interspeech 2008 incorporating SST 08

Weighted Segmental K-Means Initialization for SOM-Based Speaker Clustering

Oshry Ben-Harush[1], *Itshak Lapidot*[2], *and Hugo Guterman*[1]

[1]Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, P.O.B 653, Beer-Sheva Israel, 84105
[2]Department of Electrical and Electronics Engineering, Sami Shamoon College of Engineering, Jabotinsky 84, Ashdod, 77245 Israel
oshryb@bgu.ac.il

Abstract

A new approach for initial assignment of data in a speaker clustering application is presented. This approach employs Weighted Segmental K-Means clustering algorithm prior to competitive based learning. The clustering system relies on Self-Organizing Maps (SOM) for speaker modeling and likelihood estimation. Performance is evaluated on 108 two speaker conversations taken from LDC CALLHOME American English Speech corpus using NIST criterion and shows an improvement of approximately 48% in Cluster Error Rate (CER) relative to the randomly initialized clustering system. The number of iterations was reduced significantly, which contributes to both speed and efficiency of the clustering system.

Index Terms: Clustering, Speech, SOM, K-means, Initial Conditions.

10.21437/Interspeech.2008-4

1. Introduction

Given a conversation between R participants, a speaker clustering system deals with the problem of identifying segments in the conversation that belongs to the same speaker $\{r\}_{r=1}^{R}$ of the unknown number of speakers, with no prior knowledge about the speakers. A block diagram of an iterative clustering system is presented in Figure 1.
In order to achieve satisfactory clustering performance, a clustering system has to deal with several issues:

1. Due to the time dependence of speech signals, feature vectors have to be clustered as a series of time related vectors. Time-series clustering has many application in pattern recognition and machine learning [1] ,[2], [3] and [4].
2. Identification of speech and non speech segments or Voice Activity Detection (VAD) as a pre-processing stage for clustering; This issue is discussed in [5], [6] and [7].
3. Generation of speaker models and determining ordered speaker appearance in the conversation.

Speaker modeling can employ either parametric or non-parametric methods. Parametric models can include the Gaussian Mixture Models (GMM) [1] and [7]; non-parametric modeling techniques such as the Self Organizing Maps (SOM) [5] and [8], Linear Vector Quantization (LVQ) [9], Learning Matrix Quantization (LMQ), etc. The most popular modeling algorithms for speaker clustering are BIC [10] and GMM. In the current study SOM non-parametric modeling is used to produce speaker models. SOM models are also applied as likelihood estimators [2].
In this study the focus is placed on the initial assignment of feature vectors to R clusters, initial assignment is addressed as Initial Conditions (IC). Initial conditions form the seed for the iterative competitive learning based clustering system. The clustering algorithm was evaluated on telephone conversations taken from LDC CALLHOME American English Speech Corpus [11]. It was assumed that the number of speakers was always two.
The rest of the paper is organized as follows: section 2 describes the SOM based clustering system; sections 3 describes the Segmental K-Means algorithms which forms the basis for the Weighted Segmental K-Means algorithm presented in section 4; section 5 presents system evaluation results and conclusions appear on section 6.

2. System Description

Our speaker clustering system is based on the iterative competitive speaker clustering system presented in [6]. Conversations are framed into 20mSec frames with a 10mSec frame overlap. 12^{th} order MEL-Cepstrum features plus 12 delta features were extracted from each frame. In addition, an energy envelope was calculated every 50mSec. A threshold of 3% above the minimal energy was taken for speech/non-speech preliminary segmentation. 60 (6×10) neurons are used for each SOM model where each neuron comprises a code-word (CW) from the codebook (CB) of the speaker's model as done in [2], [6] and [12]. The clustering system presented in [6] applies segmental random initial data assignment. This paper presents an improvement to this initial assignment.

