Incremental Diarization of Telephone Conversations

Oshry Ben-Harush[†], *Itshak Lapidot*[◦], *Hugo Guterman*[†]

[†] Department of Electrical and Computers Engineering
Ben-Gurion University of the Negev, Beer-Sheva, Israel
[◦] Department of Electrical and Electronics Engineering
Sami Shamoon College of Engineering, Ashdod, Israel
oshryb@bgu.ac.il, itshakl@sce.ac.il, hugo@ee.bgu.ac.il

Abstract

Speaker diarization systems attempt segmentation and labeling of a conversation between R speakers, while no prior information is given regarding the conversation.

Most state of the art diarization systems require the full body of the conversation data prior to the application of some diarization approach. However, for some applications such as forensics, which handles vast amount of data, an on-line or incremental diarization is of high importance.

For that purpose, a two-stage incremental diarization of telephone conversations algorithm is suggested. On the first stage, a fully unsupervised diarization algorithm is applied over an initial training segment from the conversation. The second-stage is composed of time-series clustering of increments of the conversation.

Applying incremental diarization over 1802 telephone conversations from NIST 2005 SER generated an increase in diarization error of approximately 2% compared to the diarization error of an off-line diarization system.

1. Introduction

Given a conversation between R speakers, speaker diarization systems attempts clustering and labeling of temporal conversation segments to $\{S_r\}_{r=1}^{R}$ speakers and to non-speech, while no prior information is given regarding the conversation.

Conversation diarization is essential for several speech processing applications such as, conversation indexing, forensics, automatic speaker modelling and as a pre-processing stage for speaker recognition tasks. Diarization of conversations could also contribute to increased accuracy of Automatic Speech Recognition (ASR), as these systems shows improved performance while operating on a speaker-dependent mode.

Most state of the art diarization systems operate in an off-line mode, that is, all of the conversation samples must be available before the application of the diarization algorithm. Diarization is then usually applied using some hierarchical or dendogram clustering e.g., [1], [2], [3].

For several applications such as forensics and speaker recognition systems, it could be beneficial to have diarization results prior to the conclusion of the conversation. Examination of the literature shows few studies which handle on-line diarization of multi-speaker conferences or broadcast news scenarios, and most require some prior labeled data in order to train inherent models used for speech/non-speech classification, gender detection and for spawning speaker models [4], [5], [6].

In this paper, a two-stage, incremental, fully unsupervised telephone conversation diarization system is presented. The suggested incremental diarization system relies on a Self Organizing Map (SOM) based iterative diarization system previously described in [7] to perform the diarization.

On the first stage, unsupervised diarization is applied over an Initial Training Set (ITS) of the audio stream, this enables the construction of speakers and non-speech models and adaptation of the time-series clustering parameters.

On the second stage of diarization, diarization is applied over increments of the conversation using the models and HMM parameters at hand followed by an adaptation to the speaker models.

Diarization was applied over 1802 conversation from the NIST 2005 Speaker Recognition Evaluation (SER) [8]. Diarization error increases by roughly $\sim 2\%$ compared to the diarization error of the baseline system (operating over the entire conversation), this is while using initial training set length of 2Min and applying 1Min incremental diarization to the remaining 3Min.

The rest of this paper is as follows: section 2 describes the baseline diarization system. Section 3 introduces the incremental diarization algorithm. Experimentations and results are described in Section 4 and section 5 concludes this study.

2. Baseline Diarization System

Incremental diarization is accomplished by a two stage process. First, a fully unsupervised iterative diarization algorithm is applied over some Initial Training Set (ITS) of the audio stream. In this stage speakers and non-speech models are trained. On the second stage, diarization is applied over increments of the conversation using the models at hand.

A block diagram of the baseline diarization system used during the first stage of the diarization process is given in Figure 1.

