A CLUSTER-VOTING APPROACH FOR SPEAKER DIARIZATION AND LINKING OF AUSTRALIAN BROADCAST NEWS RECORDINGS

Houman Ghaemmaghami, David Dean, Sridha Sridharan

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

houman.ghaemmaghami@qut.edu.au, d.dean@ieee.org, s.sridharan@qut.edu.au

ABSTRACT

We present a clustering-only approach to the problem of speaker diarization to eliminate the need for the commonly employed and computationally expensive Viterbi segmentation and realignment stage. We use multiple linear segmentations of a recording and carry out complete-linkage clustering within each segmentation scenario to obtain a set of clustering decisions for each case. We then collect all clustering decisions, across all cases, to compute a pairwise vote between the segments and conduct complete-linkage clustering to cluster them at a resolution equal to the minimum segment length used in the linear segmentations. We use our proposed cluster-voting approach to carry out speaker diarization and linking across the SAIVT-BNEWS corpus of Australian broadcast news data. We compare our technique to an equivalent baseline system with Viterbi realignment and show that our approach can outperform the baseline technique with respect to the diarization error rate (DER) and attribution error rate (AER).

Index Terms— cluster-voting, complete-linkage clustering, speaker diarization, Viterbi realignment

1. INTRODUCTION

The task of speaker diarization is to determine *'Who spoke when?'* in a given recording [1]. This information can be useful for various applications, such as annotating and indexing multimedia data, carrying out speaker recognition of multiple speaker recordings or identifying speaker-specific speech events for improving speech recognition applications [2]. As the necessity for processing large volumes of data increases, particularly multimedia data, so too does the need for an efficient diarization approach that is capable of identifying speakers within and across multiple recordings. Dupuy et al. [3] and Viet et al. [4] extended the task of speaker diarization to cross-show speaker diarization for identifying speakers across multiple recordings. We refer to this as speaker linking, as suggested by van Leeuwen [5] and adopted by others [6, 7, 8, 9]. Speaker linking is thus the task of determining speakers across temporally-independent recordings after speaker diarization has been applied to extract speakers within each recording. For consistency with our previous work [10, 11, 12], we use the term speaker attribution to refer to the combined tasks of speaker diarization and speaker linking.

Speaker attribution has been used as a vital tool for person recognition in multimodal conditions and broadcast corpora [6, 13]. We have previously proposed and applied complete-linkage clustering to the task of diarization and speaker linking to demonstrate greater clustering efficiency and accuracy over traditional agglomerative merge and retrain techniques [14, 8]. This has led us to the development of a simple and efficient speaker attribution system with a clustering-only speaker linking module and a diarization stage that combines clustering with Viterbi segmentation [14, 11, 12]. The Viterbi segmentation stage of our diarization module can bring about inefficiencies when dealing with long recordings [15]. We thus aim to eliminate the need for Viterbi segmentation in order to achieve a highly efficient clustering-only speaker attribution system for processing large multimedia datasets with recurring speaker identities.

In this paper we propose a cluster-voting technique that can be used to combine multiple clustering decisions. We apply this technique to the task of speaker diarization by clustering multiple linear segmentations of a recording at different segment lengths. We then combine all clustering decisions at the highest possible resolution (the minimum segment length used in any of the linear segmentations) and make a collective decision based on a vote-based clustering approach. This eliminates the need for Viterbi segmentation in our diarization module and allows us to achieve a clustering-only speaker attribution system. We compare this approach to our previously proposed (and equivalent) speaker attribution system with Viterbi segmentation [11] to demonstrate a greater diarization accuracy without the need for Viterbi refinement across the SAIVT-BNEWS corpus of Australian broadcast data [11].

2. RELATION TO PREVIOUS WORK

One of the main issues with most speaker diarization systems is the lack of a simple approach that can robustly and efficiently be applied to multiple audio domains without the need for expensive agglomerative cluster merging and retraining, parameter tuning or adjusting minimum duration constraints for Viterbi realignment [2, 16]. We have previously addressed the problem of clustering efficiency for large sets of speaker segments by employing complete-linkage clustering [10, 8], however the use of Viterbi realignment in our diarization module can result in inefficiencies when processing long recordings. In this paper we propose a cluster-voting approach for taking advantage of multiple clustering decisions. We then use this technique to combine a set of clustering decisions, achieved through different length linear segmentations of a given recording, to make a more informed clustering decision without requiring Viterbi realignment to rectify incorrect clustering decisions. We show that this clustering-only approach can outperform our previously proposed system, which uses Viterbi realignment before and after segment clustering [11].

