INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Priors for Speaker Counting and Diarization with AHC

Gregory Sell, Alan McCree, and Daniel Garcia-Romero

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

Abstract

Estimating the number of speakers in an audio segment is a necessary step in the process of speaker diarization, but current diarization algorithms do not explicitly define a prior probability on this estimation. This work proposes a process for including priors in speaker diarization with agglomerative hierarchical clustering (AHC). It is also shown that the exclusion of a prior with AHC is itself implicitly a prior, which is found to be geometric growth in the number of speakers. By using more sensible priors, we are able to demonstrate significantly improved robustness to calibration error for speaker counting and speaker diarization.

Index Terms: speaker diarization, i-vector, clustering

1. Introduction

Speaker diarization is the process of grouping segments of speech from the same speaker, often referred to as determining who is speaking when. Typically, this task is performed without any knowledge of who the speakers are or even how many speakers there are. As a result, estimating the number of speakers (or clusters) is a necessary step in the diarization process.

There are numerous approaches for determining the number of speakers from a set of unclustered frames or segments of speech. In many cases, but not all, these segments are represented as i-vectors [1], a fixed-dimensional representation of an audio segment that performs well for speaker recognition and language identification. Ideally, we would partition these segments by computing the full posterior over the entire partition space, but this has been shown to become quickly infeasible for even a small number of segments [2]. Instead, a clustering algorithm in i-vector space is typically preferred. Prior work has explored K-means [3], Hierarchical Direchlet Process Hidden Markov Modelling (HDP-HMM) [4], spectral clustering [5], Variational Bayes Expectation Maximization Gaussian Mixture Modelling (VBEM-GMM) [6], mean shift [7], and agglomerative hierarchical clustering (AHC) [8].

However, the role of speaker priors has only been considered for a few of these methods. In some cases, a prior can be defined via hyperparameters (such as in HDP-HMM and VBEM-GMM), but this approach also requires costly iterative updates to converge on a solution. In other approaches, such as mean shift and AHC, the number of speakers is indirectly determined by a parameter (the bandwidth in mean shift or stopping criterion in AHC) that does not obviously relate to any particular prior. In some work, fragments of prior information were included by forcing a decision of at least two speakers (e.g. [6]), or by forcing a decision of exactly two speakers (e.g. [8]). But, otherwise, prior probabilities for number of speakers have played only a small role in speaker diarization.

In the work that follows, we define the process for including speaker priors for AHC, which is preferred for its speed and efficiency. First, in Section 2, we briefly describe the AHC diarization systems to be used in later experiments. Then, in Section 3, we describe the process for explicitly defining priors in AHC, as well as deriving the implicit prior in AHC. Section 4 presents results for measuring performance at different calibration shifts for speaker counting and speaker diarization, followed by concluding remarks.

2. AHC Diarization

AHC is a reasonable choice for speaker diarization with i-vector clustering because it is efficient and independent of initialization [8, 9] . Furthermore, unlike other clustering methods that fix the number of speakers explicitly (or fit multiple models in parallel), AHC instead merges clusters until there are no longer any pairs deemed to originate from the same speaker, essentially breaking the clustering into a series of speaker recognition decisions. The drawback to AHC is that, in order to achieve its efficiency and replicability, the process is greedy. Instead of considering all possible partitions of the segments (as discussed in [2]), AHC only considers a single partitioning at each level: merging the nearest clusters from the previous level.

I-vector clustering with AHC for diarization typically follows several steps. First, the speech is broken into short segments (roughly 1-2 seconds) based on marks from speech activity detection. Second, an i-vector is extracted for each of these segments, and a matrix of similarity scores is then computed over all i-vector pairs. AHC begins with all segments as separate clusters and then uses the score matrix to greedily merge them until a stopping criterion is met (usually either a specified number of clusters or a maximum distance between clusters).

Recent work has shown that scoring with Probablistic Linear Discriminant Analysis (PLDA) [10] provides an improvement over cosine scoring [9], and so this is the scoring process we will use in the work that follows. We will also consider two types of i-vector extraction. In the first type (called acoustic i-vectors), the acoustic features are clustered with an unsupervised universal background model (UBM), which is then used to compute the frame-level sufficient statistics for i-vector extraction [9]. In the second type (called DNN i-vectors), a DNN trained to map acoustic features to senone posteriors is used to define the zero-order statistics, which are then used in the sufficient statistics for i-vector extraction [11].