2.1. SOM Based VQ

Speaker models in this study are based on a non-parametric SOM competitive learning algorithm presented by Lapidot et al. in [5].
SOM algorithm is employed to train the non-speech model from the non-speech segments and train R models from the speech segments assigned to each of the speakers. SOM training algorithm is described in [8]. The output of a SOM training is a CB.

2.2. VQ-Based Likelihood Estimator

The description of VQ as a log-likelihood estimator can be found in [2]. Having R codebooks and L codewords per code-

Accepted after peer review of full paper

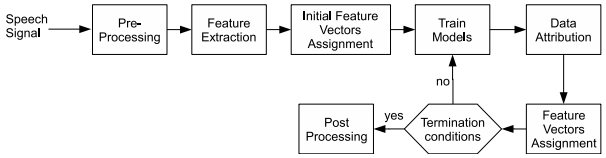

Figure 1: *Iterative Clustering System Block Diagram.*

book, the log-likelihood can be estimated under the following assumptions: for each codebook $\{CB_r\}_{r=1}^{R}$, each codeword $\{CW_r^l\}_{l=1}^{L}$ is the mean of a Gaussian probability density function (*pdf*) with unit covariance matrix. Thus, the log-likelihood of one observation can be estimated as follows: be feature vector $o_n = [o_n^1, ..., o_n^d]^T \in \mathbb{R}^d$, where T means transpose operator and $CW^l = [cw^{l,1}, ..., cw^{l,d}]^T \in \mathbb{R}^d$, then

$$L(o_n|CB_r) = -\frac{d}{2}(log(2\pi)) - \frac{1}{2}(o_n - CW_r^{l^*,n})^T(o_n - CW_r^{l^*,n}) \quad (1)$$

where

$$l^* = \arg\max_{l=1,2,...L}\{(o_n - CW_r^{l,n})^T(o_n - CW_r^{l,n})\} \quad (2)$$

The joint log-likelihood for all of the observations $\mathbf{O} = [o_1, ..., o_N] \in \mathbb{R}^{d\times N}$.

$$L(\mathbf{O}|CW_r) = -\frac{dN}{2}log(2\pi) - \sum_{n=1}^{N}(o_n - CW_r^{l^*,n})^T(o_n - CW_r^{l^*,n}) \quad (3)$$

3. Segmental K-Means Initial Assignment

Segmental K-Means Initial Conditions (SKMeansIC) of feature vectors places the foundations on which Weighted Segmental K-Means algorithm relies. Segmental and Weighted Segmental K-Means flowchart is presented in Figure 2. SKMeansIC procedure takes advantage of pauses in fluent speech to mark segments in the conversation. Segmental K-Means algorithm is as follows:

1. Perform an initial separation of speech from non-speech.
2. Mark the feature set as $\mathbf{O} = \{o_n\}_{n=1}^{N}$, non-speech segments by $\{NS_k^{l_k}\}_{k=1}^{K}$ and speech segments by $\{S_i^{l_i}\}_{i=1}^{I}$ where l_k and l_i are the lengths of the segments such that $\sum_{k=1}^{K} l_k + \sum_{i=1}^{I} l_i = N$.
3. Estimate the mean for each speech segment $\{SC_i\}_{i=1}^{I}$, where SC_i is the estimated mean of the i^{th} speech segment.
4. Apply standard K-Means clustering on $\{SC\}_{i=1}^{I}$ to calculate R centroids.
5. For all $\{SC_i \in Cluster_r\}_{i=1,...I,r=1,...R}$ assign $\{S_i \in Cluster_r\}_{i=1,...I,r=1,...R}$.