3. BASELINE SYSTEM

We use our previously proposed speaker attribution (diarization and linking) system as a baseline approach [11], which we will hereon refer to as the baseline system. The main stages of the baseline system are the speaker diarization and speaker linking modules of this system. Throughout our work, we use joint factor analysis (JFA) modeling with session compensation [8, 17], which can accommodate the problem of mismatched recording session conditions between multiple spoken recordings from the same speaker identity. We will begin by providing a brief description of our approach to speaker modeling and clustering, which forms the basis of our proposed techniques in this paper. The speaker diarization and speaker linking stages of the baseline system, which employ these techniques are then presented in this section.

3.1. Speaker modeling and clustering

At the core of every speaker diarization and linking approach these is a speaker modeling and clustering stage that carries out a major role in identifying and clustering spoken segments from the same speaker identities [2, 16]. This stage often draws heavily from recent speaker recognition research to reliably model and cluster short speaker segments. For this reason, the most common techniques employed have been either i-vector or JFA based speaker modeling [18, 10, 9, 3, 7].

We use JFA adaptation with session variability compensation using a combined-gender universal background model (UBM) [11, 19, 17]. Our UBM is trained using 512 mixture components and we use a 200-dimensional session and 200-dimensional speaker subspace. We train the speaker independent JFA hyperparameters using a coupled expectation-maximization (EM) algorithm proposed by Vogt et al. [17], however we also apply an audio-partitioning technique when training our JFA hyperparameters. To do this, the active speech in every spoken recording is partitioned into 5 second segments, with each 5 second partition assigned an independent session label. This allows for training short spoken segments from the same speaker as independent recording sessions, which do not always differ in recording session, but because of their short length the majority of the difference between them lies in their linguistic content. We believe the advantage of using this audio partitioning technique with JFA hyperparameter training is that the session subspace can be trained in such a way as to largely allow for compensation of unwanted linguistic content, while the speaker subspace would emphasize desired linguistic variations that may be useful in identifying a speaker. In addition, as speaker diarization is often carried out using short recording segments [9, 7, 11], we believe this approach is better suited to the task of speaker diarization.

After modeling speaker segments using JFA adaptation, we compare the similarity of the models using the pairwise cross-likelihood ratio (CLR) metric [20]. We have previously shown that the CLR measure provides a theoretical comparison threshold value of 0.0, which can be used reliably to carry out speaker clustering across multiple audio domains when combined with complete-linkage clustering [14, 8, 11]. We will thus use a CLR stopping threshold of 0.0 for carrying out speaker clustering throughout our work in this paper.

To achieve a clustering decision, given all pairwise CLR similarity scores between the JFA adapted models, we apply complete-linkage clustering [14]. This form of clustering is an agglomerative approach, however after each merge the CLR scores between the newly formed cluster and the remaining clusters are updated using a linkage rule and the already available scores [21]. This means that there is no need for retraining new models to represent the merged cluster and that clustering can be carried out with great efficiency [21]. In complete-linkage clustering using pairwise CLR similarity scores d, the cluster pairs with the highest similarity score are merge first, however after a merge takes place between two clusters C_i and C_j into $C_{i'} = \{C_i, C_j\}$, the CLR score between the newly formed cluster $C_{i'}$ and any remaining cluster C_x will be $d_{i'x}$ where,

$$d_{i'x} = \min(d_{ix}, d_{jx}), \tag{1}$$

which is the worst possible pairwise CLR score between any pair of elements in the two compared clusters. This provides a cautious clustering approach which is pessimistic in nature and thus reluctant to combine clusters that contain speaker segments with a pairwise CLR score lower than our stopping threshold of 0.0. We have previously shown that this approach can outperform traditional agglomerative merging and retraining techniques and state-of-the-art alternative clustering techniques for speaker clustering in the task of speaker diarization and linking [14, 8].

