vowels use five states, which is listed in table 1. In section 4, the paper will analyze why the number of the states listed in the table are most appropriated for the model.

Type	*List of Demi-syllables*	*Number of state*
Initials	"b","c","d","f","g","h","j","k","l","m","n","p","q","r","s", "t", "x","z","sh","ch","zh"	3
Single-vowels	"a","o","e","-i","i","u", "v"	3
Compound-vowels	"ai","ao","an","ang","eng", "ia","iao","ian","iang", "ing","iong","iou","ua","uo", "uai","uei","ui","uan","uen", "uang","ueng","un","ong","ve", "van","vn","ie","iu","in","ou", "ei","en","er",	5

Table 1, the number of HMM states for demi-syllables

3.2 Silence, short pause and triphone models

- Silence model and triphone model

For continuous speech, the HMM of element is not enough to describe the whole sentence. There always exist some silence or short pause in the fluent speech, especially in the head and tail of sentence and the middle of the phrases. To avoid this case, the silence and short pause model must be generated. Meantime, to get the more detailed acoustic model, triphone model is strongly recommended.

- The clustering of triphone models

In Mandarin, there are 21 initials and 40 finals. If we take silence model and short pause model into account, the total combination of them will be 242109. It is impossible to generate all of the triphone models with limited speech database. The missed models should be constructed through the existing model. The decision tree based clustering method is an efficient way to solve this question. The training of decision tree is an up to down procedure. It will stop while the similarity pass the criterion. After the tree is established, a down to up procedure is started to melt the leaves into one node while the similarity is lower to another defined criterion.

3.3 The procedure of segmentation

While doing the speech segmentation, the first step is generating the demi-syllable transcription from the pinyin sequence in the corpus. Then, the triphone models are generated through extending the demi-syllable transcription, and also trained with the corpus with the help of syllable boundaries. As mentioned above, there exist silence and short pause in the speech. In the phase of generating triphone sequence, silence model and short pause model are inserted to form the large and complicated testing chains. The final step is calculating the HMM score of the triphones with speech parameters, and searching the best path along the speech alignment within syllables. The best path of the speech alignment represents a best result of speech segmentation.

4 RESULTS AND ANALYSIS

4.1 The Results of Speech Segmentation

HMMs are trained from the speaker-dependent data of the target speaker. With the normal method, we observed that there is 11% of the elements in the database contain some gross segmentation errors (larger than 20 ms) compared to manual labeled corpus.

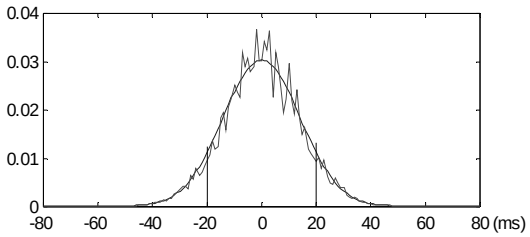

Figure 3, The speech segmentation results based on formula (3), 88% of the element boundaries are in 20ms error compared to hand-labeled corpus

The major bias was existed between plosives and vowels. The spectrum transform of plosives and fricatives is typically much smaller when compared to that of vowels. Figure 4 shows the accuracy results of the speech segmentation with the help of the syllable boundaries.

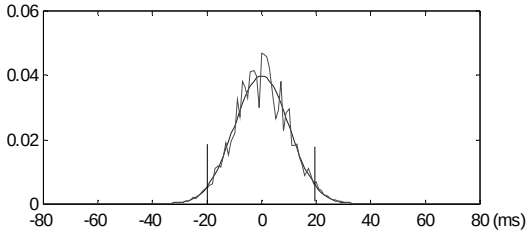

Figure 4, The accuracy of speech segmentation based on formula (4), 95% of the element boundaries are in 20ms error compared to hand-labeled corpus

Faciliated by syllable boundaries, the accuracy of the speech segmentation within 20 ms bias was improved from 88% to 95%. Furthermore, testing result shows that the accuracy of the boundaries between plosive and vowel or vowel and vowel was also improved from 81% to 92%.

