LINGUISTIC INFLUENCES ON BOTTOM-UP AND TOP-DOWN CLUSTERING FOR SPEAKER DIARIZATION

Simon Bozonnet, Dong Wang, Nicholas Evans and Raphaël Troncy

EURECOM
BP193, F-06904 Sophia Antipolis Cedex, France
{bozonnet, wangd, evans, troncy}@eurecom.fr

ABSTRACT

While bottom-up approaches have emerged as the standard, default approach to clustering for speaker diarization we have always found the top-down approach gives equivalent or superior performance. Our recent work shows that significant gains in performance can be obtained when cluster purification is applied to the output of top-down systems but that it can degrade performance when applied to the output of bottom-up systems. This paper demonstrates that these observations can be accounted for by factors unrelated to the speaker and that they can impact more strongly on the performance of bottom-up clustering strategies than top-down strategies. Experimental results confirm that clusters produced through top-down clustering are better normalized against phone variation than those produced through bottom-up clustering and that this accounts for the observed inconsistencies in purification performance. The work highlights the need for marginalization strategies which should encourage convergence toward different speakers rather than toward nuisance factors such as that those related to the linguistic content.

Index Terms— Speaker diarization, hierarchical clustering, purification, phone normalization

1. INTRODUCTION

Over the last few years speaker diarization has emerged as a dedicated and increasingly active field of research and has utility in any situation where multiple (and possibly competing, or overlapping) speakers may be expected. Speaker diarization involves identifying the number of speakers within an acoustic stream and the labeling of intervals in which each speaker is active. The problem is usually unsupervised, i.e. no a priori knowledge is available. This leads to a trial-and-error search for an optimal speaker inventory and the two dominant approaches to speaker diarization: bottom-up and top-down.

The bottom-up approach is by far the most popular and systems based on this approach have consistently achieved the best levels of performance in the NIST RT evaluations [1], e.g. [2, 3], although top-down systems also achieve competitive results [4]. While some have reported that bottom-up approaches are more robust than their top-down counterparts [5] our own work [6] shows that the two approaches give comparable results, with neither being consistently superior to the other.

One noticeable difference that we have observed in the performance of the two approaches relates to purification. Purification techniques aim to 'purify' clusters of speech from all but the dominant speaker and are reported by many to give significant improvements with bottom-up approaches [7, 8, 9]. Our experience, however, shows that performance can sometimes deteriorate when purification is applied to bottom-up strategies but that it leads to consistent improvements in top-down systems [6]. These observations led us to investigate the two diarization approaches more thoroughly and to study their relative merits.

The contribution in this paper relates to a comparison of bottom-up and top-down approaches to speaker diarization. The study shows that the two clustering approaches are similarly effective in searching for the optimal number of speakers but behave differently in discriminating between individual speakers and in normalizing nuisance variation. This paper concentrates on linguistic effects which are not explicitly related to differences between speakers. Such factors can make top-down systems more stable but less discriminative, and vice versa for bottom-up systems. We also explain why purification works well with top-down approaches but why it can degrade results when applied to bottom-up systems.

The remainder of this paper is organized as follows. Section 2 aims to formalize the problem of speaker diarization and includes an analysis of the challenges that must be addressed in practical systems. This analysis leads naturally to the bottom-up and top-down approaches which are qualitatively compared in Section 3. Empirical results reported in Section 4 aim to confirm the theory. Conclusions and directions for future work are presented in Section 5.

2. PROBLEM FORMULATION

Let O denote the parameterized audio stream. The task of speaker diarization can be formally defined as follows:

$$\begin{aligned}(\tilde{S},\tilde{G}) &= \underset{S,G}{\operatorname{argmax}}\, P(S,G|O) \\ &= \underset{S,G}{\operatorname{argmax}}\, P(S,G)P(O|S,G) \end{aligned} \quad (1)$$

where S and G are the speaker sequence and segmentation respectively and where $\tilde{S}$ and $\tilde{G}$ are their optimized counterparts representing who (S) spoke when (G). Two models are thus required to solve the optimization task: acoustic speaker models $P(O|S,G)$ and speaker turn models $P(S,G)$. The former are usually conventional Gaussian mixture models (GMMs) whereas the latter are usually omitted altogether.

There are two principle difficulties in implementing a practical speaker diarization system. First, the number of speakers is unknown and it is thus necessary to determine a speaker inventory. Second, whilst the acoustic models depend fundamentally on the speaker, they also depend on a number of other nuisance factors such as the linguistic content. In this paper we assume for simplicity that the major nuisance variation relates only to the phone class of uttered speech, which we denote as Q.

