UNSUPERVISED OPTIMAL PHONEME SEGMENTATION: OBJECTIVES, ALGORITHM AND COMPARISONS

Yu Qiao, Naoya Shimomura, and Nobuaki Minematsu

Graduate School of Frontier Sciences, the University of Tokyo
{qiao, shimo, mine}@gavo.t.u-tokyo.ac.jp

ABSTRACT

Phoneme segmentation is a fundamental problem in many 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 here is *what is the optimal segmentation*. This paper formulates the optimal segmentation problem into a probabilistic framework. Using statistics and information theory analysis, we develop three different objective functions, namely, Summation of Square Error (SSE), Log Determinant (LD) and Rate Distortion (RD). Specially, RD function is derived from information rate distortion theory and can be related to human signal perception mechanism. We introduce a time-constrained agglomerative clustering algorithm to find the optimal segmentations. 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 and indicate that our method outperforms the recently published unsupervised segmentation methods [1, 2, 3].

Index Terms— Unsupervised phoneme segmentation, Rate Distortion theory, Agglomerative clustering

1. INTRODUCTION

A number of speech analysis and synthesis applications need to divide speech signals into phonetic segments (phonemes and syllables) [4]. Both Automatic Speech Recognition (ASR) models and Text-to-Speech (TTS) systems depend on reliable segmentation for achieving good performance [5]. Unlike written language, speech signals do not include explicit spaces for segmentation. Although manual segmentation can be precise, it is usually expensive. For this reason, automatic phoneme segmentation has received much research interest [5, 6, 1, 2, 3].

The approaches to phoneme segmentation can be divided into two classes. The first class requires linguistic contents and acoustic models of phonemes. The segmentation 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 [5]. Another class of methods tries to perform phonetic segmentation without any prior knowledge, which is known as unsupervised segmentation. Our approach belongs to the second class. The unsupervised segmentation is similar to a phenomenon that an infant perceives speech [7]. Most of the previous approaches to this problem focused on detecting on the change points of speech signals and took these 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 tries to solve the phoneme segmentation problem by answering the essential question: *what kind of segmentation is optimal*. In other words, we want to find objective function to evaluate the goodness of segmentation. This is a hard problem as we have neither information on the categories of the phonemes nor prior knowledge on their acoustic models. In this paper, we formulate the segmentation problem in a probabilistic framework. Using statistics and information theory analysis, we develop three objective functions, namely, 1) Summation of Square Error, 2) Log Determinant and 3) Rate Distortion. To optimize these objective functions, we use a time constrained agglomerative clustering algorithm. We also develop an efficient implementation based on integration functions, which can largely reduce the computational time. The proposed objective functions 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, 2, 3].

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.

We 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 $x_{c_j}, x_{c_j+1}, ..., x_{e_j}$.Size of segment $|s_j| = e_j - c_j + 1$. Without any constraint, there are ${}_{n-1}C_{k-1}$ possible cases of segmentations.

For speech signals, it is natural to make the assumption that each individual phoneme is generated by an independent source. Let r_j denote a source for observed segment s_j. $R = \{r_1, r_2, ..., r_k\}$ denotes a source sequence, and $p(x|r_j)$ represents for a probability model between x and r_j. 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 sketch the deductions of several optimal objective functions for unsupervised phoneme segmentation. More details can be found in a technical report [8] [1].

[1] Available at: http://www.gavo.t.u-tokyo.ac.jp/∼qiao/optseg07.pdf

2.1. Summation of Square Error and Log Determinant

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

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

If the source sequence R is given, the above problem can be solved by Viterbi decoding or dynamic programming [6]. 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 many speech applications [4], we assume that $p(x|r_j)$ is a multi-variable normal distribution with mean m_j and covariance matrix Σ_j If s_j is known, m_j and Σ_j can be estimated as,

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

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

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

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

In information theory, the differential entropy (Chapter 9, [9]) of normal distribution $\hat{r}_j(\hat{m}_j,\hat{\Sigma}_j)$ is $\log_2((2\pi e)^d\det(\hat{\Sigma}_j))/2$. Recall 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 sequence. This is in accordance with the minimum description length principle (MDL) [10]. Because the first and the third terms of Eq. 5 do not depend on S, Eq. 5 can be reduced to the following *Log Determinant* (LD) function,

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

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

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

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

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

The above formula is the same as the objective function of k-means clustering (Chapter 3.5 [11]). However, there is an important difference that our objective is to segment sequence, while k-means aims at clustering elements without time constraint.

2.2. Rate Distortion

Consider the perception mechanism of human ears. There is a limit on the smallest spectral differences which can be perceived by human ears(Chapter 5 [12]). Human's are not sensitive to small perturbations in speech signals, that is why two linguistically identical utterances with small acoustic differences can be perceived as the same. This indicates that, for speech segmentation, we need not focus on the details of speech signals. In the next, we are going to develop a perturbation tolerance cost: *Rate Distortion* function based on the information Rate Distortion (R-D) theory (Chapter 13. [9]).

