# ENVIRONMENT AWARE SPEAKER DIARIZATION FOR MOVING TARGETS USING PARALLEL DNN-BASED RECOGNIZERS

*Maryam Najafian* [1]*, John H. L. Hansen*

Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX, USA
m.najafian@utdallas.edu[1], john.hansen@utdallas.edu

## ABSTRACT

Current diarization algorithms are commonly applied to the outputs of single non-moving microphones. They do not explicitly identify the content of overlapped segments from multiple speakers or acoustic events. This paper presents an acoustic environment aware child-adult diarization applied to the audio recorded by a single microphone attached to moving targets under realistic high noise conditions. The proposed system exploits a parallel deep neural network and hidden Markov model based approach which enables tracking of rapid turn changes in audio segments as well as capturing the cross talk labels for overlapped speech. It outperforms the state-of-the-art diarization systems without the need to prior clustering or front-end speech activity detection. [1]

***Index Terms***— Speaker Diarization, Acoustic Scene Analysis, Overlapped Speech, Deep Neural Networks

## 1. INTRODUCTION

Speaker diarization play an important role in speech technology. Broadcast news speaker diarization is composed of 5 steps. First, non-speech regions are removed using Viterbi decoding. Then, an acoustic segmentation followed by a Hierarchical Agglomerative Clustering (HAC) splits and then groups the signal into homogeneous parts according to speakers and background [1]. Next, a Gaussian Mixture Model (GMM) is trained for each cluster via the Expectation-Maximization (EM) algorithm. The signal is then re-segmented through a Viterbi decoding. The system finally performs another HAC, using the Cross-Likelihood Ratio (CLR) [2] measure and GMMs trained with the Maximum A Posteriori algorithm (MAP). Using this diarization routine, several broadcast news and meeting diarization systems have been proposed in the literature [3, 4]. Recently major improvements have been reported as a result of using i-vectors [5], bottleneck features [6], Deep Neural Network (DNN) models [7, 8], energy features [9, 10] and cluster-voting [11] for different classification purposes. Such systems can be applied as a preliminary step in Automatic Speech Recognition (ASR), when there is a need for localization of speaker segment labels. Additionally, detection of speaker specific segments enables various acoustic model adaptations for ASR which leads to a higher accuracy [12, 13, 14].

The aim of this research is to identify whether at each point in time the audio which is recorded by a moving target (child) belongs to one of the following acoustic categories: (1) primary child, (2) secondary child, (3) adult, (4) music, (5) silence and (6) crowd noise. It relies on the information captured at feature and acoustic level which helps with identification of child from adult speech [15, 16], as well as speech from non-speech. Robust speaker diarization is challenging in realistic audio environments due to time-varying acoustic noise and overlapped speech occurred during the target's movement, competing speakers, and reverberations.

This paper presents an asynchronous parallel DNN-HMM based speaker diarization system in Section 3.1, a state-of-the-art GMM-HMM diarization in Section 3.2, and a separate Threshold Optimized Combo Speech Activity Detection (TO-Combo-SAD) stage for both systems in Section 3.3. Section 4 compares the experimental results across these four systems, and in Section 5 the main conclusions are summarized.

## 2. DATA DESCRIPTION

In this study we used 7.2 hours of labeled audio recording gathered from 30 children wearing a recording unit within a childcare center [17] which is larger than that of the study carried out previously in [18, 19]. The labels are as following: primary child: speech initiated by the child wearing the recording unit; secondary child: speech originated by other children and directed at the primary child within his/her close proximity (around 4 feet); adult: speech originated by an adult and directed at the primary child within his/her close proximity; non-speech: the stream of (1) silence, (2) music, and (3) crowd noise. The average turn duration in the database is 1.7 seconds. From the manual labels, we estimated that 33%, 24%, 23%, and 20% of our speech database belongs to non-speech, adult speech, secondary child speech, and primary child speech categories respectively. In our experiments, to obtain unbiased evaluation results, we performed 5-fold cross validation and divide the data into training, validation, and test sets such that no speaker appeared simultaneously in training, validation, and test sets.