4. Weighted Segmental K-Means Initial Assignment

Weighted Segmental K-Means algorithm (WSKMeansIC) relies on SKMeansIC described in the previous section, however

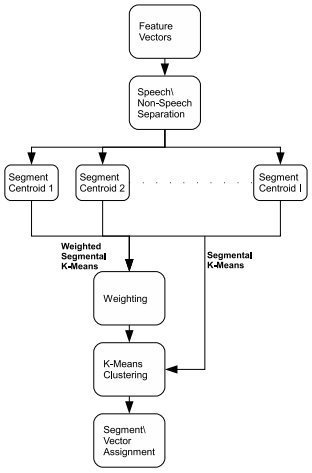

Figure 2: *Segmental and Weighted Segmental K-Means assignment flowchart.*

WSKMeansIC takes into account the length of each segment. SKMeansIC algorithm estimates the means for each segment and clusters the means using K-Means clustering algorithm. SKMeansIC does not consider segment lengths in the clustering process. This appears to be a major flaw in the algorithm due to misrepresentation of the actual distribution of feature vectors. In order to overcome this flaw, prior to K-Means clustering, we assign a weight $\{w_i\}_{i=1}^{I}$ to the mean of each segment such that $w_i = l_i$. Weighted segmental K-Means is as follows:

1. Perform steps 1-3 in the Segmental K-Means Initial Conditions procedure presented in the previous section.
2. Assign a weight $w_i = l_i$ to each of the means $\{SC_i\}_{i=1}^{I}$.
3. Mark the K-Means centroids by $\{V_r\}_{r=1}^{R}$
4. Estimate the new centroids using K-Means algorithms such that $V_r^{new} = \frac{\sum_{SC_i \in Cluster_r} w_i SC_i}{\sum_{SC_i \in Cluster_r} w_i}$
5. For all $\{SC_i \in Cluster_r\}_{i=1,...I,r=1,...R}$ assign $\{S_i \in Cluster_r\}_{i=1,...I,r=1,...R}$.

5. System Evaluation

5.1. Initial Feature Assignment.

Five methods for initial feature assignment were investigated. The initial assignment methods presented here refer to the segments identified as speech by the energy threshold mechanism. Segments identified as non-speech are automatically assigned to the non-speech model. Initial assignment methods are:

1. Random Initial Conditions (RandomIC): assigns feature vectors that belong to $\{S_i\}_{i=1}^{I}$ randomly across each of the R speakers.

2. Segmental Random Initial Conditions (SRandomIC) [6]: each of the models is assigned an equal amount of randomly selected segments from the feature segments $\{S_i\}_{i=1}^{I}$. Each model is assigned L_r speech segments $\{S_l^{L_r}\}_{r=1}^{R}$ such that $\sum_{r=1}^{R} L_r = I$ where I is the number of speech segments.
3. K-Means Initial Conditions (KMeansIC): all speech feature vectors are assigned to each model by employing K-Means algorithm to cluster the features into R clusters.
4. Segmental K-Means Initial Conditions (SKMeansIC): Initial assignment is performed using the algorithm described in section 3.
5. Weighted Segmental K-Means Initial Conditions (WSKMEansIC): Initial assignment of features is performed according to the procedure presented in section 4

5.2. Database

LDC CALLHOME American English Speech is a database of conversation sampled at 8000 Hz in a 2 channel μ-law format [11].
The two channels were summed to get a two speaker conversation and we have only used transcribed conversation between two speakers (108 conversations in total). About 10 minutes of each conversation was transcribed so we analyzed only these 10 minutes of the conversation.

5.3. Evaluation Criterion

NIST criterion [13] was employed for Cluster Error Rate (CER) calculation, this criterion takes into account the number of models with regards to the true number of speakers and the amount of falsely identified information on a time basis. The criterion can be described as:

$$\frac{\sum_{s=1}^{S}\{dur(s)\cdot(max(N_{Ref}(s),N_{Sys}(s))-N_{Correct}(s))\}}{\sum_{s=1}^{S}\{dur(s)\cdot N_{Ref}(s)\}}$$

The conversation is segmented into S segments such that:
$dur(s)$ - Duration of the segment s.
$N_{Ref}(s)$ - The number of speakers assigned to segment s.
$N_{Sys}(s)$ - The number of actual speakers in the segment s.
$N_{Correct}(s)$ - The Number of speakers assigned to segment s who actually takes part in s.
A lower NIST value means better segmentation. Due to the indexing of speakers performed with no prior information on each speaker, it is necessary to calculate NIST error criterion for each of the permutations of speaker indices (two Simultaneous speech in this research is always labled as an error).

