A RIEMANNIAN STOPPING CRITERION FOR UNSUPERVISED PHONETIC SEGMENTATION

Ciro Gracia[1]*, Xavier Anguera*[2]*, Xavier Binefa*[1]

[1]Universitat Pompeu Fabra, Department of Information and Communications Technologies, Barcelona, Spain
[2]Telefonica Research, Edificio Telefonica-Diagonal 00, Barcelona, Spain
{ciro.gracia, xavier.binefa}@upf.edu, xanguera@tid.es

ABSTRACT

With the availability of large and heterogeneous corpora of untranscribed speech we have recently seen regained interest for algorithms to perform automatic segmentation of such data into acoustically homogeneous or phonetic units. In this paper, we face the problem of phonetic segmentation under a hierarchical clustering (HC) framework. Concretely, we focus on the task of automatically estimating the optimum number of segments in speech data. For this purpose we present a Riemannian stopping criterion that is able to automatically stop the HC processing when its is closest to the underlying phonetic segmentation. We test the proposed criterion using TIMIT data and show that it outperforms previous approaches obtaining a significantly lower over-segmentation variance of 46,1% and better average R_{value} improvement of 0.14 compared to a previously proposed approach. We also show that the proposed method is robust in automatically finding the correct number of segments under data source variations.

Index Terms— speech segmentation, hierarchical clustering, cluster count estimation, Riemannian estimator.

1. INTRODUCTION

With the high availability of big online corpora of untranscribed speech it is of importance to build automatic methods to analyze and extract information from such data, regardless of the acoustic conditions and, ultimately, the language being spoken. For this reason we have recently seen a regained interest in algorithms to perform an automatic segmentation of speech data into homogeneous or phonetic units with no a priori knowledge of the spoken language or the phonetic transcription.

Currently, available Approaches to speech segmentation organize data such that each segment contains a target acoustic pattern (e.g. word, syllables or phonemes). Despite target differences, all unsupervised segmentation approaches must solve two fundamental problems: decide how many segments must be created and determine which acoustic observations belong to each segment (i.e. segment boundaries).

In this paper we focus on the estimation of the number of segments in the context of hierarchical clustering for unsupervised phonetic segmentation. As introduced in [10], a segmentation process can be approached as a hierarchical clustering (HC) problem where each segment plays the role of a cluster. Its main advantage is the possibility of building a segmentation that optimizes some given goodness criteria. Although different measures to drive the HC process have been studied in the past, their main drawback is the assumption of knowing the number of clusters as a pre-defined parameter. In [3] we proposed a improvement to this method by describing an automatic way to determine the number of segments in a speech recording. In the current paper we further improve over the previous work by obtaining a more robust automatic estimation of the number of segments under the assumption that such estimation is similar to the classical clustering problem of estimating the number of clusters.

In general, for the task of hierarchical clustering, different strategies can be taken into account for estimating the number of clusters. Some of the approaches (for example see [8]) define a specific criterion that characterizes clustering status, that is both the number of clusters and their composition. Using such criterion, the number of clusters is obtained implicitly from a decision that stops the clustering algorithm when the criterion reaches a desired value. This value is usually learned empirically from development data where the optimal number of clusters is known. Other approaches like [12] are based on the behavior (e.g. first and second moments) of the criterion value rather than on its absolute value. These methods require the exploration of all potential hypotheses before estimating the final number of clusters. In hierarchical clustering this is equivalent to generating the complete dendrogram [6].

In this paper we theoretically compare different potential criterions using the above strategies, and apply them to the task of automatic phonetic segmentation. Our objective is to obtain a criterion for the unsupervised estimation of the correct number of segments in novel acoustic data recorded in different and unknown acoustic conditions. We then propose a novel criterion which is theoretically motivated from the observation of similarities between well know model selection criterions like: GLR, BIC, Ward's Criterion [7]. Under a Gaussian assumption, these criterions can be interpreted as similar ways to compare covariance matrices. From this observation, the proposed criterion incorporates the theory of Riemannian geometry for comparing covariance matrices inside the manifold of symmetrically definite positive matrices.