R-D theory was a branch of information theory created by Shannon. It has been shown that R-D theory is related to human perception mechanism [13]. In fact, many popular audio and video compression standards such as MP3, JPEG and MPEG make use of R-D techniques [13]. 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 $E(x-y)^2 \le \epsilon$. 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 practical coding algorithm, but coding length when permitting distortion. We can model y using x with an additive Gaussian noise model: $y = x + z$, where noise $z \sim N(0,\epsilon I)$ [9]. Then, we have

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

Thus the entropy of y is bounded by $\log(2\pi e)^d\det(\epsilon I+\hat{\Sigma}_j)/2$. The rate distortion function [9] is defined as $R(\epsilon) = \min_{E(x-y)^2\le\epsilon} I(x;y)$ to represent the infinimum of rates such that bound ϵ can be achieved. We have,

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

The last line yields an upper bound for rate distortion function [2]. We use Eq. 10 to define the *Rate Distortion* (RD) function of X under segmentation S,

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

We noticed that a similar measure had been successfully used for image segmentation in computer vision recently [14]. However, we don't use the coding lengths for segmentation and mean vector. It can be proved that the segmentation by minimizng Log Determinant (Eq. 6) or Rate Distortion (Eq. 11) is invariant to orthogonal transformations. Details and proof are available in [8].

3. OPTIMIZATION ALGORITHM

In Section 2, we have developed three objective functions for segmentation: Summation of 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:

$$\min_{\{s_1,s_2,\ldots,s_k\}}\sum_{j=1}^{k} f(X,s_j), \quad (12)$$

[2]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.

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. 12 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 [11]) to optimize Eq. 12. 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

$$f(X, s_j \cup s_{j+1}) - f(X, s_j) - f(X, s_{j+1}); \qquad (13)$$

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

One of the most computationally expensive steps in the AS algorithm is to calculate variance (using Eq. 8) or covariance matrix (using Eq. 6 and Eq. 11) 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. 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:

$$G_1(i) = \sum_{k=2}^{i} x_{k-1} \qquad (G_1(1) = 0), \qquad (14)$$

$$G_2(i) = \sum_{k=2}^{i} x_{k-1} x_{k-1}^T \qquad (G_2(1) = 0), \qquad (15)$$

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:

$$m_j = \frac{1}{e_j - c_j + 1}(G_1(e_j + 1) - G_1(c_j)), \qquad (16)$$

$$\Sigma_j = \frac{1}{e_j - c_j + 1}(G_2(e_j + 1) - G_2(c_j)) - m_j m_j^T, \qquad (17)$$

$$V_j = \mathrm{Diag}(\Sigma_j), \qquad (18)$$

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 [15] 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.

Table 1. Comparison of the average absolute shift errors

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

4.1. Experiment 1: segmentation of biphone subsequences

In the first experiment, we extracted all the biphone segments by referring to the label information of TIMIT database. For each biphone segment, its central boundary is detected by minimizing the proposed objective functions. This is relatively 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 both difficult in total sequence segmentation tasks.

We did experiments to compare the performances of the following functions: 1)summation of square error (SSE), 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 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: SSE, 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 (SSE and LD-DIA). When the window size is small (20ms), the performance of SSE and RD (RD-DIA) is very near. However, the differences between SSE 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

Table 2. Recall rates of sequence segmentation

Method	SSE	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	SSE	RD	RD-DIA	ASE
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%

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, the AS algorithm with RD or SSE prefers to merge shorter segments as this will usually lead to the smaller value of Eq. 13. To verify this prediction, we did another experiment where we use a simple Average Square Error (ASE) 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 should be noted that ASE 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 (SSE, 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 use different data sets and testing protocols. Here, tolerance window size is set as 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%.

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 through statistics and information theory analysis: Summation of Square Error, Log Determinant and Rate Distortion. Especially, Rate Distortion is deduced by using information 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. Experimental results show that Rate Distortion outperforms other two objective functions. The results also indicate that our methods achieve higher recall rates than recent published methods [1, 2, 3]. It is not our main objective in this paper to develop a high recall rate segmentation method. Our main focus here is to study unsupervised phoneme segmentation by trying to answer "what is the optimal segmentation". We believe that the results can be improved if incorporating other features and using complex optimization algorithms, which will be our future work. We are also going to apply the proposed methods on the event detection problem in our universal structure study [16]. Finally, it should be noted that the methods proposed in this paper not only apply to the phoneme segmentation but also may have applications in other sequence segmentation problems.

6. 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] 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.

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

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

[8] Y. Qiao, "On Unsupervised Optimal Phoneme Segmentation," *Technical Report, The Univ. of Tokyo*, 2007.

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

[10] 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.

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

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

[13] 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.

[14] 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.

[15] 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.

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