SPEAKER DIARIZATION USING DEEP NEURAL NETWORK EMBEDDINGS

Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree

Human Language Technology Center of Excellence & Center for Language and Speech Processing
The Johns Hopkins University, Baltimore, MD 21218, USA

ABSTRACT

Speaker diarization is an important front-end for many speech technologies in the presence of multiple speakers, but current methods that employ i-vector clustering for short segments of speech are potentially too cumbersome and costly for the front-end role. In this work, we propose an alternative approach for learning representations via deep neural networks to remove the i-vector extraction process from the pipeline entirely. The proposed architecture simultaneously learns a fixed-dimensional embedding for acoustic segments of variable length and a scoring function for measuring the likelihood that the segments originated from the same or different speakers. Through tests on the CALLHOME conversational telephone speech corpus, we demonstrate that, in addition to streamlining the diarization architecture, the proposed system matches or exceeds the performance of state-of-the-art baselines. We also show that, though this approach does not respond as well to unsupervised calibration strategies as previous systems, the incorporation of well-founded speaker priors sufficiently mitigates this shortcoming.

Index Terms— Speaker diarization, deep neural networks, clustering, end-to-end learning

1. INTRODUCTION

Speaker diarization is the task of grouping segments of speech according to the speaker. It is often summarized as "who is speaking when". Since many speech processing technologies, such as in automatic speech recognition or speaker recognition, assume the presence of only one speaker, diarization can be an important front end in scenarios where the single-speaker assumption can be violated.

Many recent advances in diarization have involved extracting i-vectors from short segments of speech [1, 2, 3, 4, 5]. This is a sensible approach given the success of i-vectors for speaker recognition. Recent i-vector architectures [5] involve two disjoint generative processes: one to extract the i-vectors, and a second one to learn a probabilistic linear discriminant analysis (PLDA) scoring function [6] to decide whether two i-vectors are from the same speaker or not. The extraction of i-vectors requires a Gaussian mixture model and a factor analysis that uses a large projection matrix $\mathbf{T}$. After the i-vectors are extracted, an independent generative model is used to learn the PLDA scoring.

In this paper, we propose replacing this two-step generative process with a discriminatively trained deep neural network (DNN) that jointly learns a fixed-dimensional embedding and a scoring metric. This particular architecture has recently been shown to be effective for speaker recognition [7]. Furthermore, the results in [7] suggest that the learned embeddings are more effective relative to traditional i-vectors for shorter durations of speech. This is highly desirable for the dense segmentations used in speaker diarization (2 second segments).

After presenting a brief background on diarization, we will outline the specific architecture in our proposed diarization system. Then, we will present the results of experiments on the CALLHOME corpus, and compare the proposed system to state-of-the-art acoustic and senone i-vector diarization algorithms. The results will show that, in addition to being much simpler, the proposed method matches or exceeds the performance of the state-of-the-art baselines.

2. BACKGROUND

I-vectors were applied to speaker diarization shortly after their development for speaker recognition [8], and progress has been consistent in the years since. Early work utilizing i-vectors for segmented speech scored similarities between blocks using cosine scoring and clustered with K-means or spectral clustering [1, 2]. Other clustering algorithms on i-vectors for diarization have included Variational Bayesian GMMs [3], mean shift [4], and agglomerative hierarchical clustering (AHC) [9, 5]. The work that follows will also use AHC.

It has also recently been shown that cosine scoring can be outperformed by PLDA [5], and that the error of that calibration estimation can be reduced by incorporating speaker priors into the AHC clustering [10]. The traditional unsupervised GMM-UBM was also recently replaced with the senone partitions from a trained DNN [11], which yielded further improvements.

One shortcoming of the segmentation-based approaches above is that the resulting diarization marks will be restricted to begin and end according to the segmentation boundaries. To remedy this, a second stage of diarization, often called resegmentation, can be added. In resegmentation, the results of the clustering are used to initialize a frame-level diarization system that then iterates to refine the boundaries of speaker turns. Previously, most resegmentation was performed in the acoustic feature space with a Hidden Markov Model (HMM), but recent work has shown that resegmentation can be more effective using subspace techniques [12].

In this work, we propose a discriminatively trained DNN that jointly learns a fixed-dimensional embedding and a scoring metric. This strategy was recently applied to speaker recognition with promising results [7]. A conceptually similar DNN embedding recently showed the value of discriminative training for unsupervised speech separation of multiple overlapping speakers [13]. Like our proposed system, a DNN was trained to provide features for unsupervised clustering, resulting in a process called deep clustering. Unlike our proposed approach, deep clustering operates on each time-frequency bin with recurrent neural networks (RNNs) and learns embeddings for those bins with desirable behavior in the Euclidean space. The following section will describe our proposed system in detail.

