Towards A Phoneme Labeled Mandarin Chinese Speech Corpus

Hsin-Min Wang, Jen-Wei Kuo, and Hung-Yi Lo
Institute of Information Science, Academia Sinica, Taipei, Taiwan
{whm, rogerkuo, hungyi}@iis.sinica.edu.tw

Abstract

Phoneme level transcription of speech corpora is crucial to fundamental speech research and the increasingly interested detection-based automatic speech recognition. Currently, there is no existing phoneme-labeled Mandarin Chinese speech corpus. This paper presents our recent work towards development of such a corpus. Our goal is to label five hours of speech data selected from a Mandarin Chinese broadcast news corpus. To reduce the human effort and accelerate the labeling process, we divide the speech data into subsets and employ our recently proposed HMM/SVM-based two-stage automatic phoneme segmentation framework to obtain the initial phoneme segmentation for subsequent manual correction subset by subset. The results of experiments on the first four subsets that have been manually verified show that the cost of labeling one subset can be progressively reduced.

1 Introduction

Phoneme level transcription of speech corpora is crucially important to fundamental speech research and the increasingly interested detection-based automatic speech recognition (Lee *et al.*, 2007). However, manual phoneme segmentation of speech signals is extremely time consuming and costly. To reduce the human effort and accelerate the labeling process, we have recently proposed an HMM/SVM-based two-stage framework (Lo and Wang, 2007) for automatic phoneme segmentation. The first stage aligns a phoneme sequence of a speech utterance with its acoustic signal counterpart according to the minimum boundary error (MBE) criterion, based on MBE-trained hidden Markov models (HMMs) (Kuo and Wang, 2006). The second stage uses a support vector machine (SVM) to refine the hypothesized phoneme boundaries derived by HMM-based forced alignment.

Since there is no existing phoneme-labeled Mandarin Chinese speech corpus, we select approximately five hours of speech data from the MATBN Mandarin Chinese broadcast news speech corpus (Wang *et al.*, 2005) for further phoneme annotation. To reduce costs, we employ our HMM/SVM framework to obtain the initial phoneme segmentation for subsequent manual segmentation and verification. To do this, we divide the 5-hour speech data into 60 5-minute subsets. First, we perform conventional unsupervised maximum likelihood (ML) training of HMMs and HMM-based forced alignment on the complete set to generate the initial segmentation. When the first subset has been manually verified, it is used for supervised training of the HMMs and SVMs. To prevent over-fitting in HMM training, the remaining unverified data is used to smooth the HMM parameters. Then, based on the new HMMs and SVMs, we apply improved HMM/SVM segmentation to the remaining subsets to generate more accurate phoneme boundaries. The above training and segmentation process is repeated subset by subset until all the subsets have been manually verified. It is expected that, in this way, the accuracy of automatic segmentation can be improved stage by stage, and the overall cost of manual segmentation can be reduced.

Now the first 3 subsets have been processed completely, and the fourth subset has been processed partially. So, we evaluate the efficacy of our automatic phoneme segmentation process on them. The experiment results clearly show that the segmentation accuracy can be improved if more subsets are manually verified, i.e., the cost of labeling one subset can be progressively reduced.

The remainder of this paper is organized as follows. Section 2 presents our HMM/SVM-based two-stage automatic phoneme segmentation framework. Section 3 describes how we apply the framework for labeling the speech data selected from the MATBN Mandarin Chinese speech corpus. Section 4 details the experiment results. Finally, in Section 5, we present our conclusions.

2 HMM/SVM-based Two-stage Phoneme Segmentation

The HMM/SVM-based two-stage framework for automatic phoneme segmentation tries to imitate the human phoneme segmentation process. The first stage performs HMM-based forced alignment according to the minimum boundary error (MBE) criterion. The objective is to align a phoneme

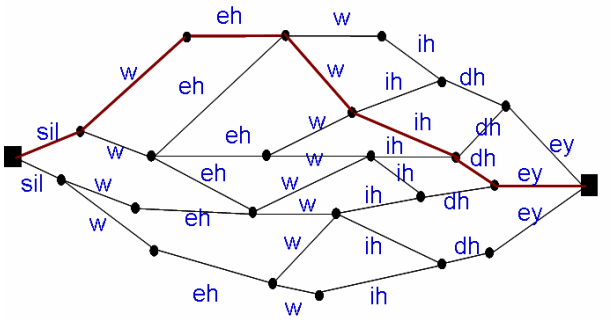

Figure 1: An illustration of the phonetic lattice for the speech utterance "where were they?".

sequence of a speech utterance with its acoustic signal counterpart based on MBE-trained HMMs. The second stage uses SVM to refine the hypothesized phoneme boundaries derived by HMM-based forced alignment, based on some discriminative features and mel-frequency cepstrum coefficients (MFCCs).

