ARTIFICIAL NEURAL NETWORK FEATURES FOR SPEAKER DIARIZATION

Sree Harsha Yella[1,2*] Andreas Stolcke[1] Malcolm Slaney[1]

[1] Microsoft Research, Mountain View, CA, USA
[2] Idiap Research Institute, CH-1920 Martigny, Switzerland
shyella@idiap.ch, anstolck@microsoft.com, malcolm@ieee.org

ABSTRACT

Speaker diarization finds contiguous speaker segments in an audio recording and clusters them by speaker identity, without any a-priori knowledge. Diarization is typically based on short-term spectral features such as Mel-frequency cepstral coefficients (MFCCs). Though these features carry average information about the vocal tract characteristics of a speaker, they are also susceptible to factors unrelated to the speaker identity. In this study, we propose an artificial neural network (ANN) architecture to learn a feature transform that is optimized for speaker diarization. We train a multi-hidden-layer ANN to judge whether two given speech segments came from the same or different speakers, using a shared transform of the input features that feeds into a bottleneck layer. We then use the bottleneck layer activations as features, either alone or in combination with baseline MFCC features in a multi-stream mode, for speaker diarization on test data. The resulting system is evaluated on various corpora of multi-party meetings. A combination of MFCC and ANN features gives up to 14% relative reduction in diarization error, demonstrating that these features are providing an additional independent source of knowledge.

Index Terms— speaker diarization, artificial neural networks, discriminative feature extraction

1. INTRODUCTION

Speaker diarization addresses the problem of "who spoke when" in a multi-party conversation. It is an unsupervised task, as there is no a-priori knowledge of the speakers or the number of speakers in a conversation [1, 2]. It has been studied in various domains such as broadcast news [3], telephone calls [4], and more recently focusing on spontaneous meeting room conversations [2, 5, 6]. The main issues in performing speaker diarization of meeting room recordings arise due to far-field audio (background noise and room reverberation) and conversational speech (short speaker turns and interruptions).

State of the art systems for speaker diarization use an agglomerative (bottom-up) clustering framework [7, 8]. These systems typically use short-term spectral characteristics, such as Mel-frequency cepstral coefficients (MFCCs) to represent the vocal tract characteristics of a speaker, as features for diarization. Recently factor-analysis based techniques, which are popular in the speaker-verification domain, have been adapted to the speaker diarization task [9]. These methods cluster i-vectors extracted from speech segments using a cosine similarity measure to provide speaker diarization output. Experiments on summed telephone channels have shown that i-vector based methods improve the performance of speaker diarization when compared to the traditional MFCC features. Another approach based on feature transforms uses linear discriminant analysis (LDA) after initial passes of diarization to obtain discriminative features [10]. However, none of these methods developed for two-party telephone conversations have so far been applied to multi-party, conference-style meetings.

In this work, we propose to use an artificial neural network (ANN) trained as a classifier to extract features for diarization. We train the ANN classifier on a related task: to decide whether two given speech segments belong to same or different speakers. We hypothesize that the hidden layers of a network trained in this fashion should transform spectral features into a space more conducive to speaker discrimination. We propose to use the hidden layer activations from the bottleneck layer of the network as a new feature for speaker diarization. We conduct experiments to evaluate the usefulness of the bottleneck features for the task of speaker diarization on various meeting-room data sets.

The paper is organized as follows. Section 2 presents a brief overview of speaker diarization system based on hidden Markov model/Gaussian mixture model (HMM/GMM) framework. Section 3 presents the method of using the proposed ANN based classifier as feature extractor for speaker diarization. Section 4 reports the experimental results on various meeting room datasets. Section 5 presents the conclusions and future directions.

*Work done while the author was an intern at Microsoft Research.

