

RESEARCH ON SEGMENTATION AND LABELING OF SPEECH CORPORA

HE Xiaodong, LIU Jian, YU Tiecheng

Speech Processing Laboratory, Institute of Acoustics, Chinese Academy of Sciences

ABSTRACT

In this paper, we suggested a Reference Sentence Alignment (RSA) method to segment and label the speech automatically based on the multiple pronunciation phoneme segmental k-means algorithm and HMM. Furthermore, based on the search path created by this method, information of pitch and energy of speech can be obtained and labeled synchronously. This segmentation and labeling strategy was applied in our "863 National Project Chinese Mandarin Speech Corpora". The accuracy more than 95% can be obtained.

1. INTRODUCTION

Segmental labeling of speech corpora is very important for speech recognition, speech understanding, and speech synthesis. It provides the essential data for speech processing. Labeling of speech corpora has a long history, a standard for labeling English prosody, ToBI, was set up firstly by Silverman *et al.*[1], and refined by Beckman *et al.*[2]. However, as to Chinese mandarin, a complete and well-accepted standard has not been built [3]. On the other hand, ToBI labeling system is very complicated and not suitable for labeling a speech corpora constructed for a specifically purpose.

Traditionally, a speech corpus is first obtained by recording, then segmented and labeled manually [3]. Several softwares can be used to help labeling speech [4]. However, labeling by human leads to some disadvantages. First, segmentation and labeling of speech corpora is a heavy workload for human and these workers were expected to have certain knowledge of speech and labeling. Second, there exist multiple pronunciations of one phoneme, which need to be distinguished, but because of the subjectivity of human, the accuracy result of manual segmentation and labeling cannot be

assured. Third, some speech information can not be obtained and labeled directly by human, such as the pitch, energy, etc.

Here we suggest an automatic speech labeling method based on the segmental k-means algorithm and HMM (Hidden Markov Model). The basic theory of HMM was first presented by Baum and his colleagues [5] in the late 1960s and 1970s and was applied for speech-processing application by Baker [6] at CMU, and by Jelinek and his colleagues at IBM [7] in the 1970s. Moreover, *k*-means algorithm is a well-known iterative procedure for clustering data. Now we applied it to speech segmentation and labeling.

In section II we describe the fundamentals of HMM and segmental *k*-means algorithm first. Then to deal with the multiple pronunciation phoneme performance, we suggest an improved algorithm Reference Sentence Alignment (RSA). Section III shows the application of this improved algorithm in Chinese mandarin segmentation and labeling. Section IV will present the practical experiment result, and a conclusion is given at section V.

2. REFERENCE SENTENCE ALIGNMENT

2.1. Segmental K-Means Segmentation

Segmental *k*-means algorithm is normally used to train HMM model in speech recognition. Assuming that we have a speech data set and the corresponding reference word string of each speech sentence. The HMM training problem can be solved by using segmental *k*-means training procedure as below:

1. Initialization:

Linearly segment each utterance into HMM states assuming no silence between two words, a single lexical pronunciation of each word, and a single model for each phoneme.

2. Estimation

All feature vectors aligned to the same state were used to estimate the model's emission probability function. In this paper we use discrete symbol densities, the feature vectors are coded by a codebook with M-codes, and the updated estimate of $b_{\lambda_j}(k)$ parameter is:

$$b_{\lambda_j}(k) = \text{number of vectors with codebook index } k \text{ in state } j \text{ of model } \lambda \text{ divided by the number of vectors in state } j \text{ of model } \lambda.$$

Updated estimate of $a_{\lambda_{ij}}$ parameter is:

$$a_{\lambda_{ij}} = \text{number of transitions from state } i \text{ to } j \text{ of model } \lambda \text{ divided by the number of transitions from state } i \text{ to any state of model } \lambda.$$

3. Segmentation

Based on this updated set of unit models, a Viterbi decoding is implemented to segment each training utterance into units and states according to their reference sentence.

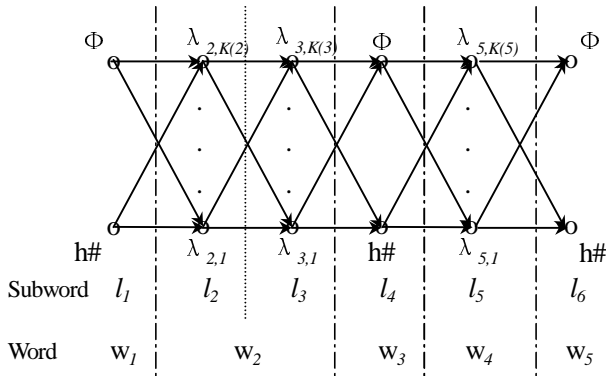
4. Iteration

Steps 2-3 are iterated until convergence.

After training, segmentation of speech is automatically obtained by Viterbi decoding in this iterative procedure.