[1]This work was conducted at CRSS-UTDallas with support from the Univ. of Kentucky. Dr. Maryam Najafian was with CRSS-UTDallas when this work was conducted. **Dr. Maryam Najafian has since moved on to the Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA – najafian@csail.mit.edu**

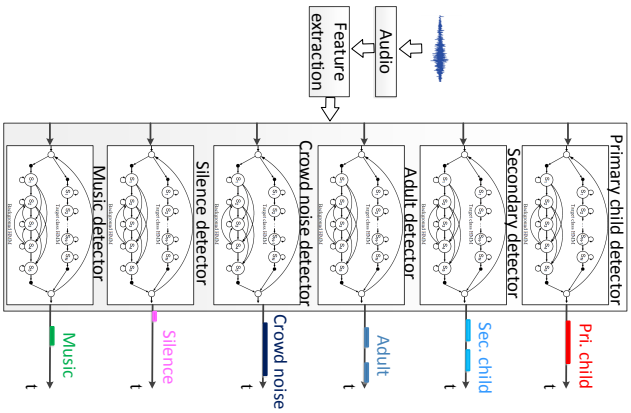


**Fig. 1**: Diarization structure

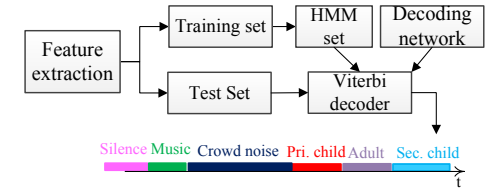


**Fig. 2**: A single detector diagram

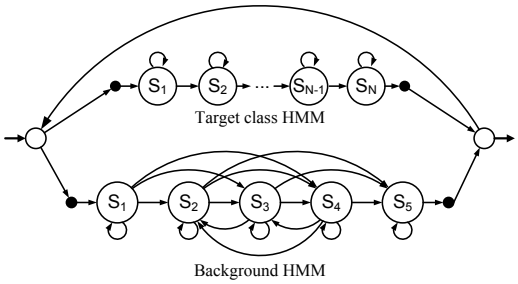


**Fig. 3**: Detector structure

## 3. SYSTEM DESCRIPTION

In this section we start by describing our proposed parallel DNN-HMM based speaker independent diarization system. Then, we present a state-of-the-art GMM-HMM based diarization toolkit which relies on the bottom-up clustering.

### 3.1. Proposed parallel DNN-HMM diarization

Figure 1 shows the a parallel array of six asynchronous HMM based detectors for identification of different sources of audio variability that might occur at overlapping time intervals. This idea was first proposed by Nabiei et. al [20] at the University of Birmingham for real-time overlapped human action recognition purpose. At any point in time the parallel structure of the detectors, enables successful detection of overlapping occurrences of six sources variability.

The Markov based acoustic modelling process, enables capture of rapid changes by modelling the time varying structure of the audio signal, plus a mechanism to relate acoustic feature vectors to Markov model states. The key process in our HMM-based detector system relies on a Viterbi decoder. Given a sequence of filter bank feature vectors $Y$ the Viterbi decoder finds the sequence of HMMs $M$ such that an approximation to the probability $p(M|Y)$ is maximized. Since $Y$ is fixed from Bayes' rule this is equivalent to finding $M$ such that $p(Y|M)P(M)$ is maximized. Since, there is no constraint on the sequence of occurrences of the target and background the probability $P(M)$ represents an open loop context free target-background language model. Using this configuration a single network is designed and the most probable path through this network is found using Viterbi decoding [20].

As shown in Figure 2, during the training stage the HMM set model parameters for each detector is trained separately for each acoustic category. During the testing stage, the Viterbi decoder unit uses the model network shown in Figure 3 to recognize occurrences of different acoustic categories within an utterance. Briefly, our Viterbi decoder works as follows [21]: At each time $t$ the decoder receives a new feature vector, $y_t$. For each state $i$ of each of its HMMs, a quantity $\alpha_t(i)$ is calculated which can be thought of as an approximation to the probability of the best explanation of data $Y = y_1, ..., y_t$ up to and including $y_t$ ending in state $i$ at time $t$. Intuitively, if the decoder is for 'primary child speech' and the $i^{th}$ state corresponds to that, then $\alpha_t(i)$ can be thought of as the probability of the best explanation of data up to time $t$. Formally $\alpha_t(i)$ is given by the recursion:

$$\alpha_t(i) = max_j \alpha_{t-1}(j) a_{j,i} b_i(y_t) \quad (1)$$

$$\rho_t(i) = argmax_j \rho_{t-1}(j) a_{j,i} b_i(y_t) \quad (2)$$

where $a_{j,i}$ is the probability of a transition from state $j$ to state $i$ and $b_i(y_t)$ is the probability of the data $y_t$ given state $i$. Note that the 'preceding' state $j$ can be in the same HMM as state $i$, or, if $i$ is an initial state, $j$ can be the final state of another HMM in the decoder. $\rho_t(i)$ provides a record from which the best explanation of the data up to time $t$ in state $i$ can be recovered. The decoding network is represented by a loop across the target and background acoustic models which simply allows any model to follow any other model. Figure 3, shows the structure of the recognition network used in our system. The insertion penalty is applied on the transition from the end of one model to the start of the next.