2. HMM/GMM BASED SPEAKER DIARIZATION SYSTEM

A HMM/GMM based speaker diarization system represents each speaker by a state of an HMM and models the state emission probabilities using GMMs. Let c_i denote the ith speaker cluster (HMM state), and b_i denotes the emission probability distribution corresponding to speaker cluster c_i. Then we model the log-likelihood $\log b_i(s_t)$ of input feature s_t for cluster c_i using a GMM as:

$$\log b_i(s_t) = \log \sum_{(r)} w_i^{(r)} N(s_t, \mu_i^{(r)}, \Sigma_i^{(r)}), \qquad (1)$$

where $N()$ is a Gaussian pdf and $w_i^{(r)}$, $\mu_i^{(r)}$ and $\Sigma_i^{(r)}$ are the weights, means and covariance matrices respectively of the rth Gaussian mixture component of cluster c_i. Clustering in an agglomerative framework starts by over-estimating the number of speaker clusters and uniformly segmenting a given audio recording. At each iterative step, we merge the clusters that are most similar. We measure the similarity between two clusters using a modified Bayesian information criterion (BIC) [11] and we merge the clusters that produce the highest BIC score. The modified BIC criterion [11] gets rid of the penalty term based on the number of parameters in the original BIC formulation, by keeping the number of parameters the same before and after the merge. The modified BIC criterion $BIC(c_i, c_j)$ for two clusters c_i and c_j is given by:

$$BIC(c_i, c_j) = \sum_{s_t \in \{c_i \cup c_j\}} \log b_{ij}(s_t) - \sum_{s_t \in c_i} \log b_i(s_t) - \sum_{s_t \in c_j} \log b_j(s_t) \qquad (2)$$

where b_{ij} is the probability distribution estimated over the combined data of cluster c_i and c_j. After each merge step, a Viterbi decoding pass segments the speech data with the new speaker cluster models. A minimum duration constraint on each state prevents rapid speaker changes. The clustering stops when no two clusters have a BIC score greater than zero.

When multiple feature streams are present, a separate GMM is estimated for each feature stream, and a weighted combination of the individual stream log-likelihoods gives the combined log-likelihood. For the case of two feature streams x and y, let $b_i^{(x)}$, $b_i^{(y)}$ denote the probability distributions estimated from streams x, y respectively for cluster c_i. The combined log-likelihood for cluster c_i is:

$$\log b_i(s_t^{(x)}, s_t^{(y)}) = w^{(x)} \log b_i^{(x)}(s_t^{(x)}) + w^{(y)} \log b_i^{(y)}(s_t^{(y)}), \qquad (3)$$

where $s_t^{(x)}$, $s_t^{(y)}$ are the feature vectors corresponding to feature streams x, y respectively, $w^{(x)}$, $w^{(y)}$ are the weights of the feature streams, such that $w^{(x)} + w^{(y)} = 1$. We estimate the weights $w^{(x)}$, $w^{(y)}$ on a held out development data set. The baseline HMM/GMM diarization system used in the current study is modelled after the state-of-the-art system developed by ICSI [7].

3. ANN FEATURES FOR SPEAKER DIARIZATION

Artificial neural networks are extensively used in supervised tasks such as speaker recognition and identification. Konig et al. [12] used a multi-layer perceptron (MLP) with five layers, trained to classify speakers, as a feature extractor. Their MLP was discriminatively trained to maximize speaker-recognition performance. They used the outputs from the second hidden layer (units of which had linear activation function) as features in a standard GMM-based speaker-recognition system. The rationale behind using hidden-layer activations as features is that the initial layers of a network that is trained to classify different speakers will transform the input features into a space more conducive to speaker discrimination, and thus make the classification task easier.

Speaker diarization is an unsupervised task and there is no a-priori information about the speakers. Therefore, in this work, we propose a neural network that is trained to classify two given speech segments (about 500 ms each) as belonging to the same or different speakers. We extract features from this network to use as a new stream in an HMM/GMM diarization model. Fig. 1 shows the architecture of the four-layer network we use in this work. We split the input layer of

