On Unsupervised Optimal Phoneme Segmentation

Yu Qiao
The Graduate School of Frontier Sciences, The university of Tokyo
http://www.gavo.t.u-tokyo.ac.jp/∼qiao/
qiao@gavo.t.u-tokyo.ac.jp

January 28, 2008

Abstract

Phoneme segmentation is a fundamental problem in a number of speech recognition and synthesis studies. Unsupervised phoneme segmentation assumes no knowledge on linguistic contents and acoustic models, and thus pose a challenge problem. The essential question behind this problem is *what is the optimal segmentation*. This paper formulates the optimal segmentation into a probabilistic framework. Using statistics and information theory analysis. we develop three optimal objective functions, namely, Mean Square Error (MSE), Log Determinant (LD) and Rate Distortion (RD). Specially, RD objective function is defined based on information Rate Distortion theory and can be related to human speech perception mechanisms. To opitimize these objective functions, we use time-constrained agglomerative clustering algorithm. We also propose an efficient method to implement the algorithm by using integration functions. We execute experiments on TIMIT database to compare the above three objective functions. The results show that Rate Distortion achieves the best performance and indicate our method outperform the recent published unsupervised segmentation methods [1, 2, 3].

1 Introduction

Many speech analysis and synthesis applications depend on segmentation to divide the speech signals into phonetic segments (phonemes and syllables) [4]. Unlike written language, speech signals do not include explicit space for segmentation. Moreover, human speech are smoothly continuous signals and do not change abruptly due to the constraints of vocal tract. All these facts make segmentation a difficult question.

In engineering speech study, Automatic Speech Recognition (ASR) models always require reliable phoneme segmentation in the initial training phases, and Text-to-Speech (TTS) systems need large speech database with phoneme

segmentation information for improving the performances. Although manual segmentation can be precise, it is heavily time and energy costly [5, 6]. Partly for this reason, phoneme segmentation has received continuous research interests. The approaches to phoneme segmentation can be divided into two classes. The first class requires the linguistic contents and the acoustic models of phonemes. The segmentation is usually converted to the alignment of speech signals with given texts. A comparison of these methods can be referred to [7]. Perhaps the most famous method of this class is the HMM-based forced alignment [8, 5].

Another class of methods tries to perform phonetic segmentation without using any prior knowledge on linguistic contents and acoustic models, which is also known as unsupervised segmentation. The approach of this paper belongs to the second class. The unsupervised segmentation is similar to the phenomenon that an infant perceives speeches [9]. Most of the previous approaches to this problem focus on detecting on the change points of speech signals and take these change points as the boundaries of phonemes. Aversano et. al [1] defined "jump function" to capture the changes in speech signals and identified the boundaries as the peaks of jump function. Dusan and Rabiner [2] detected the "maximum spectral transition" positions as phoneme boundaries . Estevan et. al [3] employed maximum margin clustering to locate boundary points.

Different from these change point detection methods, this paper try to solve phoneme segmentation problem by answering an essential question behind: *what kind of segmentation is optimal*. In other words, we want to find optimal objective functions to evaluate the goodness of segmentations. This is a hard problems as we have neither information on the categories of the phonemes nor prior knowledge on phonemes' acoustic models. It is well known that the signals within a phoneme share common characteristics. This fact also inspires us to use certain measures of the inner variance (or coherence) to evaluate the goodness of being phoneme. Formally, we will formulate the segmentation problem in a probabilistic framework. With the help of statistics and information theory, we develop three objective functions, namely, 1) Mean Square Error (MSE), 2) Log Determinant (LD) and 3) Rate Distortion (RD). Specially, RD objective function is defined based on information rate distortion theory and is related to human speech perception mechanism. We notice that an audio segmentation method, Bayesian Information Criteria (BIC) [10], can be seen as special case of our methods. To optimize the proposed objective functions, we use time constrained agglomerative clustering algorithm due to its simplicity and effectiveness. We develop an efficient implementation based on the integration functions, which can largely reduce the computational time. The proposed three measures are compared through experiments on TIMIT database. Rate Distortion achieves the highest recall rate among the three objective functions. Our rates are also better than the recent published results on unsupervised phoneme segmentation [1, 2, 3].