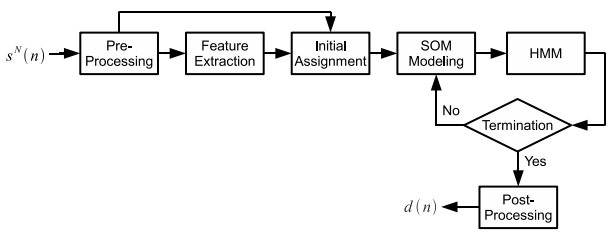

Figure 1: *Baseline diarization system.*

Assume an ITS of N samples from the audio stream $s^N(n)$. The initial training set is first pre-processed using standard pre-

emphasis filter, $f(z) = 1 - 0.95z^{-1}$. Mel Frequency Cepstral Coefficients (MFCC) features are then extracted from the ITS using 20mSec frames with 10mSec overlap between consequent frames. Twelve MFCC features are finally extracted (excluding c0) from each frame.

2.1. First-Stage Initialization

Diarization commences while there is no prior information regarding the speakers or the channel, thus, and initialization algorithm is requires. Such initialization algorithm is suggested in [7], namely Weighted Segmental K-Means initialization (WSKM).

Weighted segmental K-Means algorithm is described in Algorithm 1. Algorithm 1 only states the outline for applying WSKM, for a full description of the algorithm and relative performance comparison to other initialization algorithms, see [7].

Algorithm 1 Weighted Segmental K-Means initial Assignment (WSKMA)

Require: **ITS** $= s^N(n)$, initial training set samples. $\mathbf{O} = \{O_k\}_{k=1}^K$, a set of features extracted from the initial training set **ITS**

1: Perform an initial speech/non-speech segmentation.
2: Mark non-speech segments by $\{NS_j^{l_j}\}_{j=1}^J$ and speech segments by $\{S_i^{l_i}\}_{i=1}^I$ where l_j and l_i are the lengths of the segments such that $\sum_{j=1}^J l_j + \sum_{i=1}^I l_i = N$.
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: Assign a weight $w_i = l_i$ to each of the means $\{SC_i\}_{i=1}^I$.
5: Initialize K-Means centroids, $\{V_r\}_{r=1}^R$
6: Estimate the new centroids using K-Means algorithms such that $V_r^{new} = \frac{\sum_{SC_i \in Cluster_r} w_i SC_i}{\sum_{SC_i \in Cluster_r} w_i}$
7: For all $\{SC_i \in Cluster_r\}_{i=1,...I,r=1,...R}$ assign $\{S_i \in Cluster_r\}_{i=1,...I,r=1,...R}$.

2.2. SOM Based Vector Quantization

Speakers and non-speech models in this study are based on a non-parametric Self Organizing Map (SOM) [9]. Although speakers in the literature are almost always modeled using a statistical model, e.g. a mixture of statistical kernel functions, which is generally a Gaussian Mixture Model (GMM) [10]. For short segments, there might not be sufficient statistical data to train the speakers and non-speech models. Simpler and low cost (by means of required training data) is accomplished using SOM, while generally preserving diarization accuracy. Using SOM models, each speaker is modeled by a Code Book (CB) where each neuron in the CB is a Code Word (CW).

Given a set of feature vectors (observations) $\mathbf{O} = \{O_k\}_{k=1}^K \in \mathbb{R}^d$, an iterative algorithm for SOM training is presented in Algorithm 2.

Once speaker and non-speech model are generated, a distance or distortion measure is required in order to perform a time-series clustering of the data. Distortion measure is achieved through VQ as a likelihood estimator [11].

Having R Code Books (CB) and C Code Words (CW) in each CB, log-likelihood of the data can be estimated under the following assumption: for each CB, $\{CB_r\}_{r=1}^R$ each CW, $\{CW_l\}_{l=1}^L$ is the mean of a Gaussian probability density function (*pdf*) with a unit covariance matrix. The log-likelihood for all of the observations can be estimated for the set of feature vectors $\mathbf{O} = \{O_k\}_{k=1}^K \in \mathbb{R}^d$ and the code-book, $\mathbf{CB_r} = \{CW_r^l\}_{l=1}^L \in \mathbb{R}^d$:

$$L(\mathbf{O}|CB_r) = -\frac{dK}{2}\log(2\pi) - \\ -\sum_{k=1}^{K}(O_k - CW_r^{l^*,n})^T(O_k - CW_r^{l^*,n}) \quad (1)$$

