THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

TECHNICAL REPORT OF IEICE.

Unsupervised Phoneme Segmentation Using Mahalanobis Distance

Yu QIAO† and Nobuaki MINEMATSU†

† Grad. School of Frontier Sciences, Univ. of Tokyo, 5–1–5, Kashiwanoha Kashiwa, Chiba, 277–8561 Japan
E-mail: †{qiao,mine}@gavo.t.u-tokyo.ac.jp

Abstract One of the fundamental problems in speech engineering is phoneme segmentation. Approaches to phoneme segmentation can be divided into two categories: supervised and unsupervised segmentation. The approach of this paper belongs to the 2nd category, which tries to perform phonetic segmentation without using any prior knowledge on linguistic contents and acoustic models. In an earlier work, we formulated the segmentation problem into an optimization problem through statistics and information analysis. An objective function, summation of squared error (SSE), is developed by using Euclidean distance of cepstral features. However, it is not known whether or not Euclidean distance yields the best distance metric to estimate the goodness of segmentations. A popular generalization of Euclidean distance is Mahalanobis distance (MD). In this paper, we study whether and how MD can be used to improve the performance of segmentation. The essential problem here is how to determine the parameters (covariance matrix) for MD calculation. We deal with this problem in a learning framework and propose two criteria for determining the optimal parameters: Minimum of Summation Variance (MSV) and Maximum of Discrimination Variance (MDV). MSV minimizes the summation of variance within phonemes, while MDV maximizes the variance between phonemes and minimizes the variance within phonemes at the same time. Both of them can lead to close form solutions by using matrix calculation. We also propose an algorithm to learn the parameters without using labeled data. We carried out experiments on the TIMIT database to eveluate the proposed methods. The results indicate that the use of learning MD can increase the correct recall rates. We also found the use of power can further improve the results.

Key words Unsupervised phoneme segmentation, Optimization, Mahalanobis distance, Learning distance metric

1. Introduction

Phoneme segmentation is a basic problem in speech engineering. The objective of phoneme segmentation is to divide a speech stream into a string of phonemes. Automatic Speech Recognition (ASR) models often require reliable phoneme segmentation in the initial training phase, while Text-to-Speech (TTS) systems need a large speech database with correct phoneme segmentation information for improving the performance. Human speech is a smoothly changing continuous signal, which does not include explicit separation marks such as the spaces in written language. Moreover, human speech is smoothly continuous signal due to the temporal constraints of vocal tract motions. The difficulty of phoneme segmentation comes from the co-articulation of speech sounds, where acoustic realization of one phoneme may blend or fuse with its adjacent sounds. This phenomenon can even exist at a distance of two or more phonemes. All these facts make automatic phoneme segmentation a challenging problem.

Previous approaches to phoneme segmentation can be classified into two categories: supervised and unsupervised segmentation. In the first case, both the linguistic contents and the acoustic models of phonemes are available. Thus the segmentation problem can be reduced to align speech signals with a string of acoustic models. Perhaps the most famous approach in this category is HMM-based forced alignment [2]. The second category tries to perform phonetic segmentation using no prior knowledge on linguistic contents and acoustic models. The approach of this paper belongs to the second class. The unsupervised segmentation is similar to the situation that infants acquire spoken language [11]. Infants don't have acoustic and linguistic models for segmentation. However, psychological facts indicate that infants become able to segment speech according to acoustic difference between speech sounds and cluster speech segments into categories [8]. It is only by this procedure that infants can gradually construct the speech model of their native languages.

Most of the previous approaches to this problem focus on detecting the change points of speech stream and take these change points as the boundaries of phonemes. Aversano et.