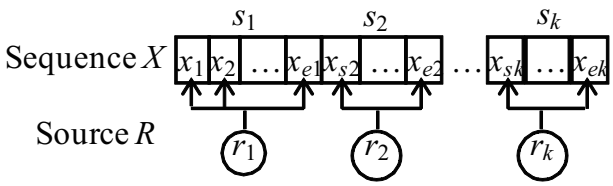

Figure 1: Diagram of Segmentation Model.

2 Formulation of Optimal Segmentation

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. 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 j-th segment.). Let $X_{c_j:e_j}$ (or X_{s_j}) represent the j-th segment $x_{c_j}, x_{c_j+1}, ..., x_{e_j}$ (Fig. 1). Size of segment $|s_j| = e_j - s_j + 1$. Without any constraint, there will be ${}_{n-1}C_{k-1}$ possible cases of segmentation.

For speech signals, it is natural to make the assumption that each individual phoneme is generated by an independent source. Let r_j denote the source for observed sequence s_j and $R = r_1, r_2, ..., r_k$ denote a source sequence. $p(x_i|r_j)$ represents the probability model of observing x_i given source r_j (Fig. 1). Thus we have,

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

In the next sections, we will deduce three optimal objective functions for unsupervised phoneme segmentation.

2.1 Mean square error and log determinant

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

$$\hat{S} = \min_S \{-\log(p(X|S, R))\} = \min_S \{\sum_{j=1}^{k} \sum_{i=c_j}^{e_j} -\log(p(x_i|r_j))\}. \tag{2}$$

If the source sequence R is known, it is not hard to see that the above problem can be solved by Viterbi decoding or dynamic programming [6]. However, in unsupervised segmentation, we have no knowledge on R and the problem becomes much more difficult. To handle this difficulty, we need to make assumptions on the source distributions r_j and to estimate the parameters of r_j.

Like most speech applications [4], we assume that r_j is a multi-variable normal distributions whose mean and covariance matrix are denoted by m_j and Σ_j. Thus,

$$p(x|r_j) = \frac{1}{(2\pi)^{d/2}\det(\Sigma_j)^{1/2}} \exp\{-\frac{1}{2}(x-m_j)^T\Sigma_j^{-1}(x-m_j)\}. \quad (3)$$

If segmentation s_j is known, the parameters m_j and Σ_j can be estimated by,

$$\hat{m}_j = \frac{1}{|s_j|}\sum_{i=c_j}^{e_j} x_i, \quad (4)$$

$$\hat{\Sigma}_j = \frac{1}{|s_j|}\sum_{i=c_j}^{e_j}(x_i-\hat{m}_j)(x_i-\hat{m}_j)^T. \quad (5)$$

Using $\hat{r}_j(\hat{m}_j, \hat{\Sigma}_j)$, Eq. 2 reduces to,

$$\begin{aligned} -\log p(X|S,\hat{R}) &= \sum_{j=1}^{k}\sum_{i=c_j}^{e_j} -\log(p(x_i|r_j)) \\ &= \sum_{j=1}^{k}\sum_{i=c_j}^{e_j} \frac{d}{2}\log(2\pi) + \frac{1}{2}(\log\det(\hat{\Sigma}_j) + (x_i-\hat{m}_j)^T\hat{\Sigma}_j^{-1}(x_i-\hat{m}_j) \\ &= \frac{nd}{2}\log(2\pi) + \frac{1}{2}\sum_{j=1}^{k}|s_j|\log\det(\hat{\Sigma}_j) + \frac{nd}{2}. \end{aligned} \quad (6)$$