We validate our proposal in two ways. First, we analyze the distribution of the errors in the estimation of the number of segments, where error is described in terms of an over-segmentation measure [11]. We focus in variance of the error distribution for the annotated utterances, especially when segmenting different instances of the same acoustic events (i.e. instances of the same words or phrases).

finally, we confirm the robustness of the proposed method by using a validation dataset for the task of blind unsupervised segmentation. We measure the error in terms of segmentation accuracy. Our results show a high Hit Rate, a low over-segmentation variance and a high R_{value} measure.

2. UNSUPERVISED ROBUST PHONETIC SEGMENTATION

In previous work [3] we proposed an automatic segmentation algorithm to partition a continuous speech signal into contiguous, non overlapping segments, that maximizes a given objective function. We review such algorithm in this section as a basis for the proposed automatic stopping criterion. Similarly to [10], in [3] we estimate the optimum segmentation for a given input utteracne in an efficient way by using a hierarchical agglomerative clustering (HC) algorithm. The algorithm takes into account the contiguous property of the segments, which means that only grouping operations resulting into contiguous segments will be taken into account. As a result of this property the algorithm complexity becomes linear in time. Formally:

Let $X = \{x_1, x_2, ..., x_n\}$ denote the sequence of feature vectors extracted from an utterance, where n is the length of X and each x_i is a d-dimensional feature vector. A segmentation that divides such sequence X into k non overlapping contiguous segments can be denoted as $S = \{s_1, s_2, ..., s_k\}$. Segments are defined as a set of continuous indices, for example: $s_j = \{c_j, c_j + 1, ..., e_j\}$, where we use c_j, e_j to represent the first and last indices in j^{th} segment and $|s_j|$ is the size of the j^{th} segment: $(e_j - c_j) + 1$

Different objective functions to drive the optimization have been explored. In [3] we found the sum of square errors (SSE) to be a good combination of simplicity, numerical stability and boundary detection accuracy. SSE criterion on S segmentation is defined as

$$SSE(X, S) = \sum_{j=1}^{k} \sum_{i=c_j}^{e_j} ||x_i - \frac{1}{|s_j|} \sum_{i=c_j}^{e_j} x_i||^2 \quad (1)$$

In the initial state, the algorithm defines only one segment s_i for each feature vector x_i present in X, Then the algorithm iteratively merges segments until it reaches the number of final segments defined a priori. The merging operation at each iteration is the one that results in the optimization (minimization in this case) of the objective function, which in is chosen to be Ward's criterion [1], also known as the minimum variance criterion. Given segments s_j , s_{j+1} and R as the segment resulting from grouping of s_j and s_{j+1}, Ward's criterion is defined as follows:

$$\Delta SSE(X, j) = SSE(X, R) - SSE(X, s_j) - SSE(X, s_{j+1}) \quad (2)$$

Experiments conducted on TIMIT corpus show that estimation of phonetic boundaries between two given phonemes based on Ward's criterion obtains a high probability of 94.32% of being closer than 40 ms from ground truth transcription. Figure 1 shows an example of how the Ward's criterion can successfully estimate the boundary between phonemes /w/ and /ao/.

By iteratively applying the Ward's criterion within a hierarchical clustering algorithm provides a way to estimate a segmentation for the input acoustic utterance. In the remainder of the paper we will describe how we define where should the HC stop creating segments, by using the Riemannian estimator.