As shown in Figure 3, each detector has a single N-state complex left-to-right HMM for modeling one specific acoustic category and a five-state complex left-to-right background HMM, which represents any data except those used by the target model. The inputs to this detector are filter bank features and the output detects whether the observed features belong to a specific acoustic class or they belong to the background model. All HMM states are associated with DNNs.

The pre-training of the DBN (i.e. Deep Belief Network, stack of Restricted Boltzmann Machines (RBMs)) is completed, using the Contrastive Divergence (CD) [22] algorithm with 1-step of Markov chain Monte Carlo sampling [23]. The first layer and the following layers of RBMs are composed of Gaussian-Bernoulli and Bernoulli-Bernouli units, respectively. The generative pre-training of the DNNs is carried out through training of a stack of RBMs. Pre-training provides a better generalization from the training data, and helps prevent falling into local minima during the fine-tuning [24].

After pre-training, a softmax layer, which contains all the training state probabilities, is added on top of the stack of the RBMs to form a pre-trained DNN. During the network fine-tuning stage the network parameters are updated by applying the error back-propagation and Stochastic Gradient Descent (SGD) [25] algorithm. In order to perform the fine-tuning, initially a GMM-HMM system needs to be trained on the feature vectors, with $K$ GMMs per state to provide a frame level alignment which will be used for minimizing the per-frame cross-entropy between the acoustic labels and the network output. During the DNN fine-tune, per-frame cross-entropy between the HMM state target posterior probabilities and network output is minimized, using mini-batch SGD.

Our DNN-HMM system is built using Kaldi [26, 27, 28] with the validation set data, the optimum number of states $N \in \{3, 4, ..., 10\}$ for each acoustic category HMM was empirically established. Initially GMM-HMMs are trained on 39 dimensional mean/variance normalized MFCCs, with 6 Gaussians per state. It results in the alignment of acoustic category states to frames for the DNN-HMM system which has 13 dimensional mean/variance normalized Mel filterbank features, spliced using a context of $\pm 10$ frames. This network uses sigmoid activation function, and 4 hidden layers (computed empirically using the validation set). Each layer within the network contains $l = 128$ hidden units (neurones) which was chosen from different values of $l \in \{64, 128, 256, 512\}$ empirically. Also, the learning rate is set to be 0.008, and the number of epochs in the pre-training process is set to be 5.

### 3.2. GMM-HMM based diarization using LIUM toolkit

In this section we describe the GMM based system designed for our application using the LIUM speaker diarization toolkit [29]. This diarization system is composed of acoustic BIC segmentation followed by BIC [30] hierarchical agglomerative clustering. The Universal Background Model (UBM) is adapted (Maximum A Posteriori MAP) for each cluster. A cluster is modeled by a GMM-HMM with 8 components. The system uses a Normalized Cross Likelihood Ratio (NCLR [31]) based on bottom-up clustering. Viterbi decoding is performed to adjust the segment boundaries. Non-speech regions are removed using GMMs. A hierarchical clustering for different acoustic classes is carried out over the clusters generated by the Viterbi decoding. The primary child, secondary child, adult classes are recognized by MAP adapting the UBM to each of these classes. The inputs to this system are 13 MFCCs with coefficient C0 as energy completed by delta coefficients. In order to identify and remove music, the audio is segmented into speech and non-speech regions using a Viterbi decoding with 5 one state HMMs, comprising of 1 model of silence, 2 models of speech (clean, over crowd noise), 1 model of crowd noise, and 1 model of music.