3. Speaker Priors for AHC

We propose explicitly inserting a prior for the number of speakers into the diarization process. However, in order to accom-

http://dx.doi.org/10.21437/Interspeech.2016-1380

plish this, we must first pose AHC as a probabilistic process. AHC is deterministic for a given matrix of scores, and so the proper approach for including a prior for number of clusters is not immediately evident. It would be possible to include this sort of information into the stopping criterion of the AHC, but there is a more direct approach. If we instead soften the hard decisions in AHC so that the probability of stopping at step k (called λ_k) is $p(\mathcal{H}_D|\lambda_k)$, and the probability of merging and continuing is $p(\mathcal{H}_S|\lambda_k)$, then priors can be naturally introduced. In this process, stopping at step k means a decision for $N-k$ speakers, if N is the total number of segments (or levels in AHC).

A decision of $N-k$ speakers requires merges for the first $k-1$ steps followed by a stop at step k.

$$p_{\#}(N-k) = \prod_{i=0}^{k-1}\Big[p(\mathcal{H}_S|\lambda_i)\Big]p(\mathcal{H}_D|\lambda_k) \quad (1)$$

By observing $p(\mathcal{H}_S|\lambda_k) = 1-p(\mathcal{H}_D|\lambda_k)$, the product of same hypotheses can be restated in terms of the prior distribution.

$$\begin{aligned}\prod_{i=0}^{k-1} p(\mathcal{H}_S|\lambda_i) &= \prod_{i=0}^{k-2}\Big[1-p(\mathcal{H}_D|\lambda_i)\Big]\Big[1-p(\mathcal{H}_D|\lambda_{k-1})\Big] \\ &= \prod_{i=0}^{k-2}\Big[1-p(\mathcal{H}_D|\lambda_i)\Big] - p_{\#}(N-k-1) \\ &= 1-\sum_{i=0}^{k-1} p_{\#}(N-i)\end{aligned}$$

Reinserting this summation into Eq. (1) defines the stopping probability at step k in terms of a speaker prior $p_{\#}$.

$$p(\mathcal{H}_D|\lambda_k) = \frac{p_{\#}(N-k)}{1-\sum_{i=0}^{k-1} p_{\#}(N-i)} \quad (2)$$

We can now introduce evidence to the process, which, in the case of diarization with AHC, is the probability that the two most similar clusters are from different speakers. So, our goal now is to determine the probability of a decision of continuing at step λ_k given evidence x_k. Using Bayes' Rule, we obtain

$$p(\mathcal{H}_D|\lambda_k, x_k) = \frac{p(x_k|\mathcal{H}_D,\lambda_k)p(\mathcal{H}_D|\lambda_k)}{\sum_{\mathcal{H}} p(x_k|\mathcal{H},\lambda_k)p(\mathcal{H}|\lambda_k)} \quad (3)$$

The diarization systems described above use PLDA scoring for determining the log likelihood ratio l_k of the two most similar clusters coming from same or different distributions.

$$L_k = e^{l_k} = \frac{p(x_k|\mathcal{H}_S,\lambda_k)}{p(x_k|\mathcal{H}_D,\lambda_k)} \quad (4)$$

By combining Eqs. (3) and (4), we are able to completely define the probability of stopping at a particular step of the AHC process given prior probabilities and PLDA likelihood ratios.

$$p(\mathcal{H}_D|\lambda_k, x_k) = \frac{p(\mathcal{H}_D|\lambda_k)}{L_k + (1-L_k)p(\mathcal{H}_D|\lambda_k)} \quad (5)$$

The posterior speaker probabilities can also be updated to include all evidence **x**, via Eq. (1).

$$p_{\#}(N-k|\mathbf{x}) = \prod_{i=0}^{k-1}\Big[1-p(\mathcal{H}_D|\lambda_i, x_i)\Big]p(\mathcal{H}_D|\lambda_k, x_k)$$

Finally, we wish to return to a deterministic process, as we began. When the stopping probabilities in Eq. (5) are forced to 0 if below 0.5, or to 1 if above 0.5, then this clustering exactly replicates the deterministic outcome from AHC. In practice, we smoothly approximate the binarization with a sigmoid function.

3.1. Implicit AHC Prior

The systems described above in Section 2 for speaker diarization with AHC clustering do not specifically employ any particular prior. However, as is often the case, the absence of a prior is implicitly a prior itself. Typically, the absence of a prior is equivalent to a flat prior, but this is not the case for AHC. Instead, the absence of prior is equivalent to a flat stopping probability at all steps of the merge (i.e. $p(\mathcal{H}_D|\lambda_k) = 0.5, \forall k$).

