Extending the Task of Diarization to Speaker Attribution

Houman Ghaemmaghami, David Dean, Robbie Vogt, Sridha Sridharan

Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia

{houman.ghaemmaghami,d.dean,r.vogt,s.sridharan}@qut.edu.au

Abstract

In this paper we extend the concept of speaker annotation within a single-recording, or speaker diarization, to a collection wide approach we call *speaker attribution*. Accordingly, speaker attribution is the task of clustering expectantly homogenous inter-session clusters obtained using diarization according to common cross-recording identities. The result of attribution is a collection of spoken audio across multiple recordings attributed to speaker identities. In this paper, an attribution system is proposed using mean-only MAP adaptation of a combined-gender UBM to model clusters from a perfect diarization system, as well as a JFA-based system with session variability compensation. The normalized cross-likelihood ratio is calculated for each pair of clusters to construct an attribution matrix and the complete linkage algorithm is employed to conduct clustering of the inter-session clusters. A matched cluster purity and coverage of 87.1% was obtained on the NIST 2008 SRE corpus.

Index Terms: speaker attribution, diarization, clustering, cross likelihood ratio, joint factor analysis

1. Introduction

We define speaker *attribution* as the task of annotating a collection of spoken audio with the identities of the speakers. Speaker attribution can be regarded as a combination of both speaker *diarization* and speaker *identification*, both of which are active areas of research [1, 2]. A speaker attribution system ultimately combines the process of attributing speech to speakers within a recording, and, determining instances of the same speaker across recordings. This paper proposes a method for conducting speaker attribution with the motivation of extending diarization and investigating the possibile proceeding strategies or methodology that would be required in order to achieve the task of speaker attribution in a robust and efficient manner.

Speaker diarization is the task of annotating an audio stream with information that associates speaker homogeneous segments of speech to their specific sources or speaker identities without *a priori* knowledge. Over the recent years, speaker diarization has become a highly active area of research as indicated by the degree of participation of the world's top research groups in the National Institute of Science and Technology (NIST) Rich Transcription (RT) evaluations [2]. The segmentation stage of speaker diarization provides labels for speaker homogeneous speech utterances within the analysed recording. The clustering stage then associates each of these segments to their specific speaker to form clusters representing speaker identities. The clustering stage commonly employs agglomerative clustering [1, 3], in which, each segment is first assigned an initial cluster and similar clusters are iteratively merged according to a distance measure until a stopping criterion is met. Most implementations have utilised model selection measures such as the Bayesian information criterion (BIC) to select an optimal model out of a number of candidate parametric models which best represents a given data set [4, 3]. Such methods have displayed high sensitivity to parameter selection and the audio domains in which they are developed and tested [4]. In most cases, significant tuning is necessary to achieve acceptable results for each target domain, thus restricting the portability of such systems across audio domains. Even when testing on recordings within the same domain, large variations in performance are quite common [4, 1]. Current systems, however, have focused on utilising speaker modelling techniques commonly employed in speaker recognition research in order to reduce computational load, achieve increased robustness and conduct speaker diarization with higher precision [5, 6, 7].

This paper extends the task of diarization to speaker attirbution by conducting inter-session clustering of expectantly homogenous utterances, in terms of speaker identity, obtained by means of diarization. The developed method in this paper utilises mean-only maximum *a posteriori* (MAP) adaptation of a combined-gender Gaussian mixture universal background model (GMM-UBM) to model the initial clusters. The normalized cross-likelihood ratio (NCLR) is then calculated as a measure of simmilarity between each pair of speakers to construct an attribution matrix [4]. Finally, the complete linkage clustering method is employed to cluster the speakers based on their similarity scores as indicated by the attribution matrix. In addition, the paper investigates incorporating speaker and session variability modelling via joint factor analysis (JFA) [5] and provides results in both cases.

Section 2 of the paper will provide the theory behind the proposed approach to speaker attribution. In Section 3 the evaluation details and experimental results are provided on a subset of the 2008 NIST Speaker Recognition Evaluation (SRE) summed channel dataset. Section 4 provides a discussion of the results.