2.1 HMM-based phoneme segmentation

2.1.1 Minimum boundary error (MBE) training

Let $\mathbf{O} = \{O^1, .., O^R\}$ be a set of training observation sequences. The objective function for MBE training can then be defined as

$$F_{MBE} = \sum_{r=1}^{R} \sum_{S_i^r \in \mathbf{\Phi}^r} P(S_i^r \mid O^r) ER(S_i^r, S_c^r), \quad (1)$$

where $\mathbf{\Phi}^r$ is a set of possible phoneme alignments for the training observation sequence O^r; S_i^r is one of the hypothesized alignments in $\mathbf{\Phi}^r$; S_c^r is the manually labeled phoneme alignment; $P(S_i^r \mid O^r)$ is the posterior probability of alignment S_i^r given O^r; and $ER(S_i^r, S_c^r)$ denotes the "boundary error" of S_i^r compared with the manually labeled phoneme alignment S_c^r. For each training observation sequence O^r, F_{MBE} gives the weighted average boundary error of all hypothesized alignments. However, Eq. (1) cannot be used directly because, in practice, $P(S_i^r \mid O^r)$ is unknown. For simplicity, we assume that the prior probability of alignment S_i^r is uniformly distributed, and the likelihood $p(O^r \mid S_i^r)$ of alignment S_i^r is governed by the acoustic model parameter set Λ. Therefore, Eq. (1) can be rewritten as

$$F_{MBE} = \sum_{r=1}^{R} \sum_{S_i^r \in \mathbf{\Phi}^r} \frac{p_\Lambda(O^r \mid S_i^r)^\xi}{\sum_{S_k^r \in \mathbf{\Phi}^r} p_\Lambda(O^r \mid S_k^r)^\xi} ER(S_i^r, S_c^r), \quad (2)$$

where ξ is a scaling factor that prevents the denominator $\sum_{S_k^r \in \mathbf{\Phi}^r} p_\Lambda(O^r \mid S_k^r)$ from being dominated by only a few alignments.

Since $\mathbf{\Phi}^r$ contains a huge number of hypothesized phoneme alignments, for efficiency, we restrict the hypothesized space $\mathbf{\Phi}^r$ to the set of alignments constructed from a phoneme lattice like the example shown in Fig. 1. The boundary error $ER(S_i^r, S_c^r)$ of the hypothesized alignment S_i^r is calculated as the sum of the boundary errors of the individual phonemes in S_i^r, i.e., $ER(S_i^r, S_c^r) = \sum_{n=1}^{N^r} er(q_n^i, q_n^c)$, where N^r is the number of total phonemes in O^r; q_n^i and q_n^c are the n-th phonemes in S_i^r and S_c^r, respectively; and $er(q_n^i, q_n^c)$ is the phoneme boundary error calculated as $\frac{1}{2} \times (|s_n^i - s_n^c| + |e_n^i - e_n^c|)$, where s_n^i and e_n^i are, respectively, the start time and end time of phoneme q_n^i; and s_n^c and e_n^c correspond to the human-labeled start time and end time, respectively.

The optimal parameter set Λ^* can be estimated by minimizing the objective function defined in Eq. (2) using the extended Baum-Welch (EB) algorithm (Povey, 2003). The detailed derivations of the re-estimation formulae for the model parameters can be found in (Kuo and Wang, 2006).

2.1.2 MBE segmentation

The MBE alignment approach is a promising realization of the *Minimum Bayes-Risk* (MBR) classifier for the automatic phoneme segmentation task. The latter can be considered as an action, $\alpha_S(O)$, taken to identify a certain alignment, S, from all the phoneme alignments of a given utterance O. Let the function $L(S, S_c)$ be the loss incurred when the action $\alpha_S(O)$ is taken, given that the true alignment is S_c. During the classification stage, we do not know the true alignment in advance, i.e., any arbitrary alignment S_j could be true. The MBR classifier is designed to select the action whose conditional risk, $R(\alpha_S \mid O) = \sum_{S_j \in \mathbf{\Phi}} L(S, S_j) P(S_j \mid O)$, is minimal, i.e., the best alignment based on the MBR criterion can be found by

$$S^* = \arg\min_S \sum_{S_j \in \mathbf{\Phi}} L(S, S_j) P(S_j \mid O). \quad (3)$$

