interspeech 2008 incorporating SST 08

Metric Learning for Unsupervised Phoneme Segmentation

Yu Qiao and Nobuaki Minematsu

Graduate School of Engineering, The University of Tokyo, Tokyo, Japan

{qiao, mine}@gavo.t.u-tokyo.ac.jp

Abstract

Unsupervised phoneme segmentation aims at dividing a speech stream into phonemes without using any prior knowledge of linguistic contents and acoustic models. In [1], we formulated this problem into an optimization framework, and developed an objective function, summation of squared error (SSE) based on the Euclidean distance of cepstral features. However, it is unknown whether or not Euclidean distance yields the best metric to estimate the goodness of segmentations. In this paper, we study how to learn a good metric to improve the performance of segmentation. We propose two criteria for learning metric: Minimum of Summation Variance (MSV) and Maximum of Discrimination Variance (MDV). The experimental results on TIMIT database indicate that the use of learning metric can achieve better segmentation performances. The best recall rate of this paper is 81.8% (20ms windows), compared to 77.5% of [1]. We also introduce an iterative algorithm to learn metric without using labeled data, which achieves similar results as those with labeled data.

Index Terms: Unsupervised phoneme segmentation, optimization, Mahalanobis distance, metric learning

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. Both automatic Speech Recognition (ASR) and Text-to-Speech (TTS) systems need correct segmentation information for improving their performances. Human speech is a smoothly changing continuous signal due to the temporal constraints of vocal tract motions, which does not include explicit separation marks such as white spaces in written language. The difficulty of phoneme segmentation also 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. Perhaps the most famous approach in this category is HMM-based forced alignment [2]. The second category tries to perform phonetic segmentation without using 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 [3]. They don't have acoustic and linguistic models. However, psychological facts indicate that infants become able to segment speech according to acoustic difference between speech sounds and cluster speech segments into categories [4]. It is by this procedure that infants can gradually construct the their spoken language models.

Most of the previous methods dealt with this problem by detecting the change points in a speech stream. Aversano et. al [5] identified the boundaries as the peaks of jump function. Dusan and Rabiner [6] detected the "maximum spectral transition" positions as phoneme boundaries. Estevan et. al [7] employed maximum margin clustering to locate boundary points. In our earlier work [1], we formulated the segmentation problem as an optimization problem by using statistics and information theory analysis, and developed a simple objective function, the Summation of Square Error (SSE) based on Euclidean distance (ED). The experimental results [1] showed that minimizing SSE by Agglomerative Segmentation (AS) algorithm can achieve better results than previous methods [5, 6, 7]. However, Euclidean distance may not be the best metric to evaluate the goodness of segmentation. [8] found that weighted cepstral distance gave better performance than ED for DTW based speech recognition. Generally speaking, for a segmentation task, a good metric should be small between two feature vectors within the same phoneme, while preserve large between two feature vectors from different phonemes. In this paper, we study how to learn a metric to improve the performance of segmentation. We limit our analysis to the metric of Mahalanobis distance form for its simpleness and linearity. 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 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 at maximizing the variance between phonemes and minimizing the variance within phonemes at the same time. We propose an algorithm to estimate parameters without using labeled sequences. The proposed methods are evaluated through experiments on the TIMIT database. The experimental results indicate that the learning metric can improve the segmentation results. We also found that the results can be further improved by incorporating power coefficients.

2. Optimal segmentation

This section describes a brief review of our previous work on optimal segmentation [1]. 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}) rep-

Accepted after peer review of full paper

resent the j-th segment $x_{c_j}, x_{c_j+1}, ..., x_{e_j}$.

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,

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

Then 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 multivariable normal distributions whose mean and covariance matrix are denoted by m_j and Σ_j. If we further fix Σ_j as an unit matrix I and only estimate mean $\hat{m}_j = 1/|s_j| \sum_{x \in s_j} x$ [1]. (The use of other covariance matrices leads to Mahalanobis distance, which will be discussed in the next sections.) We can show Eq. 2 reduces to minimize the following *Summation of Squared Error* function (SSE) [1],

$$f_{SSE}(X,S) = \sum_{j=1}^{k} \sum_{i=c_j}^{e_j} ||x_i - \hat{m}_j||^2. \quad (3)$$

The above formula is the same as the objective function of k-means clustering (Chapter 3.5 [9]). The difference is that k-means needs not consider the time constraint, which is important for phoneme segmentation problem. In [1], we introduced the Agglomerative Segmentation (AS) algorithm to find optimal segmentations, which has a time complexity of $O(n)$.

