PHONETIC BOUNDARY REFINEMENT USING SUPPORT VECTOR MACHINE

Hung-Yi Lo and Hsin-Min Wang

Institute of Information Science, Academia Sinica, Taipei, Taiwan, Republic of China
{hungyi, whm}@iis.sinica.edu.tw

ABSTRACT

In this paper, we propose using support vector machine (SVM) to refine the hypothesized phone transition boundaries given by the HMM-based Viterbi forced alignment. We conducted experiments on the TIMIT speech corpus. The phone transitions were automatically partitioned into 46 clusters according to their acoustic characteristics and the cross-validation using the training data; hence, 46 phone-transition-dependent SVM classifiers were used for phone boundary refinement. The proposed HMM-SVM approach performs as well as the recent discriminative HMM-based segmentation. The best accuracies achieved are 81.23% within a tolerance of 10 ms and 92.47% within a tolerance of 20 ms. The mean boundary distance is 7.73 ms.

Index Terms— Automatic phone alignment, support vector machine, reduced support vector machine

1. INTRODUCTION

Annotated speech corpora are indispensable to various areas of speech research, e.g., speech recognition and speech synthesis. Phoneme level annotation is especially important for fundamental speech research. However, the development of a large high-quality, manually labelled speech corpus requires lots of human effort, and is time-consuming. To reduce the human effort and speed up the labelling process, many attempts have been made to utilize automatic phone alignment approaches to provide initial phonetic segmentation for subsequent manual segmentation and verification [1, 2, 3].

The most popular method of automatic phone alignemnt is to adapt an HMM-based phonetic recognizer to align a phonetic transcription with a speech utterance. Empirically, phone boundaries obtained in this way should contain few serious errors, since HMMs in general capture acoustic properties of phones; however, small errors are inevitable because HMMs are not sensitive enough to detect changes between adjacent phones.

In this paper, we propose using support vector machine (SVM) [4, 5] to refine the hypothesized boundaries given by the HMM-based Viterbi forced alignment. As will be detailed in the following section, we adapt the reduced support vector machine (RSVM) [5] algorithm to overcome the computational difficulty of applying SVM in a task with a massive data set. In our approach, a phone-transition-dependent SVM classifier is applied to detect the true phone transition boundary around each hypothesized boundary given by the initial HMM-based segmentation. These SVM classifiers for detecting boundaries of various phone transitions are trained in advance based on multiple discriminative features in addition to MFCCs. We conducted automatic phone alignment experiments on the TIMIT speech corpus. The proposed HMM-SVM approach performs as well as the improved HMM-based segmentation [3], which used a discriminative criterion, called minimum boundary error (MBE), instead of the conventional maximum likelihood (ML) criterion for HMM training. The best accuracies achieved are 81.23% within a tolerance of 10 ms and 92.47% within a tolerance of 20 ms. The mean boundary distance is 7.73 ms.

2. SUPPORT VECTOR MACHINE

Support vector machine (SVM) has become one of the most promising learning algorithms for classification as well as regression, and has been successfully applied to many real-world pattern recognition applications. SVM finds a separating surface with a large margin between training samples of two classes in a high dimensional feature space implicitly introduced by a computationally efficient kernel mapping, and the large margin implies a better generalization ability according to the statistical learning theory [4]. The reduced support vector machine (RSVM) [5] algorithm to implement SVM was proposed in an attempt to overcome the computational difficulty as well as to reduce the model complexity in generating a nonlinear separating surface for a massive data set.

2.1. Reduced 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 $x^i \in \chi \subset R^n$ is an input vector variable and $y^i \in \{1, -1\}$ is a class label, which indicates one of the two classes, A_+ and A_-, to which the data point belongs. We represent these data points by an $m \times n$ matrix A, where the i-th row of the matrix A, A_i, corresponds to the i-th data point. We use an $m \times m$ diagonal matrix D, $D_{ii} = y_i$, to specify the