When the symmetrical zero-one function,

$$L(S, S_j) = \begin{cases} 0, & S = S_j \\ 1, & S \neq S_j \end{cases}, \quad (4)$$

is selected as the loss function, and it is assumed that the prior probability of alignment S_j is uniformly distributed, the MBR classifier is

equivalent to the conventional forced-alignment method, which picks the alignment with the maximal likelihood. When the loss function is replaced by the boundary error function, the MBR classifier becomes the MBE forced alignment approach, defined as:

$$\begin{aligned} S^* &= \arg\min_S \sum_{S_j \in \mathbf{\Phi}} ER(S, S_j) P(S_j \mid O) \\ &= \arg\min_S \sum_{S_j \in \mathbf{\Phi}} \sum_{n=1}^{N} er(q_n, q_n^j) P(S_j \mid O), \end{aligned} \tag{5}$$

where N is the number of phonemes in O; and q_n and q_n^j are the n-th phonemes in the alignments S and S_j, respectively. To simplify the implementation, we restrict the hypothesized space $\mathbf{\Phi}$ to the set of alignments constructed from the phoneme lattice shown in Fig. 1, which can be generated by a conventional beam search.

Let the ***cut*** $\mathbf{C}_n$ be the set of phoneme arcs of the n-th phoneme in the utterance. For example, in Fig. 1, there are four phoneme arcs for the second phoneme, "w", in $\mathbf{C}_2$ and six phoneme arcs for the third phoneme, "eh", in $\mathbf{C}_3$. From the figure, it is obvious that each alignment in $\mathbf{\Phi}$ will pass a single phoneme arc in each ***cut*** $\mathbf{C}_n$, n=1,2,...,N. Based on this observation, Eq. (5) can be rewritten as:

$$\begin{aligned} S^* &= \arg\min_S \sum_{n=1}^{N} \sum_{S_j \in \mathbf{\Phi}} P(S_j \mid O) er(q_n, q_n^j) \\ &= \arg\min_S \sum_{n=1}^{N} \sum_{q_{n,m} \in \mathbf{C}_n} \rho_{q_{n,m}} er(q_n, q_{n,m}), \end{aligned} \tag{6}$$

where $q_{n,m}$ is the m-th phoneme arc in $\mathbf{C}_n$; and $\rho_{q_{n,m}} = \sum_{\{S_j \in \mathbf{\Phi} \mid q_{n,m} \in S_j\}} P(S_j \mid O)$ is equivalent to the posterior probability of $q_{n,m}$ given the utterance O, which can be calculated by applying a forward-backward algorithm to the phoneme lattice. In this way, MBE forced alignment can be performed efficiently on the phoneme lattice via a Viterbi search.

2.2 Boundary refinement using SVM

For each initial boundary detected by HMM-based segmentation, several hypothesized boundaries around it are identified, and each one is examined by a phoneme-transition-dependent SVM classifier; then, the initial boundary is replaced by the most likely boundary.

2.2.1 Phoneme transition clustering

Ideally, we should be able to train an SVM classifier for each type of phoneme transition. However, this is not feasible because the training data is always limited. Maintaining a balance between the available training data and the model's complexity is critical to the training process. Furthermore, since many phoneme transitions have similar acoustic characteristics, we can partition them into clusters so that the training data can be shared and the phoneme transitions with little training data can be covered by the SVM classifiers of the categories they belong to.

For each type of phoneme transition, we gather all the feature vectors associated with the human-labeled phoneme boundaries and compute the mean vector. We then apply the K-means algorithm to cluster the phoneme transitions according to their mean vectors. Note that only phoneme transitions with enough instances are considered in this step. Finally, we assign the phoneme transitions ignored during clustering to the nearest clusters according to the Euclidean distances between their mean vectors and the cluster centers.

2.2.2 Support Vector Machine

Consider the problem of classifying data points into two classes, A_+ and A_-. We are given a training data set $\{(x_i, y_i)\}_{i=1}^m$, where $x_i \subset R^n$ is an input vector variable and $y_i \in \{1, -1\}$ is a class label that indicates which of the two classes, A_+ and A_-, it belongs to. We represent these data points by an $m \times n$ matrix A, in which the i-th row, A_i, corresponds to the i-th data point. The SVM classifier $f(x)$ is of the following form:

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i K(A_i, x) + b, \tag{7}$$

where $K(A_i, x)$ is a kernel function, and α_i and b are parameters to be trained.