al [1] identified the boundaries as the peaks of jump function. Dusan and Rabiner [3] detected the "maximum spectral transition" positions as phoneme boundaries. Estevan et. al [4] employed maximum margin clustering to locate boundary points. In our earlier work, we formulated the segmentation problem as an optimization problem by using statistics and information theory analysis [9], while the critical question is how to evaluate the goodness of segmentation. Generally speaking, a good segmentation should minimize the within-phoneme variance while maximize the between-phoneme variance. In [9], we have developed a simple objective function, the Summation of Square Error (SSE). The experimental results [9] showed that minimizing SSE by Agglomerative Segmentation (AS) algorithm can achieve better results than previous methods [1], [3], [4]. Although this objective is computationally efficient, SSE is based on Euclidean distance in cepstral space and it is not known whether or not Euclidean distance yields the best distance metric to estimate the goodness of segmentations. In fact, it was shown that the weighted cepstral distance can achieve better performance than Euclidean distance for DTW based speech recognition [12]. A popular generalization of Euclidean distance is Mahalanobis distance. In this paper, we study whether and how Mahalanobis distance can be used to improve the performance of segmentation. The essential problem here is how to determine the parameters (covariance matrix) for Mahalanobis distance calculation. We deal with this problem in a learning framework and develop two criteria for determining the optimal parameters: Minimum of Summation Variance (MSV) and Maximum of Discrimination Variance (MDV). MSV tries to minimize the summation of variance within phonemes, while MDV aims to maximize the variance between phonemes and to minimize the variance within phonemes at the same time. We propose an algorithm to estimate parameters without using labeled sequences. The performances of the proposed methods are evaluated through experiments on the TIMIT database. The experimental results indicate that the learning Mahalanobis distance can help improving the segmentation results. We also found that the results can be further improved by incorporating power coefficients.

2. Optimal Segmentation

In this section, we introduce the used notations at first, and then give a brief review of our previous work on optimal segmentation [9]. Let $X = x_1, x_2, ..., x_n$ denote a sequence of mel-cepstrum vectors calculated from an utterance, where n is the length of X and x_i is a d-dimensional vector $[x_i^1, x_i^2, ..., x_i^d]^T$. The objective of segmentation is to divide sequence X into k non-overlapping contiguous subsequences (segments) where each subsequence corresponds to a phoneme. Use $S = \{s_1, s_2, ..., s_k\}$ to denote the segmentation information, where $s_j = \{c_j, c_j + 1, ..., e_j\}$ (c_j and e_j denote the start and end indices of the j-th segment.). Let $X_{c_j:e_j}$ (or X_{s_j}) represent the j-th segment $x_{c_j}, x_{c_j+1}, ..., x_{e_j}$. Its size $|s_j|$ is $e_j - c_j + 1$.

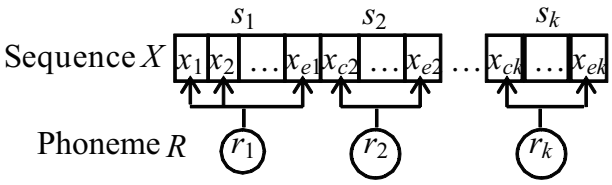

Figure 1 Diagram of Segmentation Model.

For speech signal, it is natural to make the assumption that acoustic observations of each phoneme is generated from an independent source. Let $R = \{r_1, r_2, ..., r_k\}$ denote the phoneme sequence, and $p(x_i|r_j)$ represent the probability model of observing x_i given source r_j. Thus we have,

$$p(X|S,R) = \prod_{j=1}^{k} \prod_{i \in s_j} p(x_i|r_j) = \prod_{j=1}^{k} \prod_{i=c_j}^{e_j} p(x_i|r_j). \qquad (1)$$

Using maximum likelihood estimation (MLE), the optimal segmentation can be formulated as

$$\hat{S} = \arg\min_{S} \{-\log(p(X|S,R))\}. \qquad (2)$$

Like most speech applications, we assume that r_j is a multivariable normal distributions whose mean and covariance matrix are denoted by m_j and Σ_j. If segmentation s_j is given, we can estimate the parameters by MLE. Using the estimated parameters $\hat{r}_j(\hat{m}_j, \hat{\Sigma}_j)$, Eq. 2 becomes,

$$\begin{aligned} -\log p(X|S,\hat{R}) &= \sum_{j=1}^{k} \sum_{i=c_j}^{e_j} -\log(p(x_i|r_j)) \\ &= \frac{nd}{2}\log(2\pi) + \frac{1}{2}\sum_{j=1}^{k} |s_j| \log\det(\hat{\Sigma}_j) + \frac{nd}{2}. \end{aligned} \qquad (3)$$