From the perspective of information theory, the differential entropy (Chapter 9, [11]) of normal distribution $\hat{r}_j(\hat{m}_j, \hat{\Sigma}_j)$ is $\log_2((2\pi e)^d\det(\hat{\Sigma}_j))/2$, where d is the dimensionality of $\hat{m}_j$. Remind that the entropy denotes the expectation bits to describe a random variable. Thus MLE estimation by Eq. 6 will lead to minimize the description length of the sequence. This is in concordance with the minimum description length principle (MDL) [12].[1] Because the first and the third term of Eq. 6 do not depend on S, to maximize the likelihood of Eq. 2 equals to minimize the following *Log Determinant* (LD) function,

$$LD(X,S) = \sum_{j=1}^{k}|s_j|\log\det(\hat{\Sigma}_j). \quad (7)$$

[1]Connection to largest structure: For structure representation, our objective is to find the "largest" structure where the inter distances between segments (evnets) are maximized. We can take the inter distance as mutual information. Remind the total information $I(X)$ of the sequence is fixed. The summation of the information of segmentations is $I(S) = \sum_j I(s_j)$. The mutual information of these segmentations can be approximated by $I(X|S) = I(X) - I(S)$. Therefore, the MLE estimation of Eq. 2 can also approximately find the "largest" structure.

If we fix the covariance matrix Σ as an unit matrix I and only estimate mean $\hat{m}_j = 1/|s_j| \sum_{x \in s_j} x$, Eq. 2 becomes to,

$$
\begin{aligned}
-\log p(X|S, \hat{R}) &= \sum_{j=1}^{k} \sum_{i=c_j}^{e_j} \frac{d}{2} \log(2\pi) + \frac{1}{2}(x_i - \hat{m}_j)^T (x_i - \hat{m}_j) \\
&= \frac{nd}{2} \log(2\pi) + \frac{1}{2} \sum_{j=1}^{k} \sum_{i=c_j}^{e_j} ||x_i - \hat{m}_j||^2. \qquad (8)
\end{aligned}
$$

Note only the second item is influenced by segmentation S. Thus the problem equals to minimize the following *mean square error* function (MSE),

$$
MSE(X, S) = \sum_{j=1}^{k} \sum_{i=c_j}^{e_j} ||x_i - \hat{m}_j||^2. \qquad (9)
$$

The above formula is the same as the objective function of k-means clustering (Chapter 3.5 [13]). The difference between our problem and k-means is that k-means need not consider the time constraint, which is important for phoneme segmentation.

2.2 Rate Distortion

Let us consider the mechanism of human perceiving speeches. It has been shown that the ear's perceptual mechanism places a limit on the smallest spectral differences (Chapter 5. [14]). Human don't care the small difference in speech signals, that is why two linguistically identical utterances with small different noise can be perceived as the same. This fact cannot be represented well by using mean square error (Eq. 9) or log determinant (Eq. 7). For speech segmentation, we need not focus on the details of speech signals too much. In the next, we are going to define *Rate Distortion* based on information theory (Chapter 13. [11]), which is coinciding with human perception mechanism.

R-D theory was created by Shannon in his foundational paper on information theory. It has been shown that R-D theory is related to human perception mechanism. In fact, many popular audio and video compression standards such as MP3, JPEG and MPEG make use of R-D techniques [15]. For x under Gaussian distribution $\hat{r}_j(\hat{m}_j, \hat{\Sigma}_j)$, we introduce another random variable y and allowable distance bound ϵ such that $E(x - y)^2 \leq \epsilon$. The objective of R-D is to code y with the fewest number of bits possible. Note here we don't take interest in a R-D coding algorithm, but the coding length of a segment. We can model x and y with an additive Gaussian noise model: $y = x + z$,where noise $z \sim N(0, \epsilon I)$ [11]. Then

$$
E(y - \bar{y})^2 = E(x - \bar{x})^2 + 2E(x - \bar{x})Ez + Ez^2 = \epsilon I + \hat{\Sigma}_j, \qquad (10)
$$