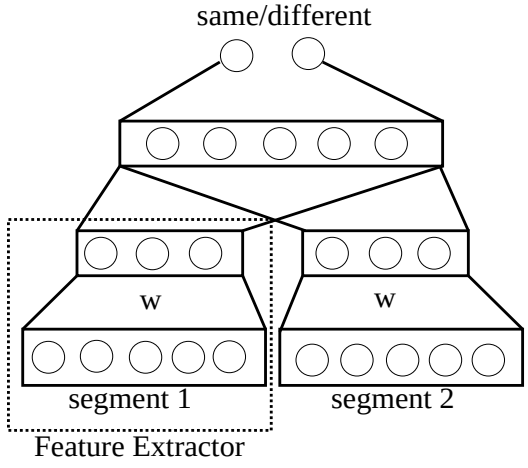

Fig. 1. *An ANN architecture to classify two given speech segments as belonging to same or different speakers. The dotted box indicates the part of the network used to generate features for diarization after the full network is trained.*

the network into two halves, left and right, to represent acoustic features belonging to the two speech segments being compared. The first hidden layer (bottleneck) is also split into two halves similar to the input layer, so each half receives input from the respective input segment i.e., the right half of the hidden layer only gets input from the right half of the input layer

Table 1. *Meeting corpus statistics as used in experiments, including numbers of distinct speakers, meeting room sites, and number of meetings used as part of train, development and test sets.*

Corpus	Speakers	Sites	Meetings		
			Train	Dev	Test
AMI	150	3	148	-	12
ICSI	50	1	-	20	55
NIST-RT	100	6	-	-	24

and the left half from the left half of the input layer. We tie the weight matrices (denoted by W in Fig. 1) connecting the right and left halves of input and hidden layers so that the network learns a single common transform for all speakers. The second hidden layer connects each half of the first hidden layer to the output layer. The output layer has two units denoting the class labels—-same or different speakers—deciding the identity/non-identity of the speakers providing the two input speech segments (segment1, segment2 in Fig. 1). All the hidden layers have sigmoid activation functions and the output layer has a softmax function to estimate the posterior probabilities of the classes (same/different). We train the network using a cross-entropy objective function.

After training the network, we use the first hidden layer activations, before applying the sigmoid function, as features for speaker diarization in a HMM/GMM system. To generate features from the network, we give a speech segment as input to one half of the input layer and extract activations at the corresponding half of the bottleneck layer. It should be noted that it does not matter to which half a speech segment is given as input to generate features since, the weight matrices connecting left and right halves of input layer to the corresponding halves in bottleneck layer are tied.

4. EXPERIMENTS AND RESULTS

We now describe our data, methodology, and experiments. As our system is based on features learned from a separate task, we report the classification performance of the feature-training system (the ANN), as well as diarization performance of the overall system.

4.1. Datasets used in experiments

Our experiments make use of meeting room recordings from various corpora: AMI [13], ICSI [14], and 2006/2007/2009 NIST-RT [15]. Table 1 summarizes the characteristics of these data sets. The AMI data set is split into train and test sets of 148 and 12 meetings, respectively. The test and train sets are disjoint in speakers. We use only speech data from the AMI train set to train the neural network classifier described in Section 3. Twenty ICSI meetings are set aside for the purpose of development and tuning, and the remaining 55 ICSI meetings form an additional test set. All NIST-RT evaluation sets (2006/2007/2009) are also used for testing.

Table 2. *Classification error rate of the ANN after training, on AMI data*

Train	Cross-validation	Test	Chance
20%	21%	35%	50%

4.2. ANN training and feature generation

We trained the ANN to classify two given speech segments as from either same or different speaker, using data from the AMI corpus. To avoid skewing the training toward particular speakers we sampled 50 utterances from each of 138 speakers. Each utterance has a duration of about 10 seconds. The cross validation (CV) set contained 10 utterances from each speaker in the training set. The AMI test set contains all the utterances from the 12 speakers which are not part of the train set (cf. Table 1).

Manual speech transcripts had been forced-aligned to the close-talking microphone recordings to obtain frame-level speaker labels. For training purposes we removed speech segments containing overlapping speech. As input features we extracted 19 MFCCs from a frame of 30 ms with a frame increment of 10 ms. The two halves of the input layer (segment 1, segment 2) each have a context of 500 ms, i.e., 51 frames. The dimensions of the two halves of the bottleneck layer (first hidden layer) is 20. The dimensions of the second hidden layer is 100 and the dimensionality of the output layer is 2, corresponding to the two classes (same/different). The network thus contains 969×2 (input), 20×2 (bottleneck), 100 (2nd hidden) and 2 (output) units.