class membership of each data point. The main goal of training is to find a classifier that can correctly predict the class label of an unseen data point. This can be achieved by constructing a nonlinear separating surface which is implicitly defined by a kernel function. In conventional SVM [4], the nonlinear kernel matrix $K(A, A') \in R^{m\times m}$ (where m is the size of the training data set) on large data sets will lead to some computational difficulties [5]. The RSVM [5], which uses a very small random subset of size $\bar{m}$ of the original m data points, where $\bar{m} << m$, can avoid these difficulties. We denote this random subset by $\bar{A}$, which is used to generate a much smaller rectangular matrix $K(A, \bar{A}') \in R^{m\times\bar{m}}$ and to replace the huge and fully dense square kernel matrix $K(A, A')$ used in conventional SVM to cut the problem size, computational time and memory usage as well as to simplify the characterization of nonlinear separating surface. We now briefly describe the reduced support vector machine formulation, which is derived from the generalized support vector machine (GSVM) [6] and smooth support vector machine [7]. The RSVM solves the following unconstrained minimization problem for an arbitrary rectangular kernel $K(A, \bar{A}')$:

$$\min_{(\bar{u},\gamma)\in R^{\bar{m}+1}} \frac{\nu}{2}\|p(e - D(K(A, \bar{A}')\bar{D}\bar{u} - e\gamma), \alpha)\|_2^2 + \frac{1}{2}(\bar{u}'\bar{u} + \gamma^2), \quad (1)$$

where the function $p(x, \alpha)$ is a very accurate smooth approximation to $(x)_+$ [7], which is applied to each component of the vector $e - D(K(A, \bar{A}')\bar{D}\bar{u} - e\gamma)$ and is defined componentwise by

$$p(x, \alpha) = x + \frac{1}{\alpha}\log(1 + e^{-\alpha x}), \alpha > 0. \quad (2)$$

The function $p(x, \alpha)$ converges to $(x)_+$ as α goes to infinity. The positive tuning parameter ν here controls the tradeoff between the classification error and the suppression of $(\bar{u}, \gamma)$.The diagonal matrix $\bar{D} \in R^{\bar{m}\times\bar{m}}$ with ones or minus ones along its diagonal to specify the membership of each point in the reduced set. A solution of this minimization program for $\bar{u}$ and γ leads to the nonlinear separating surface

$$K(x', \bar{A}')\bar{D}\bar{u} = \gamma. \quad (3)$$

Problem (1) retains the strong convexity and differentiability properties in the $R^{\bar{m}+1}$ space of $(\bar{u}, \gamma)$ for any arbitrary *rectangular* kernel. Hence we can apply the Newton-Armijo Algorithm [7] directly to solve (1) and the existence and uniqueness of the optimal solution of the minimization problem (1) are also guaranteed. In a nutshell, the RSVM can be split into two parts. First, it selects a small random subset $\{K(\cdot, \bar{A}'_1), K(\cdot, \bar{A}'_2), \cdots, K(\cdot, \bar{A}'_{\bar{m}})\}$ from the full-data basis set. The full-data set is inefficient with possibly heavy overlaps in function representation, but the conventional SVM has been using it. Secondly, the RSVM determines the best coefficients of the selected kernel functions by solving the unconstrained minimization problem (1) using the entire data set so that the surface will fit the whole data well.

3. PHONETIC BOUNDARY REFINEMENT USING SVM

The proposed SVM-based phonetic boundary refinement proceeds as follows. For each initial boundary detected by the HMM-based segmentation, several hypothesized boundaries around it are identified first; then each of which is examined by a phone-transition-dependent SVM classifier; and finally the most likely boundary is selected to replace the initial boundary. The SVM classifiers for detecting boundaries of various phone transitions are trained in advance based on multiple discriminative features in addition to MFCCs.

3.1. Useful features

In the HMM-based segmentation, each frame of the speech data is represented by a 39 dimensional MFCC-based feature vectors comprised of 12 MFCCs and log energy, plus their delta and delta-delta coefficients. In the refinement stage, each frame is represented by a 45 dimensional feature vector consisting of the above 39 MFCC-based coefficients, plus zero crossing rate, bisector frequency [8], burst degree [8], spectral entropy, general weighted entropy [9], and subband energy.

For each hypothesized boundary, the feature vectors of the left and right frames next to it, together with the symmetrical Kullback-Leibler distance (SKLD) and the spectral feature transition rate (SFTR) between the two feature vectors, are concatenated to form a 92 dimensional augmented vector. The augmented vectors are used as features to cluster the phone transitions and as the input vectors to SVM.

3.2. Phone transition clustering

Ideally, we can train a SVM classifier for each kind of phone transition. However, this is generally not feasible because the training data is always limited and some specific phone transitions might have the sparse data problem. In practical implementations, we need to partition the phone transitions into clusters according to their acoustic characteristics, such that the training data can be shared and the phone transitions with little training data can be covered by the SVM classifiers of categories to which they belong.