2.2. Reference Sentence Alignment

Figure I: Structure of RSA



In Chinese mandarin, there exist many multi- pronunciation phonemes. To distinguish them, more than one model should be built associate with the same reference phoneme. Moreover, a new search method is requested. What we suggested is a variant of Viterbi decoding, Reference Sentence Alignment (RSA). Its structure is shown in Figure I. A sentence is divided

into a sequence of subwords, at each subword position are several possible pronounce models. There maybe exist silence between two words. To align with the silence, a virtual model Φ is introduced, which create no output.

Assuming that the strict left-right discrete HMM was adopted and the observation sequence is the coded feature vectors, RSA can be implement as follows:

Using the standard notation of t to represent the test frame index, $1 \leq t \leq T$, v to represent the unit model (λ_v) index, $1 \leq v \leq V$, n to represent the HMM state index of unit model λ_v , $1 \leq n \leq N_v$, and l to represent the subword index of the sentence, $1 \leq l \leq L$. k to represent the unit index in a subword position, $1 \leq k \leq K_l$, and the model index of that unit can be retrieved as $u_{l,k}$. We define the accumulated probability $P(t,l,k,n)$ as the best score (highest probability) along a single path, at time t , which accounts for the first t observations and ends in subword l , unit index k , state n , and define the local probability $p(t,l,k,n) = b_{u_{l,k},n}(o_t)$, where o_t represent the code index of observation frame t . In order to retrieve the search result, we need to keep track of the argument that maximized the accumulated probability P . We do this via the array $\psi(t,l,k,n)$. The complete procedure for finding the best path can be presented here:

1. Initialization

$$P(1,1,k,1) = p(1,1,k,1), 1 \leq k \leq K_1$$

$$\psi(1,1,k,1) = (0,0,0,0).$$

2. Recursion

For all internal state of HMM ($n > 1$):

$$P(t,l,k,n) = \max_{n-1 \leq n' \leq n} [P(t-1,l,k,n') a_{u_{l,k},n',n}] p(t,l,k,n).$$

$$\psi(t,l,k,n) = (t-1, l, k, n^*),$$

$$n^* = \arg \max_{n-1 \leq n' \leq n} [P(t-1,l,k,n') a_{u_{l,k},n',n}] p(t,l,k,n).$$

At the model boundary ($n=1$), there are two possibilities,

If $u_{l,k}$ is the virtual model Φ :

$$P(t,l,k,n) = \max_{1 \leq k' \leq K_{l-1}} [P(t,l-1,k',N_{u_{l-1,k'}})].$$

$$\psi(t,l,k,n) = (t, l-1, k^*, N_{u_{l-1,k'}}),$$

$$k^* = \arg \max_{1 \leq k' \leq K_{l-1}} [P(t,l-1,k',N_{u_{l-1,k'}})].$$

Otherwise:

$$P(t, l, k, n) = \max (P_1, P_2) p(t, l, k, 1);$$

$$P_1 = \max_{1 \leq k' \leq K_{l-1}} [P(t-1, l-1, k', N_{u_{l-1}, k'})],$$

$$P_2 = P(t-1, l, k, 1) a_{u_{l,k}, 1, 1}.$$

if ($P_1 > P_2$)

$$\psi(t, l, k, n) = (t-1, l-1, k^*, N_{u_{l-1}, k^*}),$$

$$k^* = \arg \max_{1 \leq k' \leq K_{l-1}} [P(t-1, l-1, k', N_{u_{l-1}, k'})].$$

else

$$\psi(t, l, k, n) = (t-1, l, k, 1).$$

3. Termination

$$P^* = \max_{1 \leq k' \leq K_L} [P(T, L, k', N_{u_{L}, k'})]$$

$$\psi_T^* = (T, L, k^*, N_{u_{L}, k^*}),$$

$$k^* = \arg \max_{1 \leq k' \leq K_L} [P(T, L, k', N_{u_{L}, k'})].$$

4. Path backtracking

$$\psi_t^* = \psi(\psi_{t+1}^*), t = T-1, T-2, \dots, 1.$$

By RSA we can obtain the best segmental path in the case of existing of multi-pronunciation phoneme.

3. APPLICATION

In Chinese mandarin, according to the phonetic rule, we first choose 60 basic speech units:

(1) Totally 38 vowels:

a) 9 single vowels:

a, o, e, u, v, i, i1, i2, er

b) 13 composite vowels:

ai, ei, ao, ou,

ia, ie, iao, iu, (i family composite vowels)

ua, uo, uai, ui, (u family composite vowels)

ve. (v family composite vowels)

("iu" represent "iou", "ui" represent "uei".)

c) 16 composite nasal vowels:

an, en, ang, eng, ong,

ian, in, iang, ing, iong, (i family composite nasal vowels)

uan, un, uang, ueng, (u family composite nasal vowels)

van, vn. (v family composite nasal vowels)

("un" represent "uen", "ueng" represent "weng".)

(2) Totally 21 consonants:

a) 17 simple consonants:

b, d, g, p, t, k, z, c, s, zh, ch, sh, j, q, x, f, h.

b) 4 voiced consonants:

m, n, l, r.

(3) Totally 1 silence:

h#.

