An Unsupervised Speaker Clustering Technique based on SOM and I-vectors for Speech Recognition Systems

Hany Ahmed, Mohamed S. Elaraby, Abdullah M. Moussa
Mustafa Abdallah, Sherif M. Abdou, Mohsen Rashwan
RDI, Cairo Egypt
Cairo University, Cairo Egypt
hanyahmed@rdi-eg.com, mhmd.sl.elhady@gmail.com,
a.m.moussa@ieee.org, m.a.elhosiny@eng.cu.edu.eg
{sherif.abdou,mrashwan}@rdi-eg.com

Abstract

In this paper, we introduce an enhancement for speech recognition systems using an unsupervised speaker clustering technique. The proposed technique is mainly based on I-vectors and Self-Organizing Map Neural Network (SOM). The input to the proposed algorithm is a set of speech utterances. For each utterance, we extract 100-dimensional I-vector and then SOM is used to group the utterances to different speakers. In our experiments, we compared our technique with Normalized Cross Likelihood ratio Clustering (NCLR). Results show that the proposed technique reduces the speaker error rate in comparison with NCLR. Finally, we have experimented the effect of speaker clustering on Speaker Adaptive Training (SAT) in a speech recognition system implemented to test the performance of the proposed technique. It was noted that the proposed technique reduced the WER over clustering speakers with NCLR.

1 Introduction

Arabic Automatic Speech Recognition (ASR) is a challenging task, because of the dominance of non-diacritized text material, the several dialects, and the morphological complexity. Another factor that has a negative impact on the advance of Arabic ASR research is the lack of open resources to develop state of the art systems. During recent years, it has been shown that, in large vocabulary speech recognition systems, performance were significantly improved using speaker adaptation. Nowadays, speaker adaptation techniques are crucial in all the advanced speech recognition systems. Speaker adaptation uses data from specific speaker to move the parameters of a speaker-independent system towards a speaker dependent one.

Speaker clustering which is defined as; an unsupervised classification of voiced speech segments based on speaker characteristics (Margarita et al., 2008) is used to boost Speaker Adaptive training in ASR systems. The target of clustering is assigning a unique label to all speech segments uttered by the same speaker.

In recent years, several speaker clustering methods have been proposed, ranging from hierarchical ones, such as the bottom-up methods and the top-down ones, to optimization methods, such as the K-means algorithm and the self-organizing maps. Self-Organizing Map (SOM) is considered as a powerful tool for speaker clustering (Moattar and Homayounpour, 2012).

In this paper, we introduce a fast automatic speaker clustering technique based on SOM and I-Vectors (Dehak et al., 2011) as input features. Our proposed SOM has a feed-forward structure with a single computational layer arranged in 2 dimensions (rows and columns). Assigning correct speaker identification for each utterance can boost the adaptation performance in ASR systems. We have compared our technique with the well-known algorithm Normalized Cross Likelihood Ratio (NCLR). Speaker Clustering using SOM has notably reduced the word error rate of ASR results over both clustering using NCLR and the baseline system (were no speaker clustering performed).

The rest of the paper is organized as follows: Section 2 provides a description of the system used and explains the proposed fast automatic clustering algorithm; Section 3 describes the experimental results. The final conclusions are included in section 4.

Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 79–83,
Valencia, Spain, April 3, 2017.

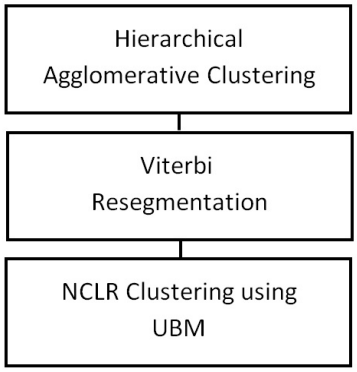

Figure 1: NCLR Block Diagram

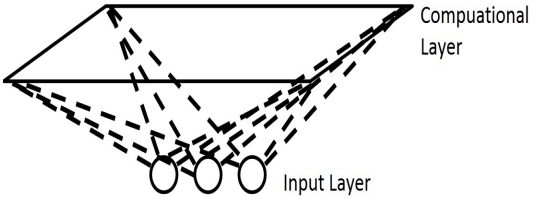

Figure 2: SOM Architecture

2 Speaker Clustering Experiments

2.1 NCLR Speaker Clustering

In order to perform speaker clustering we utilized the technique proposed in (Rouvier et al., 2013). The mentioned system uses the Normalized Cross Likelihood Ratio (NCLR) which is described in (Le et al., 2007) as: $NCLR(M_i, M_j) = \frac{1}{N_j} log(\frac{L(X_i|M_i)}{L(X_i|M_j)}) + \frac{1}{N_j} log(\frac{L(X_j|M_j)}{L(X_j|M_i)})$ The term $\frac{L(X_j|M_j)}{L(X_j|M_i)}$ measures how well speaker model M_i scores with speaker data X_j relative to how well speaker model M_j scores with its own data X_j.