3.2. Speaker diarization and linking

Our baseline attribution system can be described as having two main stages: speaker diarization and speaker linking [11]. Our speaker diarization system is based on the ICSI RT-07 diarization system by Wooters et al. [15], and the baseline method by Kenny et al. [18]. In order to carry out diarization of a recording, we first apply our implementation of the hybrid voice activity detection (VAD) and ergodic HMM Viterbi segmentation proposed for the ICSI RT-07 speaker diarization system [15]. We then linearly segment the active speech regions in the recording, using 5 second segments. Each of these 5 second segments is then modeled as a state of an ergodic HMM, using 32 component GMMs [18]. We also include the non-speech regions as a state in this HMM using a single component Gaussian and apply 3 iterations of Viterbi realignment to achieve an initial segmentation of the recording. We then apply speaker modeling and clustering to these segments (based on Section 3.1) to obtain a set of larger segments. These new segments then undergo 3 iterations of Viterbi realignment to refine their boundaries, in the same manner as before. We then repeat the modeling, clustering and Viterbi process once more to take advantage of the larger segments and to achieve a final diarization decision [11].

After diarization has been carried out on each independent recording in an analysed dataset, we are left with unique intra-speaker models as hypothesised by our speaker diarization module. We then apply a single iteration of speaker modeling and clustering, in the same manner as presented in Section 3.1, to achieve speaker linking across all recordings. This leaves us with the final speaker attribution decisions regarding *'Who spoke when?'* in each recording, and in which of the recordings.

4. PROPOSED CLUSTER-VOTING

In this section we present our proposed clustering-only speaker attribution system based on complete-linkage clustering. We do this with the aim of eliminating the need for the commonly employed Viterbi realignment in speaker diarization [15, 18, 11], which is used for refining segment boundaries but can become computationally expensive when dealing with long recordings [2, 16]. For simplicity, we will still conduct VAD using our baseline VAD module. This stage is highly efficient as the VAD HMM is constructed using only one speech and one non-speech state, which are modeled as a 2 component GMM and single Gaussian, respectively.

As discussed in Section 3.2, our baseline approach uses an ergodic HMM Viterbi realignment technique based on the ICSI-RT07 system [15] and the work by Kenny et al. [18]. In this approach each segment is modeled using a 32 component GMM and is represented by a state in an ergodic HMM with a minimum duration constraint. We rely on this technique to adjust the segment boundaries before and after carrying out modeling and clustering (Section 3.1). Applying Viterbi realignment in this manner can be computationally expensive and may require adjusting the minimum duration constraint when processing across audio domains, depending on the speaker change rate of the analysed data [2, 15]. We would achieve greater efficiency if we could apply only our modeling and clustering technique to the linear segmentation of each recording, however this would not be a reliable approach. This is because we begin with a linear segmentation (of 5 second segments) of a recording, which means that if our segment size is not small enough, we may pick up speech from more than one speaker in each segment, or a combination of non-speech and speech from a speaker, or other forms of impurities that will negatively impact our modeling and clustering stage. At the same time, if we require a linear segmentation with a higher segment purity, we would need to use shorter segments (<1 second), which may not contain enough information for reliable modeling and clustering. To overcome this problem, we propose applying our JFA modeling and complete-linkage clustering algorithm (Section 3.1) to multiple linear segmentations of a recording at varying segment lengths and then making a collective decision based on all clustering outcomes, in each linear segmentation case.

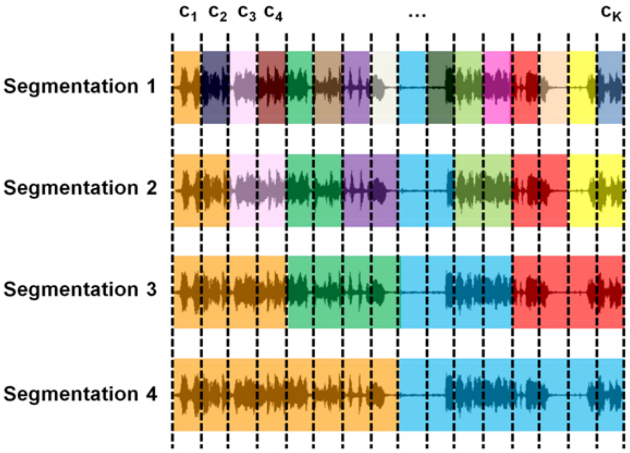

Fig. 1. Four linear segmentations of the same recording at different segment lengths, with the dashed lines representing the minimum segment length and thus the maximum resolution at which we are able to make a joint clustering decision.