3. NUMBER OF SEGMENTS ESTIMATION

A desired segmentation is one that does not incur into deletions or insertions of good segments and maximizes boundary correctness. Practically, this is a combination of a good Hit Rate (measuring the number of ground truth boundaries correctly matched) and a good over-segmentation metric (measuring the number of boundaries suggested versus the ground truth). In addition, our interest is that for close acoustic patterns like instances of the same word, differences in segmentations appear only due to differences in pronunciations, especially in terms of composing phonemes, and should not be influenced by speaker voice characteristics, stationary noises and especially temporal scale differences between utterances which would cause a different number of segments. For this reason, we believe that it is especially important not only to maximize these two metrics, but also to obtain a low over-segmentation variance. We focus the development of the Riemannian Stopping criterion described next in minimizing such over-segmentation variance.

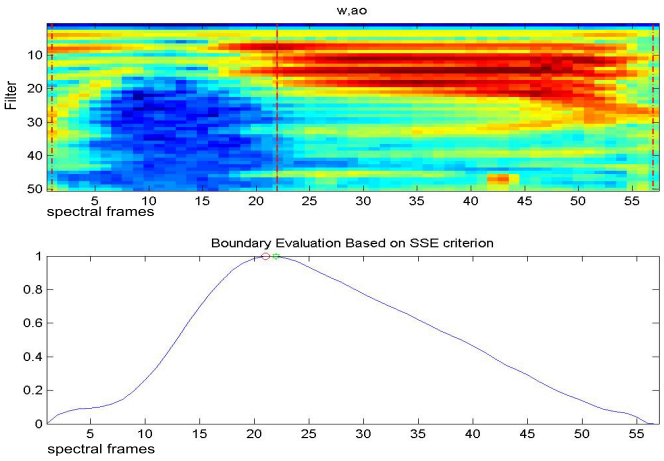

Fig. 1. Example of the estimation of a boundary between two phonemes by means of minimum variance criterion (Ward)

3.1. Riemannian Stopping Criterion

In our previous work [3], we faced the problem of extending the hierarchical clustering framework by incorporating a stopping criterion that automatically found the optimal number of segments. In such approach we used the information change rate criterion (ICR) as our stopping criterion. The ICR criterion, first published in the Diarization literature [4], is a normalized version of the log generalized likelihood ratio criterion (log GLR) designed to be the stopping criterion of a speaker clustering algorithm. Its objective is to reduce the effects caused by cluster differences in terms of number of samples. In [3] we applied ICR to the hierarchical segmentation scenario by means of averaging the comparison of each pair of neighboring segments.

By comparing the mathematical formulation used in ICR, log GLR, DeltaBIC, and Ward's criterion one can easily derive that all these methods are quite similar in structure. In general terms, all of them compare some characteristics extracted from a joint cluster (resulted from the union between compared segments) against the linear combination of characteristics extracted from the individual comparing segments. Such characteristics are all derived from the covariance matrices of the data in the segments. This points us to the Riemannian geometry as a proper way to compare such covariance matrices as it defines a positive definite covariance manifold where comparisons are more accurate. A gentle introduction to theory and practice on the Riemannian framework can be found here [9].

Let the global covariance matrix (Σ_X) be the covariance obtained from of all the data in the test, which is computed for all

data we want to segment at once by using the hierarchical clustering approach. Let also the pooled covariance matrix (Σ_S) be the covariance matrix estimated from a given hypothesis segmentation (3). Pooled covariance is a method for estimating the covariance given samples from several different considered segments where the mean may vary between segments but the true covariance is assumed to remain the same [5].

Within the hierarchical clustering approach used here for phonetic segmentation, we first build the HC dendrogram using the method described in Section 2, obtaining a set of segmentation hypotheses $D = \{S_1, ...S_n\}$. Then, for each segmentation hypothesis we obtain the pooled covariance matrix Σ_S defined above, and compare it with the global covariance matrix Σ_X by means of their geodesic distance in the covariance manifold (4). Finally, the Riemannian estimator becomes the normalized cumulative Mdist function (5) generated from the segmentations $S_t \in D$. This approach is motivated by the idea of establishing a common reference point that helps to compare different acoustic scenarios. Note that when only one segment is considered, the pooled covariance and global covariance are equal, which leads to a Riemannian estimator value of 1.