Figure 1 shows the block diagram that describes the clustering system mentioned above which we used in our experiments.

Hierarchical Agglomerative Clustering (HAC): Pre-segmented wave files are fed into HAC system which uses the BIC (Bayesian Information Criterion) measure (Chen and Gopalakrishnan, 1998).

Viterbi Resegmentation: the Viterbi uses GMM trained by Expectation Maximization (EM) to refine speaker boundaries.

NCLR Clustering: speaker models are adapted by a 512 diagonal GMM-UBM system. Afterwards NCLR is used to recombine adapted speaker models.

2.2 SPEAKER CLUSTERING USING SOM

A self-organizing map (SOM) is a type of Neural Networks. It is trained using unsupervised learning algorithm to produce map which is discrete representation of the input training samples. SOMs operate in two main modes: training and mapping. Training builds the map using input training samples, this process is called vector quantization, while mapping classifies a new input vector.

I-vector Extraction: In recent years, many approaches have been proposed to enhance speaker recognition system performance. The most popular are those based on generative models, like Gaussian Mixture Models based on Universal Background Model (GMM-UBM). Other generative models such as Eigen-voices, and the most powerful one, the Joint Factor Analysis (JFA) (Kenny et al., 2008), have been built on the success of the GMM-UBM approach.

Unlike JFA, the idea consists in finding a low dimensional subspace of the GMM super vector space, named the total variability space that represents both speaker and channel variability. The vectors in the low dimensional space are called I-vectors. In 2008 NIST speaker recognition evaluation (Wooters and Huijbregts, 2008), I-vector features were used for the first time. The I-vectors are smaller in size to reduce the execution time of the recognition task while maintaining recognition performance similar to that obtained with JFA.

SOM Clustering: Assigning correct speaker identification for each utterance can boost the SAT adaptation performance in ASR systems. For the offline decoding task, we introduce a fast automatic speaker clustering technique based on SOM and I-Vectors as input features. Our used SOM has a feed-forward structure with a single computational layer arranged in 2 dimensions (rows and columns). Each neuron is fully connected to all the source nodes in the input layer, as shown in Figure 2.

In our experiments, we construct a SOM map in which the number of rows is variable while the number of columns is forced to be 1 column.

For each utterance, a 100 dimension I-vector is calculated and considered as a spatially con-

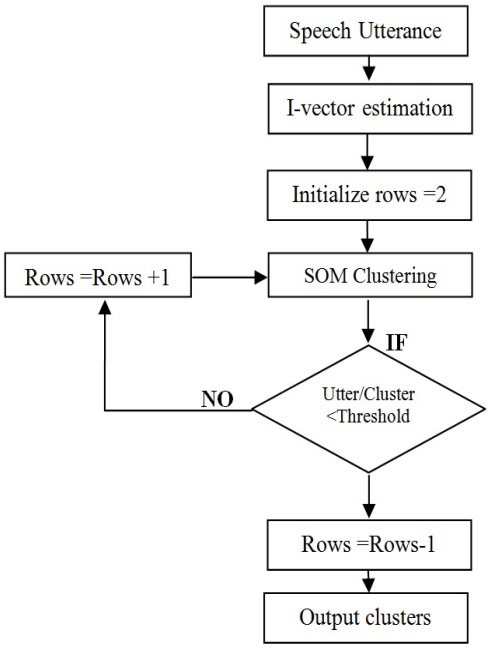

Figure 3: SOM Clustering Flow Chart

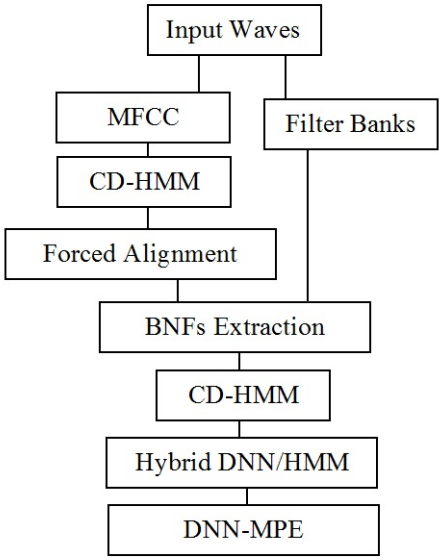

Figure 4: ASR Block Diagram

tinuous feature vector to our clustering technique. Figure 3 describes the flow chart of our proposed method.