For each phoneme transition cluster, an SVM classifier is trained by using the feature vectors associated with the true boundaries as positive training samples and the randomly selected feature vectors at least 20 ms away from the true boundaries as negative training samples. In the test phase, the feature vectors associated with the speech frames around the hypothesized boundary are examined by the associated SVM classifier. Then, the frame index associated with the feature vector with the maximum classifier output is recognized as the refined boundary.

3 The Use of HMM/SVM Segmentation in Labeling the MATBN Corpus

We have applied the HMM/SVM-based two-stage automatic phoneme segmentation framework for labeling the speech data selected from the MATBN Mandarin Chinese broadcast news corpus (Wang *et al.*, 2005).

3.1 The MATBN corpus

Training/Segmentation	SS1	SS2	SS3	SS4
Complete set unsupervised ML/ML	19.33/31.81	17.96/40.40	18.12/36.65	16.35/41.21
SS1 supervised ML/ML	NA	15.00/55.40	13.06/62.56	11.68/63.22
SS1 supervised MBE/MBE+SVM	NA	13.58/59.26	11.93/65.25	10.42/68.95
SS1+SS2 supervised MBE/MBE+SVM	NA	NA	10.61/67.34	9.71/70.59
SS1+SS2+SS3 supervised MBE/MBE+SVM	NA	NA	NA	9.35/71.79

Table 1: The results of automatic phoneme segmentation in mean boundary distance in millisecond/percentage of phoneme boundaries correctly placed within a 10 millisecond tolerance with respect to the human labeled phoneme boundaries

The MATBN Mandarin Chinese corpus contains 198 hours of broadcast news from the Public Television Service Foundation (Taiwan). The data includes orthographic transcripts and SGML tagging for annotating acoustic conditions, background conditions, story boundaries, speaker turn boundaries, and acoustic events, such as hesitations and repetitions. We select approximately five hours of speech data from the corpus for further phoneme annotation.

3.2 Strategy

We divide the 5-hour speech data into 60 5-minute subsets. First, we perform conventional unsupervised maximum likelihood (ML) training of HMMs and HMM-based forced alignment on the complete set to generate the initial segmentation. When the first subset has been manually verified, it is used for supervised training of the HMMs and SVMs. To prevent over-fitting in HMM training, the remaining unverified data is used to smooth the HMM parameters. Then, based on the new HMMs and SVMs, we apply improved HMM/SVM segmentation to the remaining subsets to generate more accurate phoneme boundaries. The above training and segmentation process is repeated subset by subset until all the subsets have been manually verified.

4 Experiments

Now the first 3 subsets have been processed completely, and the fourth subset has been processed partially. So, we can evaluate the efficacy of our automatic phoneme segmentation process on them.

4.1 Experiment setup

The acoustic models for HMM-based segmentation consist of 34 context-independent phoneme models, each represented by a 3-state continuous density HMM with a left-to-right topology. Each frame of the speech data is represented by a 39-dimensional feature vector comprised of 12 MFCCs and log energy, along with their first and second time derivatives. The frame width is 20 ms and the frame shift is 5 ms. Utterance-based cepstral variance normalization (CVN) is applied to all the training and test speech utterances.

In the SVM refinement stage, each frame of the speech data is represented by a 45-dimensional feature vector comprised of the above 39 MFCC-based coefficients, plus the zero crossing rate, bisector frequency (Lin *et al.*, 2005), burst degree (Lin *et al.*, 2005), spectral entropy, weighted entropy (Shen *et al.*, 1998), and subband energy. For each hypothesized boundary, the feature vectors of its adjacent left and right frames, together with the symmetrical Kullback-Leibler distance (SKLD) (Klabbers and Veldhuis, 2001) and the spectral feature transition rate (SFTR) (Nandasena and Akagi, 1998) between the two feature vectors, are concatenated to form a 92-dimensional augmented vector. The augmented vectors are used as features for phoneme transition clustering and as the input vectors for SVM. Given the boundary of each phoneme transition obtained by HMM-based segmentation, 11 hypothesized boundaries (extracted every 1 ms) around the initial boundary within ± 5 ms are examined by the SVM classifier associated with that specific phoneme transition. In total, 16 phoneme-transition-dependent SVMs are used. The SVM classifiers with Gaussian kernels are implemented by LIBSVM (Chang and Lin, 2001).