To formulate a solution which addresses these two challenges, we introduce the speaker inventory Δ, and let $\Gamma(\Delta)$ represent all possible speaker sequences. By omitting the speaker turn model we derive the solution from Equation 1 as follows:

$$(\tilde{S}, \tilde{G}, \tilde{\Delta}) = \underset{S,G,\Delta:S\in\Gamma(\Delta)}{\operatorname{argmax}} \sum_{Q} P(O|S,G,Q)P(Q|S)$$
$$= \underset{S,G,\Delta:S\in\Gamma(\Delta)}{\operatorname{argmax}} \sum_{Q} P(O|S,G,Q)P(Q) \quad (2)$$

where Q is naturally independent of G and is further assumed to be independent of S. Equation 2 reveals two important issues that any practical speaker diarization system must address. First, the speaker inventory Δ must be optimized together with the speaker sequence S and the segmentation G. There is no analytical solution for Δ and so a trial-and-error search is typically conducted. This leads to the two principle approaches to speaker diarization: the bottom-up approach, which searches for an optimal Δ by starting with a larger inventory before moving to a smaller inventory, and the top-down approach whose search is performed in the opposite direction. They are commonly referred to as agglomerative and divisive hierarchical clustering respectively. Second, upon the comparison of Equations 1 and 2, we note that the acoustic speaker model $P(O|S,G)$ is phone normalized. This implies that $P(O|S,G)$ must be trained with speech material containing all possible phones, otherwise Q will not be marginalized.

3. BOTTOM-UP VERSUS TOP-DOWN

The bottom-up and top-down approaches to speaker diarization are opposing strategies to determine a speaker inventory Δ. Both approaches have the potential to obtain the same inventory and thus the direction in which it is sought (top-down or bottom-up) is inconsequential; of paramount importance is how well the acoustic speaker models $P(O|S,G)$ are normalized against nuisance factors (such as Q) and how well they discriminate between speakers. In this section we compare the two processes in this respect.

3.1. Normalization and discrimination

Both bottom-up and top-down approaches rely heavily on the expectation maximisation (EM) algorithm and will converge to a local maximum of Equation 2 for a fixed size Δ. In the case where inter-speaker variation dominates over intra-speaker variation then we can assume that the local maximum corresponds to an optimal diarization on speakers, as opposed to any other acoustic class. In this situation, both bottom-up and top-down systems should provide similar diarization outputs. However, where the linguistic content bears a significant influence the local maximum may correspond to other acoustic units, such as phones Q instead of speakers S, particularly if the different speaker models are not well normalized, i.e. Q is not fully marginalized.

The top-down approach draws new speakers from a potentially well-normalized background model and usually exploits a large amount data for model adaptation. In this case linguistic nuances tends to be marginalized and the resulting models tend to be well normalized. However the speaker variation may also be normalized together with linguistic nuances. This is essential to avoid since it leads to less discriminative speaker models. The bottom-up approach, on the contrary, is more likely to converge quickly to a local maximum of Equation 2 due to the large number of small clusters that are created for initialisation, which leads to highly discriminative models at the beginning of the iterative process. However, while these models may discriminate between speakers, they may also discriminate between linguistic nuances, such as particular phone classes. In other words, speaker clusters obtained with the bottom-up approach tend to be poorly normalized. This is particularly true when short-term cepstral-based features are used, though recent work with prosodic features have potential to discourage such behavior [10].

This argument highlights the respective advantage and disadvantage of the two diarization approaches: top-down systems tend to be well normalized but less discriminative, whereas bottom-up systems are more discriminative but less normalized.

3.2. Speaker purification

No matter which approach is applied, the central idea is to maximise discrimination between speakers while normalizing non-speaker variations. For bottom-up systems, the paramount objective is to normalize non-speaker variations, in particular linguistic nuances, while for top-down approaches, the emphasis is to increase discriminability. Purification [4] is such a technique for improving cluster discrimination. By purifying the resulting models of data from other speakers, more discriminative models can be obtained and better diarization results are expected. Significant improvements have been reported with purification for both bottom-up systems [7, 8, 9] and top-down system [4], however the above analysis shows that it is likely to be more efficient with top-down approaches for which speaker purification is essential.

