

UNSUPERVISED SPEAKER CLASSIFICATION USING SELF-ORGANIZING MAPS (SOM)

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
Tel: +972-7-6472417
Fax: +972-7-6472949 (Attn. Itsik)
itsik@newton.bgu.ac.il

Abstract

An algorithm for unsupervised speaker classification using Kohonen SOM is presented. The system employs 6x10 SOM networks for each speaker and for non-speech segments. The algorithm was evaluated using high quality as well as telephone quality conversations between two speakers. Correct classification of more than 90% was demonstrated. High quality conversation between three speakers yielded 80% correct classification. The high quality speech required the use of 12th order cepstral coefficients vector. In telephone quality speech, additional 12 features of the difference of the cepstrum were required.

INTRODUCTION

Speaker recognition (identification and verification) is being used in many commercial, military and forensic applications. Usually the problem is defined as supervised classification, where *a-priori* knowledge on the speakers is available so that pre-training can be performed [1-4]. In many applications, however, no such *a-priori* knowledge is available. Unsupervised methods must be used.

Solutions to various aspects of the problem have been suggested in the literature. The application of hierarchical NN was described in [5], and HMM based systems in [6-9]. Other methods based on EM algorithm for Gaussian mixture estimation [10], and various VQ methods [11-13], were also employed.

In general, given a multi-speaker conversation, the algorithm has to estimate the number of speakers, to segment the speech signal and to assign each segment to its speaker. The problem has been also termed "speech segmentation" [10-11]. In our current application the number of speakers, R , is assumed to be known. Generally, during a conversation, it may happen that one speaker interferes with another. We assume that the speech signal does not contain such interference, namely simultaneous speech does not occur. All segments with simultaneous

speech are currently manually eliminated from the data prior to performance evaluation, (during the training process all the data was used).

We suggest here an unsupervised classification system that first makes a preliminary segmentation into speech/non-speech segments, using only "energy" threshold. The system then automatically trains $R+1$ Kohonen SOM [14]: R for the speakers and one for non-speech segments. Initial conditions are set, and then all neural networks (NN) compete among themselves until a balance is achieved.

There were four reasons why Kohonens' SOM was chosen. First, an unsupervised learning algorithm was required because of the problem definition. Second, due to short segments, multiple centroids are required to describe each speaker. Third, when we use SOM's, every SOM defines a different speaker model (or non-speech model). If we use one large network, it would be impossible, to indicate which centroids (or neurons) belongs to the same model. For this reason other unsupervised networks such as ART2 [15], or the network architecture proposed by Nissani [16], cannot be used. Fourth, every SOM is a trained code book (CB), this means that it can be used as CBs for discrete HMM that can later be used for (supervised) speaker recognition.

SYSTEM'S ARCHITECTURE

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

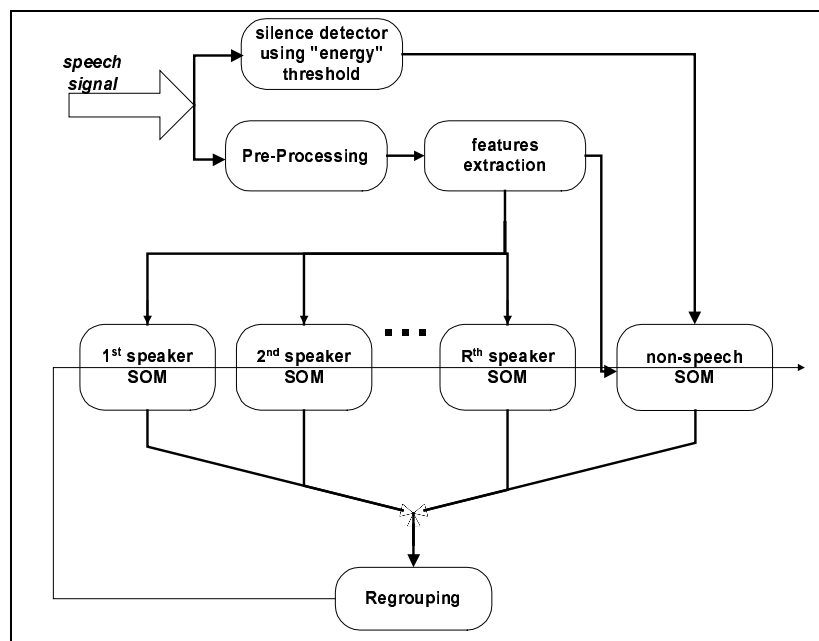


Fig. 1: General description of the unsupervised speaker classification system.