### 3.3. TO-Combo-SAD speech and non-speech detection prior to DNN/GMM based diarization

TO-Combo-SAD relies on several noise robust features that are computed at a frame level for each audio segment and the combined feature vectors are projected into a single dimension (by using Principal Component Analysis) for the speech and non-speech discrimination task [32]. This feature is efficiently obtained from a linear combination of the voicing measures, namely harmonicity, clarity, prediction gain, and periodicity. In recent studies applying a separate stage of SAD using these features was reported advantageous in classifying speech and non-speech segments using Parallel Linear Discriminant Analysis (PLDA) classifier [18, 19]. In Section 4 we compare the effectiveness of this feature level approach with GMM- and DNN-HMM based speech/non-speech classification stage introduced in Sections 3.1 and 3.2.

## 4. EXPERIMENTAL RESULTS AND DISCUSSION

The Diarization Error Rate (DER) in Equation 3, where $FA$ is the total number of non-speech frames detected as speech, $MIS$ is the total number of speech frames detected as non-speech, $ERR$ is the total number of speech frames that were detected as speech but recognized as incorrect classes, and $TOT$ is the total number of frames.

$$DER(\%) = (FA + MIS + ERR) * 100/TOT \quad (3)$$

The GMM-HMM diarization system is the baseline diarization system (Table 1). Our database was recorded in a noisy childcare center environment, which explains why a large number (18%) of secondary child data are falsely recognized as non-speech which contains a lot of distant child speech in the background . Also due to generative nature of GMM based system, a large number of classification errors occurred between the non-speech and different speech classes as well as the two primary and secondary child classes.

Table 2 compare the effectiveness of the (1) GMM, (2) feature level TO-Combo-SAD, and (3) DNN based speech and non-speech classification in four systems described in section 3. In both $B$ and $C$ systems we use TO-Combo-SAD at the first stage to discriminate between speech and non-speech segments, and this results in 9.69% $FA$ and 16.96% $MISS$ relative error reduction compared to the baseline system $A$. System $D$ comprises of unique recognizer for dif-

**Table 1**: *Confusion matrix for baseline GMM-HMM diarization*

| Baseline GMM-HMM (%) error rate, section 3.2 | Total | Adult | Prim.child | Sec.child | Non-speach |
|---|---|---|---|---|---|
| Adult | 17.45 | - | 3.04 | 5.45 | 8.95 |
| Prim. child | 21.90 | 3.60 | - | 6.90 | 11.40 |
| Sec. child | 28.18 | 4.16 | 6.06 | - | 18.00 |
| Non-speech | 30.57 | 10.38 | 8.26 | 11.92 | - |

**Table 2**: *Percentage error rate for all four systems*

| Codes | Description | Systems | FA | MISS | ERR | DER |
|---|---|---|---|---|---|---|
| *A* | Section 3.2 | GMM-HMM (baseline) | 22.08 | 8.31 | 5.97 | 36.38 |
| *B* | Section 3.3 | TO-Combo-SAD + GMM-HMM | 19.94 | 6.90 | 5.72 | 32.56 |
| *C* | Section 3.3 | TO-Combo-SAD + DNN-HMM | 19.94 | 6.90 | 3.75 | 30.59 |
| *D* | Section 3.1 | DNN-HMM | 15.38 | 3.94 | 3.54 | 22.87 |

ferent non-speech events, such as crowd noise, music, and silence as well as speech events, such as child and adult speech. Next, we compare the effect of using TO-Combo-SAD prior to DNN-HMM diarization in system $C$ with that of a parallel DNN-HMM in system $D$ for speech and non-speech classification. The results in Table 2 show that system $D$ obtains a further relative reduction of 22.86% $FA$, 42.89% $MISS$, 25.23% DER compared to system $C$.

A comparison between the confusion matrices for the proposed parallel DNN-HMM and baseline GMM-HMM diarization systems shows that using a discriminative rather than a generative modeling approach has led to a total of 41.40%, 49.31%, and 52.34% relative error reduction for adult, primary child, and secondary child classes (rows 2 to 4, column 2 of Tables 1 and 3). Moreover, considerable relative error reduction of 56.7% and 52.43% is obtained as a result of a better discrimination between the secondary child and the primary child and adult segments (row 3, columns 3 and 5). In addition to that, a comparison between the speech classes detected as non-speech (column 6) shows a major error rate reduction of 48.73%, 59.64%, and 52.43% for adult, primary and secondary child classes compared to the baseline.