$$\Sigma_S(X,S) = \sum_{j=1}^{k} \frac{|s_j|}{\sum_{j=1}^{k}|s_j|}\Sigma_{s_j} \tag{3}$$

$$Mdist(X,S) = trace(log^2(\Sigma_X^{-\frac{1}{2}}\Sigma_S\Sigma_X^{-\frac{1}{2}})) \tag{4}$$

$$Riemannian(X,S_t) = \frac{\sum_{i=1}^{t} Mdist(X,S_i)}{\sum_{i=1}^{i\in D} Mdist(X,S_t)} \tag{5}$$

By using this criterion, the optimal segmentation within the dendrogram of segmentations can be selected by thresholding the Riemannian function. As will be shown in the experimental section, such threshold can be estimated on development data and then used to robustly obtain the optimum segmentation on evaluation data.

4. EXPERIMENTS

To validate the proposed algorithm we performed experiments on the TIMIT corpus [2] using the corpus test set (168 speakers, 10 utterances/speaker) to estimate the threshold values for the stoping criteria, and the training set (462 speakers, 10 utterances/speaker) to validate the algorithms. For each utterance we computed spectral features from the speech signal using a pre-emphasis filter (factor was set to 0.97) and a 20 millisecond Hanning window with a 5 millisecond shift. The Mel Filtered spectrogram is generated by a reference software [4] using 50 triangular mel scale filters between 1Hz and 8000Hz. Instead of obtaining standard MFCC parameters through DTC, we reduce the dimensionality of the feature vectors to 24 using PCA to ensure all final dimensions are highly incorrelated. We found this crucial for the implementation of the Riemannian stopping criterion.

4.1. Segmentation evaluation measures

Phoneme segmentation evaluation focuses on the task of successfully detecting segment boundaries between phonemes. For some finite section of speech let N_{hit} be the number of boundaries correctly detected and N_{ref} be the total number of boundaries in the reference. To measure segmentation performance we use three standard measures used in the literature: hit rate, over-segmentation rate and R_{value}. Hit Rate (HR) is calculated using equation 6 and the over-segmentation (OS) rate is computed using equation 7.

$$HR = \frac{N_{hit}}{N_{ref}} \cdot 100 \tag{6}$$

$$OS = (\frac{N_f}{N_{ref}} - 1) \cdot 100 \tag{7}$$

Although HR and OS are most popular in the bibliography, recent work [11] proposes a combination of both into a single-value measure: the R_{value}, shown in equation 8.

$$R_{value} = 1 - \frac{\sqrt{(100-HR)^2 + OS^2}}{200} + \frac{\|HR - 100 - OS\|}{200\sqrt{2}} \tag{8}$$

The R_{value} is especially interesting because it solves the problem of correct evaluation of boundary matching. During boundary matching, a tolerance collar (usually expressed in millisecond) is set around reference boundaries allowing matching with a hypothesized boundary falling inside this region. Most of the literature evaluation measures require this preprocessing step and strongly depend on it but do not specify clearly how boundary matching problems are solved. In [11] an analysis of the boundary matching potential problems is presented and a solution is proposed by avoiding overlap in boundary tolerance collars. We applied this boundary matching method during the computation of the evaluation measures.

4.2. Hierarchical Clustering Segmentation Results

The first experiment perfoemed evaluates the criterion performance for the task of estimating the correct number of segments on the validation dataset. Initially, we use development data in order to estimate the stopping threshold. Then, for a given utterance, its number of segments is obtained by selecting the segmentation from the HC dendrogram where the criterion reaches the learnt optimum threshold (0.996). Given that boundary detection is given by Ward's criterion together with HC algorithm, we evaluate it by focusing into the average over-segmentation as well as its variance on the validation dataset. A low over-segmentation variance together with small bias is valuable because represents an estimation that is more robust when determining the number of segments in data.