The speech analysis was based on overlapping 15 milliseconds analysis frames, with 5 millisecond frame rate. Each frame was represented by a features vector which included the 12th order cepstral coefficients, estimated from the 12th order LPC. In the telephone data, the features vector was augmented by the 12th order first difference cepstral coefficients [1]. In addition the mean absolute value of an accumulated 50 millisecond frame was also calculated, for speech/non-speech evaluation.

Rough segmentation of speech and non-speech data was performed by thresholding the absolute value feature. The threshold level was set at three percent of the maximum, for high quality speech and one percent for telephone data. The levels were determined experimentally. The fact that higher level was required for high quality speech seems illogical. It is probably due to the fact that in the high quality data the variance of the speech amplitude is much lower than that of the telephone speech. The use of more sophisticated speech detection algorithms should be explored.

The initial conditions to the system were determined as follows: all segments, classified by the rough speech/non-speech classifier as non-speech, were used to train the non-speech network. Segments roughly classified as speech segments were randomly and equally divided and used to train the R speaker models.

Each one of the models (including the non-speech model) was a Kohonen 6x10 SOM. Each SOM was trained by the Kohonen algorithm [14]. The inputs to the SOM were the cepstrum, or cepstrum and difference cepstral coefficients. The outputs of the SOM were Euclidean distances between input vector and network's weight vectors. In each iteration, at the end of the training process, regrouping process was employed. The regrouping process was performed with a segments of 100 frames (0.5 second).

The algorithm is based on clustering the data in such a way that a total error criterion, during regrouping, is minimized.

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 in the r-th model, during iteration m:

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

In the m-th iteration, the total distance between the k-th segment and the r-th model, $D_k^{(m)}(r)$, is given in (2).

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

The k-th segment, S_k , is assigned to the model j (SOM_j) yielding minimum total error:

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

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 (3).
3. Check for termination: If termination criterion is met, exit, if not return to step 1.

It has been proved that this algorithm converges [17].

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. It is of course possible to use a less restrictive criterion which will require that two consecutive iterations will exhibit a change of no more than a given predetermined level. The use of such a criterion will reduce computation time at the expense of accuracy.

CLASSIFICATION ERROR EVALUATION

The algorithm is based on the classification of 0.5 second segments. Each segment may be assigned to one model (speaker or not-speech model) or, in transient segments, due to the finite resolution, may be common to two models or more. The definition of the classification error is clear in the non-transient segments. In case of transient segments, the correct assignment may be to either one of the correct models. Obviously, it makes sense to define classification error that takes in account a segment split between models. A linear piecewise classification error weight is used here.

Figure 2 shows 10 seconds (200 frames per second) of manually classified speech and the error weighting. The dashed lines show an example where a segment includes speech from both the first and the second speakers.

The error weighting has been defined as follows:

1. From the manual segmentation of the speech, all transient times, namely the switching times between speakers were found and denoted: $\{n_1, n_2, \dots, n_M\}$.
2. In the neighborhood of every transient time a local error weighting function, was defined as:

$$w_m(n) = \begin{cases} \left| \frac{n - n_m}{L/2} \right| & ; \quad |n - n_m| < \frac{L}{2} \\ 1 & ; \quad \text{otherwise} \end{cases} \quad (4)$$

where L is the segment's duration (L=500msec. in our case).

