社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

教師なし音素セグメンテーションの最適化に関する理論的・実験的考察

喬　　宇† 　下村　直也† 　峯松　信明†

† 東京大学大学院新領域創成科学研究科 〒 277–8561 千葉県柏市柏の葉 5–1–5
E-mail: †{qiao,shimo,mine}@gavo.t.u-tokyo.ac.jp

あらまし　音素セグメンテーションは，音声認識や音声合成における基本的な問題である。しかしながら，言語情報や音響モデルに関する知識を全く用いない教師なし音素セグメンテーションは，非常に難解な問題として挙げられる。その本質的問題は 「どうのように最適な分割を定義する か」である。本論文では，最適な分割を確率的な枠組みで定式化する。統計分析と情報理論を用いて、最適化対象として三つの目標関数を提案する： Mean Square Error (MSE), Log Determinant (LD) and Rate Distortion (RD)。特に RD 関数は、情報レート歪み理論に基づいて定義されており、人間の言語知覚メカニズムと関連性を見いだすことができる。さらに，RD 関数を用いて，最適な分割が直交変換に対して不変性をもつことを証明した。また，提案された目的関数を最適化するため、時間制約付きの agglomerative clustering アルゴリズムを使用した。そこでは、積分関数を使用することによって効率的なアルゴリズムの実装手法を提案した。 本実験では，TIMIT データベースを用いて，提案した目標関数の評価実験を行なった。 Rate Distortion が最良の音素検出性能を示し (recall rate 79.1% in 20ms tolerance windows)，それは近年発表された教師なしセグメンテーション手法 [1], [4], [5] と比較して，より良い結果を示している。

キーワード　教師なし音素的セグメンテーション, 最適化、レート歪み

Toward Optimal Unsupervised Phoneme Segmentation -A Theoretical and Experimental Investigation

Yu QIAO†, Naoya SHIMOMURA†, and Nobuaki MINEMATSU†

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

Abstract　Phoneme segmentation is a fundamental problem in speech recognition and synthesis studies. Unsupervised phoneme segmentation assumes no knowledge on linguistic contents and acoustic models, and thus poses a challenging problem. The essential question behind this problem is *how to define the optimal segmentation*. This paper formulates the optimal segmentation based on 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 by using information Rate Distortion theory and can be related to human speech perception mechanisms. And we prove that the optimal segmentation of RD is invariant to orthogonal transformation. To optimize the proposed objective functions, we use time-constrained agglomerative clustering algorithm. We also propose an efficient method to implement the algorithm by using integration functions. We carry out experiments on TIMIT database to compare the above three objective functions. The results show that Rate Distortion achieves the best performance (recall rate 79.1% in 20ms tolerance windows) and indicate that our method outperforms the recently published unsupervised segmentation methods [1], [4], [5].

Key words　unsupervised phoneme segmentation, optimization, rate distortion

1. Introduction

Many speech analysis and synthesis applications depend on segmentation to divide speech signals into phonetic segments (phonemes and syllables) [6]. Unlike written language, speech signals do not include explicit space for segmentation.

Moreover, human speech is smoothly continuous signal and does not change abruptly due to the temporal constraints of vocal tract motions. All these facts make segmentation a difficult question.

In speech engineering, Automatic Speech Recognition (ASR) models often 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 [15], [17]. Partly for this reason, phoneme segmentation has received large 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 problem is usually converted to the alignment of speech signals with given texts. Perhaps the most famous method of this class is the HMM-based forced alignment [2], [17].

Another class of method tries to perform phonetic segmentation without using any prior knowledge on linguistic contents and acoustic models. This is also known as unsupervised segmentation. The approach of this paper belongs to the 2nd class. The unsupervised segmentation is similar to the phenomenon that an infant perceives speech [14]. 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 [4] detected the "maximum spectral transition" positions as phoneme boundaries. Estevan et. al [5] employed maximum margin clustering to locate boundary points.

