Unsupervised Phoneme Segmentation Using Transformed Cepstrum Features *

◎ Yu Qiao, Naoya Shimomura, and Nobuaki Minematsu (The Univ. of Tokyo)

1 Introduction

One of the basic problems in speech engineering is phoneme segmentation, that 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, and 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. Unlike written language, speech signals don't include explicit marks, such as space, for segmentation. Moreover, there usually does not exist abrupt changes in speech signals due to the temporal constraints of vocal tract motions. The difficulty of phoneme segmentation comes from 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 divided into two categories: supervised and unsupervised segmentation. In the first case, both the linguistic contents and the 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 of this category is HMM-based force alignment [2]. The second category of method tries to perform phonetic segmentation without using any prior knowledge on linguistic contents and acoustic models. The approach of this paper belongs to the 2nd class. The unsupervised segmentation is similar to the situation that infants acquire speech [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 into a probabilistic 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). Our experimental results 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 cepstrum feature space and it is not known whether or not Euclidean distance yields the best distance metric to estimate the goodness of the segments. In fact, it was shown that the weighted cepstral distance can achieve better performance than Euclidean distance for 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 based framework and develop two criteria for determining the optimal parameters. Their performances are compared through experiments on TIMIT database. The experimental results indicate that the learning Mahalanobis distance can help to improve the segmentation results.

2 Optimal Segmentation

In this section, we introduce the notations and give a brief review of our previous work on opti-

*本研究は日本学術振興会特別研究員奨励費 (19.07078) の支援を受けて行われた.

mal 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 is $|s_j|$ is $e_j - s_j + 1$.

For speech signals, 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,

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

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

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

Like most speech applications, we assume that r_j is a multi-variable 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 $\hat{r}_j(\hat{m}_j, \hat{\Sigma}_j)$ of m_j and Σ_j, Eq. 2 becomes,

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

It can be shown that the above Equation is in coordinate with the minimum description length principle (MDL) [10]. However, in practice, this approach may have 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 $|\hat{\Sigma}| = 0$. To deal with this difficulty, we consider to fix the covariance matrix Σ as an unit matrix I and only estimate mean $\hat{m}_j = 1/|s_j| \sum_{x \in s_j} x$. In this way, Eq. 2 becomes,

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

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

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

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 phoneme segmentation.

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

3 Weighted and Transformed Cepstral Features

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 are 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. We may consider other 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,

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

If Σ is a diagonal matrix, this is equal to putty weights on cepstrum features,

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

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. 5 can be written into the SSE function on transformed cepstrum features Ax:

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

where the transformation matrix $A = \Lambda^{-1/2} U$. It is easy to examine that $A^T A = \Sigma^{-1}$.

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

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

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 featurs within the same phoneme while keeps large enough for cepstral featurs of different phonemes. In the following, we will study this problem in a learning framework. By limiting to Mahalanobis distance, the objective of learning 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) maximizes feature variance between different phonemes. Assume $|\Sigma| = 1$ to avoid scaling factors.

3.1 Criterion 1: Minimization of Summation of Variance

The first objective is to minimize the summation of variances within phonemes. Mathematically, this can be formulated as

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

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

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

Using matrix calculation, the optimal Σ can be calculated as

$$\Sigma_{MSV} = \frac{1}{|S_w|^{1/d}} S_w. \quad (12)$$

3.2 Criterion 2: Maximization of Discriminant Variance

In Eq. 10, we only consider the within phoneme variances. we consider the variance of two adjacent phoneme into the second objective function, that is, to maximize the ratio of between phoneme variances S_b to within phoneme variances S_w. Formally, we have,

$$\max_{\Sigma} \sum_{X \in D} \sum_{j=1}^{k-1} \sum_{i=c_j}^{e_{j+1}} (x_i - \hat{m}_{j,j+1})^T \Sigma^{-1} (x_i - \hat{m}_{j,j+1})$$
$$\min_{\Sigma} \sum_{X \in D} \sum_{j=1}^{k} \sum_{i=c_j}^{e_j} (x_i - \hat{m}_j)^T \Sigma^{-1} (x_i - \hat{m}_j), \quad (13)$$

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 the numerator of Eq. 13). This is because, for phoneme segmentation, the same phoneme may appear more than one time in a single sequence.

Define between-phoneme variance matrix as

$$S_b = \sum_{X \in D} \sum_{j=1}^{k-1} \sum_{i=c_j}^{e_{j+1}} (x_i - \hat{m}_{j,j+1})(x_i - \hat{m}_{j,j+1})^T. \quad (14)$$

[6] showed an solution of Eq. 13 as

$$\Sigma_{MDV} = \frac{1}{|S_b^{-1} S_w S_b^{-1}|^{1/d}} S_b^{-1} S_w S_b^{-1}, \quad (15)$$

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. 14 of the utterances in TIMIT database. We found that the main energy is located in the diagonal for all three matrices. Fig. 1 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 small the weight of corresponding feature is.

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.

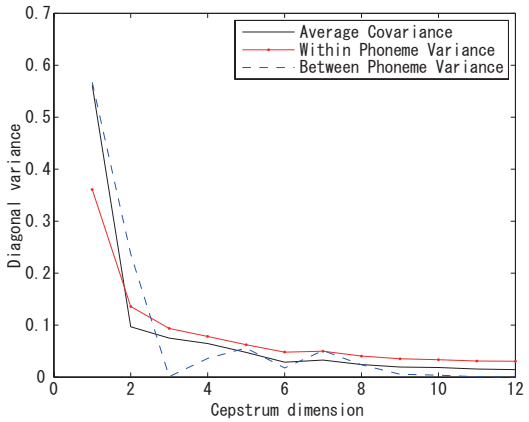

Fig. 1　Variance of different dimensions .

Table 1　Recall rates of sequence segmentation

Method	ED	MD	MSV	MDV
20ms	76.8%	73.6%	77.7%	77.6%
30ms	87.1%	86.3%	88.2%	87.9%
40ms	92.4%	92.9%	93.7%	93.5%

For each sentence, we calculate the spectral features from speech signals by 16ms Hamming windows with 1ms shift, and then transform spectral features into 12 mel-cepstrum coefficients (excluding the power coefficient).

We make comparisons between Euclidean distance (ED) and the classical Mahalanobis distance (MD) (Eq. 9) and the Mahalanobis distance using learning parameters MSV (Eq. 12) and MDV (Eq. 15) for segmentation. 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. Among all 4,620 sentences, we randomly select 56 sentences for learning the covariance matrix of MSV and MDV. And the other sentences are used for evaluation. 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,

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

The results are summarized in Table 1. We can find that classical Mahalanobis Distance does not lead to better performance than Euclidean distance, while Mahalanobis Distance using learning parameters achieves higher recognition rates than ED. Among all the methods, MSV has the best results.

5　Conclusions

This paper addresses the unsupervised segmentation problem by using learning Mahalanobis distance. We develop two optimization criteria, namely, MSV and MDV. MSV minimizes the summation of variance withing phonemes, and MDV tries to maximize the ratio of the variance between phonemes to the variance within phonemes. Both these criteria can lead to close form optimal solutions by using matrix calculation. We carried out experiments on the TIMIT database to compare the proposed methods. The results indicate that the use of learning Mahalanobis distance can improve the segmentation performance.

参考文献

[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 T-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 and et. al. Segmentation of speech: Child's play? *Interspeech*, pages 1953–1957, 2007.

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