Speaker Diarization with I-Vectors from DNN Senone Posteriors

Gregory Sell, Daniel Garcia-Romero, Alan McCree

Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD, USA
{gsell,dgromero,alan.mccree}@jhu.edu

Abstract

Motivated by recent gains in speaker identification by incorporating senone posteriors from deep neural networks (DNNs) into i-vector extraction, we examine similar enhancements to speaker diarization with i-vector clustering. We examine two DNNs with different numbers of senone targets in combination with a diagonal or full covariance universal background model (UBM) in the context of the multilingual corpus CALLHOME. Results show that the larger DNN with a full covariance UBM gives the best performance. The improvements appear to have a strong dependence on number of speakers in a conversation, and a lesser dependence on language. Overall, when combined with resegmentation, the proposed system improves CALLHOME performance to 10.3% DER.

Index Terms: Speaker diarization, clustering, deep learning

1. Introduction

Speaker diarization is the process of segmenting speech into the sections spoken by each unique person in the conversation. Since most speech technologies assume only one speaker in a given utterance (and, in fact, sometimes exploit that assumption, such as with speaker adaption for automatic speech recognition (ASR)), diarization is a valuable front-end process in the event that multiple speakers are present.

Speaker diarization is similar to speaker identification in some regards, since both tasks determine if two segments were spoken by the same person or by different people. Due to this similarity, most current speaker diarization systems utilize i-vectors, a speech representation common in speaker identification, in order to cluster short segments of speech. However, unlike in speaker identification, there is no enrollment data to define speaker identities for diarization. Furthermore, the number of speakers is typically unknown, though there are presumably only a few speakers in any given conversation. As a result, unsupervised clustering has become an effective approach for speaker diarization, since it requires no prior knowledge of speaker identity and can manage a reasonable number of speakers [1, 2, 3].

Recently, significant progress has been made in speaker identification by incorporating senone posteriors from deep neural networks (DNNs) into the i-vector extraction process [4, 5]. Since speaker identification and speaker diarization share many similarities, there is reason to believe this approach will improve diarization results as well.

In this paper, we explore the use of i-vectors extracted with DNN senone posteriors for speaker diarization. We examine the performance of several system variations on the CALLHOME conversational telephone speech (CTS) corpus, and consider the clustering results in terms of number of speakers as well as spoken language. Overall, we find that, when combined with resegmentation, the use of DNNs for i-vector extraction can improve a state-of-the-art diarization system from 11.5% DER to 10.3%.

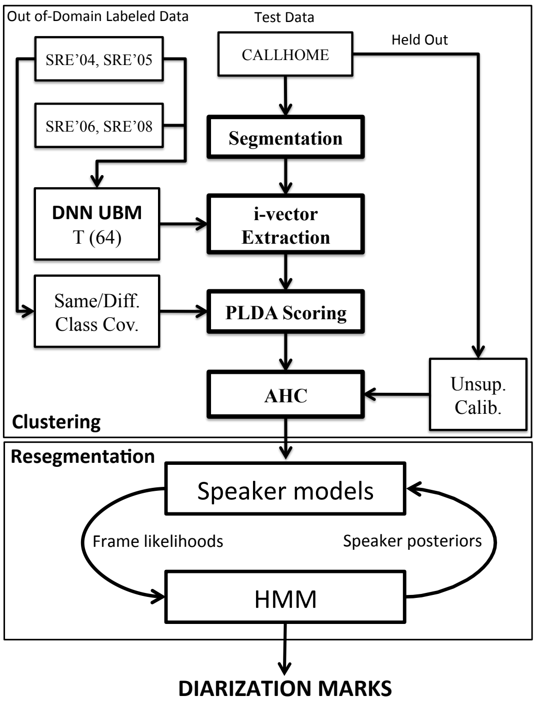

Figure 1: System diagram for the proposed diarization system. This system is identical to that in [10] except the substitution of an unsupervised UBM with a DNN UBM.

2. Background

2.1. Diarization

I-vectors are commonly used in speaker recognition [6], and, more recently, have become common in speaker diarization as well [7, 8, 1, 2, 3]. In diarization, i-vectors are extracted for short blocks of speech, usually on the order of 1-2 seconds. These i-vectors are then clustered using an unsupervised method, such as agglomerative hierarchical clustering (AHC) [3], Variational Bayesian Gaussian Mixture Models [1], or mean shift [2]. These clustering algorithms are typically followed by or iterated with a resegmentation algorithm that refines the transition boundaries, either with in the feature space [9] or in the factor analysis subspace [10].