All in all, the proposed system achieved high diarization accuracy and it is capable of detecting overlapped speech. Since there were not enough examples of overlapped speech we were not able to train a separate model for the overlapped speech in the GMM-HMM diarization baseline system. We had a total of 14 minutes containing examples of overlapped speech from the adult, primary and secondary classes. The proposed DNN-HMM based system managed to identify the different classes involved in overlapped speech with 64.7% diarization accuracy. From these 35.3% DERs, the total of 23.5% was due to miss classification of primary child as secondary child and visa versa, the rest was due miss-classification between adult and child classes.

## 5. CONCLUSIONS

This paper describes a parallel DNN-HMM based diarization system for the data recorded from a moving target under a high noise condition. The proposed system attempts to address five main challenges. The direct application of our system is in diarization tasks containing overlapped speech from multiple resources, under noisy condition, where the number of classes are known and each class represent a group of speakers based on their age or distance from the microphone.

(1) The overlapped speech problem is addressed using a parallel set of independent recognizers for each audio class enables the system to identify whether the audio segment belongs to multiple classes (Sources of cross talk).

(2) The speech activity detection under high background noise during child or adult speech segments is another main challenge. This is addressed by allocating a unique model with complex HMM structure to different non-speech classes such as crowd noise, silence, and music. This has led to major reduction in speech and non-speech miss-classification (30.31% $FA$ and 52.58% $MIS$ error reduction).

(3) To track and detect rapid turn takings among speakers we used parallel HMMs which are capable of detecting occurrences of different speech and non-speech events simultaneously; in addition to that our technique doesn't rely on expensive agglomerative cluster merging and retraining which is unable to effectively capture rapid speaker turn takings and requires an additional stage of Viterbi alignment.

(4) To reduce the errors occurred during the feature modeling stage, we used a discriminative rather than a generative modelling strategy by replacing GMMs with DNNs during the estimation of the HMM state output probabilities. Using the proposed system has resulted in 37.11% relative DER reduction across all groups.

(5) Instead of relying on TO-Combo-SAD or other low dimensional features (e.g. energy, zero-crossing rate, periodicity and formant information) that don't perform well under high time varying crowd noise conditions, we use acoustic level models of the non-speech distribution which can generalize well after training. In a DNN-HMM based system using a parallel speech vs non-speech recognizers rather that TO-Combo-SAD has resulted in 25.24% relative DER reduction.

**Table 3**: *Confusion matrix for proposed DNN-HMM diarization*

| Proposed DNN-HMM (%) error rate, section 3.1 | Total | Adult | Prim.child | Sec.child | Non-speach |
|---|---|---|---|---|---|
| Adult | 10.22 | - | 1.86 | 3.77 | 4.59 |
| Prim. child | 11.1 | 2.60 | - | 3.90 | 4.6 |
| Sec. child | 12.68 | 1.52 | 2.62 | - | 8.56 |
| Non-speech | 21.30 | 7.28 | 5.38 | 8.61 | - |

## 6. ACKNOWLEDGMENT

Authors wish to thank our collaborators, Dr. Beth S. Rous, Dr. Dwight Irvin, and Ying Luo, at the University of Kentucky for their financial support as well as work in establishing the protocol and collecting and organizing the child database used in this study.

## 7. REFERENCES


[1] Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Luc Gauvain, "Multistage speaker diarization of broadcast news," *ASLP*, vol. 14, no. 5, pp. 1505–1512, 2006.

[2] Douglas A Reynolds, Elliot Singer, Beth A Carlson, Gerald C O'Leary, Jack McLaughlin, and Marc A Zissman, "Blind clustering of speech utterances based on speaker and language characteristics.," in *ICSLP*, 1998.

[3] Xavier Anguera Miro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker diarization: A review of recent research," *ASLP, IEEE Transactions on*, vol. 20, no. 2, pp. 356–370, 2012.

[4] Deepu Vijayasenan and Fabio Valente, "Diartk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings.," in *INTERSPEECH*, 2012, pp. 2170–2173.

[5] Gregory Sell, Alan McCree, and Daniel Garcia-Romero, "Priors for speaker counting and diarization with ahc," *INTERSPEECH*, pp. 2194–2198, 2016.

[6] Linxue Bai, Peter Jancovic, Martin Russell, and Philip Weber, "Analysis of a low-dimensional bottleneck neural network representation of speech for modelling speech dynamics," *INTERSPEECH*, 2015.

[7] Rosanna Milner and Thomas Hain, "DNN-Based Speaker Clustering for Speaker Diarisation," *INTERSPEECH*, pp. 2185–2189, 2016.