where $\bar{y}$ and $\bar{x}$ are the expected value of y and x, respectively. Thus the entropy of y is bounded by $\log(2\pi e)^d \det(\Sigma I + \hat{\Sigma}_j)/2$. R-D theory defines a rate distortion

function $R(\epsilon) = \min_{E(x-y)^2 \leq \epsilon} I(x;y)$ to represent the infimum of rates such that bound ϵ can be achieved.We have,

$$\begin{aligned} I(x;y) &= h(y) - h(z) \\ &\leq \frac{1}{2}\log(2\pi e)^d \det(\epsilon I + \hat{\Sigma}_j) - \frac{1}{2}\log(2\pi e)^d \det(\epsilon I) \\ &= \frac{1}{2}\log\det(I + \hat{\Sigma}_j/\epsilon) \end{aligned} \tag{11}$$

The last line yields a upper bound for rate distortion functions. [2] We use Eq. 11 to define the following rate distortion (RD) function of X under segmentation S,

$$RD(X,S) = \sum_{j=1}^{k} |s_j| \log\det(I + \hat{\Sigma}_j/\epsilon). \tag{12}$$

We also noticed that a similar measure had been successfully used for image segmentation in vision field recently [16]. But different from their methods, we don't use the coding lengths for segmentation and for mean vector.

2.3 Invariance to orthogonal transformation

In this Section, we will prove that the segmentation by optimizing log determinant of Eq. 7, and rate distortion of Eq. 12 is invariant to orthogonal transformation.

Theorem 1 *Consider two sequences $X = x_1, x_2, ..., x_n$ and $X' = x'_1, x'_2, ..., x'_n$ where $x'_i = Ax_i + b$ (A denotes a full-rank $d \times d$ transformation matrix and b represents a translation vector). By minimizing the LD and RD objective functions defined by Eq. 7 and Eq. 12, X and X' will have the same segmentations.*

Proof 1) At first, we prove the theorem for Eq. 7. Let u_j and Σ_j denote the mean and covariance of X_{s_j} and u'_j and Σ'_j denote the mean and covariance of X'_{s_j}. It is easy to examine that

$$\begin{aligned} u'_j &= Au_j + b, \\ \Sigma'_j &= A\Sigma_j A^T. \end{aligned} \tag{13}$$

Under segmentation S, we have

$$\begin{aligned} LD(X', S) &= \sum_{j=1}^{k} |s_j| \log\det(\hat{\Sigma}'_j) \\ &= 2d\log\det(A) + \sum_{j=1}^{k} |s_j| \log\det(\hat{\Sigma}_j) \\ &= 2d\log\det(A) + LD(X,S). \end{aligned} \tag{14}$$

[2] The upper bound by Eq. 11 still holds when x is not gaussian. Roughly speaking, this is because gaussian variables are mostly difficult to code.

In the above equation the first term $2d \log\det(A)$ is a constant which does not depend on S. Therefore, X and X' will have the same optimal MLE segmentation (Eq. 7).

2) In the next, we prove the theorem for Eq. 12. Apply eigen-decomposition on covariance matrix $\hat{\Sigma}_j = U^T D U$, where U is the matrix of eigenvectors and D is a diagonal matrix composed by the eigen values $\lambda_1, \lambda_2, ..., \lambda_m$. Then,

$$\log\det(I + \hat{\Sigma}_j/\epsilon) = \sum_{k=1}^{d} \log(1 + \lambda_k/\epsilon). \tag{15}$$

It is easy to see that the above RD objective function only depends on the eigen values of the covariance matrixes. Also according to Eq. 13, the orthogonal transformation will not change the eigen values. Thus

$$\begin{aligned} RD(X', S) &= \sum_{j=1}^{k} |s_j| \log\det(I + \hat{\Sigma}'_j/\epsilon) \\ &= \sum_{j=1}^{k} |s_j| \log\det(I + \hat{\Sigma}_j/\epsilon) \\ &= RD(X, S). \end{aligned} \tag{16}$$

Therefore, X and X' will have the same optimal segmentation by Eq. 12.