Where

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

Algorithm 2 Self Organizing Map Training

1: Initialization
- Set the size of the CB $\rightarrow C$
- Initialize reference vectors $\mathbf{v^0} = \{v_c^0\}_{c=1}^C$
- Set small and positive learning coefficients α and γ, and the "winner" neuron neighborhood E^j.
- Set the number of SOM training iterations J

2: **for** $j = 1, ..., J$ **do**
3: Randomly choose an observation $O(k_r)$
4: Find the "winner" neuron

$$v_{c^*}^j = \min_c ||O(k_r) - v_c^j||^2 \quad \forall c = 1, ..., C$$

5: Update the "winner" neuron $v_{c^*}^j$ and it's neighbor neurones $E_{c^*}^j$:

$$v_c^{j+1} = v_c^j + \alpha^j\left[O(k_r) - v_c^j\right] \quad i \in E_{c^*}^j$$
$$v_c^{j+1} = v_c^j \quad i \notin E_{c^*}^k$$

6: Decrease the learning coefficient $\alpha^{j+1} = \alpha^j - \epsilon$
7: Decrease the neighborhood radius $E^{j+1} = E^j - \gamma$
8: **end for**

2.3. HMM Time-Series Clustering

Time series clustering in this study is conducted using a modified Hidden Markov Model (HMM).

Hidden Markov Model is a statistical finite-state machine characterized entirely by model parameters $\lambda = (A, B, \pi)$, where A is the state transition probabilities matrix, B states the observation likelihood (emission) matrix and π states the initial probabilities for each state.

Modifications to the HMM described in [12] are required in order to implement a physical restriction over speaker turns. In order to provide some constraints for HMM parameter estimation, a minimum duration τ is enforced over speaker turns, it is assumed that once speaker r has commenced speaking, he/she would continue speaking for at least τ seconds.

Such HMM can be described using a hyper-state transition matrix. Assume each speaker and non-speech forms a hyper-state in a HMM as shown in Figure. 2.

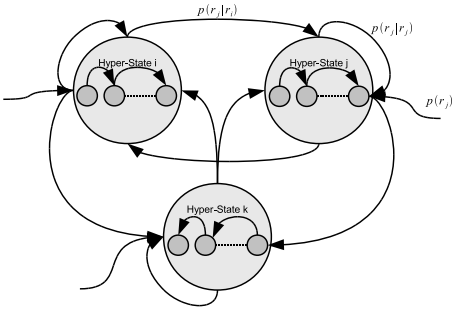

Figure 2: *Hyper state HMM.*

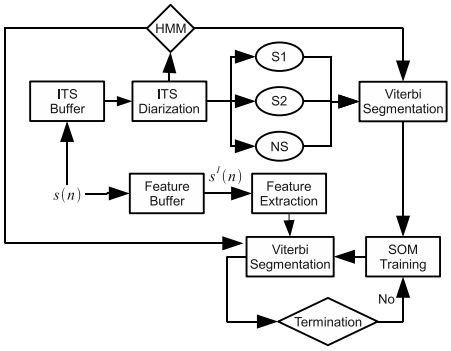

Figure 3: *Incremental diarization system.*

Hyper-state transition matrix A is a block matrix:

$$A = \begin{pmatrix} a_{r_1,r_1} & a_{r_1,r_2} & \cdots & a_{r_1,r_R} \\ a_{r_2,r_1} & a_{r_2,r_2} & \cdots & a_{r_2,r_R} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r_R,r_1} & a_{r_R,r_2} & \cdots & a_{r_R,r_R} \end{pmatrix}_{R\tau \times R\tau} \quad (3)$$

With diagonal elements, $\{a_{r_i,r_i}\}_{i=1}^{R}$ which are hyper-state transition matrices and off-diagonal elements, $\{a_{r_i,r_j}\}_{i=1,j=1,i\neq j}^{R,R}$ which are inter-hyper-state transition matrices.

Construction of the observation matrix B is through the use of SOM as a likelihood estimator discussed in the previous subsection. Initial probabilities are uniformly set as $\frac{1}{R}$.