Plugging this into Eq. (1) shows that the implicit AHC prior grows geometrically with the number of speakers m, which is almost certainly a poor choice.

$$p_{\#}(m) = 2^{(m-N)}, \text{ where } m = N-k \quad (6)$$

While the ideal prior will vary for different corpora, it is difficult to imagine a circumstance where the most likely number of speakers is equal to the number of 1-2 second segments in the duration of the conversation. This offers promise to the potential for improvement from using Eq. (5) to include a more sensible prior.

3.2. Calibration Error with Priors

Score calibration interprets likelihood ratios as probabilities [12]. In [9], unsupervised calibration from in-domain, unlabeled data was used successfully for PLDA score calibration for AHC diarization. However, this process required advance access to the evaluation data (without labels) in order to learn the parameters, which may not always be possible.

While true logistic regression calibration learns both a scale and shift (or bias) parameter in log-space, in the case of the unsupervised calibration in [9], only a shift parameter was learned. In this sense, the process determines a threshold rather than a full calibration function. So, for simplicity, we will only consider the shift parameter here (referred to as β), which can be easily factored into Eq. (5).

$$p(\mathcal{H}_D|\lambda_k, x_k) = \frac{p(\mathcal{H}_D|\lambda_k)}{e^{-\beta}L_k + (1-e^{-\beta}L_k)p(\mathcal{H}_D|\lambda_k)} \quad (7)$$

Note that, since β is a linear shift to l_k, it must be exponentiated and multiplied with L_k here.

The relationship in Eq. (7) between the threshold, the prior stopping probability, and the stopping probability with evidence will drive the subsequent experiments.

4. Experiments with Speaker Priors

In order to explore the role of speaker priors in diarization, we examine several experiments with the LDC CALLHOME corpus, a collection of multi-lingual telephone data with conversations between 2-7 speakers. The true distribution of number of speakers is shown in Fig. 1 (the oracle prior).

The acoustic and DNN diarization systems were trained as described in [9] and [11], respectively. Also, the stopping probabilities in Eq. (5) were roughly binarized with a sigmoid function centered at 0.5 and with a scaling parameter of 1000. Other sigmoid scaling parameters were also considered, but the stability of a sharper sigmoid was found to be preferable, likely due to

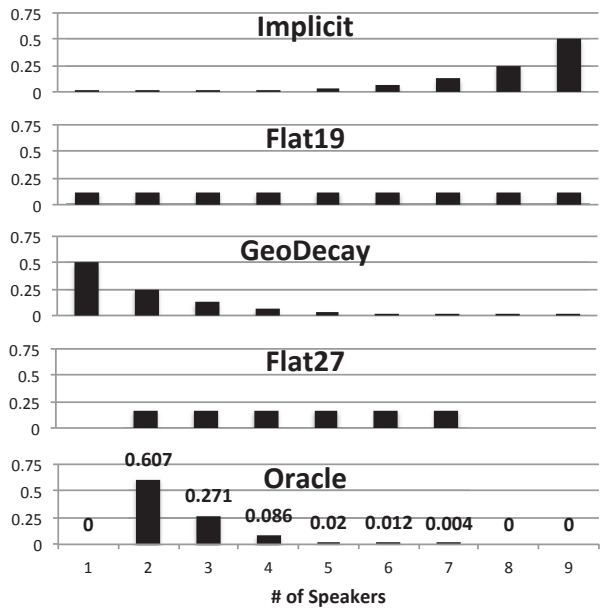


Figure 1: *Prior distributions used in the experiments. The oracle distribution is the true distribution of CALLHOME.*

the lack of a scaling in the calibration. The number of speakers was then selected according to the maximum in the posterior.

4.1. Expected Posterior with Calibration Error

A first experiment in examining priors was to consider the effect of calibration error in the absence of evidence, which was achieved by setting all likelihood ratios to 1 in Eq. (7) (i.e., $L_k = 1, \forall k$). Without evidence, the calibration error only serves to modify the prior, and so this first experiment examines the effect of calibration on the prior itself.

In this experiment and others to follow, we considered five different priors:

- **Implicit** - The implicit prior (Eq. (6)).
- **Flat19** - Flat prior probability across 1-9 speakers.
- **GeoDecay** - Geometric decay $p_{\#}(m) = 2^{-m}$.
- **Flat27** - Flat prior probability across 2-7 speakers (the oracle range of outcomes for the corpus).
- **Oracle** - The oracle CALLHOME prior.

These speaker priors are converted to a prior stopping probability at each level via Eq. (2) and then inserted into Eq. (5) to compute stopping probabilities. In all cases, more than 9 speakers was set to zero probability for simplicity. Note that the last two priors (flat27 and oracle) utilize some degree of oracle information. Each of these distributions is shown in Fig. 1.