3. Metric learning for segmentation

The SSE objective (Eq. 3) is based on simple Euclidean distance, where each dimension of cepstrum features is treated equally and the correlations 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 another covariance matrices. 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,

$$f_{MD}(X,S) = \sum_{j=1}^{k} \sum_{i=c_j}^{e_j} (x_i - \hat{m}_j)^T \Sigma^{-1} (x_i - \hat{m}_j). \quad (4)$$

If Σ is a diagonal, this is equal to weighted cepstrum features,

$$f_w(X,S) = \sum_{j=1}^{k} \sum_{i=c_j}^{e_j} \sum_{q=1}^{d} w_q (x_i^q - \hat{m}_j^q)^2, \quad (5)$$

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. 4 can be written into a SSE form with transformed features Ax:

$$f_{MD}(X,S) = \sum_{j=1}^{k} \sum_{i=c_j}^{e_j} ||Ax_i - A\hat{m}_j||^2, \quad (6)$$

where the transformation matrix $A = \Lambda^{-1/2} U$. It is easy to examine that $A^T A = \Sigma^{-1}$. The formulation of Eq. 6 allows us to use the Agglomerative Segmentation (AS) algorithm [1] to optimize the objective function Eq. 4 .

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

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x_i - m)(x_i - m)^T, \quad (7)$$

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. Here the question is how to estimate covariance matrix Σ. Suppose there exists a set of training utterances D with labeled phoneme boundaries. In the next, we will 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

$$\min_{\Sigma} MSV(D,\Sigma) = \min_{\Sigma} \sum_{X \in D} \left[\sum_{j=1}^{k} \sum_{i=c_j}^{e_j} (x_i - \hat{m}_j)^T \Sigma^{-1} (x_i - \hat{m}_j) \right], \quad (8)$$

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

$$S_w = \sum_{X \in D} \sum_{j=1}^{k} \sum_{i=c_j}^{e_j} (x_i - \hat{m}_j)(x_i - \hat{m}_j)^T. \quad (9)$$

In the following, we deduce the optimal solution for Eq. 8. Remind $A^T A = \Sigma^{-1}$, Eq. 8 can be written into

$$MSV(D,\Sigma) = \text{Tr}(AS_w A^T), \quad (10)$$

where "Tr" denotes the trace of a matrix.

Since $|A^T A| = 1$, we have the Lagrangian function of Eq. 8 as follows,

$$L(A,\lambda) = \text{Tr}(AS_w A^T) + \lambda(|A^T A| - 1). \quad (11)$$

Calculating the derivative of Eq. 11 to A, we have

$$\frac{\partial L(A,\lambda)}{\partial A} = \frac{\partial \text{Tr}(AS_w A^T)}{\partial A} + \frac{\partial \lambda(|A^T A| - 1)}{\partial A} = 2AS_w + 2\lambda |A^T A| A^{-T} = 0. \quad (12)$$

Since $A^T A = \Sigma^{-1}$, the optimal covariance matrix of Eq. 8 can be calculated by,

$$\Sigma_{MSV} = \frac{1}{|S_w|^{1/d}} S_w. \quad (13)$$

3.2. Criterion 2: maximization of discriminant variance

The 2nd criterion will simultaneously take account of the variance of within and between adjacent phonemes, that is to maximize the between phoneme variances and minimize the within phoneme variances. Formally,

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

$$\min_{\Sigma} \sum_{X\in D}\sum_{j=1}^{k}\sum_{i=c_j}^{e_j} (x_i-\hat{m}_j)^T\Sigma^{-1}(x_i-\hat{m}_j), \quad (15)$$

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. 15. 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 D as

$$S_b = \sum_{X\in D}\sum_{j=1}^{k-1}\sum_{i=c_j}^{e_{j+1}} (x_i-\hat{m}_{j,j+1})(x_i-\hat{m}_{j,j+1})^T. \quad (16)$$

Then, Eq. 14, 15 can be reduced to,

$$\max_{\Sigma} \mathrm{Tr}(AS_bA^T), \quad (17)$$

$$\min_{\Sigma} \mathrm{Tr}(AS_wA^T). \quad (18)$$

This is a multi-objective problem. We need to convert it to a single objective one. Basically, there are two choices. One is based on the subtraction of trace

$$\min_{\Sigma}\{\mathrm{Tr}(AS_wA^T) - \alpha\mathrm{Tr}(AS_bA^T)\} \quad (19)$$

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

$$\max_{\Sigma} \frac{\mathrm{Tr}(AS_bA^T)}{\mathrm{Tr}(AS_wA^T)}. \quad (20)$$

Eq. 19 can be optimized by using the same techniques of MSV,

$$\Sigma_{MDV-ST} = \frac{S_w - \alpha S_b}{|S_w - \alpha S_b|^{1/d}}. \quad (21)$$

However, there is no close form solution for Eq. 20. [10] showed an approximate answer for Eq. 20 as

$$\Sigma_{MDV-RT} = \frac{S_b^{-1}S_wS_b^{-1}}{|S_b^{-1}S_wS_b^{-1}|^{1/d}}. \quad (22)$$

3.3. Metric learning without labeled data

In Section 3.1 and 3.2, we assume there is a set of data with labeled boundary information for estimating Σ. 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 this Section, we will develop an Iterative Segmentation Algorithm (ISA) which doesn't need any labeled data. The ISA uses SSE to initiate segmentation S, and then iteratively update S and Σ. Details of ISA are given in Algorithm 1.