5.4. Results

The system described in section 2 was employed in evaluation of the above mentioned five methods for initial feature assignment.
Figure 3 presents the average CER of the conversations using all five initial conditions algorithms as a function of the number of iterations. From Figure 3 it can be seen that WSKMeansIC contributes to the best performance of the clustering system, followed by SKMeansIC and SRandomIC. The poorest clustering results derive from using KMeansIC and RandomIC. The obvious conclusion from the above comparison is that the assignment of continuous segments is preferable over the assignment of single feature vectors. Amongst the single feature vector assignment methods, KMeansIC presents the poorest performance after five iterations of the system. KMeansIC starts off with a lower CER than RandomIC, however, RandomIC converges faster and to a lower CER. This fact requires an in depth research, however, from listening to several conversations it appears that K-Means assignment algorithm assigns the vectors to voiced and unvoiced clusters and thus, the initial assignment actually damages clustering performance.

To emphasize the contribution to clustering performance of each of the initial assignment methods, we examine the CER after one iteration of the clustering system. Table 1 presents the CER using each of the initial assignment algorithms

Table 1: *Cluster Error Rate (CER) after a single iteration.*

Initial Condition algorithm	CER
RandomIC	49.49%
KMeansIC	48.89%
SRandomIC	47.42%
SKMeansIC	27.41%
WSKMeansIC	23.12%

Further examination of Figure 3 shows that the CER line flattens around 5 iterations, that is, on average, 5 iterations will suffice for the convergence of the clustering algorithm. It is worthwhile to compare the clustering error after the fifth iteration using each of the suggested algorithms. It is also interesting to examine the clustering error before any training took place, i.e., just after the initial assignment of feature vectors. Table 2 presents the CER for the initial assignment and for the fifth iteration of the clustering system.

Table 2: *Cluster Error Rate (CER) for initial assignment and for the fifth iteration.*

Initial Condition algorithm	CER after initial assignment	CER after the fifth iteration
RandomIC	66.24%	34.37%
KMeansIC	62.98%	38.47%
SRandomIC	66.00%	31.39%
SKMeansIC	61.08%	24.51%
WSKMeansIC	59.02%	20.05%

Examination of Table 2 and Figure 3 shows that prior to any iteration of the clustering system, the maximal CER reduction by using WSKMeansIC was of about 7% (or a relative improvement of 11%). After five iterations, the maximal CER reduction is about 18% (or a relative improvement of about 48%). SKMeansIC and WSKMeansIC contributes to the fastest convergence rate and the lowest CER from all of the initial assignment methods explored in this work.

6. Conclusions

In this work the influence of the initial assignment of feature vectors on the performance of a speaker clustering system was studied.
Five initial assignment algorithms were compared: random as-