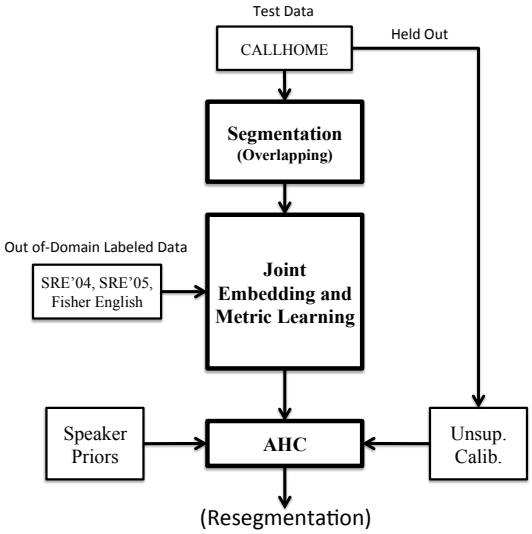
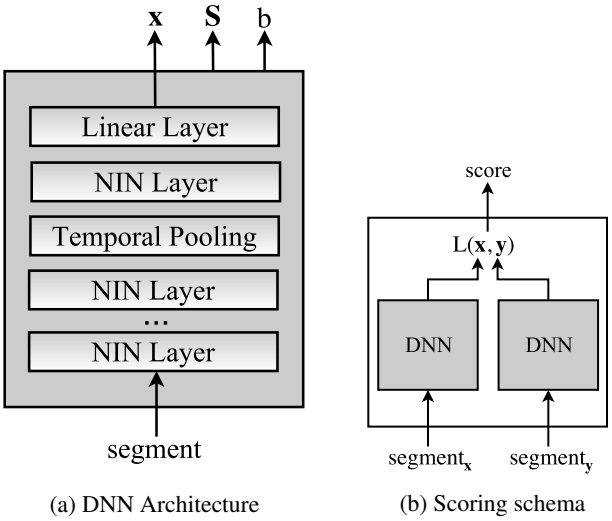


Fig. 1. System diagram for the diarization system presented here. This architecture is simpler than that in [5], with the i-vector extraction and PLDA steps merging into the joint embedding and metric learning, and with UBM and $\mathbf{T}$ entirely removed.



(a) DNN Architecture (b) Scoring schema

Fig. 2. Diagram of the DNN and scoring method.

3. DIARIZATION SYSTEM

Our approach begins with a temporal segmentation into 2 second segments. These segments are embedded into a fixed-dimensional vector using a DNN. The DNN is trained to jointly learn the embeddings and a scoring metric to discriminate between pairs of embeddings (same-speaker vs different-speaker pairs classification). The embeddings and the scoring metric are then projected into a conversation-dependent space using PCA. The conversation-dependent PCA adapts the scoring metric to the unique characteristics of the conversation. The projected segment embeddings are then clustered with the scoring metric using AHC. A threshold learned on unlabeled data (no speaker labels) is used to stop the clustering [5]. The resulting diarization is further refined using VB resegmentation [12]. A system diagram laying out this process is shown in Fig. 1, and each of these modules will be discussed in detail below.

3.1. Temporal Segmentation

We employ 2 second segments with 500ms of overlap with its preceding and following segment (leading to a total of 1 second of overlap). This denser sampling allows for maintaining up to 2 second segments while providing the same number of samples as segmentation at half that length. In [5] we showed that this denser sampling improves our clustering.

3.2. Joint Learning of Embedding and Similarity Metric

3.2.1. Overview

The proposed system uses a feed forward DNN with a temporal pooling layer to extract embeddings from variable-length acoustic segments. The network is based on the architecture recently introduced in [7] for speaker recognition. It is implemented using the nnet3 neural network library in the Kaldi Speech Recognition Toolkit [14].

3.2.2. Features

The features are 40 dimensional MFCCs from 40 mel-spaced filters with a frame-length of 25ms (normally referred to as high-resolution MFCCs within Kaldi recipes). To allow the DNN to compensate for energy variations, volume perturbation was applied to all cuts using a gain from a uniform distribution between 1/8 and 2. No mean or variance normalization was applied.