Theorem 1 has significant meaning for our structure study [17, 18]. The structure representation need to divide input sequences to several events (segments). It is hoped that the input sequences under different transformations can be divided into the same way. With Theorem 1, we can achieve invariant segmentation for orthogonal transformation by optimizing Eq. 7 or Eq. 12.

We can also show that the RD objective based segmentation is invariant to transformation $Y_j = X_{s_j:e_j} C$, where $X_{s_j:e_j}$=$[x_{s_j}, x_{s_j+1}, ..., x_{e_j}]$ and C is a $n \times n$ full rank matrix. We can calculate the covariance matrix of Y_j as,

$$\hat{\Phi}_j = \frac{1}{|s_j|} \sum_{i=c_j}^{e_j} C^T (x_i - \hat{u}_j)^T (x_i - \hat{u}_j) C. \tag{17}$$

It is not hard to examine that $\hat{\Phi}_j$ and $\hat{\Sigma}_j$ share the same eigen values. So they have the same rate distortion.

2.4 Relation to BIC

In this section, we will show that Bayesian Information Criterion (BIC) [10], which has been developed for audio segmentation task, can be seen as the special cases of our variance functions. Audio segmentation aims at dividing audio stream into homogenous segments such as speech, music and laughter etc. The

BIC detects the boundary between two segments $S = S_1, S_2$ by maximizing the following equation [10]:

$$\max_{s_1,s_2}\{\frac{|s_1|+|s_2|}{2}\log\det\Sigma - \frac{|s_1|}{2}\log\det\Sigma_1 - \frac{|s_2|}{2}\log\det\Sigma_2 - \lambda P\}, \quad (18)$$

where Σ, Σ_1 and Σ_2 are the covariance matrices for segments $s_1 \cup s_2$, s_1 and s_2 respectively, and $P = (d + d(d+1)/2)\log(|s_1| + |s_2|)/2$. If we consider to divide sequence X into two segments, the first term and the last term of Eq. 18 will be const. And the problem reduces to that of Log Determinant based segmentation (Eq. 7).

3 Optimization Algorithm

In Section 2, we have developed three objective functions for segmentation: Mean Square Error (Eq. 9), Log Determinant (Eq. 7) and Rate Distortion (Eq. 12). The next problem is how to minimize these objective functions. It is not hard to see that all the three functions can be written into the following form:

$$\min_{\{s_1,s_2,\ldots,s_k\}}\sum_{j=1}^{k} f(X, s_j), \quad (19)$$

where $f(X, s_j)$ can be seen as a function to represent the inner variance (or coherence) of segmentation X_{s_j}.

Perhaps the quickest idea to optimize Eq. 19 for a sequence is to use dynamic programming (DP). However, the direct use of DP needs time cost $O(n^2k)$, where n is the length of sequence and k is the number of segments. This makes it impractical for our problem, as an utterance of sentence may contain several thousands of frames. In this paper, we use an agglomerative clustering algorithm (Chapter 3.2 [13]) to optimize Eq. 19. The algorithm works in a bottom-up manner. It begins with each frame as a segment and merge some frames into larger segments successively in a greedy way. The algorithm can be solved in time $O(n)$. Details are as follows.

Algorithm 1 Agglomerative Segmentation (AS) Algorithm

1: **INPUT** sequence $X = (x_1, x_2, ..., x_n)$ and the number of segments k.
2: **Initialize** segmentations as $S = \{s_j = j\}_{j=1}^{n}$, $t = n$.
3: **while** $t > k$ **do**
4: find index j', which minimizes the following equation

$$f(X, s_j \cup s_{j+1}) - f(X, s_j) - f(X, s_{j+1}); \tag{20}$$

5: merge $s_{j'}$ and $s_{j'+1}$ into a single segment;
6: $t = t - 1$.
7: **end while**
8: **OUTPUT** segmentation S.