Figure (2) shows the comparison between Riemannian and ICR criterions by means of their over-segmentation cumulative density function (c.d.f). This representation describes the distribution of the over-segmentation absolute values, allowing for an easy comparison between methods. As we can see on the figure, the Riemannian approach significantly outperforms ICR criterion. As an example: The probability of obtaining a number of segments estimation with an over-segmentation between +-20% is 0.936 for the Riemannian against 0.83 for the ICR criterion. Overall, the over-segmentation variance for the Riemannian criterion is 46,10% lower than using the ICR criterion approach. The average evaluation measures for both methods are presented in table 1. These results show that both approaches have aproximately the same mean values. Despite of that, a closer look at figure 3 reveals that Riemannian segmentations mostly outperform ICR in all the experiments. This is due to the Riemannian lower over-segmentation variance which promotes that worst scenarios become closer to the average scenario. This explains why the average R_{value} for Riemannian criterion slightly outperform ICR (1.6% of relative improvement), as R_{value} penalty for the over-segmentation error is severe.

[4] Dan Ellis. Lab Rosa Matlab Audio Processing Examples

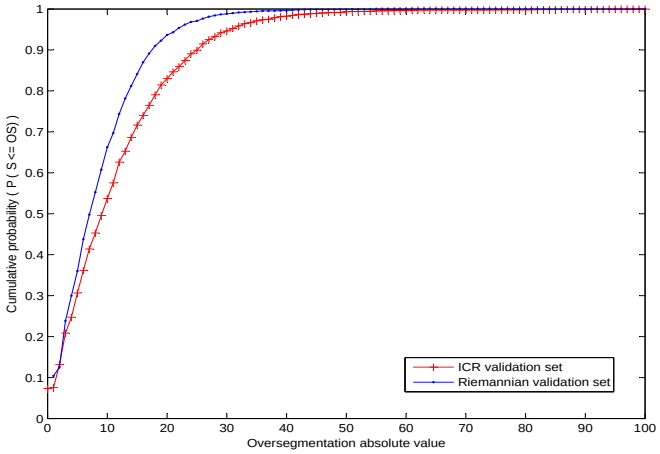

Fig. 2. over-segmentation c.d.f for RIM and ICR criterions on validation experiments

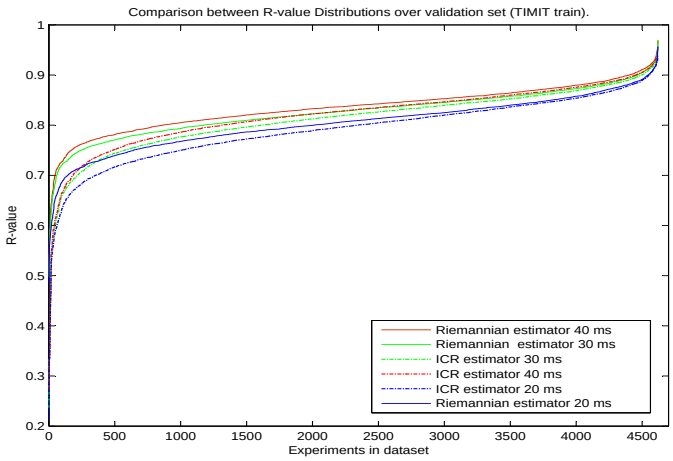

Fig. 3. Visual comparison of the experiments in validation dataset: R-value evaluations in ascending order for both estimators

The second experiment focuses on the estimation of the number of segments for isolated words. Segmentation on isolated words is interesting because words can suffer from intraclass variability in pronunciation and high temporal stretching. A robust word segmentation is especially interesting for posterior processing focused on learning the structure of word classes. We isolated all the word from TIMIT training set and then segmented them using both estimators. A total of 2349 different words were extracted, with a total of 12073 word instances. Figure 4 shows the words over-segmentation average cumulative distribution for both criterions. Again, we can see that Riemannian criterion consistently outperforms ICR, obtaining a 52% lower error variance.