Different from these change point detection methods, this paper tries to solve phoneme segmentation problem by answering the essential question behind: *what kind of segmentation is optimal*. In other words, we want to find 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. Formally, we will formulate the segmentation problem in a probabilistic framework. Using 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 can be related to human speech perception mechanism. To optimize the proposed objective functions, we use time constrained agglomerative clustering algorithm. We develop an efficient implementation based on the integration

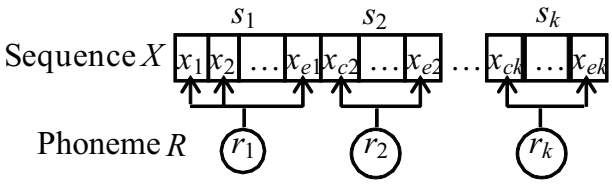

図 1 Diagram of Segmentation Model.

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 recently published results on unsupervised phoneme segmentation [1], [4], [5].

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 $R = \{r_1, r_2, ..., r_k\}$ denote the phoneme sequence, and $p(x_i|r_j)$ represent the probability model of observing x_i given source r_j (Fig. 1). Thus we have,

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

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

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

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 [15]. However, in unsupervised segmentation, we have no knowledge on R . To handle this difficulty, we need to make assumptions on the source distributions r_j. Like most speech applications [6], 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 known, the parameters m_j and Σ_j can be estimated by,

$$\hat{m}_j = \frac{1}{|s_j|}\sum_{i=c_j}^{e_j} x_i, \tag{3}$$

$$\hat{\Sigma}_j = \frac{1}{|s_j|}\sum_{i=c_j}^{e_j} (x_i - \hat{m}_j)(x_i - \hat{m}_j)^T. \tag{4}$$

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

$$\begin{aligned} -\log p(X|S,\hat{R}) &= \sum_{j=1}^{k}\sum_{i=c_j}^{e_j} -\log(p(x_i|r_j)) \\ &= \sum_{j=1}^{k}\sum_{i=c_j}^{e_j} \frac{d}{2}\log(2\pi) + \frac{1}{2}(\log\det(\hat{\Sigma}_j) + (x_i-\hat{m}_j)^T\hat{\Sigma}_j^{-1}(x_i-\hat{m}_j) \\ &= \frac{nd}{2}\log(2\pi) + \frac{1}{2}\sum_{j=1}^{k}|s_j|\log\det(\hat{\Sigma}_j) + \frac{nd}{2}. \end{aligned} \tag{5}$$

From the perspective of information theory, the differential entropy (Chapter 9, [3]) 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. 5 will lead to minimize the description length of the speech sequence. This is in concordance with the minimum description length principle (MDL) [13]. Because the first and the third term of Eq. 5 do not depend on S, to maximize the likelihood of Eq. 2 equals to minimize the following *Log Determinant* (LD) function,

$$LD(X,S) = \sum_{j=1}^{k}|s_j|\log\det(\hat{\Sigma}_j). \tag{6}$$

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,

$$\begin{aligned} -\log p(X|S,\hat{R}) &= \sum_{j=1}^{k}\sum_{i=c_j}^{e_j} \frac{d}{2}\log(2\pi) + \frac{1}{2}(x_i-\hat{m}_j)^T(x_i-\hat{m}_j) \\ &= \frac{nd}{2}\log(2\pi) + \frac{1}{2}\sum_{j=1}^{k}\sum_{i=c_j}^{e_j}||x_i-\hat{m}_j||^2. \end{aligned} \tag{7}$$

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

$$MSE(X,S) = \sum_{j=1}^{k}\sum_{i=c_j}^{e_j}||x_i-\hat{m}_j||^2. \tag{8}$$

The above formula is the same as the objective function of k-means clustering (Chapter 3.5 [8]). The difference between our problem and k-means is that k-means needs not consider the time constraint, which is important for phoneme segmentation.