2.3 Clustering Results

We have run our experiments on the Development data of the Multi Genre Broadcast MGB-2 challenge described in (Ali et al., 2016). The data consists of 10 hours of Aljazeera TV Broadcast. Table 1 illustrates the results of the NCLR clustering algorithm verses the proposed SOM technique. The metric used to measure the systems' performance is the Speaker Error Rate (SER) defined in (Anguera, 2006) as:

$$SER = \frac{\Sigma_{s=1}^{S}(max(N_{ref(s)},N_{hyp(s)})-N_{correct(s)})}{\Sigma_{s=1}^{S} dur(s).N_{ref}}$$

where S is the total number of segments where both reference and hypothesis segment agree on same speaker and N_{ref} is the number of speakers in segment s.

Table 1 shows SER of the proposed SOM technique verses the NCLR technique.

Metric	**SOM**	**NCLR**
SER	4.96%	5.42%

Table 1: SER of SOM clustering vs SER of NCLR

3 Automatic Speech Recognition (ASR) System

We integrate the results of both Clustering techniques we experimented with Automatic Speech Recognition System (ASR).

3.1 ASR System Components

Figure 4 describes the block diagram of the ASR system.

MFCCs HMM System: Context Dependent HMM CD-HMM is trained over 500 hours of the MGB Challenge training data. The model obtained is used to force align training data.

Bottleneck Features (BNFs) : DNNs have proved to be effective in Acoustic modeling (Hinton et al., 2012). DNNs can be used either as the main classifier for acoustic modeling or as a prior step to extract acoustic features then train the main classifier. We used both mentioned techniques. After aligning all the given waves, a Deep Neural Network consists of five 1500-neuron hidden layers, with a 40 neuron Bottleneck layer, was trained on top of filter banks. 40 dimensional BNFs were extracted from the BNF layer and used to train SAT/HMM-DNN system.

Hybrid DNN-HMM and DNN-MPE: Finally the 40 BNFs are extracted to train a hybrid DNN-HMM followed by the discriminative training for DNN (Vesel et al., 2013).

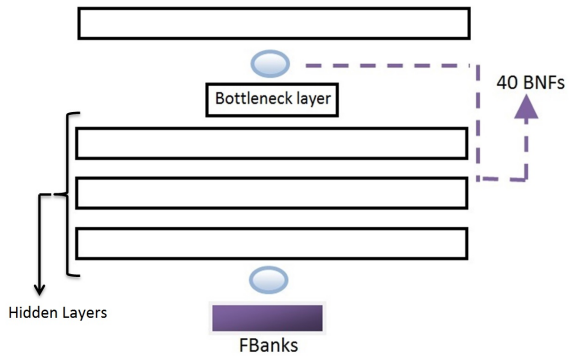

Figure 5: Bottleneck Features Extraction

In our experiments we have used the Kaldi toolkit (Povey et al., 2011) for acoustic modelling.

3.2 Impact of clustering on ASR

In our experiments, it has been proven that using SOM improves the system performance by reducing WER.

Using the SOM to specify new speaker labels for development data and replicating the decoding process on different ASR system, the following enhancements have been verified; First, for the SI-tandem ASR system trained on top of BNFs + fMLLR features, the SOM clustering has given a relative reduction in WER by 3.05% over the tandem BNFs +fMLLR (SI) where all speaker segments per episode were given the same label. Moreover, the mentioned system has given an absolute reduction of 1.16% over the tandem BNFs +fMLLR (SI) integrated with NCLR (SI-NCLR) clustering technique.

Table 2 shows the Speaker Independent (SI) results of the GMM-HMM tandem system trained over BNFs and fMLLR features.

Experiment	**WER**
Tandem (BNFs + fMLLR) (SI)	37.91
Tandem (BNFs + fMLLR) (SI-NCLR)	37.41
Tandem (BNFs + fMLLR) (SI-SOM)	**36.75**

Table 2: Tandem GMM-HMM Speaker-Independent results

Second, for the SD-tandem ASR system trained on top of BNFs + fMLLR features, the SOM clustering has given a relative reduction in WER by 4.22% over the tandem BNFs +fMLLR (SD). In addition, the mentioned system has given an absolute reduction of 1.52% over the tandem BNFs +fMLLR (SD) (SD-NCLR).