2.2. I-vectors from Senone Posteriors

Previously, the universal background model (UBM) utilized for i-vector extraction was represented with unsupervised GMMs that partition the feature space. Recent approaches substitute the unsupervised partitioning with supervised DNN senone posteriors. This approach leverages transcribed speech data in order to provide more consistent content-based labels for the acoustic features. Improving the labels should improve speaker discriminability, as the partitioning of the feature space will be based on the spoken content rather than only acoustic similarity. This approach has been shown to be effective for speaker identification [4, 5], even with domain mismatch [11].

3. Diarization with Senone Posteriors

In order to incorporate the new i-vectors into speaker diarization, we begin with an existing speaker diarization system based on i-vector clustering [10], but instead use i-vectors extracted with senone posteriors from DNNs rather than unsupervised GMMs. A system diagram for this new process is shown in Fig. 1. Note that this is identical to the system diagram from [10], except that the unsupervised UBM has been replaced with a supervised DNN UBM.

Though not reflected in the system diagram, we also found that the cut-dependent PCA from [3] removed more energy for these i-vectors, and so, instead of projecting to 3 dimensions in all cases, here, we projected to the appropriate dimension to keep 50% of the energy.

We tested two DNNs with 5 hidden layers of dimension 5000 with p-norm nonlinearities ($p = 2$) and an input/output dimension ratio of 10 [12]. Both were trained on 1200 hours of speech from the Fisher English corpus. However, the DNNs differed in number of target senones. A set of 7591 senones was used for the larger DNN (called ENG7591), and a set of 2186 senones was used for the smaller DNN (called ENG2186). For each DNN, we trained a diagonal covariance UBM and a full covariance UBM using data from NIST SRE04, SRE05, SRE06, and SRE08. Each output senone is associated with a component in the UBM mixture.

4. Experiments

4.1. Data

We evaluated each system using the CALLHOME corpus, a CTS collection of calls between familiar speakers. The corpus includes six languages: Arabic, English, German, Japanese, Mandarin, and Spanish. All conversations are presented in a single channel sampled at 8kHz, and the number of speakers in each conversation varies from 2 to 7 (and the majority between 2 and 4).

The CALLHOME corpus has been used to evaluate several of the systems discussed in Section 2. The resulting error rates are collected in Table 1.

4.2. Performance Metrics

We measured performance with Diarization Error Rate (DER), the standard metric for diarization. DER combines missed speech, mislabeled non-speech, and incorrect speaker assignment. However, we followed the typical practice of using oracle SAD marks in order to evaluate our diarization method independent of any particular SAD algorithm, and, as a result, only incorrect speaker labeling factors into the DERs measured in this work. Also, as is standard, our DER tolerated errors within 250ms of a speaker transition and ignored overlapping (multi-speaker) segments in scoring.

Method	DER
Castaldo et al [13]	13.7
*Shum et al [1]	14.5
Senoussaoui et al [2]	12.1
Sell and Garcia-Romero [10]	11.5
Proposed	**10.3**

Table 1: DERs for several systems on CALLHOME. The (*) reflects that the results for Shum et al were estimated from plots displaying results per speaker.

4.3. Results

4.3.1. DNN Size

The clustering performance (without resegmentation) of each system is shown in Fig. 2 in comparison to the baseline unsupervised UBM clustering process [3]. The larger DNN outperforms the smaller in almost all cases, indicating a finer partitioning of the feature space is advantageous. Interestingly, in five-speaker conversations, the smaller DNN is slightly better for both UBM covariance types, but this trend is not replicated in any other case.

4.3.2. UBM Covariance

Fig. 2 also clearly suggests that using a full covariance UBM improves over diagonal covariances, and, in fact, the diagonal UBM covariances perform below the baseline in most cases. It is possible that the highly non-linear partitioning of the DNN requires full covariance matrices in order to see any advantage at all.

4.3.3. Number of Speakers

It is clear in Fig. 2 that the DNN systems degrade in the presence of increasing numbers of speakers faster than the baseline unsupervised UBM system, as the relative performance between the two is highly dependent on number of speakers. In the two speaker case, all DNN systems outperform the baseline. For three speakers, all but the smallest DNN with diagonal UBM covariances outperforms the unsupervised UBM system. For four speakers, only the largest DNN with full UBM covariances performs as well as the baseline, and for five speaker or above, all DNN systems are worse. It is unclear what is causing this strong dependence on number of speakers. Given the small number of conversations with five or more speakers, it is possibly an anomaly, but this requires further analysis before diarization systems with DNNs should be used for applications with large numbers of speakers.