4. EXPERIMENTAL WORK

In this section we present our experimental work which aims to confirm the behaviour of the two approaches outlined above. We briefly describe the two experimental systems in Section 4.1 and datasets in Section 4.2. Diarization results are reported in Section 4.3 before experiments to assess differences in phone normalization and cluster purity are reported in Sections 4.4 and 4.5 respectively.

4.1. Experimental systems

Both our bottom-up and top-down systems were implemented with the same ALIZE software toolkit [11] and development approaches (e.g. pre-processing algorithms, parameter optimization, etc.). Each system starts with a common speech activity (SAD) component which is based upon a two-state hidden Markov model (HMM). The two states represent speech and non-speech events respectively and are 32-component GMM models trained on appropriate external data using an EM/ML algorithm [4]. Iterative Viterbi decoding and model re-estimation are applied to adapt the models to the prevailing ambient conditions.

Segmentation and clustering is then performed according to the bottom-up or top-down scenario. Both rely on a common HMM strategy where each state aims to characterize a single speaker and the state transitions represent speaker turns. Our bottom-up system is an agglomerative hierarchical clustering (AHC) strategy with a sequential EM algorithm based on the approach in [3]. Clustering is controlled according to the Information Change Rate (ICR) [12] and a Ts stopping criterion [13] is used to stop cluster merging. Our top-down system is a divisive hierarchical clustering (DHC) approach based on an evolutive HMM strategy and is exactly as described in [4]. Purification [4] is optionally applied before a common MAP

based re-segmentation with feature normalization is applied to the outputs of each system.

4.2. Datasets

Our experimental systems were optimized on a development dataset of 23 conference meetings from the NIST RT'04, '05 and '06 evaluations. Performance was then assessed on independent RT'07 and RT'09 evaluation datasets and on a separate corpus containing 19 hours of televised, French-language Grand Echiquier (GE) chat/debate television shows [14]. There is no overlap between development and evaluation datasets and in all cases no prior knowledge is available, except an approximate idea of the number of speakers which is used solely in the case of the bottom-up system. This is only so that the system is initialized with a number of clusters that exceeds the likely number of true speakers. Only results obtained on the evaluation datasets are reported here.

4.3. Diarization performance

Diarization Error Rates (DERs) for the four different systems are illustrated in Table 1. Results are presented with (OV) and without (NOV) the scoring of overlapping speech. Since it is the default scoring metric in the NIST RT evaluations we concentrate only on the former. Performance for the bottom-up system is illustrated on row 3 of Table 1. DER scores of 23.8%, 19.1% and 33.7% are obtained on the RT'07, RT'09 and GE datasets respectively. We note a large difference in performance between meeting and TV-show domains. This is mainly due to the higher number of (often relatively inactive) speakers in the case of TV-shows (average of 13 speakers cf. 5 for meeting data).

Performance for the top-down system is given on row 5 of Table 1. DERs of 18.3%, 26.0% and 40.4% are obtained on the three datasets respectively and thus indicate an inconsistency in the comparative performance of top-down and bottom-up approaches: top-down performance is superior for the RT'07 dataset whereas bottom-up performance is superior for the RT'09 and GE datasets. Our hypothesis is that this discrepancy is accounted for by factors that are unrelated to differences between speakers. This argument is explained further in Sections 4.4 and 4.5. First though, we investigate the impact of purification on both system outputs.

The performance of bottom-up and top-down systems with purification is illustrated on rows 4 and 6 of Table 1 respectively. For the bottom-up approach we note that for the RT'07 dataset, even if there is a slight improvement in performance with purification (22.7% DER cf. 23.8%) there is a significant degradation in performance for the RT'09 dataset and a smaller degradation for the GE dataset. For the top-down system, however, performance consistently improves upon the application of purification (bottom two rows) for all three datasets. These observations support our conjecture proposed in Section 3 that (i) the clusters identified by top-down systems are less discriminative and thus require purification, and (ii) those produced with the bottom-up systems are less well normalized against phone variation and that this cannot always be improved upon through purification.