Table 3 shows the Speaker Dependent (SD) results of the GMM-HMM tandem system over BNFs and fMLLR features.

Experiment	**WER**
Tandem (BNFs + fMLLR) (SD)	36.00
Tandem (BNFs + fMLLR) (SD-NCLR)	34.96
Tandem (BNFs + fMLLR) (SD-SOM)	**34.48**

Table 3: Tandem GMM-HMM Speaker-Dependent results

Third, for the hybrid DNN/HMM system trained on top of fMLLR +BNFs with Sequence Discriminative training criterion (DNN/HMM-MPE), the SOM clustering gave a relative reduction of 3.84% in WER of the hybrid system that used no clustering technique. In comparison with the hybrid system where NCLR clustering was applied, the SOM gave a relative reduction in WER of 1.87%.

Table 4 shows the final results of the hybrid DNN-HMM trained with Minimum phoneme error rate criterion (MPE).

Clustering Technique	**WER**
Tandem DNN/HMM-MPE (SD)	27.8
Tandem DNN/HMM-MPE (SD-NCLR)	27.24
Tandem DNN/HMM-MPE (SD-SOM)	**26.73**

Table 4: Hybrid DNN/HMM-MPE results

It is noticeable that the performance of the hybrid system improved after using the proposed clustering technique.

4 Conclusion

In this paper, we have proposed an algorithm for automatic speaker clustering based on Self-Organizing Map. The performance of the new algorithm has been compared with a well-known technique of speaker clustering (Normalized Cross Likelihood Ratio). The experimental results on Multi Genre Broadcast data have shown noticeable reduction in Speaker Error Rate. The clustering algorithm has been integrated with state of art Automatic Speech Recognition techniques to boost Speaker adaptive training performance. It

is experimentally verified that the proposed technique achieved notable reduction in word error rate compared to the traditional tandem system. In addition, the proposed algorithm attained a reduction in word error rate in comparison with the reduction attained by NCLR clustering technique.

Acknowledgments

The authors would like to show their gratitude to Professor Hassanin M. Al-Barhamtoshy of King Abdulaziz University (KAU) for his support in the project. This research was done by the computational power support provided by King Abdulaziz University.

References

A. Ali, P. Bell, J. Glass, Y. Messaoui, H. Mubarak, S. Renals, and Y. Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. *Proc. Spoken Language Technology Workshop (SLT)*.

X. Anguera. 2006. *Robust speaker diarization for meetings*. Ph.D. thesis, Universitat Politcnica de Catalunya.

S. Chen and P. Gopalakrishnan. 1998. Speaker, environment and channel change detection and clustering via the bayesian information criterion. *Proc. of DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, Virginia, USA, February,pp. 127132*.

N. Dehak, P. Kenny, R. Dehak, and P. Ouellet. 2011. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio Speech and Language Processing 19(4):788 - 798*.

G. Hinton, D. Yu L. Deng, G. E Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N Sainath, and B. Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine 29, 8297*.

P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. 2008. A study of inter-speaker variability in speaker verification. *IEEE Transactions on Audio, Speech and Language Processing 16(5):980 - 988*.

V. Le, O. Mella, and D. Fohr. 2007. Speaker diarization using normalized cross likelihood ratio. *Interspeech (Vol. 7, pp. 1869-1872).*

K. Margarita, V. Moschou, and C. Kotropoulos. 2008. Speaker segmentation and clustering. *Signal processing 88(5), 1091-1124.*

M. H. Moattar and M. M. Homayounpour. 2012. A review on speaker diarization systems and approaches. *Speech Commun. 54, 10651103*.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz. 2011. The kaldi speech recognition toolkit. *IEEE 2011 workshop on automatic speech recognition and understanding IEEE Signal Processing Society,*.

M. Rouvier, G. Dupuy, E. Khoury P. Gay, T. Merlin, and S. Meignier. 2013. An open-source state-of-the-art toolbox for broadcast news diarization. *Interspeech, Lyon (France), 25-29 Aug*.

K. Vesel, A. Ghoshal, L. Burget, and D. Povey. 2013. Sequence-discriminative training of deep neural networks. *Interspeech (pp. 2345-2349)*.

C. Wooters and M. Huijbregts. 2008. The icsi rt07s speaker diarization system. *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*.