Text Speech and Dialogue Workshop, September 1998

VALIDITY CRITERION FOR UNSUPERVISED SPEAKER RECOGNITION

Itshak Voitovetsky , Hugo Guterman , and Arnon Cohen

Department of Electrical and Computer Engineering
Ben-Gurion University of the Negev
P.O.B. 653, Beer-Sheva, 84105
Israel
{itsik, hugo, arnon}@ee.bgu.ac.il

Abstract. It is often required to perform automatic segmentation of multi-speaker conversations, without having any prior knowledge on the speakers. Given the conversation signal, it is desired to estimate the number of speakers, segment the signal to single speaker segments and label each segment. A method for the determination of the number of speakers participating in a dialogue is presented in this paper. Multiple Self-Organizing Maps (SOMs) are used for clustering, with each SOM representing a cluster. At the end of each stage, a validity criterion (to determine the number of speakers) is calculated for different numbers of SOMs based clustering. Several experiments with dialogues of 2 and 3 speakers were conducted. For high quality speech, the number of speakers was correctly estimated. In telephone quality speech 2 out of eight files were estimated to have three (rather than two) speakers.

1 Introduction

Speaker recognition has many applications in commercial, military and forensic areas. When *a-priori* knowledge about the speakers is available, pre-training can be performed and the problem is defined as supervised classification [1], [2], [3]. However, in many cases, no such *a-priori* knowledge is available (e.g. eavesdropping or conference transcription). In these cases, unsupervised methods must be applied.

Different approaches for unsupervised speaker recognition (otherwise known as speech segmentation) such as: Dendrogram [4], HMM based systems [5], [6], EM algorithm for Gaussian mixture estimation [7], [8] and various VQ methods [9], [10], have been suggested in the literature.

In general, given a multi-speaker conversation, it is needed to estimate the number of speakers and to segment the speech signal. Then, each segment must be assigned to a speaker. Recently a novel Unsupervised Classification System (UCS) has been demonstrated [11]. The proposed UCS automatically trains R+1 Kohonen's Self-Organizing Maps (SOM) [12], R for the speakers and one for non-speech

segments. Provided with *a-priori* information on the number of speakers the system demonstrated a success separation rate of, at least, 90% and 80% for 2 and 3 speakers respectively (for high quality speech).

This paper describes an algorithm that estimates the number of speakers in a conversation. To this end a general validity criterion has been defined. The proposed UCS has been tested with a Hebrew conversation database. The algorithm was tested with high quality and telephone quality dialogs.

2 Unsupervised Segmentation: System's Architecture

The general block diagram of the system is shown in figure 1.

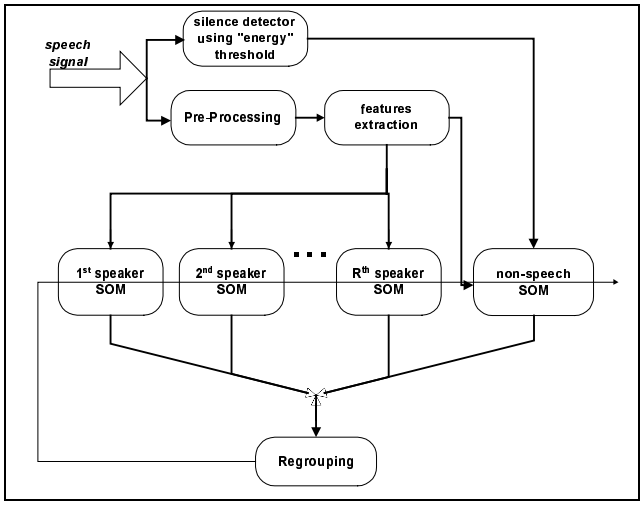

Fig. 1. General description of the unsupervised speaker classification system.

The speech analysis was based on overlapping 15-milliseconds analysis frames, with 5-millisecond frame rate. The features vector, that represented each frame, included 12^{th} order cepstral coefficients, estimated from the 12^{th} order LPC and an additional 12^{th} order first difference cepstral coefficients [1]. The mean absolute value of an accumulated 50-millisecond frame, for speech/non-speech detection, was calculated as well. The mean of the absolute value feature was used for preliminary speech/non-speech segmentation.

The initial conditions of the system were determined as follows: all segments, classified by the coarse speech/non-speech classifier as non-speech, were used to train the non-speech network. The remaining speech segments were randomly and equally divided and used to train the R speaker models. Each of the models (including the non-speech model) was a Kohonen 6x10 SOM and trained using Kohonen's algorithm [12]. The segmentation was performed by an iterative procedure, with regrouping at the end of each iteration. The grouping process was performed with

segments of 100 frames (total of 0.5 second). The clustering algorithm minimized the total distortion during the regrouping process.