The partition can be determined based on either prior knowledge [10] or statistical learning [1]. In this paper, we use a data-driven clustering approach as follows:

1. For each specific phone transition case, we gather all augmented vectors associated with the human-labelled phone boundaries, and compute the mean vector.

2. For each one of the three phone transition classes, namely *sonorant to non-sonorant*, *sonorant to sonorant*, *non-sonorant to non-sonorant*, we apply the K-means algorithm to cluster the phone transitions according to their mean vectors. Note that only the phone transitions with enough instances are considered in this step. The number of clusters is determined according to the cross-validation accuracy that the resulting SVM classifiers achieve in the training data.

3. We assign the phone transitions, which are ignored in Step 2 due to sparse instances, to the nearest clusters according to the distances between their mean vectors and the cluster centers.

3.3. Input vector to SVM

For each partition subset, two discriminative features, namely discriminative weighted entropy and discriminative subband energy, are believed to be more specialized to each partition subset. The discriminative weighted entropy is computed by

$$H = -\sum_{i=1}^{N} w_i^e p_i \log p_i, \tag{4}$$

where w_i^e is a weight vector and p_i is the element of power spectrum which is normalized to satisfy $\sum_{i=1}^{N} p_i = 1$. The weight vector of each partition subset is trained by linear SVM using the vectors $p \log p$ extracted from the right frames next to the true boundaries as positive samples and those from the left frames as negative samples. The goal of training the weight vector is to maximize the variation of the weighted entropy feature close to the true boundary. The discriminative subband energy is computed by:

$$E_{sub} = E_j \,\big|_{\arg\max_j F_j}, \tag{5}$$

where E_j, $j = 1, \ldots, 9$, is pre-defined subband energy and the weight score F_j is:

$$F_j = \frac{\mu_j^+ - \mu_j^-}{\sigma_j^+ - \sigma_j^-}, \tag{6}$$

where μ_j and σ_j are the mean and standard deviation of the j-th subband energy for the training samples of the positive or negative class.

After these two parameters are determined, the general weighted entropy and the subband energy extracted in section 3.1 are replaced by the discriminative weighted entropy and the discriminative subband energy in the input vectors to SVM in both training and testing phases.

3.4. Boundary recognition using SVM

For each phone transition subset, a SVM classifier is trained by the RSVM algorithm for boundary detection using the augmented vectors associated with the true boundaries as the positive training samples and the randomly selected augmented vectors at least ± 20 ms away from the true boundaries as the negative training samples. Gaussian kernel with the weighted Euclidean distance $K(x, z') = e^{-\gamma \|w(x_i - x_j)\|_2^2}$ is applied, and the weight is used to emphasize the more important and discriminative features. In the testing phase, the augmented vectors associated with the speech frames around the hypothesized boundary are examined by the SVM classifier according to the partition to which the phone transition belongs, and the frame index associated with the augmented vector with the maximum classifier output is recognized as the refined boundary.

4. EXPERIMENT RESULTS

4.1. Experiment setup

Our experiments were conducted on the TIMIT acoustic-phonetic continuous speech corpus. TIMIT contains a total of 6,300 sentences, comprised of 10 sentences spoken by each of 630 speakers from 8 major dialect regions in the United States. The TIMIT suggested training and testing sets contain 462 and 168 speakers, respectively. We discard the dialect sentences (SA1 and SA2 utterances) and utterances with phones shorter than 10 ms. The resulting training set and testing set contain 3,696 sentences and 1,312 sentences, respectively.

The acoustic models for HMM-based segmentation consist of 50 context-independent phone models, each represented by a 3-state continuous density HMM (CDHMM) 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, and their delta and delta-delta coefficients. 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 testing speech. The acoustic models were trained on the training speech according to the human-labelled phonetic transcriptions and boundaries by the Baum-Welch algorithm using the ML criterion with 10 iterations.

By using the cross-validation on the TIMIT training data, the number of phone transition cluster is 20 in the *sonorant to non-sonorant* class, 16 in the *sonorant to sonorant* class, and 10 in the *non-sonorant to non-sonorant* class. As a result, 46 SVM classifiers are used. In the refinement phase, given the boundary of each phone transition obtained by the HMM-based segmentation, 16 hypothesized boundaries extracted every 5 ms around the initial boundary within ± 40 ms will be examined by SVM.