2.2 Rate Distortion

Let us consider the mechanism of human perceiving speech. It has been shown that the ear's perceptual mechanism places a limit on the smallest spectral differences (Chapter 5. [16]). Human don't care the small difference in speech signals, that is why two linguistically identical utterances with small acoustic differences can be perceived as the same. This fact cannot be represented well by using mean square error (Eq. 8) or log determinant (Eq. 6). 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. [3]), 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 [12]. 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 $E(x-y)^2 \leqq \epsilon$. 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 $z \sim N(0, \epsilon I)$ [3]. Then

$$E(y-\bar{y})^2 = E(x-\bar{x})^2 + 2E(x-\bar{x})Ez + Ez^2 = \epsilon I + \hat{\Sigma}_j, \tag{9}$$

where $\bar{y}$ and $\bar{x}$ are the expected value of y and x, respectively. Thus the entropy of y is bounded by $\log_2((2\pi e)^d \det(\epsilon I + \hat{\Sigma}_j))/2$. R-D theory defines a rate distortion function $R(\epsilon) = \min_{E(x-y)^2\leqq\epsilon} I(x;y)$ to represent the infimum of rates such that bound ϵ can be achieved. We have,

$$\begin{aligned} I(x;y) &= h(y) - h(z) \\ &\leqq \frac{1}{2}\log(2\pi e)^d\det(\epsilon I + \hat{\Sigma}_j) - \frac{1}{2}\log(2\pi e)^d\det(\epsilon I) \\ &= \frac{1}{2}\log\det(I + \hat{\Sigma}_j/\epsilon) \end{aligned} \tag{10}$$

The last line yields a upper bound for rate distortion functions. (注1) We use Eq. 10 to define the following rate distortion (RD) function of X under segmentation S,

$$RD(X,S) = \sum_{j=1}^{k}|s_j|\log\det(I + \hat{\Sigma}_j/\epsilon). \tag{11}$$

We also noticed that a similar measure had been successfully

(注1) : The upper bound by Eq. 10 still holds when x is not Gaussian. Roughly speaking, this is because gaussian variables are mostly difficult to code.

used for image segmentation in vision field recently [9]. 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. 6, and rate distortion of Eq. 11 is invariant to orthogonal transformations.

[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. 6 and Eq. 11, X and X' will have the same segmentations.

Proof 1) At first, we prove the theorem for Eq. 6. 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

$$\begin{aligned} u'_j &= Au_j + b, \\ \Sigma'_j &= A\Sigma_j A^T. \end{aligned} \tag{12}$$

Under segmentation S, we have

$$\begin{aligned} LD(X', S) &= \sum_{j=1}^{k} |s_j| \log\det(\hat{\Sigma}'_j) \\ &= 2d \log\det(A) + \sum_{j=1}^{k} |s_j| \log\det(\hat{\Sigma}_j) \\ &= 2d \log\det(A) + LD(X, S). \end{aligned} \tag{13}$$

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. 6).

2) In the next, we prove the theorem for Eq. 11. Apply eigen-decomposition on covariance matrix $\hat{\Sigma}_j = U^T DU$, 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,

$$\log\det(I + \hat{\Sigma}_j/\epsilon) = \sum_{k=1}^{d} \log(1 + \lambda_k/\epsilon). \tag{14}$$

It is easy to see that the above RD objective function only depends on the eigen values of the covariance matrices. Also according to Eq. 12, the orthogonal transformation will not change the eigen values. Thus

$$\begin{aligned} RD(X', S) &= \sum_{j=1}^{k} |s_j| \log\det(I + \hat{\Sigma}'_j/\epsilon) \\ &= \sum_{j=1}^{k} |s_j| \log\det(I + \hat{\Sigma}_j/\epsilon) \\ &= RD(X, S). \end{aligned} \tag{15}$$

Therefore, X and X' will have the same optimal segmentation by Eq. 11.

Theorem 1 has pratical meaning for our structure study [10], [11]. 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. 6 or Eq. 11.