Given an HMM λ, segmentation of the conversation is accomplished by finding the optimal state transition path through the use of the Viterbi algorithm [12].

Once such an optimal state transition path is found, an adaptation of SOM models is performed. This segmentation and model adaptation process is iterated until converges (which is empirically found to be five iterations).

3. Incremental Diarization

Once the ITS is fully processed, $\mathbf{M} = \{M_r\}_{r=1}^{R+1}$ speaker and non-speech models as well as a tuned HMM are available for the incremental diarization stage.

Given an increment of the conversation $s^i(n)$, segmentation is first applied using the models, $\mathbf{M} = \{M_r\}_{r=1}^{R+1}$, and HMM parameters at hand. Models are then adapted in accordance with the segmentation generated by the Viterbi algorithm (using Algorithm 2), followed by a re-segmentation stage. Three experimentations are conducted, involving one, two and three iterations of this procss.

A block diagram of the incremental diarization system is given in Figure 3.
Where $S1, S2$ and NS are speaker 1, speaker 2 and non-speech models respectively. Note that HMM parameters are not updated following initial estimation and are used for both the first and the second stage of diarization.

4. Experimentations and Results

The suggested incremental diarization was applied over 1802, 5Min length recordings extracted from the NIST 2005 Speaker Recognition Evaluation (SRE) [8]. Recordings are of two speaker conversations recorded using a two microphone channel (4-Wire) at a sampling frequency of 8kHz, the channels are summed and normalized in order to generate a single channel audio stream (2-Wire). Twelfth order MFCC features are extracted from each audio stream (excluding c0).

Three experiments are conducted, each involving an increasing number of incremental diarization iterations ranging from one to three.

4.1. Diarization Error Measurement

Diarization error is generally measured using the Diarization Error Rate (DER) measure as defined by the NIST Rich Transcription evaluation [13]. Diarization error rate measures the fraction of the time not attributed correctly to either one of the speakers or non-speech.
Assume segments in the segmented conversation $\mathbf{C} = \{C_s\}_{s=1}^{S}$, then the DER is measured using equation 4:

$$DER = \frac{\sum_{s=1}^{S} dur(C_s) \cdot (max(N_r(C_s), N_h(C_s)) - N_c(s))}{\sum_{s=1}^{S} dur(C_s) \cdot N_r} \quad (4)$$

Where:

- $N_r(C_s)$ states the number of speakers in segment C_s stated by the reference diarization
- $N_h(C_s)$ states the number of speakers in segment C_s stated by the hypothesized diarization
- $N_c(C_s)$ states the number of speakers in segment C_s that were correctly assigned by the diarization system.

For telephone conversations diarization, only two speakers exist, however, two speakers conversing at once, that is, overlapped speech, must also be taken into account. The suggested diarization system does not currently handle overlapped speech. Segments labeled as overlapped speech by the reference diarization are always in an error state (current diarization system assigns segments to one of two speakers or to non-speech). That is, the error incurred by overlapped speech is added to the overall DER. Moreover, non-speech is also taken as one of the models while evaluating DER.

4.2. Results

Diarization error rate as a function of ITS length and Incremental segment Length (IL) are given in Figure 4 for three learning iterations of the incremental stage. Diarization error rate for ITS of 2Min and IL of 10, 30 and 60Sec are marked in black circles.

The ”Optimal” diarization error achieved using all of the conversation data for diarization is $\sim 20\%$. Incremental di-

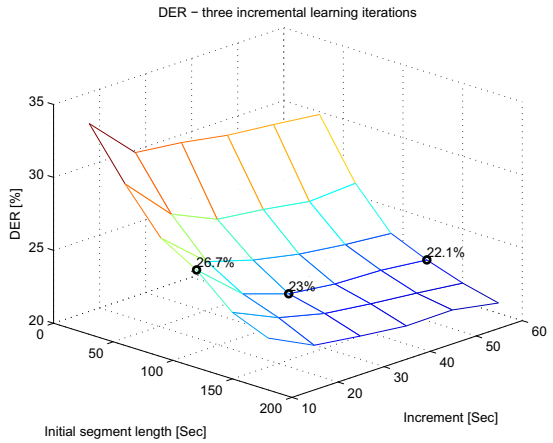

Figure 4: *Incremental diarization results.*

arization seems to approach the suggested "optimal" bound as the ITS is increased. Increment length also affect diarization error; using 30Sec length increments seems a good compromise between resolution and diarization performance.