2. Speaker attribution

The task of speaker attribution can be represented as a joint process of speaker diarization and speaker identification. Assuming the initial set of N recordings $X = \{X_i; i = 1, ..., N\}$ were obtained from a corpus of spoken audio, with each recording split into segments $X_i = \{x_{im}; m = 1, ..., M_i\}$ by the diarization process, the challenge is to determine the corresponding set of S responsibility vectors r_{im} over a set of S speakers of interest, where r_{ims}=1 if speaker s is responsible for (i.e. speaking in) segment m and $r_{ims} = 0$ otherwise. The responsibility vector r_{im} corresponding to segment m in recording i has only one non-zero entry corresponding to the responsible speaker. We assume that each segment of observed data, x_{im} has been produced by a single speaker, indicating that the accurate speaker segmentation has been performed in the diarisation stage.

This section provides a brief summary of the theory associated with the proposed speaker attribution system based on the stated assumptions. The MAP adaptation based GMM-UBM [8] approach for cluster modelling is explained as well as cluster modelling with speaker and session variability modelling using JFA [9, 5]. The derivation of the normalized cross-likelihood ratio (NCLR) as a similarity measure is provided [10]. Finally, the complete linkage clustering method utilised for the proposed attribution task is presented.

2.1. GMM-UBM cluster modelling

One of the main issues with speaker clustering techniques commonly utilised in speaker diarization systems is the high computational load associated with iterative model training and merging [3, 1]. In addition, speaker diarization systems are typically tuned and associated with a single recording session. Such methods are not applicable to the task of speaker attribution, where an initial high volume of large inter-session clusters are typically expected. For this reason, the proposed speaker attribution system utilises a GMM-UBM based approach to model initial clusters. This was implemented as two distinct systems using, 1- a mean-only MAP adaptation of the UBM and 2- a JFA based speaker and session variability modelling approach.

2.1.1. MAP system

The MAP system employs a mean-only MAP adaptation of the UBM [8]. To do this, each of the initial clusters are used to adapt the means of the UBM and obtain a cluster model represented by the common UBM and, a speaker and session dependent GMM mean supervector offset, unique to cluster s and session i:

$$g(x) = \sum_{c=1}^{k} w_c N(x; \mu_c, \Sigma_c), \tag{1}$$

$$m_i(s) = [{\mu_1}^T, ..., {\mu_k}^T]. \tag{2}$$

2.1.2. JFA system

The JFA based attribution system employs joint factor analysis to conduct speaker and session variability modelling as proposed in [9]. To do this, first a constrained offset of the speaker dependent GMM mean supervector is introduced:

$$m_i(s) = m(s) + Ux_i(s), \tag{3}$$

where $m(s)$ is the speaker/cluster dependent, session independent, GMM mean supervector, $x_i(s)$ is the low-dimensional representation of the variability in session i, and U is the low-rank transformation matrix from the session variability subspace to the GMM mean supervector space.

The method described in [11] can be employed to obtain the speaker/cluster dependent supervector $m(s)$. However, in order to also achieve speaker variability modelling the joint factor analysis method in [9] was utilised. To do this $m(s)$ is represented as:

$$m(s) = m + Vy(s) + Dz(s), \tag{4}$$

where m is the speaker and session independent GMM mean supervector. $y(s)$ is the speaker factors and represents the parameters of the speaker in the specified subspace with a standard normal distribution. V is a low-rank transformation matrix from the speaker variability subspace to the GMM mean supevector space and $Dz(s)$ is used to model the residual variability that is not captured by the speaker subspace.

D can be estimated using the method in [8] which states that D must satisfy the following constraint:

$$I = \tau D^T \Sigma^{-1} D, \tag{5}$$

where τ is the relevance factor and Σ is a diagonal matrix of the components' covariance matrices Σ_c in (1).

It can be seen that in order to achieve speaker and session variability modelling the full joint factor model and thus the speaker independent *hyperparameters* V, U, D, m and Σ, must be estimated which was achieved using the method proposed in [9].