4.1. Combining decisions by cluster-voting

To carry out speaker diarization of a recording using cluster-voting (CV), we first conduct multiple linear segmentations of a recording and then apply our modeling and clustering approach to each segmentation to achieve multiple clustering decisions for the same recording. We use $n = 1, \ldots, N$ linear segmentations, each with the respective segment length of nL, where L is the minimum segment length and the length used in our first segmentation case. In our N^{th} linear segmentation case we thus use a segment length of NL. Figure 1 displays four linear segmentations of the same audio recording. In this case, $N = 4$ and each segmentation case is labeled accordingly. In Figure 1 we have indicated the linear segmentation of the recording using different colours for each segment in each segmentation scenario. In addition, we have used dashed lines to indicate the minimum segment length at which we can make a clustering decision. We refer to this as our *decision resolution*.

From Figure 1, it can be seen that the first segmentation case may have more pure segments but less reliability for modeling and clustering due to its shorter segment lengths. As the number of segmentation scenarios and the employed segment length increases, we should be able to achieve more reliable models, however because of the inevitably introduced impurities within each segment this may not always be the case. It is therefore vital to combine the multiple clustering decisions that are made in each segmentation case.

Figure 2 displays the outcome of our speaker modeling and clustering process applied to each of the segmentation scenarios in Figure 1, as well as the ground truth diarization label for this recording. We use a unique colour to show every unique speaker/cluster within each segmentation scenario after clustering. It can be seen that the decisions made in Segmentation 2 and 3 can provide a good diarization estimate compared to the ground truth labels, while Segmentation 1 and 4 appear to display more errors. We use our proposed cluster-voting (CV) approach to combine these decisions. To do this, we first obtain a pairwise score v between segments. This pairwise score is the count of the number of times that two segments are clustered together, at our decision resolution, across all N segmentation scenarios. For example, from Figure 2 $v(c_1, c_2) = 3$. This is because c_1 and c_2 share a cluster 3 times (all cases except Segmentation 1) out of the 4 clustering decisions made in each segmentation scenario. We refer to this pairwise score as the voting score, thus the name cluster-voting (CV). As every pair of segments can only share a cluster once in every segmentation scenario, we will always have a maximum possible pairwise score of N votes for N linear segmentation cases. We propose using these pairwise voting scores to carry out a final complete-linkage clustering of all segments, which will provide us with a final clustering decision at our decision resolution. As linkage clustering requires a distance, or dissimilarity score, we can conduct clustering using v', where $v'(c_i, c_j) = N - v(c_i, c_j)$. In our approach we use a stopping criterion of $v < 2$ votes for stopping our complete-linkage clustering. This means that we require at least two of our clustering decisions to agree in order to merge a pair of segments, all while segments with higher voting scores will be merged with a higher priority within our clustering hierarchy. We believe this is a justified stopping criterion, as normally we would rely on a single clustering decision (at least 1 vote), but we are now requiring at least two decisions to agree before conducting a merge. This allows us to take advantage of the multiple clustering decisions.

5. EVLUATION RESULTS

We carry our speaker attribution using our proposed CV technique across the SAIVT-BNEWS dataset [11]. We report on the performance of our approach and our baseline attribution system using the standard diarization error rate (DER) metric [22], as well as the cluster purity (CP) and cluster coverage (CC) metrics [12]. In order to report on the errors associated with attributing speaker between independent recordings we compute the DER metric across all recordings, while taking into account the unique global identity of the speakers appearing in each file. We call this the attribution error rate (AER) to avoid confusion with the within-recording diarization errors that are reflected using DER.