The objective function for the ANN was cross entropy; training used error back propagation and stochastic gradient descent for 25 epochs. For ANN training and performance evaluation, an equal number of same- and different-speaker speech segment pairs was sampled, making chance error rate 50%. After training, the classification performance (error rate) was as shown in Table 2. Despite not having seen any of the test speakers in training, the network did perform much better than chance on the unseen speakers. Test set error was roughly half-way between training and chance error rates.

After training the network, we obtain new features for the test data by feeding 500ms (50 frames) of acoustic features around a given frame to one half of the input-bottleneck layer portion of the ANN (see Fig. 1). The output values, before the sigmoid non-linearity, were fed as feature vectors to the HMM/GMM diarization system.

4.3. Speaker diarization evaluation

We performed speaker diarization experiments on different test sets to evaluate the usefulness of the features obtained

Table 3. *Speaker error rates obtained on various test sets for different feature streams. ANN denotes the bottleneck features obtained from the neural net classifier and ANN + MFCC denotes the multi-stream combination.*

Data-set	MFCC	ANN	ANN + MFCC
AMI	25.1	32.0	21.5
ICSI	20.6	25.8	18.4
RT-06	14.1	32.5	13.9
RT-07	11.3	25.3	11.8
RT-09	16.8	25.9	18.7

from the neural network classifier, comparing performance to that of the standard 19-dimensional MFCCs typically used for speaker diarization. We also combined the bottleneck features with the MFCCs in a multi-stream fashion as described in Section 2 to exploit any complementary information present in the two feature streams. We fixed the weights when combining these two streams to 0.9 for the MFCC stream and 0.1 for the bottleneck features, based on experiments on the development subset of the ICSI corpus (cf. Table 1). We report performance using the diarization error rate (DER), the standard evaluation metric used in the NIST-RT evaluation campaigns [15]. DER is the sum of speech/non-speech error and speaker error, measured as a percentage of total speaker time. Speech/non-speech segmentation is typically handled by a preprocessing step (known as speech or voice activity detection) to the diarization algorithm. In order to focus evaluation on the speaker clustering aspect of the diarization task, we used the reference speech/non-speech segmentation in all our experiments. The DER in our experiments therefore consists entirely of speaker errors.

Table 3 reports the speaker error rates obtained for various feature streams: MFCC, bottleneck features from an ANN classifier (ANN), and the multi-stream combination of MFCC and bottleneck features (MFCC + ANN). We see that, on their own, bottleneck features do not work as well as MFCC features. However, when the ANN features are combined with MFCCs in a multi-stream system, the speaker error reduces from 25.1% (MFCC) to 21.5% (ANN + MFCC) on the AMI test set and from 20.6% (MFCC) to 18.4% (ANN + MFCC) on the ICSI test set.

The results on the NIST-RT data sets (RT-06, RT-07, RT-09) are less promising. The bottleneck features do not decrease the error even when combined with the MFCC features. We hypothesize that this is because the NIST-RT datasets were collected from a multitude of sites, encompassing a variety of acoustic environments and recording equipment. The ANN, while learning a notion of speaker identity, may have learned to ignore nuisance factors as they occurred in the AMI meetings, but not necessarily as found in other environments. [1]

[1]While the AMI corpus was itself collected at three different sites, the general setup and recording equipment was standardized.

Table 4. *Speaker errors obtained on AMI and ICSI datasets for matched and mismatched training conditions.*

Train	Test	MFCC	ANN + MFCC	Rel. change
AMI	AMI	25.1	21.5	-14.3%
AMI	ICSI	20.6	18.4	-10.7%
ICSI	ICSI	20.6	15.1	-26.7%

It stands to reason that the ANN features could be trained to perform better on the RT data given matched training data. Unfortunately, no such data was available for the various NIST-RT sites. To further investigate the effect of train/test mismatch we ran an additional experiment using the ICSI corpus, where we did have spare data that could be used for training. While the AMI-trained features did improve the diarization error on ICSI data, we trained a second ANN on the non-test portion of the ICSI corpus, with results as shown in the last row of Table 4. The relevant results from training on AMI data also listed for comparison.