Since the large DNN (ENG7591) with full UBM covariances consistently provides the best clustering performance, the remaining analysis will only consider this combination.

4.3.4. Language

It is possible that adding a supervised component trained on English data and using English senone labels will introduce an undesired language-dependence into the diarization system. To explore this possibility, performance of the ENG7591 DNN sys-

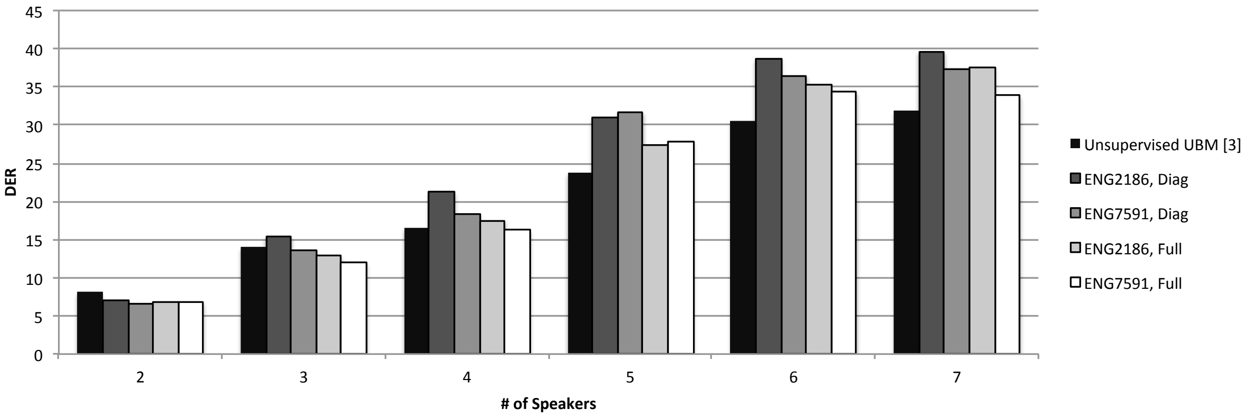

Figure 2: DER from the clustering for each DNN system compared to the clustering performances using an unsupervised UBM (from [3]). Performance is broken down by number of speakers in the call.

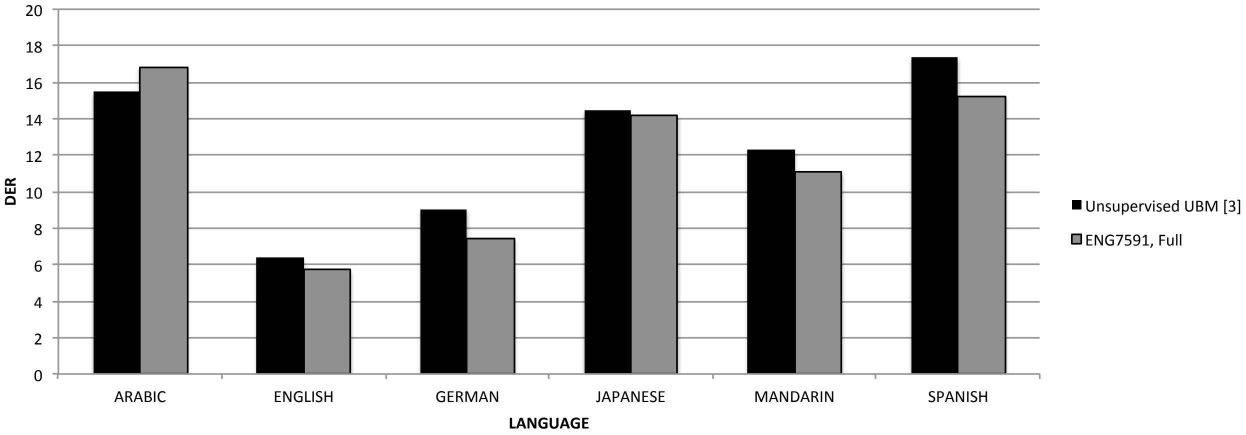

Figure 3: Clustering DER for the ENG7591 DNN with full UBM covariances for each language as compared to a baseline system with an unsupervised UBM (from [3])

tem with full UBM covariances (clustering only) is shown for each language in CALLHOME, along with performance by the baseline clustering algorithm, is shown in Fig. 3.

