DEEP NEURAL NETWORKS FOR COCHANNEL SPEAKER IDENTIFICATION

Xiaojia Zhao[1]*, Yuxuan Wang*[1] *and DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH, USA
{zhaox, wangyuxu, dwang}@cse.ohio-state.edu

ABSTRACT

Speaker identification (SID) in cochannel speech, where two speakers are talking simultaneously over a single recording channel, is a challenging problem. Previous studies address this problem in the anechoic environment under the Gaussian mixture model (GMM) framework. On the other hand, cochannel SID in reverberant conditions has not been addressed. This paper studies cochannel SID in both anechoic and reverberant conditions. We explore deep neural networks (DNNs) for cochannel SID and propose a DNN-based recognition system. Evaluation results demonstrate the proposed DNN-based system outperforms the two state-of-the-art cochannel SID systems in both anechoic and reverberant conditions and various target-to-interferer ratios.

Index Terms— Cochannel speaker identification, reverberation, deep neural network, Gaussian mixture model, target-to-interferer ratio

1. INTRODUCTION

To separate speech signals from multiple talkers, one can place microphones at different locations and take advantage of the time and intensity differences of the recordings. The task, however, becomes considerably more challenging with a single microphone. Cochannel speech is such a case where two speakers are recorded in a single communication channel. Unlike a conversation, the speakers are not aware of each other, creating large amounts of overlapping speech.

Cochannel speech separation is a challenging problem. Supervised methods [13, 16] usually assume that the speaker identities are available in order to utilize the speaker models. Other work conducts cochannel speaker identification (SID) as a front-end for separation, or jointly with separation. Compared to cochannel speech recognition, one advantage of cochannel SID is that it only needs a subset of homogenous speech segments to infer speaker identities. Such segments are called usable speech [10]. How to group usable speech across time into two streams is deemed as a sequential grouping problem. Shao and Wang jointly search all the grouping hypothesis and speaker candidates to get the optimal one [18, 19]. Mowlaee *et al.* propose to treat cochannel SID and separation as an iterative process [11]. Later they improve the performance by fusing adapted GMM and Kullback-Leibler divergence scores [12]. Hershey *et al.* get the best speech recognition performance thanks in part to excellent performance of cochannel SID and separation [7]. Their SID system first creates a short list of most probable speaker candidates. The top speaker is then paired with the rest for expectation-maximization (EM) based gain estimation. The output is the speaker pair whose gain adapted model maximizes the likelihood of the test utterance. Their system achieves the average SID accuracy of better than 98%. Li *et al*. take a very similar SID approach [9]. It adds a few constraints to the generation of the short list. The top speaker model is directly combined with each of the rest and the combined models are used for SID directly without the EM step. The refined system yields an accuracy greater than 99%. These two may be regarded as the state-of-the-art cochannel SID methods.

Due to the excellent performance of deep neural networks (DNNs) in many tasks, researchers begin to study how to incorporate DNN in speaker recognition [2, 4, 17]. However, DNN has not been utilized in cochannel SID to our knowledge. State-of-the-art cochannel SID performance is reported on the speech separation challenge (SSC) corpus [7, 9]. This corpus [3], however, was tailored for robust speech recognition rather than speaker recognition. The relative small vocabulary and common words between training and testing reduce the difficulty of the SID task [22]. In this study, we employ a speaker recognition evaluation (SRE) dataset of the National Institute of Standards and Technology (NIST). We propose the first DNN-based cochannel SID system working in both anechoic and reverberant conditions. It trains a frame level multi-class DNN classifier that outputs the posterior probability of a frame being dominated by each speaker. Frame level decisions are integrated to make the final decision.

The rest of the paper is organized as follows. In Sect. 2, we formulate the cochannel SID problem and describe the proposed system. Sect. 3 describes the currently dominant GMM-based approach. Model training is discussed in Sect. 4, followed by evaluation and comparison in Sect. 5. We conclude this paper in Sect. 6.

2. DNN-BASED COCHANNEL SID

We formulate cochannel SID as a discriminative learning problem, where we directly learn a mapping from cochannel observations to the corresponding speak identities. Specifically, we treat cochannel SID as a multi-class classification problem and employ DNN as the learning machine. To our knowledge, this is the first study of DNN-based cochannel SID.s

Figure 1 shows the schematic diagram of the proposed DNN-based system. It trains a DNN using frame level features. The output layer has the same number of nodes as speakers. Only the two nodes corresponding to the underlying speakers have non-zero training labels. During testing, the frame level output is aggregated across time to generate the final output.