2.2. Normalized cross likelihood ratio (NCLR)

After obtaining the cluster models using the GMM-UBM approach the developed speaker attribution system computes the normalized cross likelihood ratio (NCLR) as a similarity measure between each pair of adapted cluster models. The NCLR was selected as the preferred choice of measure as it has been shown to be a robust and efficient measure in obtaining a similarity measure between adapted speaker/cluster models [4]. The NCLR between two cluster models M_i and M_j is presented in [4] as:

$$NCLR = \frac{1}{N_i} log \frac{p(x_i|M_j)}{p(x_i|M_B)} + \frac{1}{N_j} log \frac{p(x_j|M_i)}{p(x_j|M_B)} \tag{6}$$

where in the context of speaker attribution, N_i and N_j represent the number of samples, or observations, associated with the adapted cluster models M_i and M_j, respectively. $p(x|M)$ denotes the likelihood of the data x given cluster model M, and M_B represents the UBM.

In [10], it is shown that the GMM likelihood function can be calculated as:

$$\begin{aligned} logp(x|M) = \sum_{c=1}^{k} (N_c log \frac{1}{(2\pi)^{\frac{F}{2}} |\Sigma_c|^{\frac{1}{2}}}) - \frac{1}{2} tr(\Sigma^{-1} S) \\ + Z^* \Sigma^{-1} F + \frac{1}{2} Z^* N \Sigma^{-1} Z, \end{aligned} \tag{7}$$

where N, F and S represent zeroth, first and second order statistics of the cluster segment x calculated using model M, respectively. Z is a representation of the sum of the speaker/cluster and channel supervectors. Σ denotes the covariance of the speaker independent UBM and Σ_c is the diagonal covariance matrix of mixture component c.

It is shown in [10], that the first two terms in (7) are only dependent on the cluster segment x. This is while the last two terms display dependency on, not only x, but also the cluster model. Furthermore, the first two terms will cancel out during NCLR computation and thus the likelihood of segment x given model M becomes:

$$logp(x|M) = Z^* \Sigma^{-1} F + \frac{1}{2} Z^* N \Sigma^{-1} Z, \tag{8}$$

where F and N for each cluster were obtained over each component, c, of the UBM and F was centralized on the UBM mean mixture components, m_c.

$$F = \sum_{n=1}^{N} p(c|x_n, M_B)(x_n - m_c) \tag{9}$$

The NCLR was thus computed using (6) and (8), where in (6) N_i and N_j represent the sum of occupancy counts associated

with the features of each cluster i and j, respectively, in each component, c, of the UBM.

The proposed system utilises the NCLR measure between each pair of initial clusters to conduct a similarity check and in turn conduct clustering. It can be seen from (6) that a large NCLR value corresponds to a higher similarity. In order to prepare for the clustering process an attribution matrix was thus constructed using the NCLR measure with s rows and s columns, where s denotes the number of initial clusters.

2.3. Complete linkage clustering

After obtaining the attribution matrix containing the NCLR values, the complete linkage clustering method is utilised to cluster similar speakers based on the furthest distance measure [12]. The NCLR attribution matrix is treated as containing distance measures by the linkage algorithm, hence high scores are selected by the clustering method to conduct speaker attribution.

The s by s attribution matrix, A, is first zeroed along the diagonal and converted to a square symmetric vector ω of length $(\frac{s^2-s}{2})$. The inconsistency coefficients, $\tilde{\omega}$, are then calculated for each element present in ω and clustering is conducted using a selected cutoff value:

$$\tilde{\omega} = \frac{(d-\mu)}{\sigma_d}, \tag{10}$$

where $\tilde{\omega}$ is the inconsistency coefficient, d is the number of links included in the calculation, μ is the mean of the lengths and σ_d is the standard deviation of the links included in the computation of $\tilde{\omega}$.

3. Evaluation

Evaluation of the proposed speaker attribution system was conducted using 691 excerpts from the summed channel mixer-3 NIST SRE 2008 test data. Each recording contained summed telephone speech from two speakers. A total of 1382 initial speakers/clusters, each of length ≈ 5 minutes were tested. The true number of speaker identities present in the testing database was equal to 751 speakers.