Interestingly, it is immediately evident that performance varies with language for both systems, though the reasons for these differences are not obvious. The dependence could source from some cultural conversational tendency (if conversations in a language tend to have longer pauses between speaker turns, for example). It is also possible that the unsupervised UBM system already had language dependence from the predominately-English SRE data used in training. It is also worth noting that the distribution of number of speakers in a conversation varies between each language, and so some of the DER variation can be attributed to this difference. However, these distributions do not fully explain the results. Japanese, for example, has one of the higher DERs among these languages but is roughly 75% two-speaker conversations.

Comparing the performance of the two systems, there are two important results. The first is that, while English improves in performance with the use of the DNN, the improvement is not greater in magnitude than for other languages. German, for example, makes a larger improvement, both relative and absolute. This would seem to contradict the concern that using an English-trained DNN would disproportionately benefit diarization of English conversations.

However, the second important result from the comparative DERs in Fig. 3 is that the proposed system performs worse on Arabic, despite improving all other languages (though the Japanese improvement is marginal). This difference could also be caused by some cultural characteristic that correlates with language, or it could be that the English senones are somehow fundamentally worse for Arabic. Understanding this result would require further study, but it suggests that using English senones for i-vector extraction can have a negative impact on diarization in particular languages.

Entering this study, it was reasonable to suggest that using English senones might improve error rates in English at the cost of performance in other languages. Instead, it appears, at least within this small language group, that English is improved similarly to other languages, and only one language (Arabic) sees reduced performance.

4.3.5. With Resegmentation

Pairing the best performing clustering system (ENG7591 with full UBM covariances) with a Variational Bayes resegmentation algorithm shown to be effective for diarization [10] yields even further gains, shown in Table 2 both in terms of number of speakers and overall. Not only is overall performance improved by 1.2% DER, but performance is now improved for all numbers of speakers except five-speaker conversations. The comparative improvement due to resegmentation is especially pronounced for six- and seven-speaker conversations, but it is important to note that there are only 6 and 2 such conversations, respectively, so these scores are highly subject to random variations.

5. Conclusion

It has been shown that, as in speaker identification, using senone labels from DNNs for i-vector extraction improves speaker diarization. Comparing to a similar system with an unsupervised UBM shows that the clustering improvement diminishes as the number of speakers in the conversation increases. Analysis by language shows that, of the six languages studied, all but Arabic improve, and English does not improve by a greater magnitude than other languages, suggesting a low level of language dependence. When combined with resegmentation, the overall DER is reduced from a baseline of 11.5% to 10.3%.

# Spkrs	From [10]	ENG7591+VB
2	6.4	4.7
3	11.2	10.0
4	14.2	13.4
5	22.3	26.0
6	27.8	24.1
7	31.9	28.9
Overall	11.5	10.3

Table 2: DER performance for the proposed system (ENG7591 DNN with full UBM covariances and Variational Bayes resegmentation) as compared to the system from [10]. The proposed system improves performance for all cases except five speakers.

6. References

[1] S. H. Shum, N. Dehak, R. Dehak, and J. 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.

[2] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. 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.

[3] G. Sell and D. Garcia-Romero, "Speaker Diarization with PLDA I-Vector Scoring and Unsupervised Calibration," in *Proceedings of the IEEE Spoken Language Technology Workshop*, 2014.

[4] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A Novel Scheme for Speaker Recognition Using a Phonetically-Aware Deep Neural Network," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2014.

[5] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep Neural Networsk for Extracting Baum-Welch Statistics for Speaker Recognition," in *Proceedings of Odyssey*, 2014.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[7] S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, and J. Glass, "Exploiting Intra-Conversation Variability for Speaker Diarization," in *Proceedings of Interspeech*, 2011.

[8] S. Shum, N. Dehak, and J. Glass, "On the Use of Spectral and Iterative Methods for Speaker Diarization," in *Proceedings of Interspeech*, 2012.

[9] P. Kenny, D. Reynolds, and F. 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] G. Sell and D. Garcia-Romero, "Diarization Resegmentation in the Factor Analysis Subspace," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2015.

[11] D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey, "Improving Speaker Recognition Performance in the Domain Adaptation Challenge Using Deep Neural Networks," in *Proceedings of the IEEE Spoken Language Technology Workshop*, 2014.

[12] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving Deep Neural Network Acoustic Models Using Generalized Maxout Networks," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2014.

[13] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream-Based Speaker Segmentation Using Speaker Factors and Eigenvoices," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2008.