It can be shown that the above Equation is in accordance with the minimum description length principle (MDL) [10]. However, in practice, this approach may raise a problem, a phoneme usually only consists of a small number of frames, which makes it difficult to estimate reliable covariance matrix $\hat{\Sigma}$. Especially, when the number of frames is less than d, the covariance matrix is singular and $|\hat{\Sigma}| = 0$. Moreover, the calculation of matrix determinant is computationally expensive. To circumvent this difficulty, we fixed the covariance matrix Σ as an unit matrix I and only estimated mean $\hat{m}_j = 1/|s_j| \sum_{x \in s_j} x$ [9]. The use of other covariance matrix leads to Mahalanobis distance, which will be discussed in the next sections. In this way, Eq. 3 becomes,

$$-\log p(X|S,\hat{R}) = \frac{nd}{2}\log(2\pi) + \frac{1}{2}\sum_{j=1}^{k}\sum_{i=c_j}^{e_j}||x_i - \hat{m}_j||^2. \tag{4}$$

Note only the second item is influenced by segmentation S. Thus the problem is equal to minimizing the following *Summation of Squared Error* function (SSE),

$$f_{SSE}(X,S) = \sum_{j=1}^{k}\sum_{i=c_j}^{e_j}||x_i - \hat{m}_j||^2. \tag{5}$$

The above formula is the same as the objective function of k-means clustering (Chapter 3.5 [7]). The difference between our problem and k-means is that k-means needs not consider the time constraint, which is important for our phoneme segmentation.

In [9], we introduce the Agglomerative Segmentation (AS) algorithm, which begins with each frame as a segment and iteratively merges two consecutive segments into one in a greedy fashion. The algorithm has a time complexity of $O(n)$. We also proposed an efficient implementation of this algorithm by using integration functions.

3. Segmentation using Mahalanobis distance

The SSE objective Eq. 5 is based on simple Euclidean distance, where each dimension of cepstrum features is treated equally and the correlation between these features is ignored. However, in real problems, the cepstrum features can be correlated and different features may have different weights for segmentation. The Euclidean distance comes from the use of I as covariance matrix in Eq. 3. We may consider another covariance matrix. Let Σ denote a full rank covariance matrix. Euclidean distance $||x_i - x_j||^2$ can be generalized to Mahalanobis distance $(x_i - x_j)^T\Sigma^{-1}(x_i - x_j)$.

In this way, we can define a Mahalanobis distance based objective function as follows,

$$f_{MD}(X,S) = \sum_{j=1}^{k}\sum_{i=c_j}^{e_j}(x_i - \hat{m}_j)^T\Sigma^{-1}(x_i - \hat{m}_j). \tag{6}$$

If Σ is a diagonal matrix, this is equal to weight the cepstrum features,

$$f_w(X,S) = \sum_{j=1}^{k}\sum_{i=c_j}^{e_j}\sum_{q=1}^{d} w_q(x_i^q - \hat{m}_j^q)^2, \tag{7}$$

where w_q denotes the weight of q-th cepstrum feature. If Σ is not diagonal, we can apply eigen-decomposition on it : $\Sigma = U^T\Lambda U$, where U consists of the eigen vectors and Λ is a diagonal matrix whose diagonal components are the eigen values. Then, Eq. 6 can be written into the SSE function on transformed cepstrum features Ax:

$$f_{MD}(X,S) = \sum_{j=1}^{k}\sum_{i=c_j}^{e_j}||Ax_i - A\hat{m}_j||^2, \tag{8}$$

where the transformation matrix $A = \Lambda^{-1/2}U$. It is easy to examine that $A^TA = \Sigma^{-1}$. The formulation of Eq. 8 allows us to use the Agglomerative Segmentation (AS) algorithm [9] to optimize the objective function Eq. 6 .

In classical Mahalanobis distance, Σ is estimated as the covariance matrix of the total data

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)(x_i - m)^T, \tag{9}$$

where mean $m = \sum_{i=1}^{n} x_i/n$. However, this calculation only considers the statistical characteristics of the whole data. We are more interested in a distance metric which is small enough for cepstral features within the same phoneme while keeps large enough for cepstral features of different phonemes. In the following, we will study this problem in a learning framework. By limiting to Mahalanobis distance, the objective of learning parameters is to estimate covariance matrix Σ. Suppose there exists a set of training utterances D with labeled phoneme boundaries. We are going to develop two criteria which minimize the feature variance within the same phoneme and (or) maximize feature variance between different phonemes. Assume $|\Sigma| = 1$ to avoid scaling factors.