4.2 Speech segmentation with various numbers of HMM states

To make sure if the number of HMM states is essentially important for speech segmentation, the various numbers of HMM states are tested, which is shown in figure 5. The x-axes means the vector of HMM state number in following order, (initial, single-vowel, compound-vowel).

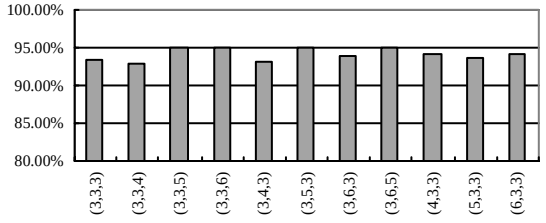

Figure 5, Accuracy of speech segmentation according to different amount of HMM state

It is very interesting that we cannot find that the accuracy of speech segmentation was sensitive to the amount of HMM state too much, though, it was proved that increased number of HMM state can improve the final results of speech segmentation a little, especially when the state number of compound-vowel was increased. Nevertheless, good context coverage and consistent segmentation by HMMs can partly overcomes the drawback of an imperfect automatic segmentation when compared to manual segmentation. To acquire the balance of computing cost and good results, we finally use (3,3,5) as the state number of initials, single-vowels and compound-vowels.

4.3 Which is the essential amount of the corpus for training

Normally the size of corpus used for speech synthesis or speech recognition is extremely large. Is it necessary that all of the database should be labeled with syllable boundary first? It is important for us to know how many corpus with syllable boundaries are essential for initial HMM training. The figure 6 shows the trend of the accuracy of element segmentation of the whole speech corpus related to variational amount of training corpus with the labeling of syllable boundaries.

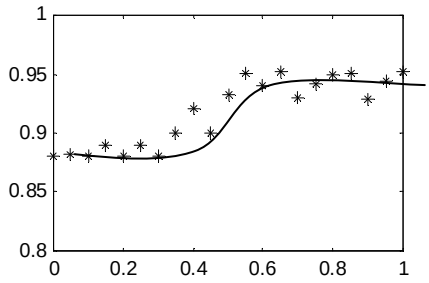

Figure 6, The accuracy of speech segmentation according to variational amount of training corpus with syllable boundaries (x-axes means the percent of the corpus which is labeled with syllable boundaries, y-axes denotes the accuracy of speech segmentation of the whole corpus)

The result shows that the accuracy increased rapidly from 30% to 60%. It means that it is not necessary to label the corpus with syllable boundaries beyond 60% of the corpus.

5 CONCLUSIONS

The paper describes a new method to segment the speech in demi-syllable level for Mandarin with HTK, which is based on syllable boundaries. The testing result shows that the new method improves the segmentation results a lot. It also shows good quality when the final automatic

segmentation results are used for TTS system directly. The method was also proved to be very useful while we tried to do the speech segmentation in phoneme level for German. It also improves the acoustic model of the German TTS system Papageno. Future work will be focused on how to integrated the work described in the paper into the speaker self-adaptive system and the speech conversion system.

References

[1] James R. Glass and Victor W. Zue., "Multi-level acoustic segmentation of continuous speech", ICASSP, p429-432, 1988

[2] Ronald Cole and Lily Hou, "Segmentation and broad classification of continuous speech", ICASSP, p453-456, 1988

[3] Kaichiro Hatazaki, Yasuhiro Komori, Takeshi Kawabata, and Kiyohiro Shikano, "Element segmentation using spectogram reading knowledge", ICASSP, p393-396, 1989

[4] T. Svendsen and F. Soong. "On the automatic segmentation of speech signals", ICASSP, p341-344, 1987

[5] Ljolje, A., Riley, M.D. "Automatic segmentation and labeling of speech", Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 Page(s): 473 -476 vol.1, 1991

[6] Pawate, B.I., Dowling, E. "A HMM-based approach for segmenting continuous speech", Signals, Systems and Computers, 1992. 1992 Conference Record of The Twenty-Sixth Asilomar Conference on, Page(s): 1105 -1110 vol.2, 1992

[7] van Hemert, J.P. "Automatic segmentation of speech", Signal Processing, IEEE Transactions on, Volume: 39 4, Page(s): 1008 –1012, April 1991