Throughout our experiments we use 19 mel-frequency cepstral coefficient (MFCC) features including the zeroth order coefficient with deltas and feature warping [23], which are extracted using a 32 ms Hamming window with a 10 ms window shift. It must also be noted that when we carry out JFA model adaptation, we use the

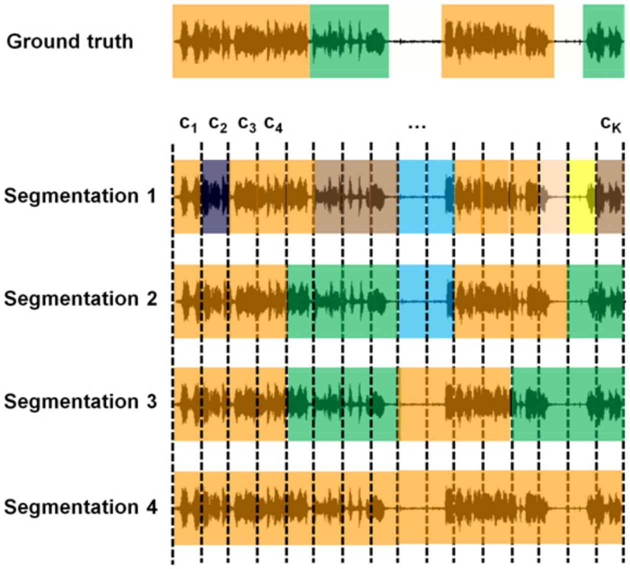

Fig. 2. Result of JFA modeling and complete-linkage clustering applied to each of the segmentation scenarios in Figure 1, as well as the ground truth diarization label for the example recording.

zeroth and first order Baum-Welch statistics for each segment [8]. This is very convenient as we only need to use our UBM to extract these statistics at our decision resolution (minimum segment length used in the first linear segmentation scenario). We can then simply sum these statistics to obtain the segment statistics for the next linear segmentation scenario or a final speaker cluster [12]. Finally, we use $N = 5$ segmentation scenarios and a decision resolution of $L = 1$ second, providing us with a total of 5 clustering decisions at 5 segmentation scenarios to apply CV. This is because we use 5 second segments to apply linear segmentation in our baseline approach and believe that larger segments would be too impure to consider for clustering, however we aim to investigate this in later studies.

5.1. SAIVT-BNEWS evaluation corpus

We employ the SAIVT-BNEWS evaluation corpus [11], which is a publically available collection of Australian broadcast television recordings. This dataset contains 55 videos of news programs with inter-related topics that allow for recurring identities across multiple recordings and session conditions. This corpus is available with a full set of reference diarization labels that can be used for speaker diarization and linking evaluations [11]. In addition, the SAIVT-BNEWS corpus contains a large variety of speakers; reporters, presenters, politicians, elderly people and children. It also contains overlapping speaker segments and music played during broadcast programs. The recordings in this dataset range from 47 seconds to 5 minutes and 47 seconds in recording length and contain from 1 to a maximum of 9 unique speakers within each recording, with a total of 92 unique speakers across the set of recordings in the entire corpus.

5.2. Experimental results

Table 1. Performance evaluation of the baseline system (with Viterbi realignment) compared to our proposed cluster-voting (CV) system that eliminates the need for Viterbi segmentation or refinement.

Diarization	**DER%**	**CP%**	**CC%**	**Speakers**
Baseline	13.1	80.7	92.9	162
cluster-voting	12.4	84.6	91.7	211
Attribution	**AER%**	**CP%**	**CC%**	**Speakers**
Baseline	35.7	74.4	75.3	67
cluster-voting	32.1	77.8	75.1	83

Table 1 displays the performance of our baseline system with Viterbi realignment (Section 3), against our proposed cluster-voting (CV) system, evaluated over the SAIVT-BNEWS corpus. It can be seen that our proposed CV system performs better than the baseline approach for the task of speaker diarization. As the speaker linking stage of the baseline approach and CV system are the same, it can be said that the improvements to the speaker diarization stage have been carried through to achieve an overall relative improvement of approximately 10% in AER, for the ultimate task of speaker attribution. This is because the CV system is able to achieve a higher CP metric in the task of diarization, thus finding more pure intra-recording speaker clusters across the set of recordings, which can then be linked with greater accuracy to achieve improvements with respect to the AER metric. Finally, it can be seen that our proposed CV system also finds a more accurate number of globally unique speakers across the evaluation corpus, when compared to the baseline approach (for reference there are 92 unique speakers in the SAIVT-BNEWS dataset).