3.1 Criterion 1: Minimization of Summation Variance

The first criterion is to find matrix Σ, which minimizes the summation of variances within phonemes. Mathematically, this can be formulated as

$$\min_{\Sigma} MSV(D,\Sigma) = \min_{\Sigma}\sum_{X\in D}\left[\sum_{j=1}^{k}\sum_{i=c_j}^{e_j}(x_i - \hat{m}_j)^T\Sigma^{-1}(x_i - \hat{m}_j)\right], \tag{10}$$

where $\hat{m}_j$ is the mean of the j-th segment in utterance X. Define within-phoneme variance matrix of utterance set D

$$S_w = \sum_{X\in D}\sum_{j=1}^{k}\sum_{i=c_j}^{e_j}(x_i - \hat{m}_j)(x_i - \hat{m}_j)^T. \tag{11}$$

In the following, we deduce the optimal solution for Eq. 10. Remind $A^TA = \Sigma^{-1}$, Eq. 10 can be written into

$$\begin{aligned} MSV(D,\Sigma) &= \sum_{X\in D}\sum_{j=1}^{k}\sum_{i=c_j}^{e_j}\mathrm{Tr}(A(x_i - \hat{m}_j)(x_i - \hat{m}_j)^TA^T) \\ &= \mathrm{Tr}(AS_wA^T), \end{aligned} \tag{12}$$

where "Tr" denotes the trace of a matrix.

Since $|A^TA| = 1$, we have the Lagrangian function as follows,

$$L(A,\lambda) = \mathrm{Tr}(AS_wA^T) + \lambda(|A^TA| - 1). \tag{13}$$

Calculating the derivative of Eq. 13 to A, we have

$$\frac{\partial L(A,\lambda)}{\partial A} = \frac{\partial \mathrm{Tr}(AS_wA^T)}{\partial A} + \frac{\partial \lambda(|A^TA|-1)}{\partial A}$$
$$= 2AS_w + 2\lambda|A^TA|A^{-T} = 0. \quad (14)$$

Therefore,

$$S_w = -\lambda(A^TA)^{-1}. \quad (15)$$

Remind $|A^TA| = 1$ and $A^TA = \Sigma^{-1}$, the optimal covariance matrix can be calculated as

$$\Sigma_{MSV} = \frac{1}{|S_w|^{1/d}}S_w. \quad (16)$$

3.2 Criterion 2: Maximization of Discriminant Variance

In Eq. 10, we only consider the within phoneme variances. The second criterion tries to take account of the variance of two adjacent phoneme, that is, to maximize the between phoneme variances and to minimize the within phoneme variances. Formally,

$$\max_{\Sigma} \sum_{X\in D}\sum_{j=1}^{k-1}\sum_{i=c_j}^{e_{j+1}} (x_i - \hat{m}_{j,j+1})^T\Sigma^{-1}(x_i - \hat{m}_{j,j+1}) \quad (17)$$

$$\min_{\Sigma} \sum_{X\in D}\sum_{j=1}^{k}\sum_{i=c_j}^{e_j} (x_i - \hat{m}_j)^T\Sigma^{-1}(x_i - \hat{m}_j), \quad (18)$$

where $\hat{m}_{j,j+1}$ is the mean of the j-th and the $j+1$-th segment in X. It is noted that we only consider the between variances of two adjacent phonemes in Eq. 18. This is because, for phoneme segmentation, the same phoneme may appear more than one time in a single sequence, and for segmentation problem the difference of adjacent phonemes are most important.

Define between-phoneme variance matrix of utterance set D as

$$S_b = \sum_{X\in D}\sum_{j=1}^{k-1}\sum_{i=c_j}^{e_{j+1}} (x_i - \hat{m}_{j,j+1})(x_i - \hat{m}_{j,j+1})^T. \quad (19)$$

Using the same technique of Eq. 12, we can reduce Eq. 17, 18 to,

$$\max_{\Sigma} \mathrm{Tr}(AS_bA^T), \quad (20)$$

$$\min_{\Sigma} \mathrm{Tr}(AS_wA^T). \quad (21)$$

Eq. 17, 18 is a multi-objective problem. The usual approach to a multi-objective problem is to convert it to a single objective one.