This research was supported in part by an AFOSR grant (FA9550-12-1-0130). We would like to thank the Ohio Supercomputer Center for providing computing resources.

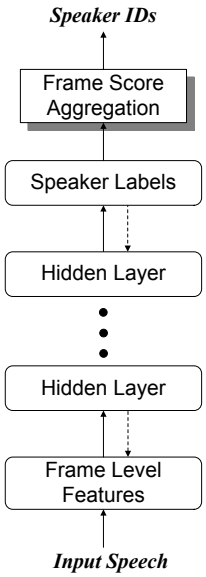

Figure 1. Schematic diagram of the proposed DNN based cochannel SID system

We use frame level log-spectral features as input. To encode temporal context, we splice a window of 11 frames of features to train the DNN. The training target of the DNN is the true speaker identities. We use soft training labels where the two underlying speakers each have a probability of generating the current frame. The sum of their probabilities equals one, whereas the other speakers have zero probabilities. We compare frame level energy of two speakers and use their ratio for the soft labels. More specifically, we construct the ideal binary mask (IBM) [20], and frame level energy of each speaker is calculated from the mixture cochleagram according to the IBM.

The DNN employed in our study is a deep multilayer perceptron. The DNN uses three hidden layers, each having 1024 sigmoidal hidden units. The standard backpropagation algorithm coupled with dropout regularization (dropout rate 0.2) is used to train the network. No unsupervised pretraining is used, as we have sufficient labeled data. We use the adaptive gradient descent along with a momentum term as the optimization technique. A momentum rate of 0.5 is used for the first 5 epochs, after which the rate increases to 0.9. We use a softmax output layer and cross-entropy as the loss function. The training data is discussed in Section 4.

3. GMM-BASED COCHANNEL SID

In this section, we present the currently dominant GMM-based cochannel SID framework. This introduction serves to contrast our DNN-based approach, and describe the algorithms used for later comparisons.

Given an observation O, the goal of cochannel SID is to get the two underlying speakers $\hat{\lambda}_a$ and $\hat{\lambda}_b$ that generate the observation. This can be formulated as searching for the speaker pair with the highest posterior probability.

$$\hat{\lambda}_a, \hat{\lambda}_b = \arg\max_{\lambda_a, \lambda_b} P(\lambda_a, \lambda_b | O)$$

$$= \arg\max_{\lambda_a, \lambda_b} \frac{p(O|\lambda_a, \lambda_b) P(\lambda_a, \lambda_b)}{p(O)}$$

$$= \arg\max_{\lambda_a, \lambda_b} p(O|\lambda_a, \lambda_b) . \quad (1)$$

We apply the Bayes formula to convert the posterior probability to the likelihood of a joint distribution of two speakers, with the assumption that all speaker pairs are equally probable. $p(O)$ is not dependent on speakers and can thus be dropped from the calculation. The question now becomes how to calculate likelihoods of a joint distribution. Shao and Wang have introduced a variable g, to (1), to assign each speech segment to one of the two speaker sources [18, 19]. The derivation is shown as follows.

$$\hat{\lambda}_a, \hat{\lambda}_b = \arg\max_{\lambda_a, \lambda_b} p(O|\lambda_a, \lambda_b)$$

$$= \arg\max_{\lambda_a, \lambda_b} \sum_g p(O, g|\lambda_a, \lambda_b)$$

$$\approx \arg\max_{\lambda_a, \lambda_b} \left(\max_g p(O, g|\lambda_a, \lambda_b) \right)$$

$$= \arg\max_{\lambda_a, \lambda_b} \left(\prod_{X \in S} \max\left(p(X|\lambda_a), p(X|\lambda_b) \right) \right) . \quad (2)$$

Here X denotes a speech segment, S the set of all segments, and g an assignment vector of the same length as S. Each element of g is a binary label that assigns the corresponding segment to a speaker. The integration over all assignments is approximated as a *max* operation, assuming that the optimal assignment dominates the summation. By assuming that segments are independent, the problem reduces to finding the best assignment for each segment and the likelihood of the utterance is the multiplication of segment likelihoods. The speaker pair with the highest likelihood is the SID output. The corresponding optimal assignment also gives a solution to the cochannel separation problem by organizing segments into two groups. In other words, this approach jointly performs cochannel SID and separation, so we name it joint SID & separation (JSS).