3.2.3. Neural Network Architecture

The network, illustrated in Figure 2a, consists of five hidden layers, a temporal pooling layer and an affine output layer. Short-term temporal context is incorporated into the first four layers of the network using a time-delay architecture similar to [15]. Suppose T is the total number of frames in a segment and t, where $0 \leq t \leq T$, is the index of some frame. The input layer splices together feature frames $[t-1, t+1]$. At layers two, three and four, activations at frames $[t-2, t+1]$, $\{t-3, t, t+3\}$ and $\{t-3, t, t+3\}$ are spliced together, leading to a total context of $[t-9, t+8]$ at the fourth layer. The temporal pooling layer aggregates the output of the fourth layer over the total length of the input segment $[0, T]$, computes its average, and propagates it to a fifth hidden layer. Finally, this is passed to an affine layer that outputs the embedding $\mathbf{x}$ (400 dimensions for our experiments). The symmetic matrix $\mathbf{S}$ and offset b are constant outputs of the network learned jointly with the embeddings, and are used in the scoring metric in Equation 2.

The hidden layer activations are a type of recently introduced network-in-network (NIN) nonlinearity [16]. The nonlinearity is a mapping from a d_i-dimensional input to a d_o-dimensional output. Within the component, n micro neural networks [17] with tied parameters project the input to a d_h-dimensional space. A micro neural network consists of a stack of three rectified linear units interspliced with affine layers. The NIN configuration $\{n = 50, d_i = 150, d_h = 1000, d_o = 500\}$ is used in this work, which results in a model with 460K parameters.

System	Unsupervised Calibration		Speaker Priors		Oracle Calibration	
	Clustering	+VB Refine	Clustering	+VB Refine	Clustering	+VB Refine
Acoustic	13.5	11.5	13.6	11.2	13.3	11.0
Senone	12.9	10.3	13.8	10.9	12.6	10.2
Embeddings	14.9	13.7	12.8	9.9	12.6	10.3

Table 1. DER results for the proposed embeddings as well as the two baselines with several calibration strategies. The results are shown at the threshold determined with unsupervised calibration, with geometrically decaying speaker priors, and with oracle calibration for optimal cluster DER. All systems are shown with their clustering DER as well as DER after VB resegmentation [12]. Unlike for the baselines, unsupervised calibration struggles to find a near-optimal threshold for the embeddings, but the embeddings performs best with priors and comparably to the best system at oracle calibration. Note that since the oracle calibration optimizes only cluster DER, the resegmentation at that calibration is not best performing, as resegmentation of the embedding system after applying speaker priors yields an overall best score of 9.9.

3.2.4. Training

The probability of embeddings $\mathbf{x}$ and $\mathbf{y}$ belonging to the same speaker is modeled by the logistic function in Equation 1. Equation 2 defines the distance between two embeddings (Figure 2b illustrates its use) and is similar to the discriminative PLDA training criteria in [18, 19]. Let P_{diff} and P_{same} be the set of all different-speaker and same-speaker pairs, respectively. The objective function (Equation 3) optimizes the two-class cross-entropy to discriminate between same-speaker and different-speaker pairs. Since there are many more pairs in the set P_{diff} than in P_{same}, we introduce a constant K so that each set has the same weight in the objective function.

$$Pr(\mathbf{x}, \mathbf{y}) = \frac{1}{1 + e^{-L(\mathbf{x},\mathbf{y})}} \tag{1}$$

$$L(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y} - \mathbf{x}^T\mathbf{S}\mathbf{x} - \mathbf{y}^T\mathbf{S}\mathbf{y} + b \tag{2}$$

$$E = -\sum_{\mathbf{x},\mathbf{y}\in P_{\text{same}}} ln\left(Pr(\mathbf{x}, \mathbf{y})\right) - K\sum_{\mathbf{x},\mathbf{y}\in P_{\text{diff}}} ln\left(1 - Pr(\mathbf{x}, \mathbf{y})\right) \tag{3}$$

Training examples are organized as pairs of segments, each belonging to the same speaker and extracted from the same recording. Segments consists of 2 seconds of speech with no overlap. Minibatches are formed by picking N pairs, such that no two pairs are from the same speaker. Combining segments across pairs results in an additional $N(N+1)$ different-speaker pairs. In our training recipe, we use minibatches of size $N = 16$. All $2N$ segments in the minibatch are propagated through the DNN to produce corresponding embeddings $\mathbf{x}_1$, $\mathbf{x}_2$, ... $\mathbf{x}_{2N}$ and constant outputs b and $\mathbf{S}$, and passed to the objective function. Derivatives are computed with respect to these quantities and backpropagated to the network for parallelized stochastic gradient descent [20]. To ensure that a wide range of speakers and segments are compared, the training examples are shuffled after each iteration.