Basically, there are two choices. One is based on subtraction of trace

$$\min_{\Sigma}\{\mathrm{Tr}(AS_wA^T) - \alpha\mathrm{Tr}(AS_bA^T)\} \quad (22)$$

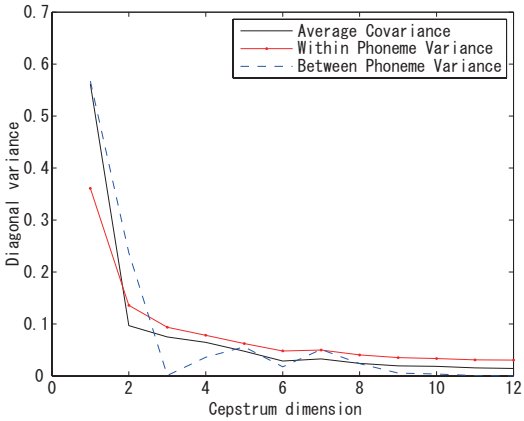

Figure 2 Variance of different dimensions .

where α is a coffecient; the other is based on ratio of trace,[1]

$$\max_{\Sigma} \frac{\mathrm{Tr}(AS_bA^T)}{\mathrm{Tr}(AS_wA^T)}. \quad (23)$$

Eq. 22 can lead to a close form solution

$$\Sigma_{MDV\_ST} = \frac{\mathrm{Tr}(AS_wA^T) - \alpha\mathrm{Tr}(AS_bA^T)}{|\mathrm{Tr}(AS_wA^T) - \alpha\mathrm{Tr}(AS_bA^T)|^{1/d}}. \quad (24)$$

However, there is no close form solution for Eq. 23. [6] showed an approximate answer for Eq. 23 as

$$\Sigma_{MDV\_RT} = \frac{1}{|S_b^{-1}S_wS_b^{-1}|^{1/d}}S_b^{-1}S_wS_b^{-1}. \quad (25)$$

It can be seen that the estimation of Σ depends on within-phoneme matrix S_w and between-phoneme matrix S_b. We calculated global covariance matrix Σ by Eq. 9, within-phoneme matrix S_w by Eq. 11 and between-phoneme matrix S_b by Eq. 19 of the utterances in the TIMIT database. We found that the main energy is located in the diagonal for all three matrices. Fig. 2 shows the diagonal components of them (the summation are normalized to one). It can be seen that generally the variance decreases as dimension index increases, however the curve of S_b shows a vibration pattern. The curve of S_w decreases slowly than that of Σ. Usually, the larger the variance is, the smaller the weight of its corresponding feature is.

3.3 Fully unsupervised approach

In Section 3.2 and 3.1, we assume there is a set of data with labeled boundary information for estimating the covariance matrix Σ and develop two criteria: Minimize Summation of Variance and Maximize Discriminant Variance. However, there are two limitations, 1) a set of labeled data must be available for learning the optimal matrix, and 2) once Σ is learned it is fixed and cannot adapt to the new data. In

[1] Someone may suggest to use trace ratio $\max_{\Sigma}\mathrm{Tr}(\frac{AS_bA^T}{AS_wA^T})$ as a criteria, which is widely adopted in linear discriminant analysis (LDA). However, it can be proved that trace ratio is invariant to Σ.

Algorithm 1 Iterative Segmentation Algorithm

1: **INPUT** A set of utterance $D = \{X\}$, the number of segments k_X for each utterance X and the maximum iteration number T.
2: **Initialize** Σ^0 as an unit matrix I and iteration index $t = 0$.
3: **while** Not Convergence and $t < T$ **do**
4: For each utterance X, calculate its segmentation S_X^t by minimizing Eq. 6 (Σ is set as Σ^t).
5: Calculate S_w^t based on segmentations S_X^t and Eq. 11.
6: Update Σ^t by using Eq. 16[2].
7: $t = t + 1$.
8: **end while**
9: **OUTPUT** segmentation S_X^t.

this Section, we are going to develop a fully unsupervised approach to deal with this problem, where Σ is learned from unlabeled utterances.

Given utterance data for segmentation, our problem is somehow similar to egg and chicken: if the segmentation is known, we can estimate an optimal Σ by MSV or MDV criteria; and if a better Σ is given, we can estimate a better segmentation by minimizing Eq. 6. Based on this observation, we introduce the following Iterative Segmentation Algorithm (ISA) which iteratively updates segmentation and Σ.