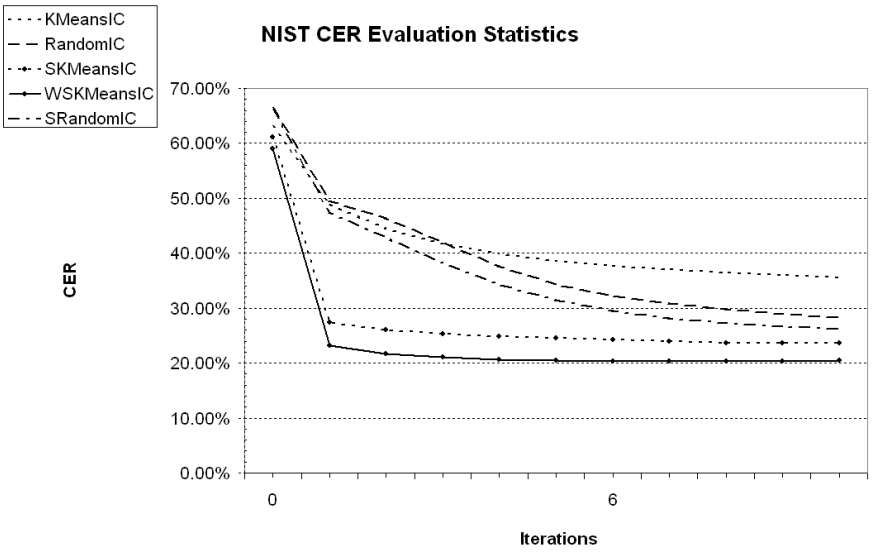

Figure 3: *Cluster Error Rate as a function of iteration.*

signment of feature vectors, random assignment of feature segments, K-Means assignment of feature vectors, K-Means assignment of feature segments and Weighted K-Means assignment of feature segments.
Weighted Segmental K-Means assignment performs much better and significantly fewer number of iterations were required.
Weighted Segmental K-Means assignment achieved lower CER while requiring fewer number of iterations. The WSKMeansIC based assignment produced an average CER of 20.05% for five iterations, compared to the 38.47% CER of the K-Means initial assignment algorithm, which presents the poorest CER of the five initial assignment methos. Table 3 presents the relative improvement of CER of the initial assignment algorithms compared to the RandomIC.

Table 3: *Relative Cluster Error Rate (CER) improvement of four IC methods compared to RandomIC after five iterations of the clustering system*

Initial Condition algorithm	CER Improvement [%]
KMeansIC	-11%
SRandomIC	8%
SKNeansIC	29%
WSKMeansIC	42%

KMeansIC does not improve the CER when compared to RandomIC, KMeansIC is assumed to cluster short acoustic events such as voiced/unvoiced rather than speakers. WSKMeansIC achieves the highest CER improvement, the other initial assignment algorithms never achieve the same CER as WSKMeansIC.

7. References

[1] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *IEEE Speech Recognition and Understanding Workshop*, 2003.

[2] I. Lapidot, "SOM as a likelihood estimator for speaker clustering," in *Eurospeech*, Switzerland, 2003.

[3] Y. Luan and C. T. Li, "Unsupervised clustering of gene expression time series with conditional random fields," in *IEEE International Conference on Digital Ecosystems and Technologies*, 2007, pp. 571–576.

[4] Y. Xiong and D. Y. Yeung, "Time Sereis clustering with ARMA Mixtures," Hong Kong University of Science and Tehnology, Tech. Rep., may 2003.

[5] I. Lapidot, H. Guterman, and A. Cohen, "Unsupervised speaker recognition based on competition between self-organizing-maps," in *IEEE Trans. on NN*, vol. 13, no. 4, July 2002, pp. 877–887.

[6] I. Lapidot, "Unsupervised speaker recognition," Ph.D. dissertation, Ben-Gurion University of the Negev, September 2001.

[7] L. Rabiner, *Fundamentals of speech recognition*. Prentice Hall, 1993.

[8] T. Kohonen, "The self organizing map," *Proceeding of the IEEE*, vol. 78, no. 9, September 1990.

[9] S. Nakamura and T. Akabane, "A neural speaker model for speaker clustering," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 1991, pp. 853–856.

[10] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia, USA, February 1998.

[11] "Liguistic data consortium," LDC97S42, Catalog, 1997, available: http://www.ldc.upenn.edu/Catalog.

[12] I. L. (Voitovetsky), H. Guterman, and A. Cohen, "Unsupervised speaker classification using self organizing maps (SOM)," *Neural Networks for Signal Processing VII Proceedings of the 1997 IEEE Workshop*, pp. 578–587, July 2002.

[13] "The rich transcription spring 2003 (rt-03s) evaluation plan," in *The EARS 2003 Evaluation Plan*, February 2003.