Let $d_{n,k}^{(m)}(r)$, be the Euclidean distance between the n-th-vector of the k-th segment ($v_{n,k}$) and the closest centroid ($c_{n,k}^{(m)}(r)$) in the r-th model, during iteration m:

$$d_{n,k}^{(m)}(r) = d\left(v_{n,k}, c^{(m)}(r)\right) = \left(v_{n,k} - c_{n,k}^{(m)}(r)\right)^T \left(v_{n,k} - c_{n,k}^{(m)}(r)\right) \quad (1)$$

In the m-th iteration, the total distance between the k-th segment and the r-th model, $D_k^{(m)}(r)$, is given as:

$$D_k^{(m)}(r) = \sum_{n=1}^{100} d_{n,k}^{(m)}(r) \quad (2)$$

The k-th segment S_k is assigned to the model j that yields the minimum total error:

$$j = \arg\min_{r=0,\ldots,R} \left\{ D_k^{(m)}(r) \right\} \Rightarrow S_k \in SOM_j \quad (3)$$

Hence an iteration of the process is defined by:

1. Retrain the models with the new clusters achieved by the previous iteration.
2. Regroup the data using equation 3.
3. Test for termination: If the termination criterion is met, exit, if not return 1.

At the end of this iterative procedure, the system provides $R+1$ models, for the R speakers and for non-speech data. The data is segmented and labeled as required.

The termination criterion used here was based on the regrouping. Termination was declared when two consecutive iterations showed no change in the clusters.

3 Validity Criterion

In good clustering the intra-cluster distance should be small while inter-cluster distance should be large. It is therefore logical to define the validity of a given partition to be proportional to the ratio between clusters' intra-distances and inter-cluster distance.

Lets R be the number of clusters, $R_1 \le R \le R_2$. The estimated number of clusters will be the one that minimizes a certain validity criterion.

Let: M_r - the number of segments that belong to the r-th cluster.

n_m - the number of vectors in the m-th segment.

$d_n(r)$ - the distance between n-th vector and SOM_r.

$Dcb_n(r,p)$ - the distance between SOM_r and SOM_p for the n-th vector.

If $c_n(r)$ is the closest centroid of the SOM_r to the n-th vector, and $c_n(r,p)$ is the closest centroid of SOM_p to $c_n(r)$, then:

$$\mathrm{Dcb_n(r,p)} = \left[\left(c_n(r) - c_n(r,p)\right)^T \left(c_n(r) - c_n(r,p)\right) \right]^{1/2} \quad (4)$$

Then, the contribution of the r-th cluster to the validity coefficient, Q_r^R, will be define as:

$$\mathrm{Q_r^R} = \frac{1}{M_r} \sum_{m \in r} \frac{1}{n_m} \sum_{n=1}^{n_m} \frac{\mathrm{d_n}(r)}{\sum_{p=1,\ldots,R \; p \neq r} M_p \, \mathrm{Dcb_n(r,p)}} \quad (5)$$

The validity coefficient of the R clusters partition will be a sum of all the contributions, Q_r^R:

$$\mathrm{Q^R} = \sum_{r=1}^{R} \mathrm{Q_r^R} \quad (6)$$

One of the most popular ways to reduce the number of clusters is multi-level dendrogram cutting [4], [8]. In this method the algorithm starts with a large number of clusters. At each stage the algorithm finds the two closest clusters and merges them. The process continues until the final number of clusters is achieved. The assumption of this method is that if two clusters have been merged they can not be separated again.

In this work, clusters are not merged but rather one cluster is reduced. Two ways were checked for cluster reduction. The first is the "knock one cluster out" method. Each time a cluster was knocked out, regrouping process of its segments was applied, and total segmentation distortion was calculated. The cluster that was removed is the one whose removal caused the smallest distortion. The second method was to choose the one with the minimum speech duration and regroup its segments. There were no differences in the validity or clustering results, but the second method proved to be much faster.

4 The Data Base

The Hebrew database consisted of 12 files with two speakers, 3 files with three speakers all of high quality speech dialogue, and 12 telephone dialogues (two speakers). More details about the database can be found in [11].

When the number of speakers was known, all high quality dialogs converged and yielded over 90% correct segmentation for 2 speakers and at least 80% for 3 speakers. Eight out of 12 telephone dialogs converged as the high quality speech. The other 4 dialogs did not converge. When the segmentation algorithm was applied, with the assumption that the number of speakers is three, 3 of the files got 2 clusters representing the two speakers with about 15% segmentation error. The third cluster contained some data from both speakers, and data of simultaneous speech, breathe

sounds, coughs and other interference. The fourth file had one good cluster, one of both speakers and third that contained simultaneous speech segments.

5 Validity Results

All the files were tested with clusters range between 2 and 6. The results of 2 and 3 speakers, high quality speech, are shown in Fig. 2. All the validity functions have their minimum in the correct place. Telephone conversation results are presented in Fig. 3. Fig. 3a shows the validity function of the 8 conversations that had converged well when the number of speakers was known. In six cases the validity minimum was correctly 2, in the other two cases it was located at 3. Despite the error in the validity the segmentation for 2 speakers was as in the case where the number of speakers was known. For other four files (Fig. 3b) the validity decision estimated 3 speakers.