The diarization reference labels for the dataset were obtained in [5] using speech recognition on each channel. In this work, they were utilised to obtain speaker homogeneous clusters for the two speakers from each of the 691 recordings. Hence, obtaining the initial 1382 clusters, excluding doubletalk regions. This was carried out to avoid introducing diarization errors into the attribution task in order to exclusively investigate the errors associated with speaker attribution in the case of a perfect diarization system. Finally, 12 Mel-frequency Cepstral coefficient (MFCC) features including the zeroth order coefficient with deltas and feature warping [13] were extracted for each of the 1382 initial clusters.

3.1. Hyperparameter training

3.1.1. UBM

A combined gender UBM was trained on the entire NIST SRE 2004 data with a selection of Switchboard II, phase 2 and 3 data to increase the diversity. The UBM was trained using 512 mixture components, 12 MFCC features, including the zeroth order coefficient, with deltas and feature warping [13].

3.1.2. JFA hyperparameters

The speaker and session subspaces were estimated using a coupled EM algorithm based on the UBM supervector space. A 50-dimensional session subspace and 200-dimensional speaker subspace were trained in this manner using telephone data from the Switchboard II, SRE 2004 and SRE 2005.

3.2. Evaluation metrics

In order to conduct the evaluation of the proposed speaker attribution system the cluster purity and coverage metrics were employed. To obtain these measures, each cluster is first assigned a speaker identity. The cluster is analysed and the speaker with the most number of samples within the cluster is selected as the dominant speaker of that cluster and the cluster is labeled by that speaker's identity.

Cluster purity, C_p, refers to the ratio of the total number of correctly clustered samples from speaker/cluster i, N_i, in its labeled cluster, to the total number of samples available in that cluster C_{total}:

$$\%C_p = 100(\frac{N_i}{C_{total}}). \tag{11}$$

Cluster coverage is a complementary metric to cluster purity and refers to the "best" coverage of a speaker's samples in a single cluster. That is, for each speaker i, the cluster containing the most number of samples, $max(N_i)$, for that speaker, is selected and the ratio of this value to the total number of data samples for that speaker, N_{total} is calculated to obtain cluster coverage, C_c:

$$\%C_c = 100(\frac{max(N_i)}{N_{total}}). \tag{12}$$

The average value over the set of speaker specific C_p and C_c measures can be obtained to represent the cluster purity and coverage of the evaluated speaker attribution systems. In the following sections C_p and C_c will refer to the average cluster purity and coverage percentages, respectively, over all speakers.

3.3. Results

The MAP attribution system was first evaluated, followed by the JFA attribution system. Table 1 displays the average cluster purity and coverage achieved using each of the developed attribution systems. The table displays the C_p and C_c percentages obtained using the minimum difference of the two metrics, or, the true number of speakers as two operating points used for evaluation. It can be seen that a slight improvement of C_p and C_c rates can be achieved using JFA to conduct speaker and session variability modelling.

Figure 1 displays the relationship between cluster purity C_p and cluster coverage C_c. It can be seen that the JFA based speaker attribution system performs consistently better, covering a larger area under the curve.

4. Discussion

The results obtained using the proposed JFA based speaker attribution system display a 4.8% relative improvement at the minimum accuracy difference operating point (OP) compared to the results achieved by the MAP based system. The results in Table 1 display a higher C_p and C_c achieved using the JFA based speaker attribution system. The JFA system performs consistently better. It can be seen, however, that when utilising the minimum difference of the two metrics as the OP, the MAP

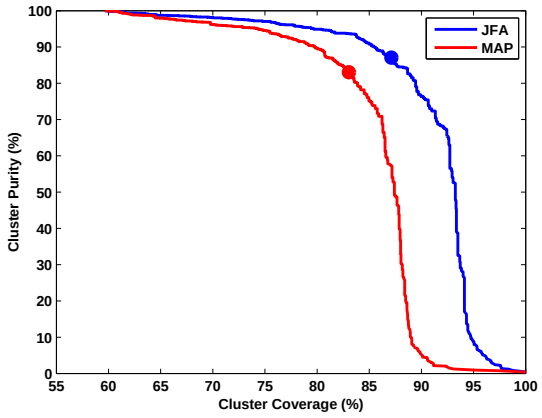

Figure 1: *Cluster purity C_p versus cluster coverage C_c for the proposed MAP and JFA based speaker attribution systems. The operating point corresponding to the minimum difference of the C_p and C_c metrics is displayed on each plot.*