[1]Readers 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 optimal segmentation S_X^t. (Σ is set as Σ^t.)
5: Calculate S_w^t based on segmentations S_X^t.
6: Update Σ^t by using MSV.
7: $t = t + 1$.
8: **end while**
9: **OUTPUT** segmentation S_X^t.

The ISA is somewhat near 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 [4]. This discriminant ability resembles the 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 [11] 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. The agglomerative segmentation (AS) algorithm [1] 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) and calculate the recall rates for comparison. Due to the space limitation, the evaluation results on other criteria, such as F-measure, are omitted. However, it is noted that evaluation results based on F-measure show the same conclusion as that on recall rate.

4.1. Experiment 1: segmentation by metric learning

In 1st experiment, we make comparisons between Euclidean distance (ED), classical Mahalanobis distance (MD) (Eq. 7), and learning Mahalanobis distance with parameters Σ estimated by MSV (Eq. 13), MDV-ST (Eq. 21) and MDV-RT (Eq. 22) 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.

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%

Table 2: Recall rates using unlabeled data

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%

4.2. Experiment 2: metric learning from unlabeled data

In 2nd experiment, we use iterative segmentation algorithm descried in Section 3.3 to calculate matrix Σ and segmentation from unlabeled data. 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 also 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 segmentation cost function. Let o_i denote a power coefficient at i-th frame. Basically, there are two methods to incorporate power. One is to augment cepstrum vector x_i into a new vector $\mathbf{x}_i = [x_i, o_i]$. The other is to consider power and cepstrum independently,

$$f_p(X,S) = \sum_{j=1}^{k} \sum_{i=c_j}^{e_j} \{(x_i - \hat{m}_j)^T \Sigma^{-1} (x_i - \hat{m}_j) + \beta (o_i - \hat{o}_j)^2\}, \quad (23)$$

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. In our experiments , $\beta = 1$.

We conducted experiments to compare the two different methods, where Σ is estimated by MSV (Eq. 13) and MDV-RT (Eq. 22) due to their good performance. The results are shown in Table 3, where 'P1' denotes the augmented vector method and 'P2' denotes the method of Eq. 23. We find that the using of power features can 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 of argument feature vector).

4.4. Comparisons with other methods

We make comparisons 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 [6], with the same database, the authors showed a detected rate of 84.5%, and among them 89% are within 20ms. So their rate is 0.845$\times$ 0.89=75.2%. Moreover, our insertion rate is 20.9%, which is lower than 28.2% shown by [6]. [7] used the testing part of TIMIT database with less number of sentences (1,344) and showed a recall rate of 76.0%. In [5], the authors obtained an recall rate of 73.6% from a subset of TIMIT database (480 sentences). The best recall rate in our previous work [1] is 77.5%. Moreover, unlike the best method in [1], 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], we do not make use of linguistic contents and acoustic models in unsupervised segmentation.

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%

5. Conclusions

This paper investigates how metric learning can improve the performance of unsupervised phoneme segmentation. We develop two optimization criteria for metric learning, 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 the parameters of MSV from unlabeled data. The experimental results on the TIMIT database show that the use of learning metric can improve segmentation performance. The ISA with unlabeled data achieves similar recall rates as those with labeled data. We also find that the segmentation results can be further improved by incorporating power coefficient. Compared with our previous work [1], the recall rate is increased to 81.8% from 77.5%. Finally, it is noted that the proposed criteria MSV and MDV can have other applications more than segmentation.

6. References

[1] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised Optimal Phoneme Segmentation: Objectives, Algorithm and Comparisons," *Proc. ICASSP*, pp. 885–888, 2008.

[2] F. Brugnara and et. al, "Automatic segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication*, vol. 12, no. 4, pp. 357–370, 1993.

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

[4] P.K. Kuhl, "Early language acquisition: cracking the speech code," *Nature Rev. Neurosc.*, vol. 5, no. 11, pp. 831–843, 2004.

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

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

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

[8] Y. Tohkura, "A weighted cepstral distance measure for speech recognition," *IEEE ASSP*, vol. 35, no. 10, pp. 1414–1422, 1987.

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

[10] T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *IEEE PAMI*, vol. 18, no. 6, pp. 607–616, 1996.

[11] J.S. Garofolo and et. al, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *NIST, Gaithersburgh, MD*, 1988.