The Iterative Segmentation Algorithm is similar to the mechanism of infants speech acquisition. Psychological researches indicate that infants do not have acoustic models of the phonemes of their native languages, but they have the ability to discriminate sounds [8]. This discriminant ability resembles the distance metric we used for segmentation, which enable infants to preliminarily segment speech signals. Then the infants can adapt their sound discriminant ability based on the segmentation results. This procedure is considered to repeat during the infants build acoustic models of their native languages.

4. Experiments

We use the training part from the TIMIT American English acoustic-phonetic corpus [5] to evaluate and compare the proposed objective functions. The database includes 4,620 sentences from 462 American English speakers of both genders from 8 dialectal regions. It includes more than 170,000 boundaries, totally. The sampling frequency is 16kHz. For each sentence, we calculate the spectral features from speech signals by using 16ms Hamming windows with 1ms shift, and then transform spectral features into 12 mel-cepstrum coefficients.

[2]Although it is possible to estimate Σ by Eq. 25 or Eq. 24, we experimentally found this does not lead to good segmentation results in iterative segmentation algorithm. This is partly because that Eq. 25 and Eq. 24 are too sensitive to the segmentations.

Table 1 Recall rates using ED, MD and learning MD

Method	ED	MD	MSV	MDV-RT	MDV-ST
20ms	76.8%	73.6%	77.7%	77.6%	77.2%
30ms	86.7%	86.3%	88.2%	87.9%	88.1%
40ms	92.4%	92.9%	93.7%	93.5%	93.8%

For all the following experiments, the agglomerative segmentation (AS) algorithm [9] is used to find the optimal segmentation. The stop number of the AS algorithm is set as the number of phonemes in a sentence. For each method, we count how many ground truth boundaries are detected within a tolerance window (20∼40ms). The recall rate is adopted as a comparison criterion,

$$\text{Recall rate} = \frac{\text{number of boundaries detected correctly}}{\text{total number of ground truth boundaries}}.$$

4.1 Experiment 1: segmentation using Mahalanobis distance

In the first experiment, we make comparisons between Euclidean distance (ED), classical Mahalanobis distance (MD) (Eq. 9), and learning Mahalanobis distance with parameters Σ estimated by MSV (Eq. 16), MDV-ST (Eq. 24) and MDV-RT (Eq. 25) for segmentation. In classical MD, the covariance matrix is calculated for each utterance. Among all 4,620 utterances, we randomly select 56 sentences for learning the covariance matrix of MSV, MDV-ST and MDV-RT.

The results are summarized in Table 1. We can find that classical MD does not lead to better performance than Euclidean distance, while MD using learning parameters (MSV, MDV-RT and MDV-ST) can improve the recall rates compared to ED and classical MD. Among all these methods compared, MSV has the best results. But the results of MSV, MDV-RT, and MDV-ST are very near.

4.2 Experiment 2: segmentation with unsupervised learning MD

In the first experiment, the covariance matrix is learned from a set of utterances with labeled boundaries. In this Section, we make no use of the labeled utterances. Covariance matrix Σ are estimated by iterative segmentation algorithm descried in Section 3.3. The segmentation results are summarized in Table 2. It can be seen that we only need to execute iterative segmentation algorithm for a few iterations (2 or 3) to obtain good segmentation results. The increase of iteration number does not lead to significant improvements of recall rates. It can be seen that the unsupervised learning MD can achieve comparable results with supervised learning MD in Section 4.1.

4.3 Experiment 3: incorporation of power

In the above two experiments, we only made use of cepstral coefficients and did not consider power coefficient. In the next, we take account of power coefficient into the seg-

Table 2 Recall rates using unsupervised learning MD

Iteration t	0	1	2	3	10
20ms	76.8%	76.9%	77.4%	77.6%	77.9%
30ms	86.7%	87.8%	87.2%	87.8%	87.9%
40ms	92.4%	93.6%	92.7%	93.4%	93.3%

Table 3 Recall rates using Power

Method	MSV+P1	MSV+P2	MDV-RT+P1	MDV-RT+P2
20ms	79.0%	81.4%	80.2%	81.8%
30ms	89.3%	90.0%	89.4%	89.8%
40ms	94.4%	94.3%	94.2%	94.0%

mentation cost function. Let o_i denote a power coefficient at i-th frame. Basically, there are two methods to incorporate power. One is to argument cepstrum vector x_i into a new vector $\mathbf{x}_i = [x_i, o_i]$, then the same analysis for cepstrum vector can be applied. The other is to consider power and cepstrum independently. In this way, the objective function Eq. 6 bceomes:

$$f_{MD}(X,S) = \sum_{j=1}^{k} \sum_{i=c_j}^{e_j} \{(x_i - \hat{m}_j)^T \Sigma^{-1} (x_i - \hat{m}_j) + \beta (o_i - \hat{o}_j)^2\}, \tag{26}$$

where $\hat{o}_j$ is the average power of the j-th segment and β is a constant to take the balance between cepstrum and power.