Table 1: *Results obtained for each system on the 1382 initial clusters at the C_p and C_c minimum difference and the true speaker count operating points.*

System	Operating point	C_p (%)	C_c (%)	Speakers
MAP	$C_p \approx C_c$	83.04	83.04	750
MAP	Speaker=751	83.10	83.04	751
JFA	$C_p \approx C_c$	87.08	87.12	774
JFA	Speaker=751	85.53	87.38	751

based attribution system obtains 750 attributed clusters of the true 751 speakers/clusters. This is while the JFA based attribution system obtains 774 attributed clusters at this OP. It can be concluded that the use of JFA for speaker and session variability modelling improves the task of attribution, however, short utterances are no longer attributed to their correct clusters. This brings about a minor decrease in C_c but a higher C_p, at cluster counts slightly larger than the true number of speakers/clusters, can be achieved.

From Figure 1, it can be seen that the JFA based speaker attribution system preforms consistently better compared to the MAP based system, with a larger coverage of area under curve. This indicates that prior to the employment of speaker and session variability modelling, clusters existed that were not attributed to their speaker. This was observed to be due to cases where speakers with identical sessions were attributed to the same speaker cluster, which was resolved through utilisation of the joint factor modelling compensation. The remaining error was mainly due to short length utterances that could not be modeled with great accuracy and would thus be attributed to the false cluster. Finally, the minimum accuracy difference OP has been marked in Figure 1 to display the point of minimum difference between the C_p and C_c rates achieved by each system.

5. Conclusion

In this paper the task of speaker attribution was defined as an inter-session clustering of expectantly homogenous utterances obtained by means of speaker diarization. A speaker attribution system was proposed and evaluated to conduct a study of this task. Two variations of the proposed system, a MAP based approach and a JFA based system, were evaluated. The MAP based system used a mean-only MAP adaptation of a combined-gender UBM to model output clusters of an ideal diarization system. The JFA approach also conducted speaker and session variability modelling. The normalized cross-likelihood ratio (NCLR) was used as a similarity measure to perform attribution using the complete linkage clustering algorithm. It was demonstrated that the use of JFA is benficial to the task of attribution achieving a 4.8% relative improvement at the minimum accuracy difference OP when incorporated in the system. The system was evaluated on the 2008 NIST SRE data and was shown to achieve a cluster purity and coverage of up to 87.1%.

6. Acknowledgements

This research was supported by an Australian Research Council (ARC) Linkage Grant No: LP0991238.

The authors would like to thank Patrick Kenny for kindly providing the ASR transcriptions for the NIST 2008 SRE data.

7. References

[1] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1557 –1565, 2006.

[2] (2007) The NIST rich transcription website. http://www.nist.gov/speech/tests/rt/.

[3] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, 2003, pp. 411 – 416.

[4] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multistage speaker diarization of broadcast news," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1505 –1512, 2006.

[5] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 6, pp. 1059 –1070, 2010.

[6] F. Valente, P. Motlicek, and D. Vijayasenan, "Variational Bayesian speaker diarization of meeting recordings," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 4954 –4957.

[7] N. Bassiou, V. Moschou, and C. Kotropoulos, "Speaker diarization exploiting the eigengap criterion and cluster ensembles," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 2134 –2144, 2010.

[8] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," in *Digital Signal Processing*, 2000, p. 2000.

[9] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech 2008*, 2008, pp. 853–856.

[10] O. Glembek, L. Burget, N. Dehak, N. Brummer, and P. Kenny, "Comparison of scoring methods used in speaker recognition with joint factor analysis," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 0, pp. 4057–4060, 2009.

[11] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Comput. Speech Lang.*, vol. 22, pp. 17–38, January 2008. [Online]. Available: http://portal.acm.org/citation.cfm?id=1287851.1288082

[12] A. Jain, A. Topchy, M. Law, and J. Buhmann, "Landscape of clustering algorithms," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 1, 2004, pp. 260 – 263 Vol.1.

[13] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *A Speaker Odyssey, The Speaker Recognition Workshop*, June 18-22 2001, pp. 213–218.