Li *et al*. have proposed a two stage algorithm that produces state-of-the-art performance in the SSC corpus [5, 9]. The first stage ranks speakers according to their posterior probabilities given the observation. The posterior probability of each speaker λ given X is calculated as follows.

$$p(\lambda|X) = \frac{p(X|\lambda)P(\lambda)}{\sum_m p(X|\lambda_m)P(\lambda_m)} \quad (3)$$

where m is the speaker index. $P(\lambda)$ and $P(\lambda_m)$ are prior probabilities. Assuming that all the speakers are equally probable, the priors can be eliminated. Frame level posterior probabilities are aggregated across time to obtain utterance level probabilities. Speakers are ranked based on the aggregated scores. The top ten speakers are kept for the second stage where the top speaker is combined with each of the remaining nine. The composite GMMs are used for standard speaker recognition to get the best speaker pair. We point out that the composition operates on a per frame basis.

Li *et al.*'s two stage algorithm is a fine-tuned version of Hershey *et al*.'s SID system [7]. Overall, the two systems yield the best performance in the SSC corpus with Li *et al*.'s average performance around 1% higher.

4. MODEL TRAINING

In this study, we deal with both anechoic and reverberant test conditions. For the anechoic condition, we use anechoic data to train GMMs and DNNs. However, such models do not generalize well to

reverberant conditions. Thus, we directly model speakers in the reverberant environments.

The degree of reverberation is typically indicated by *reverberation time* (T_{60}), the time taken for a direct sound to attenuate by 60 dB [8]. Reverberation is modeled as a convolution between a *room impulse response* (RIR) and a direct sound signal. An RIR characterizes a specific reverberant environment and is determined by factors such as the geometry of the room, and locations of sound sources and receivers.

Assuming no knowledge of test reverberant conditions, we simulate *N* representative reverberant training conditions covering a plausible range of T_{60}. Our previous study has shown that this technique has reasonable generalization [23]. We prepare training data in each of the *N* conditions. GMMs are trained using single speaker data while DNNs are trained with cochannel data mixed at different TIRs. Details are given in the next section.

5. EVALUATION AND COMPARISON

5.1. Experimental Setup

We randomly select 100 speakers from the 2008 NIST SRE dataset (short2 part of the training set). The telephone conversation excerpt of each speaker is roughly 5 minutes long. Large chunks of silence in the excerpt are removed. Then we divide the recording into 5 s pieces. Two pieces with the highest energy are used for tests in order to provide sufficient speech information. The rest is used for training. Overall each speaker has about 20 training utterances. More details of the evaluation corpus can be found in [23].

A Matlab implementation of the image method of Allen and Berkley is used to simulate room reverberation [1, 6]. We focus on the T_{60} range up to 1s that covers realistic reverberant conditions [8]. Three rooms are simulated to obtain 3 training T_{60}'s: 300, 600 and 900 ms. For each T_{60}, we generate 5 RIRs by randomly positioning the source and receiver while keeping their distance fixed at 2 m. Each training utterance is convolved with the 5 RIRs of each room to create reverberant training data. Seven rooms are simulated to obtain 7 test T_{60}'s from 300 ms to 900 ms with a step size of 100 ms. We randomly generate 3 pairs of RIRs at each T_{60} where each pair provides one RIR for the target and one for the interferer. In total there are 21 pairs of test RIRs. Note that the RIRs are different between training and testing even when they are generated with the same T_{60}.

DNNs are trained using cochannel training data. Instead of one DNN per speaker, we train a universal DNN for all the speakers. We include training data from every speaker pair for a complete coverage. For anechoic conditions, we create 10 anechoic cochannel utterances per speaker pair at 3 TIRs (-5, 0 and 5 dB). In total, there are 4950 speaker pairs and 49500 cochannel training utterances per TIR. For reverberant conditions, we create 10 reverberant cochannel utterances at each of the 3 T_{60}'s and 3 TIRs. In total, there are 49500 cochannel training utterances per TIR and per T_{60}.

For JSS, we extract 22-dimensional MFCC as speaker features. Speaker models are adapted from a 1024-component universal background model (UBM) trained by pooling training data from all the speakers [15]. For Li *et al.*, we extract 64-dimensional log-spectral features for GMM training. Specifically, a 64-channel gammatone filterbank is employed as the front-end. The filter output is converted to cochleagram [21]. We take the log operation on the cochleagram to get the features. For anechoic conditions, a 256-component GMM is trained for each speaker [14]. Another 256-component GMM is trained using the reverberant training data by convolving the anechoic training data with the RIRs at 3 T_{60}'s.

Cochannel test set covers all possible speaker pairs. For each pair, we create two anechoic utterances and two reverberant utterances at -5, 0 and 5 dB TIRs. There are totally 9900 anechoic test utterances and 9900 reverberant test utterances per TIR. Each reverberant cochannel test utterance is created using a randomly selected RIR pair from the 21 RIR pair library.