3.1 Fast implementation

The most time-costly computation in the AS algorithm is to calculate the variance (when using Eq. 9) or covariance matrix (when using Eq. 7 and Eq. 12) for a segment. This computation must repeat many times until the algorithm terminates. In fact, we need not directly use the summation form of Eq. 4, Eq. 9 and Eq. 5 to calculate mean, variance and covariance every time. There is a more efficient way. We can calculate the following integration functions firstly:

$$G_1(i) = \sum_{k=2}^{i} x_{k-1} \qquad (G_1(1) = 0), \tag{21}$$

$$G_2(i) = \sum_{k=2}^{i} x_{k-1} x_{k-1}^T \qquad (G_2(1) = 0), \tag{22}$$

where $i = 1, 2, ..., n+1$. Note $G_1(i)$ is a vector and $G_3(i)$ is a matrix. Then the mean m_j, variance V_j and covariance matrix Σ_j of segment X_{s_j} ($s_j = (c_j, ..., e_j)$) can be calculated by:

$$m_j = \frac{1}{e_j - c_j + 1}(G_1(e_j + 1) - G_1(c_j)), \tag{23}$$

$$\Sigma_j = \frac{1}{e_j - c_j + 1}(G_2(e_j + 1) - G_2(c_j)) - m_j m_j^T, \tag{24}$$

$$V_j = \mathrm{Diag}(\Sigma_j), \tag{25}$$

where 'Diag' denotes the diagonal of a matrix. In this implementation, the integration functions only need to be calculated once at the beginning. After that, mean, variance and covariance can be estimated without summation operations.

3.2 Fast implementation

The most time-costly computation in the AS algorithm is to calculate the variance (when using Eq. 9) or covariance matrix (when using Eq. 7 and Eq. 12)

for a segment. This computation must repeat many times until the algorithm terminates. In fact, we need not directly use the summation form of Eq. 4, Eq. 9 and Eq. 5 to calculate mean, variance and covariance every time. There is a more efficient way. We can calculate the following integration functions firstly:

$$G_1(i) = \sum_{k=2}^{i} x_{k-1} \qquad (G_1(1) = 0), \tag{26}$$

$$G_2(i) = \sum_{k=2}^{i} x_{k-1} x_{k-1}^T \qquad (G_2(1) = 0), \tag{27}$$

where $i = 1, 2, ..., n+1$. Note $G_1(i)$ is a vector and $G_2(i)$ is a matrix. Then the mean m_j, variance V_j and covariance matrix Σ_j of segment X_{s_j} ($s_j = (c_j, ..., e_j)$) can be calculated by:

$$m_j = \frac{1}{e_j - c_j + 1}(G_1(e_j + 1) - G_1(c_j)), \tag{28}$$

$$\Sigma_j = \frac{1}{e_j - c_j + 1}(G_2(e_j + 1) - G_2(c_j)) - m_j m_j^T, \tag{29}$$

$$V_j = \mathrm{Diag}(\Sigma_j), \tag{30}$$

where 'Diag' denotes the diagonal of a matrix. In this implementation, the integration functions only need to be calculated once at the beginning. After that, mean, variance and covariance can be estimated without summation operations.

4 Experiments

We use the training part from the TIMIT American English acoustic-phonetic corpus [19] 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 16ms Hamming windows with 1ms shift, and then transform spectral features into 12 mel-cepstrum coefficients (excluding the power coefficient). We design the following two experiments to evaluate and compare the three types of objective functions. Comparisons with other methods are also given at last.

4.1 Experiment 1: segmentation of biphone subsequences

In the first experiment, we extracted all the biphone subsequences by referring to the label information of TIMIT database. The segmentation in biphone subsequence is relatively a simple problem. We can easily find the global optimal boundary and calculate the shift error between the detected boundary and the ground truth boundary, which are both difficult in total sequence segmentation tasks.