4.4. Phone normalization

As argued above, we hypothesize that bottom-up systems are relatively more likely than top-down systems to converge to sub-optimal local maximums of Equation 2. These are likely to correspond to nuisance variation such as that related to the linguistic content. In order to confirm this hypothesis we computed and compared the distribution of phones within each cluster of the diarization output. This is obtained through an automatic phone alignment using the ground-truth word-level transcriptions. The phone distribution is computed for each cluster according to the fraction of speech time attributed to each phone. Then the average inter-cluster distance D is computed for each file as follows:

$$D = \binom{N}{2}^{-1} \sum_{n=1}^{N} \sum_{m=n+1}^{N} D_{\text{KL2}}(C_n||C_m),$$

where N is the size of the speaker inventory Δ, i.e. the number of clusters, and where the binomial coefficient $\binom{N}{2}$ is the number of unique cluster pairs. $D_{\text{KL2}}(C_n||C_m)$ is the symmetrical Kullback-Leibler (KL) distance between the phone distributions for clusters C_n and C_m, defined as:

$$D_{\text{KL2}}(C_n||C_m) = \frac{1}{2}\Big(D_{\text{KL}}(C_n||C_m) + D_{\text{KL}}(C_m||C_n)\Big)$$

where $D_{\text{KL}}(C_n||C_m)$ is the KL divergence of C_n from C_m. We note that the symmetrical KL metric has been used for the segmentation and clustering of broadcast news [15].

In the case of good phone normalization we expect the average inter-cluster distance to be small since the clusters should have the same phone distribution, while higher average inter-cluster distances may indicate a higher degree of convergence toward phones, or other acoustic classes, rather than toward speakers.

The mean and variance of the average inter-cluster distance for the RT'07 and RT'09 datasets are illustrated in Table 2. Results for the GE dataset are not included since there are no ground-truth word-level transcriptions for this dataset. For the baseline bottom-up system the average inter-cluster distances are 0.17 and 0.14 for the two datasets respectively. When purification is applied these figures fall to 0.13 and 0.12 thereby indicating a slight improvement in phone normalization in both cases. The average inter-cluster distances are consistently lower for the top-down system where they fall from 0.11 and 0.10 to 0.07 and 0.08 with purification. Considering the variances in columns 4 and 5 of Table 2, we note a consistent decrease in all cases: reductions in the mean are accompanied by reductions in the variation. These results suggest that, as predicted, the clusters identified with the top-down system are better normalized against phone variation than those identified with the bottom-up system.

4.5. Cluster purity

The results presented in Section 4.4 do not account for why results deteriorate significantly when purification is applied to the RT'09 dataset. To explain this behavior we analyzed the average speaker purity in each system output. The speaker purity is defined as the percentage of data in each cluster which are attributed to the most dominant speaker. Columns 2 and 3 of Table 3 present the average cluster purities for the RT'07 and RT'09 datasets. For the RT'07 dataset purification leads to marginal improvements: 1.6% absolute improvement for the bottom-up system and 2.3% for the top-down system. However, for the RT'09 dataset performance is different for bottom-up and top-down systems. While cluster purity improves by 2.3% for the top-down system, purity deteriorates by 4% for the bottom-up system.

Since a larger number of clusters will naturally lead to higher purities it is necessary to consider the number of clusters in each case to properly appreciate the resulting effects of purification on diarization

System	RT‘07		RT‘09		GE	
	OV	NOV	OV	NOV	OV	NOV
Bottom-up	23.8	20.8	19.1	13.5	33.7	29.0
Bottom-up+Pur.	22.7	19.6	27.0	21.8	33.9	29.1
Top-down	18.3	15.0	26.0	21.5	40.4	36.0
Top-down+Pur.	17.8	14.4	21.1	16.0	38.5	33.9

Table 1. DERs with (OV) and without (NOV) the scoring of overlapping speech, with and without purification (Pur.).

System	Mean		Variance	
	RT‘07	RT‘09	RT‘07	RT‘09
Bottom-up	0.17	0.14	0.167	0.013
Bottom-up + Pur.	0.13	0.12	0.017	0.005
Top-down	0.11	0.10	0.006	0.004
Top-down + Pur.	0.07	0.08	0.001	0.002

Table 2. Inter-cluster phone distribution distances.

performance. The number of clusters in each system output is illustrated in columns 4 and 5 of Table 3 in which the last row indicates the true number of speakers. All systems over-estimate the number of speakers and purification always reduces their number. When coupled with increases in average purity, then improved diarization performance should be expected. For the bottom-up system and the RT‘09 dataset the decrease in the number of clusters when purification is applied is negligible, whereas the purity also decreases. This can only result in poorer diarization performance.