5.2. Performance on the SSC Corpus

The state-of-the-art cochannel SID systems of Hershey *et al.* and Li *et al.* have reported performance on the SSC corpus. This corpus consists of 17000 training utterances from 34 speakers. Each training utterance is created following a fixed grammar: *command*, *color*, *preposition*, *letter*, *number*, and *adverb*. Each of the six positions has a small number of word choices. The cochannel test set of the SSC corpus comprises six TIRs from -9 dB to 6 dB. There are 600 test utterances for each TIR. Every test utterance is mixed from clean test utterances of two speakers. Note that the clean utterances follow the same grammar and share the same vocabulary as the training utterances.

We evaluate our proposed system on this dataset in order to make a direct comparison. Table I gives the SID results of the proposed system and competing systems. As can be seen, our implementation of Li *et al.*'s two stage system achieves the same average performance as their paper. The proposed DNN-based system yields the best results, although the performance gain is probably not significant. As the results are nearly perfect, there is not much room to improve and we can conclude that the proposed system work comparably well.

5.3. Performance on NIST SRE Dataset with 50 speakers

First we test on a subset of 50 speakers with 1225 speaker pairs, to be roughly comparable with the SSC corpus in terms of speaker number. We create two cochannel utterances for each pair at each of 3 TIRs, −5 dB, 0 dB and 5 dB. In total, there are 2450 test trials per TIR. The performance is given in Table II. As shown in the table, there is a substantial drop of performance compared to the SSC corpus, confirming that the SSC corpus is rather easy for cochannel SID evaluation. For this dataset, JSS outperforms Li *et al*. by an average of 4.3%. We also evaluate the DNN-based cochannel SID system, which further outperforms the best competing system by a large margin (almost 13%).

Next we test in the reverberant conditions, and the results are shown in Table III. As can be seen, the performances of all the methods degrade in the reverberant conditions. JSS drops by about 30%. Li *et al*.'s is slightly more robust, but still drops by more than 20%. In addition, the proposed DNN-based system continues to perform the best, outperforming JSS by more than 19% and Li *et al*.'s system by 14%.

5.4. Performance on NIST SRE Dataset with 100 speakers

The SID task becomes more challenging as the number of speakers (classes) increases. To quantify cochannel SID dependency on the number of speakers, we have performed cochannel SID evaluation by increasing the number of speakers from 50 to 100, quadrupling the number of classes to 4950. As in the previous results, the default DNN configuration (3 hidden layers with 1024 nodes each) outperforms the best competing system. With the increase of speaker size as well as training data size, we have also explored a few different DNN configurations. As we increase the number of units from 1024 to 2048 for each hidden layer, the SID performance improves by around 4.5%. There is a slight improvement as we expand the number of hidden layers from 3 to 5 without changing the hidden layer size, for either 1024 or 2048 hidden units. Further enlargement of the DNN size is expected to improve the performance even more, but at the expense of substantially increased computational complexity.

Table I: SID accuracy (%) on SSC corpus.

Method	−9 dB	−6 dB	−3 dB	0 dB	3 dB	6 dB	Avg.
Reported Performance of Hershey *et al.*	96.5	98.1	98.2	99.0	99.1	98.4	98.2
Reported Performance of Li *et al.*	97.3	98.8	99.5	99.7	99.7	98.8	99.0
JSS	89.8	95.3	98.0	98.5	98.2	96.8	96.1
Li *et al.*	96.7	99.0	99.5	99.7	100.0	99.2	99.0
DNN	98.3	99.5	100	99.8	100	99.0	99.4

Table II: SID accuracy (%) on anechoic NIST SRE dataset with 50 speakers

Method	−5 dB	0 dB	5 dB	Avg.
JSS	82.24	83.51	80.12	81.96
Li *et al.*	77.02	79.96	75.84	77.61
DNN	94.12	96.90	92.69	94.57

Table III: SID accuracy (%) on reverberant NIST SRE dataset with 50 speakers

Method	−5 dB	0 dB	5 dB	Avg.
JSS	51.43	53.76	49.51	51.57
Li *et al.*	55.02	59.35	56.37	56.91
DNN	70.86	75.31	66.29	70.82

Table IV: SID accuracy (%) on reverberant NIST SRE dataset with 100 speakers

Method	−5 dB	0 dB	5 dB	Avg.
JSS	39.59	41.76	38.70	40.02
Li *et al.*	43.58	47.12	43.58	44.76
DNN (1024 by 3)	52.67	59.78	52.58	55.01
DNN (1024 by 5)	54.13	61.31	54.33	56.59
DNN (2048 by 3)	56.91	64.76	56.99	59.55
DNN (2048 by 5)	57.32	64.82	57.52	59.89

6. CONCLUDING REMARKS

This paper has a number of novel contributions. Our first contribution lies in the introduction of DNN for cochannel SID. Our proposed DNN system substantially outperforms the state-of-the-art SID methods, which are GMM-based. Secondly, we address cochannel SID in reverberant conditions, a topic that has not been studied before.