The impact of learning iterations during the incremental stage can be seen in Table 1 for one, two and three learning iterations using ITS length of 2Min and IL of 10, 30 and 60Sec. For comparison, 23.9% DER was achieved by the diarization system for 120 Sec ITS and no model adaptation.

Table 1: Learning iterations and incremental DER

	10Sec IL	30Sec IL	60Sec IL
1 Iter	24.0%	23.0%	22.7%
2 Iter	25.2%	22.8%	22.5%
3 Iter	26.7%	23.0%	22.1%

It seems that incremental iterations only improves diarization error while given sufficiently long incremental segments. For high resolution incremental diarization, a low number of adaptation iterations are to be used.

Previous experimentations in online diarization system described in [14] provides DER of about 24% while using the models trained from an ITS of 120Sec which were then used to perform diarization over the entire conversation while no further adaptation is applied. It seems that incremental diarization seems a good compromise between on-line and off-line diarization.

5. Conclusion

Incremental diarization is implemented through a two-stage diarization algorithm. In the first stage, unsupervised, iterative diarization is applied over some initial training set extracted from the conversation in order to produce speakers and non-speech models as well as tuned HMM parameters. The second stage of diarization consists of applying the diarization algorithm over increments of the conversation.

Applying the diarization system over 1802 conversations extracted from the NIST 2005 speaker recognition evaluation while using 120Sec ITS length and IL of 60Sec provided 22.1% DER. This is roughly 2.1% higher than the lower bound attained by applying first stage diarization over the entire conversation and 1.9% better than the on-line diarization.

The diarization system suggested does not require any a-priori given information regarding the speakers or the environmental conditions/channel. No other parameter is required to be set prior to the application of the diarization system, these properties makes the suggested diarization system highly robust and scalable.

6. References

[1] V. Gupta, P. Kenny, P. Ouellet, G. Boulianne, and P. Dumouchel, "Combining Gaussianized/Non-Gaussianized Features to Improve Speaker Diarization of Telephone Conversations," *Signal Processing Letters, IEEE*, vol. 14, no. 12, pp. 1040–1043, November 2007.

[2] C. Costin and M. Costin, "New attempts in sound diarization," in *Soft Computing Applications, 2009. SOFA '09. 3rd International Workshop on*, September 2009, pp. 71–76.

[3] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1557–1565, August 2006.

[4] K. Markov and S. Nakamura, "Never-ending learning system for on-line speaker diarization," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, December 2007, pp. 699–704.

[5] D. Lilt and F. Kubala, "Online speaker clustering," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, May 2004, vol. 1, pp. I–333–6 vol.1.

[6] T. Koshinaka, K. Nagatomo, and K. Shinoda, "Online speaker clustering using incremental learning of an ergodic hidden Markov model," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 4093–4096.

[7] O. Ben-Harush, I. Lapidot, and H. Guterman, "Weighted Segmental K-Means Initialization for SOM-Based Speaker Clustering," in *INTERSPEECH 2008*, 2008.

[8] "NIST Speaker Recognition Evaluation, http://www.itl.nist.gov/iad/mig/tests/sre/," .

[9] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, August 2002.

[10] A.P. Dempster, N.M. Laird, D.B. Rubin, and Others, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 138, 1977.

[11] I. Lapidot, "SOM as likelihood estimator for speaker clustering," in *EUROSPEECH 2003*, 2003.

[12] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, August 1989.

[13] "NIST Rich Transcription evaluation, website: http://www.nist.gov/speech/tests/rt," .

[14] O. Ben-Harush, I. Lapidot, and H. Guterman, "Online Diarization of Telephone Conversations," in *Odyssey 2010*, 2010.