We conducted experiments to compare the two different methods, where Σ is estimated by MSV (Eq. 16) and MDV-RT (Eq. 25) in Section 4. 1 due to their good performance. The results are shown in Table 3, where 'P1' denotes the first method to incorporate power and 'P2' denotes the second. We find that the using of power features can significantly improve the recall rates about 3-5 percents. The second method to incorporate power (treat power and cepstrum independently) usually achieves better results than the first method (use argument feature vector).

4. 4 Comparisons with other methods

We make comparisons of our results with other published results. Tolerance window size is set as 20ms, since it is most widely used. Our best recall rate is 81.8% shown in Table 3. In [3], with the same database, the authors showed a detected rate of 84.5%, and among them, as 89% are within 20ms. So their rate is 0.845× 0.89=75.2%. Moreover, here our insertion rate is 20.9%, which is lower than 28.2% shown by [3]. [4] used the testing part of TIMIT database, which includes less number of sentences (1,344) than we used. When their over-segmentation equals zero, the correct detection rate in their experiments corresponds to our recall rate. In this case, our result is better than theirs (76.0%) [4]. In [1], the authors use a subset of TIMIT database which contains 480 sentence and showed a recall rate 73.6%. We had obtained a recall rate 77.5% in a previous work [9], which is lower than 81.8% in this paper. Moreover, unlike the method in [9], we do not need to calculate the determinant of covariance matrix for each possible segmentation which is computationally expensive. Although our results are still lower than those of the HMM-based segmentation methods [2], our methods do not make use of prior knowledge such as linguistic contents and acoustic models.

5. Conclusions

This paper addresses the unsupervised phoneme segmentation problem by using Mahalanobis distance. We develop two optimization criteria, namely, minimization of summation of variance (MSV) and maximization of discriminant variance (MDV). We deduce the optimal solutions of MSV and MDV by using matrix calculation. We also propose an iterative segmentation algorithm (ISA) to learn covariance matrix of MD calculation without using labeled data. We carried out experiments on the TIMIT database. The results show that the use of learning MD can improve the segmentation performance. The MD learned with unlabeled data by ISA can achieve similar recall rates as the MD learned with labeled data. We also find that the segmentation results can be further improved by incorporating power coefficient.

References

[1] G. Aversano and et. al. A new text-independent method for phoneme segmentation. *IEEE Midwest Sym. on Cir. and Sys.*, pages 516–519, 2001.

[2] F. Brugnara and et. al. Automatic segmentation and labeling of speech based on Hidden Markov Models. *Speech Communication*, 12(4):357–370, 1993.

[3] S. Dusan and L. Rabiner. On the Relation between Maximum Spectral Transition Positions and Phone Boundaries. *INTERSPEECH*, pages 17–21, 2006.

[4] Y. P. Estevan, V. Wan, and O. Scharenborg. Finding Maximum Margin Segments in Speech. *ICASSP*, pages 937–940, 2007.

[5] J.S. Garofolo and et. al. Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database. *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, 1988.

[6] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Trans. PAMI*, 18(6):607–616, 1996.

[7] A.K. Jain and R.C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1988.

[8] P.K. Kuhl. Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, 5(11):831–843, 2004.

[9] Y. Qiao, N. Shimomura, and N. Minematsu. Unsupervised Optimal Phoneme Segmentation: Objectives, Algorithm and Comparisons. *ICASSP (accepted)*, 2008.

[10] J. Rissanen. A Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics*, 11(2):416–431, 1983.

[11] O. Scharenborg, M. Ernestus, and V. Wan. Segmentation of speech: Child's play? *In Proc. of Interspeech*, pages 1953–1957, 2007.

[12] Y. Tohkura. A weighted cepstral distance measure for speech recognition. *IEEE Trans. ASSP*, 35(10):1414–1422, 1987.