Since this is the first study of applying DNN to cochannel SID, therefore there is likely room for future improvement. For instance, training features and labels can be systematically examined, and DNN architecture can be optimized. With the excellent performance of cochannel SID, we believe that the use of DNN represents a promising direction to pursue noise robust SID, reverberation robust SID, and speaker verification tasks.

7. RELATION TO PRIOR WORK

The work presented here has focused on the cochannel SID problem. Previous studies on this topic focus on GMM-based approaches in the anechoic condition. DNN has not been studied for this problem, and there is no previous work on cochannel SID in reverberant conditions. Our study addresses this problem in both anechoic and reverberant conditions by introducing a DNN-based approach.

8. REFERENCES

[1] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943-950, 1979.

[2] K. Chen and A. Salman, "Learning speaker-specific characteristics with a deep neural architecture," *IEEE Transactions on Neural Networks,* vol. 22, no. 11, pp. 1744-1756, 2011..

[3] M. Cooke and T. Lee, "Speech separation and recognition competition," 2006 [Online]. Available: http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm

[4] S. Garimella and H. Hermansky, "Factor analysis of auto-associative neural networks with application in speaker verification," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 24, no. 4, pp. 522-528, 2013.

[5] Y. Guan and W. Liu, "A two-stage algorithm for multi-speaker identification system," in *Proc. International Symposium on Chinese Spoken Language Processing*, 2008, pp. 161-164.

[6] E.A.P. Habets, "Room impulse response generator," 2010 [Online]. Available: http://home.tiscali.nl/ehabets/rir_generato r.html

[7] J. Hershey, S. Rennie, P. Olsen, and T. Kristjansson, "Super human multi-talker speech recognition: A graphical model approach," *Computer Speech & Language*, vol. 24, pp. 45–66, 2010

[8] H. Kuttruff, *Room Acoustics*. New York, NY: Spon, 2000.

[9] P. Li, Y. Guan, S. Wang, B. Xu and W. Liu, "Monaural speech separation based on MAXVQ and CASA for robust speech recognition," *Computer Speech & Language*, vol. 24, pp. 30–44, 2010.

[10] J. M. Lovekin, R. E. Yantorno, K. R. Krishnamachari, D. S. Benincasa and S. J. Wenndt, "Developing usable speech criteria for speaker identification," in *Proc. ICASSP*, 2001, pp. 421–424.

[11] P. Mowlaee, R. Saeidi, Z. Tan, M. Christensen, P. Fränti, and S. Jensen, "Joint single-channel speech separation and speaker identification," in *Proc. ICASSP,* 2010, pp. 4430–4433.

[12] P. Mowlaee, R. Saeidi, M. Christensen, Z. Tan, T. Kinnunen, P. Fränti and S. Jensen, "A joint approach for single-channel speaker identification and speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 9, pp. 2586-2601, 2012.

[13] A. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1766-1776, 2007.

[14] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91-108, 1995.

[15] D.A. Reynolds, T.F. Quatieri, and and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[16] S. T. Roweis, "One microphone source separation," in *Proc. NIPS*, 2000, pp. 793-799.

[17] M. Senoussaoui, N. Dehak, P. Kenny, R. Dehak and P. Dumouchel, "First attempt of boltzmann machines for speaker verification," in *Proc. Odyssey, The Speaker and Language Recognition Workshop*, 2012.

[18] Y. Shao and D.L. Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," in *Proc. ICASSP*, 2003, pp. 205-208.

[19] Y. Shao and D.L. Wang, "Model-based sequential organization in cochannel speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 289-298, 2006.

[20] D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, P. Divenyi, Ed. Norwell, MA: Kluwer Academic, 2005, pp. 181-197.

[21] D.L. Wang and G.J. Brown, Eds., *Computational Auditory Scene Analysis*: *Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley-IEEE, 2006.

[22] X. Zhao, Y. Shao, and D.L. Wang, "CASA-based robust speaker identification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 5, pp. 1608 – 1616, 2012.

[23] X. Zhao, Y. Wang and D.L. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 4, pp. 836-845, 2014.