3.3. Clustering with Prior Specification

It was recently shown that prior probabilities for the number of speakers m can be incorporated in AHC diarization [10]. In that work, the use of a reasonable prior was shown to reduce the negative effects of calibration error. In the experiments that follow, for the proposed system and for the two baselines, AHC will use either no specified prior (which was shown in [10] to correspond to a geometric growth in number of speakers), or, when explicitly stated, a geometrically decaying prior $p_{\#}(m) = 2^{-m}$, which was demonstrated to be a sensible choice with good performance.

4. EXPERIMENTS

4.1. Data

The DNN was trained using approximately 10K cuts taken from Fisher English, SRE04, and SRE05 data. We make use of the speaker labels to produce the minibatches. For the two baseline systems, the UBM and $\mathbf{T}$ matrix were trained using 37K cuts from SRE04, 05, 06 and 08. Their PLDA scoring was trained on the same 10K utterances as the DNN. Also, the senone i-vector system used 1600 hours from Fisher English corpus to train the DNN to classify senones [11].

We evaluated our systems using the CALLHOME corpus, which is a CTS collection between familiar speakers. Within each conversation, all speakers are recorded in a single channel. There are anywhere between 2 and 7 speakers (with the majority of conversations involving between 2 and 4), and the corpus also is distributed across six languages: Arabic, English, German, Japanese, Mandarin, and Spanish.

4.2. Performance Metrics

We evaluated our methods with Diarization Error Rate (DER), a common metric for diarization[1]. In its purest form, DER combines all types of error (missed speech, mislabeled non-speech, incorrect speaker cluster), but, as is currently the practice when reporting on CALLHOME, we used oracle SAD marks. As a result, only incorrect speaker labeling factors into the DER. Also, as is typical, our DER tolerated errors within 250ms of a speaker transition and ignored overlapping segments in scoring.

System	UBM or Senone-DNN	Extractor	Scoring	Total
Acoustic	0.04	1.3	0.004	1.34
Senone	15.6	9.7	0.004	25.3
Embeddings	-	0.38	0.08	0.46

Table 2. Number of parameters per component and their total (units are in millions).

[1]The scoring software is available at `www.itl.nist.gov/iad/mig/tests/rt/2006-spring/code/md-eval-v21.pl`

4.3. Results

Table 1 summarizes the DER results for the two baseline systems and the proposed architecture. The results are shown at the threshold determined with unsupervised calibration, with geometrically decaying speaker priors, and with oracle calibration for optimal cluster DER. All systems are shown with their clustering DER as well as DER after VB resegmentation [12]. Unlike for the baselines, our current implementation of unsupervised calibration struggles to find a near-optimal threshold for the new system, and therefore, the overall DER is worse. However, as seen in the las two columns, the oracle stopping threshold indicates that the DNN embeddings can achieve better results than the acoustic i-vector system, and match the performance of the much more computationally expensive senone i-vector system (see Table 2). Also, initializing the VB resegmentation with these clusterings results in improved DER for all systems.

In our recent work [10], we have shown that the use of a reasonable prior can be effective in mitigating the negative effects of calibration error (i.e., sub-optimal AHC stopping threshold selection). In particular, a geometrically decaying prior for the number of speakers was shown to be a sensible way to widen the region of optimal performance (i.e., reduce the sensitivity to sub-optimal threshold selection). Ideally, we would like for this process to result in a system with optimal performance using an AHC threshold at 0 (which is the optimal threshold for a perfectly calibrated system, see [5] for more details). Figure 3 shows the performance of the three systems for a range of calibration shifts. Although the region of good performance is quite wide for all systems, the performance of the DNN system seems to benefit more from this particular use of priors. Moreover, initializing the VB refinement stage with these clusters produces our overall best result of 9.9 DER (see the middle columns in Table 1). Note that since the clustering and VB resegmentation are independent processes, it is possible to get a final performance that is better than when we start from the oracle calibration threshold.

In the future, we plan to study the different behavior of these systems to prior specification. Also, we plan to modify the unsupervised calibration approach (currently a 2-mixture GMM with tied covariances) to better fit the scores distribution produced by the DNN.