5. CONCLUSIONS

In this paper we apply hierarchical clustering for phonetic segmentation and propose a novel criterion for the estimation of the number of segments in acoustic data. The proposed criterion takes advantage of Riemannian geometry for selecting the most adequate segmentation from the HC dendrogram. Overall, the proposed method outperforms previous approaches obtaining consistently better segmentations, lower error variance and becoming more robust to data source variations.

Stop est	R-value 20 ms		R-value 30 ms		R-value 40 ms
ICR	79.08%		81.16%		81.98%
Riem.	80.37%		82.57%		83.45%
Stop est	HR	OS	precision	recall	F-score
ICR	80.24%	-1.1%	82.0%	80.2%	80.6%
Riem.	80.34%	-1.9%	82.4%	80.3%	81.1%

Table 1. Average evaluation Result from validation dataset comparing ICR and Riemannian criterions

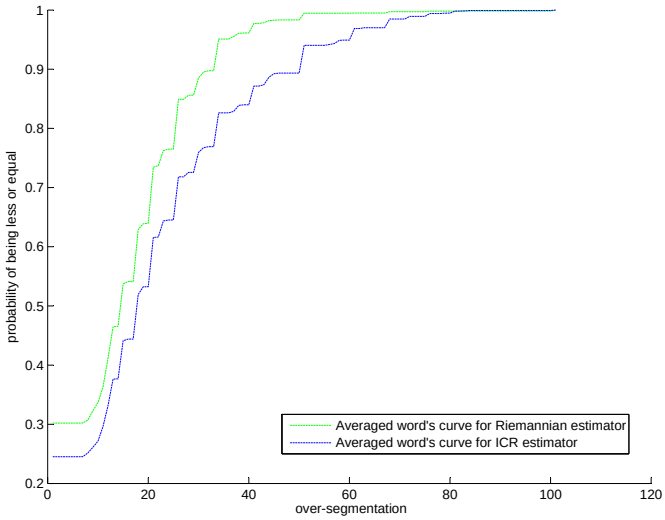

Fig. 4. c.d.f for RIM and ICR estimators showing oversegmentation distributions on isolated words

6. ACKNOWLEDGMENTS

This work was partially funded by the Spanish MITC under the "Avanza" Projects *Ontomedia* (TSI-020501-2008-131) and *Consumedia* (IPT-2011-1015-430000).

7. REFERENCES

[1] V. Batagelj. Generalized ward and related clustering problems. *Classification and related methods of data analysis*, pages 67–74, 1988.

[2] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. *NTIS order number PB91-100354*, 1993.

[3] C. gracia Pons & Xavier Binefa. On hierarchical clustering for speech phonetic segmentation. In *19th European Signal processing conference*, 2011.

[4] K. J. Han, S. Kim, and S. S. Narayanan. Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization. *IEEE Transactions on Audio, Speech & Language Processing*, 16:1590–1601, 2008.

[5] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis (6th Edition)*. Prentice Hall, 2007.

[6] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

[7] V. Le, O. Mella, D. Fohr, et al. Speaker diarization using normalized cross likelihood ratio. *proceeding of INTERSPEECH 2007*, pages 1869–1872, 2007.

[8] D. Pelleg and A. W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 727–734, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[9] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *Int. J. Comput. Vision*, 66(1):41–66, Jan. 2006.

[10] Y. Qiao, N. Shimomura, and N. Minematsu. Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3989–3992, 2008.

[11] O. J. Räsänen, U. K. Laine, and T. Altosaar. An improved speech segmentation quality measure: the r-value. In *Interspeech*, pages 1851–1854, 2009.

[12] S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pages 576 – 584, nov. 2004.