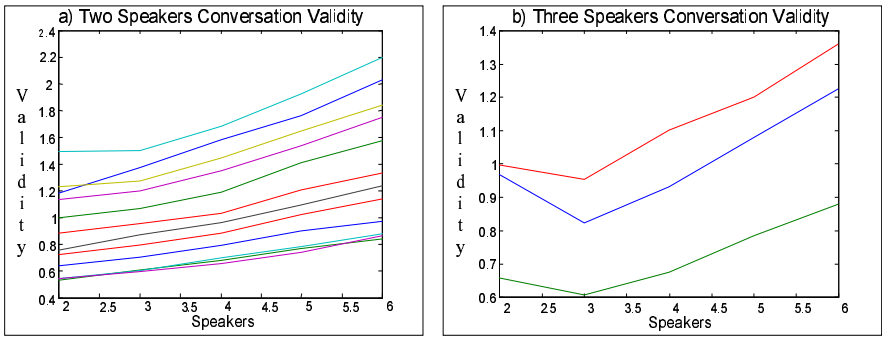

Fig. 2. Validity of high quality conversations, a) two speakers, b) three speakers

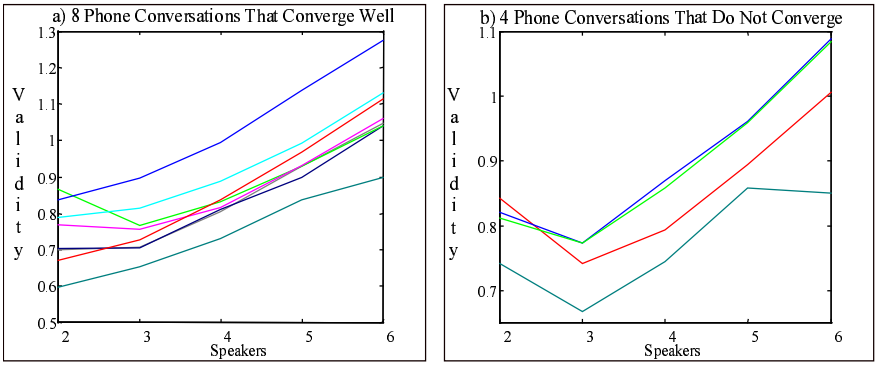

Fig. 3. Validity of telephone quality conversations, a) 8 well converged files, b) 4 files that needed additional cluster

6 Conclusions

A validity function for time series clustering, applied to unsupervised speaker recognition, was presented. The results show that for high quality conversation (with 2 and 3 speakers) the validity correctly estimates the number of speakers. For telephone quality data two out of the eight files were wrongly estimated. Due the narrow band of the channel, poorer signal to noise ratio and larger intra-speaker variability, errors in telephone quality conversation can appear more often than in high quality speech. Results presented in Fig. 3b indicate that even though there were only two speakers the data actually included three clusters. Close examination of the clusters revealed that the third cluster contained a mix of both speakers, simultaneous speech and other sources of interference.

The validity criterion that was presented here was tested only on two and three speakers. Hence, it is necessary to further test this criterion so that its confidence can be established.

References

1. Song F. K. and Rosenberg A. E.: On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Trans. Acoust., Speech, Signal Processing 36 (1998) 871-879
2. Furui S.: Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust., Speech, Signal Processing ASSP-29 (1981) 254-272
3. Xu L., Oglesby J., and Mason J. S.: The optimization of perceptually-based features for speaker identification. ICASSP-89 1 (1989) 520-523
4. Kuhn M. H.: Speaker recognition accounting for different voice conditions by unsupervised classification (cluster analysis). IEEE Trans. Syst., Man, Cybern. SMC-10 (1980) 54-57
5. Wilcox L., Chen F., Kimber D., and Balasubramanian V.: Segmentation of speech using speaker identification. ICASSP-94 1 (1994) 161-164
6. Cohen A. and Lapidus V.: Unsupervised speaker segmentation in telephone conversation. The Nineteenth Convention of Electrical and Electronics Engineers in Israel (1996) 102-105
7. Siu M. -H., Yu G., and Gish H.: An unsupervised, sequential learning algorithm for the segmentation of speech waveform with multiple speakers. ICASSP-92 2 (1992) 189-192
8. Solomonoff A., Mielke A., Schmidt M., and Gish H.: Clustering speakers by their voices. ICASSP-98 2 (1998) 575-560
9. Sugiyama M., Murakami J., and Watanabe H.: Speech segmentation and clustering based on speaker features. ICASSP-93 2 (1993) 395-398
10. Cohen A. and Lapidus V.: Unsupervised text independent speaker classification. The Eighteenth Convention of Electrical and Electronics Engineers in Israel. (1995) 3.2.2 1-5
11. Voitovetsky I., Guterman H., and Cohen A.: Unsupervised speaker classification using self-organizing maps (SOM). Neural Networks for Signal Processing VII Proceedings of the 1997 IEEE Workshop (1997) 578-587
12. Kohonen T. K.: The self-organizing map. Proc. IEEE 78 (1990) 1464-1480