The analysis of the speech data shows that voiced simple consonants appear in the practical speech, so at least two models need to be built for each simple consonant, 17 additional speech units for simple consonants are listed below:

bv, dv, gv, pv, tv, kv, zv, cv, sv, zhv, chv, shv, jv, qv, xv, fv, hv.

To segment and label speech, a HMM is built for each unit, then we implement the automatic segment and labeling of speech corpora as follows:

First we train the 60 basic units' HMM by classical segmental k-means training, then based on the spectral analysis, each of those simple consonant utterance can be divided into two classes. Therefore, we can obtain two initial HMMs of each of that simple consonant by isolated unit HMM training. Replace the initialization step of the classical segmental k-means training by these initial HMMs and replace the Viterbi decoding in segmentation step by RSA, we do the improved segmental k-means procedure once more and the final segmental path is obtained ultimately.

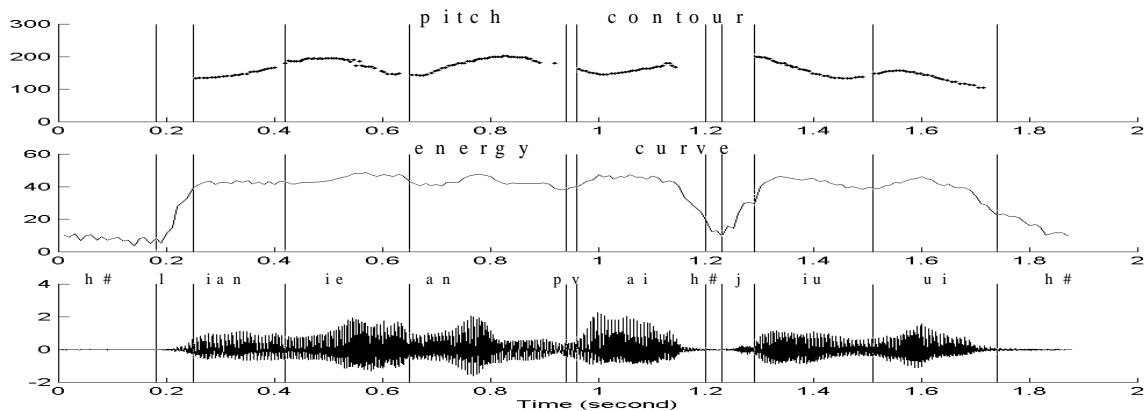
In the meantime, some other information of speech such as energy and pitch can be also obtained and labeled. Upon the segmentation path, the pitch extraction process can be applied only on those voiced units to get their pitch contours. Here, for example, pitch extraction is based on the de-emphasized LPC residual signal [8].

After automatic segmentation and labeling, a manual checking and correction is needed to obtain the final speech segmentation and labeling result.

4. EXPERIMENTS

The automatic speech segmentation and labeling strategy described above was applied on the "863 National Project Chinese Mandarin Speech Corpora", which includes 52000 sentences from 50 male and 50 female speakers. LPC-CEP

Figure II: The automatic segmentation and labeling result



feature and discrete HMM were adapted in the implementation of RSA. We also calculated and labeled the energy and pitch synchronously. Experimental results indicated that a high segmentation and labeling accuracy of more than 95% was obtained. One labeling sample is shown in figure II. A detailed experiment result is expressed in table I.

Table I: segmentation and labeling experiments result

Error Range	D<10ms (1 frame)	D<20ms (2 frames)	D<30ms (3 frames)
Correct Rate	95.2%	97.8%	99.4%

D means the acceptable error range between real segmental point and the automatic segmental point.

5. CONCLUSION

This paper provides a new speech search and alignment method attempt on automatic segmentation and labeling of speech corpora. The experiments indicated that a high segmentation and labeling accuracy more than 95% was achieved. Although a manual correction of that result is still needed, it can reduce the workload remarkably. Furthermore, unlike segmentation and labeling by human, it avoids the error induced by the subjectivity of human and gives a statistical correct result. In the meantime, it can label some useful speech information such as pitch and energy synchronously.

REFERENCE

[1] K. Silverman, at el. "ToBI: a standard for labeling

English prosody", *ICSLP92 Proceedings*, Vol. 2, pp.867-870, Canada, Oct. 1992.

- [2] M.E. Beckman, at el. "The ToBI annotation conventions", ToBI release 2.0, New Mexico Institute of Mining and Technology.
- [3] F.Y. Mo, at el. "speech database collection and labeling of Chinese text-speech translation system", *NCMMSC98 Proceedings*, Vol. 1, pp.369-372, China, Jul. 1998.
- [4] *Speech Filing System User Manual*. London University. London, UK.
- [5] L.E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Stat.*,37: pp.1554-1563, 1966.
- [6] J.K. Baker, "The dragon system-An overview," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-23(1):pp.24-29, Feb. 1975.
- [7] F. Jelinek, "A fast sequential decoding algorithm using a stack," *IBM J. Res. Develop.*, 13: pp.675-685, 1969.
- [8] J.D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans.*, Vol.AU-20, no.5, pp.367-377, Dec. 1972