3. Sum all the local weighting functions and subtract $(M - 1)$:

$$g(n) = \sum_{m=1}^M w_m(n) - (M - 1) \quad (5)$$

4. The general weighting function will be:

$$w(n) = \begin{cases} g(n) & ; g(n) > 0 \\ 0 & ; \text{otherwise} \end{cases} \quad (6)$$

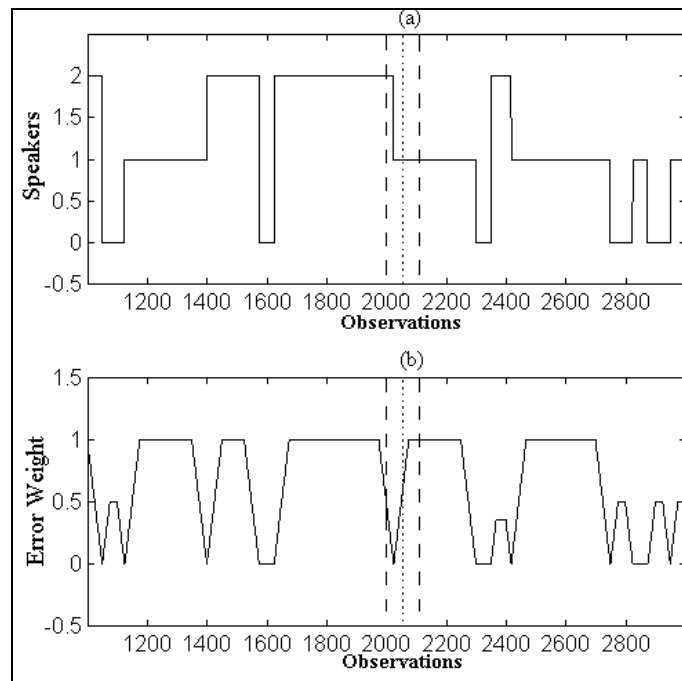


Fig 2: Error weighting function (“0”-non-speech, “1”- speaker A, “2”- speaker B).
a) 10 seconds of manual segmentation, b) Weighted error function.

THE DATA BASE

The Hebrew data base, used for the evaluation of the system, consisted of 9 files with two speakers, 3 files with three speakers, all of high quality speech dialogue, and 12 telephone dialogues. The duration of the high quality speech files were 72-180 seconds per file. Telephone files duration were about two minutes per file. The high quality speech (7.8kHz bandwidth) was sampled in an acoustic room, at 16KHz sampling rate and 12Bit resolution. Five males and one female participated in the conversations. One of the speakers took part in all dialogues. The telephone quality dialogues were recorded from the telephone line. The data

was filtered by a 3.8KHz low pass filter and sampled at 8KHz sampling rate, with 12Bit resolution. Twenty four male speakers participated in the dialogues.

RESULTS

The algorithm was evaluated with the small data base described above. All high quality conversations between two speakers, where tested on segments of half second, without overlapping. It was found that most of the errors between automatic and manual segmentation were due to transitional segments and the relatively poor resolution of the system. Table 1 shows a sample of the results for high quality speech.

TABLE 1: CONFUSION MATRIX (HIGH QUALITY SPEECH)

	weighted error		
	A	B	NS
A	93.5	0	1.6
B	4.3	94.9	3.4
NS	2.2	5.1	95.0
Total Error [%]	5.6		

For high quality speech, using the part of the data base with two males conversation, the error was between 5.5% and 6.0%. For male/female dialogue the error was only 4.3%.

The algorithm was also evaluated with, two speakers, telephone quality speech. When 12th order cepstrum features were used and half a second segments without overlapping were employed, classification results were very poor (11-47% weighted error). Augmenting the features vector with 12 delta-cepstrum features and 75% overlapping - the results improved significantly. Eight (out of the total of 12) conversations significantly improved their classifications (approximately 6% weighted error). The confusion matrix is presented in table 2. Other 4 (out of the 12) did not converge. These four files were examined by a human listener. The files were found to be of very low quality. One of the files was judged by the listener as having three, rather than two speakers.

We have tried to apply a 3 speakers (plus non-speech) networks to these files. The non-speech segments were all well classified. Two of the four files converged into two separate clusters (with error of about 15%) and one extra cluster that contained segments of both speakers. The third file had one good cluster, one cluster containing segments of both speakers and one extra cluster that contained simultaneous speech. The last file converged into two good clusters and one extra cluster that contained breath sounds, coughs and other interferences.