We did experiments to compare the performances of the following functions: 1)mean square error (MSE), 2) log determinant estimated by diagonal covariance matrix (LD-DIA), 3)log determinant estimated by full covariance matrix (LD), 4) rate distortion estimated by diagonal covariance matrix (RD-DIA), 5)rate distortion estimated by full covariance matrix (RD). To avoid the singular problem of covariance matrix, the minimum length of a segment is set as 18ms. The R-D distance bound ϵ (Eq. 12) is set as 0.05. The Absolute Shift Error (ASE) between the detected boundary and the ground truth are calculated for each subsequence. The average ASEs of the five methods are shown in Table. 1. We can find that RD has the least ASE among all the compared objectives.

4.2 Experiment 2: segmentation of sentences

In the second experiment, we examine the proposed objective functions on the sequence segmentation tasks. The agglomerative segmentation (AS) algorithm introduced in Section 3 is used. We set the stop number k of the AS algorithm as the number of phonemes in the sentence. The AS algorithm starts with one frame in each segmentation. When the number of frames of a segmentation is less than 12, the covariance matrix of the segmentation will be singular and its determinant will be zero. This fact prohibits us to use LD. So we execute experiments on the other four methods: MSE, LD-DIA, RD, and RD-DIA. 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 2. We can find that rate distortion based measures (RD and RD-DIA) always outperform other measures (MSE and LD-DIA). When the window size is small (20ms), the performance of MSE and RD (RD-DIA) is very near. However, the differences between MSE and RD (RD-DIA) increase when the tolerance windows enlarge. We think the reason mostly comes from the AS-algorithm. The reliable calculation of covariance matrix for RD (RD-DIA) requires enough number of frames in a segment. However, this requirement cannot be satisfied at the beginning phase of the AS algorithm, when the segments are small. Moreover, when using RD, the AS algorithm with RD or MSE prefers to merge shorter segments as this will usually lead to the smaller value of Eq. 20. To verify this prediction, we did another experiment where we use a simple Average Mean Square Error (AMSE) function $f_m(X, s)$ for pre-segmentation. $f_m(X, s) = \sum_{j \in s}(x_j - \bar{x})^2/|s|$, where mean $\bar{x} = \sum_{j \in s} x_j/|s|$. It

Table 1: Comparison of the average absolute shift errors

Method	MSE	LD	LD-DIA	RD	RD-DIA
Error(ms)	16.6	18.8	17.8	**15.1**	16.0

Table 2: Recall rates of sequence segmentation

Method	MSE	LD-DIA	RD	RD-DIA
20ms	**76.7%**	70.4%	76.1%	**76.7%**
30ms	86.7%	83.5%	**88.5%**	87.8%
40ms	92.4%	90.6%	**94.7%**	93.6%

Table 3: Recall rates with pre-segmentation

Method	MSE	RD	RD-DIA	AMSE
20ms	77.1%	77.1%	**77.5%**	72.5%
30ms	86.8%	**89.0%**	88.1%	80.5%
40ms	92.3%	**94.9%**	93.7%	85.3%

should be noted that AMSE has a poor performance if we use it thoroughly (Last column, Table 3). Here we just used it to do pre-segmentation until the number of segments reaches five times of the number of phonemes in a sentence. The pre-segmentation is done in the same way for all the compared methods (MSE, RD and RD-DIA). The results are shown in Table 3. We can find that the recall rates can be improved with such a simple pre-segmentation. It is should be noted that this is just a rough test. One may improve the results by using better cost functions and schemas for pre-segmentation.

4.3 Comparisons with other methods

It is not easy to directly compare our method with other unsupervised segmentation methods, since many authors use different data sets and testing protocols. Here, we assume that tolerance window size is 20ms, since we found that it is most widely used. In [2], 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%, which is lower than ours 77.5%. Moreover, our insertion rate is 20.9%, which is lower than 28.2% shown by [2]. [3] 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% [3]. In [1], the authors use a subset of TIMIT database which contains 480 sentence and showed a recall rate 73.6%. In [20], the authors also run experiments on TIMIT database and listed the detection rates for different types of boundary. The average detection rate is 73.8%. Although our recall rates are still lower than the HMM-based segmentation methods [8, 5], our methods don't make use of prior knowledge such as linguistic contents or acoustic models and don't need a training procedure.