4.2 Experiment results

Table 1 shows the experiment results. "Complete set unsupervised ML/ML" denotes the conventional unsupervised HMM-based segmentation. "SS1 supervised MBE/MBE+SVM" means that the HMM/SVM framework uses HMMs (MBE-trained) and SVMs trained with the first subset (SS1), while "SS1 supervised ML/ML" denotes its conventional HMM-based segmentation counterpart using supervised ML-trained HMMs. Since the models are trained with the first subset, they can be tested on the following subsets, i.e., SS2 to SS4. Comparing the results in rows 2 (Complete set unsupervised ML/ML), 3 (SS1 supervised ML/ML), and 4 (SS1 supervised MBE/MBE+SVM) of Table 1, we observe that the supervised trained HMMs outperform the unsupervised trained HMMs and our HMM/SVM-

Training/Segmentation	Mean Boundary Distance	%Correct marks (distance ≦ tolerance)			
		≦5ms	≦10ms	≦15ms	≦20ms
Complete set unsupervised ML/ML	16.35	29.23	41.21	54.03	67.14
SS1 supervised ML/ML	11.68	38.68	63.22	77.14	84.61
SS1 supervised MBE/MBE+SVM	10.42	43.92	68.94	80.84	88.62
SS1+SS2 supervised MBE/MBE+SVM	9.71	47.43	70.59	82.58	89.46
SS1+SS2+SS3 supervised MBE/MBE+SVM	9.35	48.45	71.79	83.82	89.83

Table 2: The results of evaluation on SS4 in percentage of phoneme boundaries correctly placed within different tolerances with respect to the human labeled phoneme boundaries

based segmentation outperforms the conventional HMM-based segmentation, which uses supervised trained HMMs. Comparing the results in rows 4, 5, and 6, we observe that the segmentation accuracy can be improved if more subsets are manually verified. In other words, the cost of labeling one subset can be progressively reduced.

Table 2 shows the detailed results of evaluation on the fourth subset (SS4) in percentage of phoneme boundaries correctly placed within different tolerances with respect to the human labeled phoneme boundaries. From the fifth row of the table, we observe that, based on the HMMs and SVMs trained with the first two subsets (10 mininutes of labeled training speech), the HMM/SVM framework can achieve a mean boundary distance of less than 10ms and an accuracy of near 90% within a tolerance of 20ms. The results show that a small amount of labeled data can be very useful for improving the segmentation accuracy.

5 Conclusion

In this paper, we have explored the use of our recently proposed HMM/SVM-based two-stage automatic phoneme segmentation framework in phoneme labeling of the speech data selected from a Mandarin Chinese speech corpus that has orthographic transcripts. We divided the speech data into subsets and employed automatic phoneme segmentation to obtain the initial phoneme segmentation for subsequent manual correction subset by subset. The preliminary results of evaluation on the first four subsets that have been manually verified are rather promising. They demonstrate that the segmentation accuracy can be improved if more subsets are manually verified, i.e., the cost of labeling one subset can be progressively reduced. The annotation work is ongoing and the results will be made available at a future time.

6 Acknowledgements

This work was supported by the National Science Council of Taiwan under Grant: NSC 96-2221-E-001-003.

References

E. Klabbers and R. Veldhuis. 2001. Reducing audible spectral discontinuities. IEEE *Trans. on Speech and Audio Processing*, 9(1):39-51.

J. W. Kuo and H. M. Wang. 2006. A minimum boundary error framework for automatic phonetic segmentation. *ISCSLP2006*.

C. H. Lee, M. A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B. H. Juang, and L. R. Rabiner. 2007. An overview on automatic speech attribute transcription (ASAT). *InterSpeech2007*.

C. C. Chang and C. J. Lin. 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

C. Y. Lin, J. S. R. Jang, and K. T. Chen. 2005. Automatic segmentation and labeling for Mandarin Chinese speech corpora for concatenation-based TTS. *International Journal of Computational Linguistics and Chinese Language Processing*, 10(2):145-166.

H. Y. Lo and H. M. Wang. 2007. Phonetic boundary refinement using support vector machine. *ICASSP2007*.

A. C. R. Nandasena and M. Akagi. 1998. Spectral stability based event localizing temporal decomposition. *ICASSP1998*.

D. Povey. 2003. *Discriminative Training for Large Vocabulary Speech Recognition*. Ph.D. Thesis, University of Cambridge.

J. L. Shen, J. W. Hung, and L. S. Lee. 1998. Robust entropy-based endpoint detection for speech recognition in noisy environments. *ICSLP1998*.

H. M. Wang, B. Chen, J. W. Kuo, and S. S. Cheng. 2005. MATBN: A Mandarin Chinese broadcast news corpus. *International Journal of Computational Linguistics and Chinese Language Processing*, 10(2):219-236.