High quality, three speakers conversation files were processed with features and segmentation similar to the ones used in telephone quality speech. The results were not as good as for two speakers case (19.5% classification error).

TABLE 2: TELEPHONE CONVERSATION, CONFUSION MATRIX.

	weighted error		
	A	B	NS
A	93.6	2.4	1.8
B	0.9	94.0	4.6
NS	5.5	3.5	93.6
Total Error [%]	6.2		

Figure 3 shows a 10 seconds of segmented speech. It can be seen that except for very short segments of non-speech at the beginning and at 8 seconds, there is an agreement between the manual and automatic segmentation.

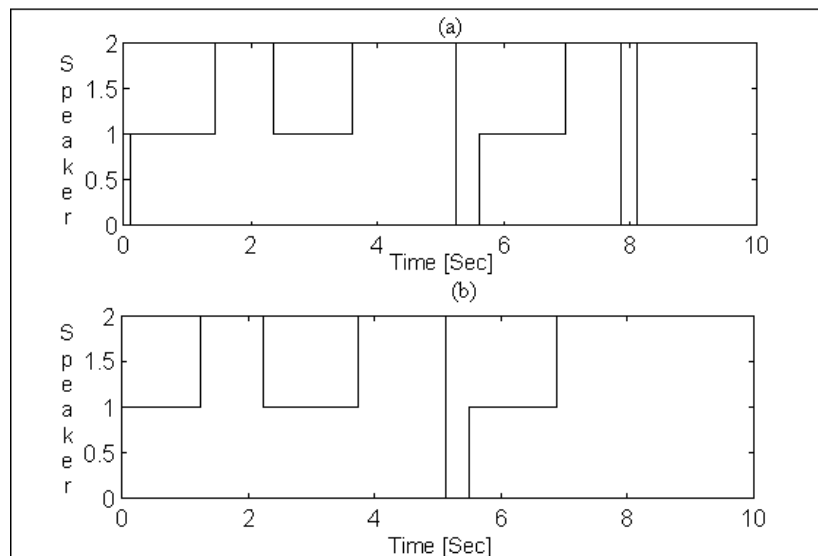


Fig. 3: 10 second classification of telephone conversation.
 (“0”-non-speech, “1”- speaker A, “2”- speaker B).
 a) manual segmentation, b) SOM networks segmentation

Figure 4 shows an example of the system’s convergence as a function of iteration number. In case of two speakers, thirty to forty iterations are needed for convergence for one minute of telephone conversation. For two minutes of data , 50-65 iterations are usually needed. However, after 20 iterations, the system

usually yields results close to optimal. In practice, about 20 iterations will be needed.

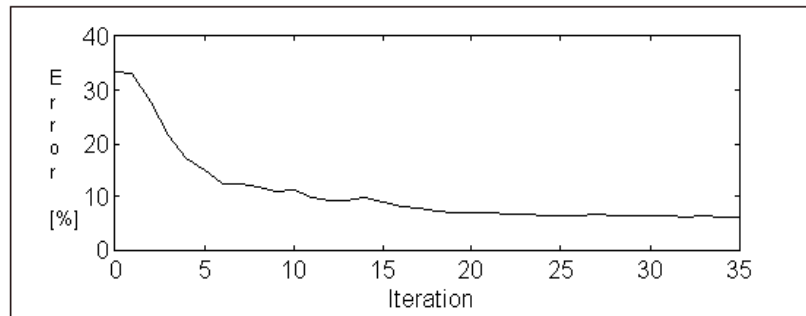


Fig. 4: The weighted error as a function of iteration number, An example for convergence determination.

CONCLUSIONS

A new architecture for unsupervised speaker classification was presented, using Kohonen SOM. With two speakers, and cepstral coefficients, both high quality and most telephone quality conversations, yielded classification errors of about 6%.

The same data base was used with an unsupervised algorithm based on HMM [9]. The results of the two algorithms are compatible. More work is required in order to determine, with sufficient statistical significance, whether one algorithm is more accurate than the other. The computation time difference between the two must also be further examined before a conclusive comparison can be made.

The algorithm achieves better results for conversations between male and female. This result is not surprising, because of the differences in the voice characteristics of the sexes.