5 Conclusions

This paper proposes a class of optimal segmentation methods for unsupervised phoneme boundary detection. We formulate the segmentation problem in a probabilistic framework, and develop three objective functions for segmentation based on statistic and information theory analysis: Mean Square Error (MSE), Log Determinant (LD) and Rate Distortion (RD). Especially, RD function is deduced from Rate Distortion theory and can be related to human audio perception mechanism. We introduce an agglomerative segmentation algorithm to find the optimal segmentation and show how to implement the algorithm in an efficient way. Extensive experiments are executed to compare the three objective functions. The results show that RD function outperform other two objective functions. The theories and methods proposed in this paper not only apply to the phoneme segmentation methods but also may have applications in other sequence segmentation problems.

Finally, it should be noted that it is not our objective to develop a high recall rate segmentation methods in this paper. The objective is to formulate the phoneme segmentation in an optimal fashion by trying to answer what is the optimal segmentation. We have found through limited experiments that the recall rates can be improved by using better optimization algorithms and incorporating other features. The exploration on this will be our further works.

We are going to apply the proposed methods on the event detection problems in our structure study [17, 18]. The proposed methods have some preferable characteristics for structure study. 1) All the three objective functions aim at minimizing the inner variance of segments (events). Roughly speaking, they also maximizes the intra-difference of different events to make the structure large. 2) Theorem 1 shows that segmentations of optimizing Eq. 7 or Eq. 12 are invariant to orthogonal transformations.

6 Acknowledgements

I would like to thank Naoya Shimomura for his preparation of the experimental data.

References

[1] G. Aversano, A. Esposito, and M. Marinaro, "A new text-independent method for phoneme segmentation," *IEEE Midwest Symposium on Circuits and Systems*, pp. 516–519, 2001.

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

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

[4] S. Furui, *Digital Speech Processing, Synthesis and Recognition*, Marcel Dekker, 2001.

[5] DT Toledano, LAH Gomez, and LV Grande, "Automatic phonetic segmentation," *IEEE Trans. on SAP*, vol. 11, no. 6, pp. 617–625, 2003.

[6] T. Svendsen and F. Soong, "On the automatic segmentation of speech signals," *ICASSP*, pp. 77–80, 1987.

[7] J. Adell, A. Bonafonte, JA Gomez, and MJ Castro, "Comparative study of Automatic Phone Segmentation methods for TTS," *ICASSP*.

[8] 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.

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

[10] S. Chen and PS Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127–132, 1998.

[11] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley-Interscience New York, 2006.

[12] J. Rissanen, "A Universal Prior for Integers and Estimation by Minimum Description Length," *The Annals of Statistics*, vol. 11, no. 2, pp. 416–431, 1983.

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

[14] J.V. Tobias, *Foundations of modern auditory theory*, Academic Press, 1970.

[15] A. Ortego and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 23–50, 1998.

[16] Y. Ma, H. Derksen, W. Hong, and J. Wright, "Segmentation of Multivariate Mixed Data via Lossy Coding and Compression," *IEEE Trans. on PAMI*, vol. 29, no. 9, pp. 1546–1562, 2007.

[17] N. Minematsu, "Yet another acoustic representation of speech sounds," *Proc. ICASSP*, pp. 585–588, 2004.

[18] N. Minematsu, "Mathematical Evidence of the Acoustic Universal Structure in Speech," *Proc. ICASSP*, pp. 889–892, 2005.

[19] 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.

[20] P. Micallef and T. Chilton, "Automatic Identification of Phoneme Boundaries Using a Mixed Parameter Model," *Fifth European Conference on Speech Communication and Technology*, 1997.