6. CONCLUSION

We proposed a clustering-only approach for the task of speaker diarization with the objective of eliminating the need for computationally expensive Viterbi segmentation and realignment. We presented a cluster-voting (CV) technique for taking advantage of multiple clustering decisions, made across multiple linear segmentation scenarios, to achieve a combined clustering decision that is more accurate and more efficient than a single decision that is then refined using Viterbi realignment. Throughout our approach we use JFA model adaptation and complete-linkage clustering to model and cluster speaker segments. This makes our approach highly efficient and thus ideal for dealing with large sets of multiple spoken recordings with recurring speaker identities. We aim to further investigate the benefits of our proposed CV approach for large scale speaker diarization and linking in future studies.

7. ACKNOWLEDGEMENTS

This research was supported by an Australian Research Council (ARC) Linkage Grant (No: LP130100110).

8. REFERENCES

[1] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," Tech. Rep., IBM TJ Watson Research Center, Yorktown Heights, NY, 1998.

[2] Xavier Anguera Mir, Simon Bozonnet, Nicholas W. D. Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech & Language Processing*, pp. 356–370, 2012.

[3] Grégor Dupuy, Mickael Rouvier, Sylvain Meignier, and Yannick Estève, "I-vectors and ILP clustering adapted to cross-show speaker diarization.," in *INTERSPEECH*, 2012.

[4] T. Viet-Anh, L. Viet Bac, C. Barras, and L. Lamel, "Comparing multi-stage approaches for cross-show speaker diarization.," in *INTERSPEECH*. 2011, pp. 1053–1056, ISCA.

[5] D. A. Van Leeuwen, "Speaker linking in large data sets," in *Odyssey2010*, Brno, Czech Republic, June 2010, pp. 202–208.

[6] H. Bourlard, M. Ferras, N. Pappas, A. Popescu-Belis, S. Renals, F. McInnes, P. Bell, S. Ingram, and M. Guillemot, "Processing and linking audio events in large multimedia archives: The eu inevent project," in *Proceedings of SLAM 2013*, 2013.

[7] M. Ferras and H. Bourlard, "Speaker diarization and linking of large corpora," in *IEEE SLT Workshop 2012*, Dec., pp. 280–285.

[8] H. Ghaemmaghami, D. Dean, and S. Sridharan, "Speaker linking using complete-linkage clustering," in *SST2012*, 2012.

[9] C. Vaquero, A. Ortega, and E. Lleida, "Partitioning of two-speaker conversation datasets," in *Interspeech 2011*, August 28-31 2011, pp. 385–388.

[10] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Extending the task of diarization to speaker attribution," in *Interspeech2011*, Florence, Italy, August 2011.

[11] H. Ghaemmaghami, D. Dean, and S. Sridharan, "Speaker attribution of Australian broadcast news data," in *SLAM2013*, Marseille, France, August 2013, pp. 72–77, Sun SITE Central Europe.

[12] H. Ghaemmagami, D. Dean, and S. Sridharan, "An iterative speaker re-diarization scheme for improving speaker-based entity extraction in multimedia archives," in *Interspeech*, 2014.

[13] A. Giraudel, M. Carr, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The REPERE Corpus : a multimodal corpus for person recognition," in *Proceedings of LREC'12*, Istanbul, Turkey, may 2012.

[14] H. Ghaemmaghami, D. Dean, R. Vogt, and S. Sridharan, "Speaker attribution of multiple telephone conversations using a complete-linkage clustering approach," in *IEEE ICASSP2012*, march 2012, pp. 4185 –4188.

[15] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Multimodal Technologies for Perception of Humans*. Springer Berlin / Heidelberg, 2008.

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

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

[18] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal STSP*, vol. 4, no. 6, pp. 1059 –1070, 2010.

[19] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," .

[20] C. Barras, X. Zhu, S. Meignier, and J. Gauvain, "Multistage speaker diarization of broadcast news," *IEEE ASLP*, vol. 14, no. 5, pp. 1505 –1512, 2006.

[21] A.K. Jain, A. Topchy, M.H.C. Law, and J.M. Buhmann, "Landscape of clustering algorithms," in *Proceedings of ICPR2004*, 2004, vol. 1, pp. 260 – 263 Vol.1.

[22] "The NIST rich transcription website," http://www.nist.gov/speech/tests/rt/, 2007.

[23] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Odyssey2001*, June 18-22 2001, pp. 213–218.