5. CONCLUSIONS

Even though the performance of bottom-up and top-down approaches to speaker diarization is generally comparable they potentially exhibit different behavior in the face of nuisance factors that are unrelated to different speakers, such as that related to the linguistic content. While bottom-up approaches are more discriminative they tend to produce clusters which are less well normalized against such variation and are thus more likely than their top-down counterparts to converge to other acoustic units that are unrelated to differences between speakers. The latter tend to produce clusters which are better normalized but less discriminative. This explains why performance can sometimes degrade when purification is applied to clusters obtained in bottom-up systems. Future work should focus on enhanced purification algorithms for bottom-up systems and approaches that are generally more robust to nuisance factors such as the linguistic content.

6. ACKNOWLEDGMENTS

This work was partially supported by the French Ministry of Industry (Innovative Web call) under contract 09.2.93.0966, ‘Collaborative Annotation for Video Accessibility’ (ACAV) and by the ‘Adaptable Ambient Living Assistant’ (ALIAS) project funded through the joint national Ambient Assisted Living (AAL) programme.

System	Cluster Purity (%)		No. Clusters	
	RT‘07	RT‘09	RT‘07	RT‘09
Bottom-up	80.6	79.2	7.0	7.0
Bottom-up + Pur.	82.2	75.2	5.8	6.9
Top-down	81.8	79.1	5.0	6.0
Top-down + Pur.	84.1	81.4	4.8	5.3
Ground-truth	100.0	100.0	4.4	5.4

Table 3. Average cluster purity and number of clusters.

7. REFERENCES

[1] NIST, “The NIST Rich Transcription 2009 (RT’09) evaluation,” http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf, 2009.

[2] C. Wooters and M. Huijbregts, “The ICSI RT07s speaker diarization system,” in *Lecture notes in Computer Science - Multimodal Technologies for Perceptionof Humans*, Fiscus Stiefelhagen, Bowers, Ed. 2008, vol. 4625/2008, pp. 509–519, Springer.

[3] H. Sun, B. Ma, S. Z. K. Khine, and H. Li, “Speaker diarization system for RT07 and RT09 meeting room audio,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, March 2010.

[4] S. Bozonnet, N. W. D. Evans, and C. Fredouille, “The LIA-EURECOM RT‘09 Speaker Diarization System: enhancements in speaker modelling and cluster purification,” in *Proc. ICASSP*, Dallas, Texas, USA, March 14-19 2010.

[5] S.E. Tranter and D.A. Reynolds, “An overview of automatic speaker diarization systems,” *IEEE TASLP*, vol. 14, no. 5, pp. 1557–1565, 2006.

[6] S. Bozonnet, N. Evans, C. Fredouille, D. Wang, and R. Troncy, “An integrated top-down/bottom-up approach to speaker diarization,” in *Proc. Interspeech, to appear*, September 2010.

[7] X. Anguera, C. Wooters, and J. Hernando, “Purity algorithms for speaker diarization of meetings data,” in *Proc. ICASSP*, May 2006.

[8] X. Anguera, C. Wooters, and J. Hernando, “Frame purification for cluster comparison in speaker diarization,” in *Second Workshop on Multimodal User Authentication (MMUA)*, 2006.

[9] H. Sun, T. L. Nwe, B. Ma, and H. Li, “Speaker diarization for meeting room audio,” in *Proc. Interspeech’09*, September 2009.

[10] G. Friedland, O. Vinyals, Yan Huang, and C. Muller, “Prosodic and other long-term features for speaker diarization,” *IEEE TASLP*, vol. 17, no. 5, pp. 985–993, July 2009.

[11] J.-F. Bonastre, F. Wils, and S. Meignier, “ALIZE, a free toolkit for speaker recognition,” in *Proc. ICASSP’05*, Philadelphia, USA, March 2005, vol. 1, pp. 737–740.

[12] K.J. Han, S. Kim, and S.S. Narayanan, “Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization,” *IEEE TASLP*, vol. 16, no. 8, pp. 1590–1601, Nov. 2008.

[13] T. H. Nguyen, E. S. Chng, and H. Li, “T-test distance and clustering criterion for speaker diarization,” in *Proc. Interspeech*, Brisbane, Australia, 2008.

[14] S. Bozonnet, F. Vallet, N. W. D. Evans, S. Essid, G. Richard, and J. Carrive, “A multimodal approach to initialisation for top-down speaker diarization of television shows,” in *Proc. EUPSICO 2010, 18th European Signal Processing Conference, August 23-27, 2010, Aalborg, Denmark*, 08 2010.

[15] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, “Automatic segmentation, classification and clustering of broadcast news audio,” in *Proc. DARPA Speech Recognition Workshop*, 1997, pp. 97–99.