3. Optimization Algorithm

In Section 2., we have developed three objective functions for segmentation: Mean Square Error (Eq. 8), Log Determinant (Eq. 6) and Rate Distortion (Eq. 11). 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:

$$\min_{\{s_1, s_2, ..., s_k\}} \sum_{j=1}^{k} f(X, s_j), \tag{16}$$

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. 16 for a sequence is to use dynamic programming (DP). However, the direct use of DP needs time cost $O(n^2 k)$, 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 [8]) to optimize Eq. 16. 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

$$f(X, s_j \cup s_{j+1}) - f(X, s_j) - f(X, s_{j+1}); \tag{17}$$

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 time-costly computation in the AS algorithm is to calculate the variance (when using Eq. 8) or covariance matrix (when using Eq. 6 and Eq. 11) for each segment. This computation must repeat many times until the algorithm

terminates. In fact, we need not directly use the summation form of Eq. 3, Eq. 8 and Eq. 4 to calculate mean, variance and covariance every time. There is a more efficient way. We can calculate the following integration functions firstly:

$$G_1(i) = \sum_{k=2}^{i} x_{k-1} \qquad (G_1(1) = 0), \tag{18}$$

$$G_2(i) = \sum_{k=2}^{i} x_{k-1} x_{k-1}^T \qquad (G_2(1) = 0), \tag{19}$$

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:

$$m_j = \frac{1}{e_j - c_j + 1}(G_1(e_j + 1) - G_1(c_j)), \tag{20}$$

$$\Sigma_j = \frac{1}{e_j - c_j + 1}(G_2(e_j + 1) - G_2(c_j)) - m_j m_j^T, \tag{21}$$

$$V_j = \mathrm{Diag}(\Sigma_j), \tag{22}$$

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 [7] to evaluate 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. For each biphone segment, its middle boundary was detected by the minimizing the proposed objective functions. The task is simple. We can easily find the global optimal boundary and calculate the shift error between the detected boundary and the ground truth boundary, which are difficult in sequence segmentation with multiple phonemes.

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. 11) 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 found that RD has the least ASE among all the compared objective functions.

表 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

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. But 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 an 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. 17. 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 is noted that AMSE has a poor

表 2 Recall rates of sequence segmentation

Method	MSE	LD-DIA	RD	RD-DIA
20ms	78.3%	72.3%	77.8%	**78.4%**
30ms	87.9%	84.9%	**89.7%**	89.0%
40ms	93.2%	91.6%	**95.4%**	94.4%

表 3 Recall rates with pre-segmentation

Method	MSE	RD	RD-DIA	AMSE
20ms	78.6%	78.8%	**79.1%**	74.0%
30ms	88.0%	**90.1%**	89.2%	81.7%
40ms	93.2%	**95.5%**	94.4%	86.3%

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 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 used different data sets and testing protocols. We assume that tolerance window size is 20ms, since it is most widely used. In [4], 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× 0.89=75.2%, which is lower than ours 79.1%. Moreover, our insertion rate is 20.9%, which is lower than 28.2% in [4]. [5] 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 79.1%, while theirs is 76.0% [5]. In [1], the authors use a subset of TIMIT database containing 480 sentence and showed a recall rate 73.6%. Although our recall rates are still lower than the HMM-based segmentation methods [2], [17], 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 outperforms the 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. We are going to apply the proposed methods on the event detection problems in our structure study [10], [11].

6. Acknowledgements

The first author would like to thank the Japan Society for the Promotion of Science (JSPS) for the support (P07078).

文 献

[1] G. Aversano, A. Esposito, and M. Marinaro. 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] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley-Interscience New York, 2006.

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

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

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

[7] 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.

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

[9] Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of Multivariate Mixed Data via Lossy Coding and Compression. *IEEE Trans. on PAMI*, 29(9):1546–1562, 2007.

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

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

[12] A. Ortego and K. Ramchandran. Rate-distortion methods for image and video compression. *IEEE Signal Processing Magazine*, 15(6):23–50, 1998.

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

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

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

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

[17] DT Toledano, LAH Gomez, and LV Grande. Automatic phonetic segmentation. *IEEE Trans. on SAP*, 11(6):617–625, 2003.