We find that, as expected, the performance on ICSI test set is much improved with matched training data, with the relative error reduction going from 10.7% to 26.7%. This relative reduction surpasses the result on the AMI test set, which is likely due to the fact that, unlike for the AMI data, there are shared speakers in the training and test portions of the ICSI corpus.[2]

5. CONCLUSIONS AND FUTURE WORK

We have developed a speaker diarization framework that uses ANNs as trainable acoustic feature extractors. The ANN is first trained to classify pairs of speech segments as belonging to the same or different speakers, while forcing the raw MFCC features to undergo a shared transform via a bottleneck layer. The learned transform can then be applied to unseen data to generate features that are combined with baseline MFCCs as input to a standard agglomerative clustering diarization system. We find that the resulting system reduces speaker error substantially (11–14% relative) when trained on data that is reasonably matched to the test data (AMI or ICSI test data when trained on AMI speakers not seen in testing). With some speakers seen in training (as when training and testing on ICSI meetings) the reduction is more dramatic. Our method thus provides an effective way to adapt a diarization system to available training data without requiring specific knowledge of the speakers present in testing, something that the standard GMM/HMM diarization framework does not allow.

In future work, we plan to explore variants of the framework presented here. First, we arrived at the network dimen-

[2]Speaker overlap was unavoidable in the ICSI train/test sets since a small number of speakers occur in a large number of meetings.

sions using only prior experience with similar ANN applications, and need to systematically optimize input window size and layer dimensions for our task. We also plan to investigate deeper and recurrent neural net architectures as feature extraction networks.

6. REFERENCES

[1] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, pp. 1557–1565, Sept. 2006.

[2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, pp. 356–370, Feb. 2012.

[3] D. Moraru and al., "The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation," in *ICASSP*, vol. 1, pp. 373–376, 2004.

[4] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, pp. 1059–1070, Dec 2010.

[5] X. Anguera, *Robust speaker diarization for meetings*. PhD thesis, Universitat Politecnica de Catalunya, 2006.

[6] D. Vijayasenan, *An Information Theoretic Approach to Speaker Diarization of Meeting Recordings*. PhD thesis, Ecole polytechnique fédérale de Lausanne, December 2010.

[7] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Multimodal Technologies for Perception of Humans. International Evaluation Workshops CLEAR 2007 and RT 2007* (R. Stiefelhagen, R. Bowers, and J. Fiscus, eds.), vol. 4625 of *Lecture Notes in Computer Science*, (Berlin), pp. 509–519, Springer, 2008.

[8] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 7, pp. 1382–1393, 2009.

[9] S. Shum, N. Dehak, E. Chuangsuwanich, D. A. Reynolds, and J. R. Glass, "Exploiting intra-conversation variability for speaker diarization," in *Interspeech*, (Florence, Italy), pp. 945–948, 2011.

[10] I. Lapidot and J.-F. Bonastre, "Integration of LDA into a telephone conversation speaker diarization system," in *Electrical Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of*, pp. 1–4, 2012.

[11] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *Signal Processing Letters, IEEE*, vol. 11, pp. 649–651, August 2004.

[12] Y. Konig, L. Heck, M. Weintraub, and K. Sonmez, "Nonlinear discriminant feature extraction for robust text-independent speaker recognition," in *Proceedings RLA2C Workshop: Speaker Recognition and its Commercial and Forensic Applications*, (Avignon, France), pp. 72–75, Apr. 1998.

[13] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction* (S. Renals and S. Bengio, eds.), vol. 3869 of *Lecture Notes in Computer Science*, pp. 28–39, Springer, 2006.

[14] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. ICASSP*, vol. 1, (Hong Kong), pp. 364–367, Apr. 2003.

[15] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The Rich Transcription 2007 meeting recognition evaluation," in *Multimodal Technologies for Perception of Humans. International Evaluation Workshops CLEAR 2007 and RT 2007* (R. Stiefelhagen, R. Bowers, and J. Fiscus, eds.), vol. 4625 of *Lecture Notes in Computer Science*, pp. 373–389, Berlin: Springer, 2008.