The proposed HMM-SVM approach was compared with the improved HMM_{MBE}-based segmentation [3]. The MBE discriminative training approach was applied to manipulate the above ML-trained HMMs with 10 more iterations.

Table 1. The percentage of phone boundaries correctly placed within different tolerances with respect to their associated human-labelled phone boundaries.

Methods	Mean Boundary Distance (ms)	Accuracy %				
		$< 5ms$	$< 10ms$	$< 20ms$	$< 30ms$	$< 40ms$
HMM_{ML}	9.73	46.85	71.53	89.17	94.62	97.16
HMM_{ML+MBE}	7.79	58.73	80.15	92.09	95.93	97.89
HMM_{ML}-SVM	7.82	58.18	81.19	92.47	96.05	97.78
HMM_{ML+MBE}-SVM	7.73	58.25	81.23	92.46	96.08	97.95

4.2. Experiment results

Table 1 shows the percentage of phone boundaries correctly placed within different tolerances with respect to their associated human-labeled phone boundaries. The second row represents the results of the ML-trained HMM forced alignment, and the third row comes from the MBE-trained HMM forced alignment. The fourth row is the performance of the SVM-based refinement based on the initial boundaries given by the ML-trained HMM forced alignment, and the fifth row is the performance of the SVM-based refinement based on the initial boundaries given by the MBE-trained HMM forced alignment. We observe that the proposed HMM_{ML}-SVM approach performs as well as the discriminative HMM_{MBE}-based segmentation. However, the SVM-based refinement can only slightly improve the segmentation accuracy given the initial boundaries provided by the HMM_{MBE}-based segmentation. It seems that the refinement system doesn't benefit much from a more accurate initial alignment. The best accuracies achieved are 81.23% within a tolerance of 10 ms and 92.47% within a tolerance of 20 ms. The mean boundary distance is 7.73 ms.

5. CONCLUSIONS

SVM has been successfully applied in many applications, but it is less widely applied in speech processing research. In this paper, we have presented a SVM-based boundary refinement approach to improve the HMM-based forced alignment for automatic phonetic segmentation. The preliminary experiment results on the TIMIT corpus show that the proposed HMM-SVM approach performs as well as the improved HMM-based segmentation, which used a minimum boundary error (MBE) criterion for discriminative HMM training. Although the current SVM-based refinement system seems not able to benefit from a more accurate initial alignment given by the HMM_{MBE} forced alignment in our experiments, a more accurate segmentation is expectable if a more comprehensive investigation into the acoustic features and the characteristics of various phone transitions can be carried out to improve the phone-transition-dependent SVM classifiers.

6. ACKNOWLEDGMENTS

This work was supported in part by the National Science Council, Taiwan, under Grant: NSC95-2221-E-001-035.

7. REFERENCES

[1] K. S. Lee, "MLP-based phone boundary refining for a tts database," *IEEE Trans. on Speech and Audio Processing*, vol. 14, pp. 981–989, 2006.

[2] J. Keshet, S. Shalev-Shwartz, Y. Singer, and D. Chazan, "Phoneme alignment based on discriminative learning," in *Proc. Interspeech*, 2005.

[3] J.-W. Kuo and H.-M. Wang, "Minimum boundary error training for automatic phonetic segmentation," in *Proc. Interspeech*, 2006.

[4] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[5] Y.-J. Lee and O. L. Mangasarian, "RSVM: Reduced support vector machines," in *Proc. SDM*, 2001.

[6] O. L. Mangasarian, "Generalized support vector machines," in *Advances in Large Margin Classifiers*, 2000.

[7] Y.-J. Lee and O. L. Mangasarian, "SSVM: A smooth support vector machine," *Computational Optimization and Applications*, vol. 20, pp. 5–22, 2001.

[8] C.-Y. Lin, J.-S. Roger Jang, and K.-T. Chen, "Automatic segmentation and labeling for mandarin chinese speech corpora for concatenation-based TTS," *Computational Linguistics and Chinese Language Processing*, vol. 10, no. 2, pp. 145–166, 2005.

[9] J.-L. Shen, J.-W. Hung, and L.-S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," in *Proc. ICSLP*, 1998.

[10] E.-Y. Park, S.-H Kim, and J.-H Chung, "Automatic speech synthesis unit generation with MLP based post-processor against auto-segmented phoneme errors," in *Proc. IJCNN*, 1999.