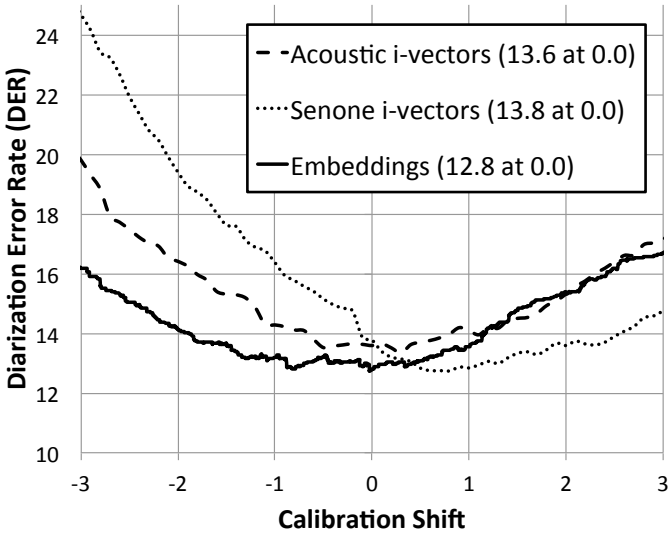


Fig. 3. A comparison of DER for the two baselines and the proposed embeddings for a range of calibration shifts. A geometrically decaying prior is used for number of speakers for all three systems, and, as a result, the optimal performance range is widened and centered.

5. CONCLUSION

In this work, we have presented a discriminatively trained DNN that replaces the two-step generative process of i-vector based diarization systems. The proposed architecture simultaneously learns a fixed-dimensional embedding for acoustic segments of variable length and a scoring metric. Results on the CALLHOME conversational telephone speech corpus demonstrate that, in addition to streamlining the diarization architecture, the proposed system matches or exceeds the performance of state-of-the-art baselines. We also show that, although this approach does not respond as well to our current unsupervised calibration strategy as previous systems, the incorporation of well-founded speaker priors addresses this shortcoming. Using the resulting DNN clustered segments to initialize the VB resegmentation produces our best results.

6. REFERENCES

[1] Stephen Shum, Najim Dehak, Ekapol Chuangsuwanich, Douglas Reynolds, and Jim Glass, "Exploiting Intra-Conversation Variability for Speaker Diarization," in *Proceedings of Interspeech*, 2011.

[2] Stephen Shum, Najim Dehak, and Jim Glass, "On the Use of Spectral and Iterative Methods for Speaker Diarization," in *Proceedings of Interspeech*, 2012.

[3] Stephen H. Shum, Najim Dehak, Réda Dehak, and James R. Glass, "Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–28, October 2013.

[4] Mohammed Senoussaoui, Patrick Kenny, Themos Stafylakis, and Pierre Dumouchel, "A Study of the Cosine Distance-Based Mean Shift for Telephone Speech Diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 217–27, January 2014.

[5] Gregory Sell and Daniel Garcia-Romero, "Speaker Diarization with PLDA i-vector Scoring and Unsupervised Calibration," in *Proceedings of the IEEE Spoken Language Technology Workshop*, 2014.

[6] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE International Conference on Computer Vision (ICCV)*, Rio de Janeiro, 2007.

[7] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Proceedings of the IEEE Spoken Language Technology workshop (SLT)*, 2016 (submitted).

[8] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–98, May 2011.

[9] Patrick Kenny, Douglas Reynolds, and Fabio Castaldo, "Diarization of Telephone Conversations using Factor Analysis," *IEEE Journal of Special Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–70, December 2010.

[10] Gregory Sell, Alan McCree, and Daniel Garcia-Romero, "Priors for Speaker Counting and Diarization with AHC," in *Proceedings of Interspeech*, 2016.

[11] Gregory Sell, Daniel Garcia-Romero, and Alan McCree, "Speaker Diarization with i-vectors from DNN Senone Posteriors," in *Proceedings of Interspeech*, 2015.

[12] Gregory Sell and Daniel Garcia-Romero, "Diarization Resegmentation in the Factor Analysis Subspace," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[13] John R. Hershey, Jonathan Le Roux, Zhuo Chen, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2016.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, et al., "The Kaldi speech recognition toolkit," in *Proceedings of the Automatic Speech Recognition & Understanding (ASRU) Workshop*, 2011.

[15] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proceedings of Interspeech*, 2015.

[16] P. Ghahremani, V. Manohar, D. Povey, and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs," in *Interspeech 2016*. IEEE, 2016.

[17] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[18] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011.

[19] O. Glembek, L. Burget, N. Brummer, O. Plchot, and P. Matejka, "Discriminatively trained i-vector extractor for speaker verification," in *Interspeech*, 2011.

[20] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of deep neural networks with natural gradient and parameter averaging," *CoRR*, vol. abs/1410.7455, 2015.