For conversations with 3 speakers, classification error is about 20%. The errors appear between the speakers. The non-speech model yields about 5% error, similar to the two speaker case. Note that the data duration in two and three speakers experiments were approximately the same. More data may be needed in order to improve the three speakers results.

The current algorithm assumes that the number of speakers is known. We are currently in the process of developing a validity algorithm which will estimate the number of speakers participating in the conversation. In addition we are working on increasing the resolution of the algorithm.

A pre-processing algorithm will be developed to detect incidents of simultaneous speech, to allow automatic removal of such incidents.

REFERENCE

- [1] F. K. Song and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," **IEEE Trans. Acoust., Speech, Signal Processing**, vol. 36, no. 6, pp. 871-879, June 1988.
- [2] A. Cohen and I. Froind, "On text independent speaker identification using a quadratic classifier with optimal features," **Speech Communication**, vol. 8, no. 1, pp. 35-44, March 1989.
- [3] S. Furui, "Cepstral analysis technique for automatic speaker verification," **IEEE Trans. Acoust., Speech, Signal Processing**, vol. ASSP-29, no. 2, pp. 254-272, April 1981.
- [4] L. Xu, J. Oglesby, and J. S. Mason, "The optimization of perceptually-based features for speaker identification," **ICASSP-89**, vol. 1, pp. 520-523, 1989.
- [5] M. Zaki, A. Ghalwash, and A. A. Elkouny, "CNN: a speaker recognition system using a cascaded neural network," **International J. of Neural Systems**, vol. 7, no. 2, pp. 203-212, May, 1996.
- [6] L. Wilcox, F. Chen, D. Kimber, and V. Balasubramanian, "Segmentation of speech using speaker identification," **ICASSP-94**, vol. 1, pp. 161-164, 1994.
- [7] J. O. Olsen, "Separation of speakers in audio data," **Proc. of 4th European Conference on Speech Communication and Technology**, vol. 1, pp. 355-358, September, 1995.
- [8] A. Cohen, and V. Lapidus, "Unsupervised speaker segmentation in telephone conversation," **The Nineteenth Convention of Electrical and Electronics Engineers in Israel**, pp. 102-105, 1996.
- [9] A. Cohen, and V. Lapidus, "Unsupervised, text independent, speaker classification," **ICSPAT-96**, pp. 1745-1749, 1996.
- [10] M. -H. Siu, G. Yu, and H. Gish, "An unsupervised, sequential learning algorithm for the segmentation of speech waveform with multiple speakers," **ICASSP-92**, vol. 2, pp. 189-192, 1992.
- [11] M. Sugiyama, J. Murakami, and H. Watanabe, "Speech segmentation and clustering based on speaker features," **ICASSP-93**, vol. 2, pp. 395-398, 1993.
- [12] T. Kosaka, S. Matsunaga, and S. Sagayama, "Speaker-independent speech recognition based on tree-structured speaker clustering," **Computer Speech and Language**, vol. 10, no. 1, pp. 55-74, January, 1996.
- [13] A. Cohen, and V. Lapidus, "Unsupervised text independent speaker classification," **The Eighteenth Convention of Electrical and Electronics Engineers in Israel**, pp. 3.2.2 1-5, 1995.
- [14] T. K. Kohonen, "The self-organizing map," **Proc. IEEE**, vol. 78, no. 9, pp. 1464-1480, September, 1990.

Neural Networks for Signal Processing VII
Processing of the 1997 IEEE Workshop, pp. 578-587, September 24-26 1997

- [15] G. A. Carpenter and S. Grossberg, "ART 2: self-organization of stable category recognition codes for analog input patterns," **Applied Optics**, vol. 26, no. 23, pp. 4919-4930, December 1987.
- [16] D. N. Nissani, "An unsupervised hyperspheric multi-layer feedforward neural network model," **Biol. Cybern.**, vol. 65, no. 6, pp. 441-450, 1991.
- [17] I. Voitovetsky, H. Guterman, and A. Cohen, "Training algorithm convergence of multiple code book system," **Internal Report**, Ben-Gurion University of the Negev, Dep. of Electrical and Computer Engineering, 1997.