[8] Hongjun Su, Shufang Tian, Yue Cai, Yehua Sheng, Chen Chen, and Maryam Najafian, "Optimized extreme learning machine for urban land cover classification using hyperspectral imagery," *Frontiers of Earth Science*, pp. 1–9, 2016.

[9] Jinyi Zou, Wei Li, Chen Chen, and Qian Du, "Scene classification using local and global features with collaborative representation fusion," *Information Sciences*, vol. 348, 2016.

[10] Mengyuan Liu, Hong Liu, Chen Chen, and Maryam Najafian, "Energy-based global ternary image for action recognition using sole depth sequences," *3D Vision*, 2016.

[11] Houman Ghaemmaghami, David Dean, and Sridha Sridharan, "A cluster-voting approach for speaker diarization and linking of Australian broadcast news recordings," in *ICASSP*. IEEE, 2015, pp. 4829–4833.

[12] M. Najafian and M. Russell, "Modelling accents for automatic speech recognition," *EUSIPCO*, 2015.

[13] M. Najafian, S. Safavi, A. Hanani, and M. Russell, "Acoustic model selection using limited data for accent robust speech recognition," *EUSIPCO*, pp. 1786–1790, 2014.

[14] M. Najafian, A. DeMarco, S. Cox, and M. Russell, "Unsupervised model selection for recognition of regional accented speech," *INTERSPEECH*, pp. 2967–2971, 2014.

[15] Saeid Safavi, Maryam Najafian, Abualsoud Hanani, Martin Russell, and Peter Jančovič, "Comparison of speaker verification performance for adult and child speech," in *WOCCI*, 2014.

[16] Saeid Safavi, Maryam Najafian, Abualsoud Hanani, Martin J Russell, Peter Jancovic, and Michael J Carey, "Speaker recognition for children's speech," in *INTERSPEECH*, 2012, pp. 1836–1839.

[17] Maryam Najafian and John HL Hansen, "Speaker independent diarization for child language environment analysis using deep neural networks," *SLT*, pp. 1–7, 2016.

[18] Maryam Najafian, Saeid Safavi, Philip Weber, and Martin Russell, "Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems," in *ODYSSEY*, 2016, pp. 132–139.

[19] Maryam Najafian, Dwight Irvin, Ying Luo, Beth S Rous, and John HL Hansen, "Employing speech and location information for automatic assessment of child language environments," *SPLINE*, pp. 65–69, 2016.

[20] Roozbeh Nabiei, Maryam Najafian, Manish Parekh, Peter Jancovic, and Martin Russell, "Delay reduction in real-time recognition of human activity for stroke rehabilitation," *SPLINE*, pp. 70–74, 2016.

[21] Mark Gales and Steve Young, "The application of hidden markov models in speech recognition," *Foundations and trends in signal processing*, vol. 1, no. 3, pp. 195–304, 2008.

[22] Geoffrey E Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[23] Geoffrey E. Hinton, "A practical guide to training restricted Boltzmann machines," in *Neural Networks: Tricks of the Trade - Second Edition*, pp. 599–619. 2012.

[24] Ruslan Salakhutdinov and Geoffrey E Hinton, "Deep Boltzmann machines," in *AIS*, 2009, pp. 448–455.

[25] Léon Bottou, "Online learning and stochastic approximations," *Online learning in neural networks*, p. 25, 1998.

[26] George Saon and Hagen Soltau, "A comparison of two optimization techniques for sequence discriminative training of deep neural networks," in *ICASSP*, 2014, pp. 5567–5571.

[27] Maryam Najafian, Saeid Safavi, John HL Hansen, and Martin Russell, "Improving speech recognition using limited accent diverse British English training data with deep neural networks," *MLSP*, 2016.

[28] Maryam Najafian, *Acoustic model selection for recognition of regional accented speech*, Ph.D. thesis, University of Birmingham, 2016.

[29] Sylvain Meignier and Teva Merlin, "LIUM SpkDiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, 2010.

[30] Matthew A Siegler, Uday Jain, Bhiksha Raj, and Richard M Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *DARPA*, 1997.

[31] Viet Bac Le, Odile Mella, Dominique Fohr, et al., "Speaker diarization using normalized cross likelihood ratio.," in *INTERSPEECH*, 2007, vol. 7, pp. 1869–1872.

[32] Seyed Omid Sadjadi and John HL Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *Signal Processing Letters, IEEE*, vol. 20, no. 3, pp. 197–200, 2013.