Figure 2 shows the cross-entropy of the modified prior and the true CALLHOME oracle prior as a function of calibration shift as a means of measuring the difference between the distributions. A negative shift indicates a tendency toward a smaller number of speakers, while a positive shift encourages more speakers. The implicit prior has the largest cross-entropy at all positive shifts and for small magnitude negative shifts as well. For large negative shifts, the geometric decay prior is worst, because it already gives a great deal of mass to one speaker, and so

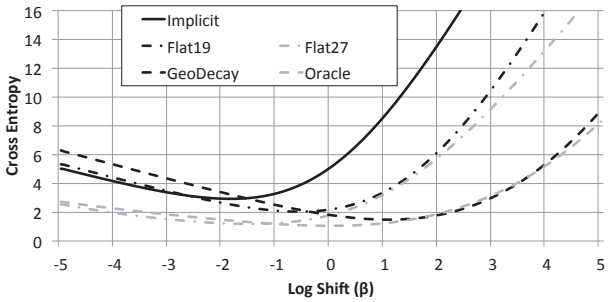


Figure 2: *Cross-entropy of various priors and the oracle prior as a function of calibration shift. Priors in gray utilize oracle information.*

a negative shift quickly feeds this tendency. The flat27 and oracle priors, on the other hand, are comparatively stable for large negative shifts, because they give no mass to a single speaker.

However, for positive shifts, it is clear that priors that give less weight to more speakers (here, the oracle and geometric decay priors) are more stable. But, for these shifts, even a flat prior is significantly preferable to the implicit prior.

4.2. Speaker Counting with Calibration Error

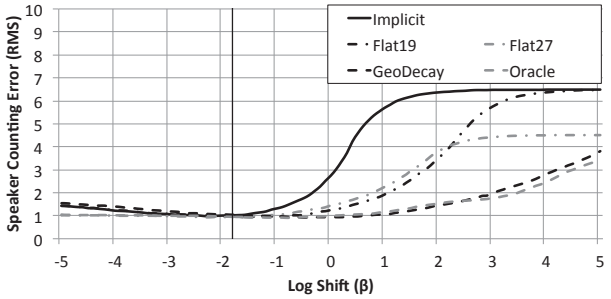
For a second experiment, we explored speaker counting on the CALLHOME corpus as a function of calibration shift. In this case, AHC diarization was used to estimate the number of speakers for a particular conversation, and error was measured as the root-mean-square (RMS) of the absolute error. This error RMS for the acoustic and DNN diarization systems is plotted for the five different priors in Fig. 3. It is worth noting that the unsupervised calibration algorithm determined a calibration threshold of -1.88 for the acoustic i-vectors and -1.40 for the DNN i-vectors, which are marked in Fig. 3

A first observation is that the acoustic and DNN systems behave very similarly. Also, it is noteworthy that for negative log shifts of larger magnitude (which corresponds to estimating fewer speakers), most priors behave identically. The only true difference at the negative end of the graph is that the flat27 prior and oracle prior do not increase in error for very large negative shifts, due to zero probability for single speaker decisions.

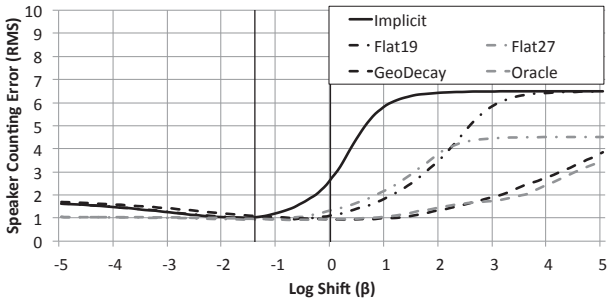
However, the priors show significantly different behaviors for positive log shifts. The implicit prior increases in error at a lower calibration shift than any other prior, and it saturates at maximum error quickly. This observation aligns with the derivation from Section 3.1 that the implicit prior of AHC is geometric growth, and so it would be expected to predict more speakers for a lower calibration shift than other priors.

The errors for the flat priors increase for positive shifts, though the climb is much slower than for the implicit prior. Also, the exclusion of speakers greater than 7 from the flat27 prior leads to a lower maximum error, but the two flat priors are otherwise nearly identical in error. The geometric decay and oracle priors both stay at a lower error for a wider range of shifts, and also climb slowest.

In general, it is easy to see that speaker counting is more robust to threshold variation for increasingly accurate priors. Improving the prior also improves the optimal performance, but the difference is reasonably small. However, the range of scores for which the performance is near optimal significantly widens as the prior improves. It is also worth noting that, for this exper-

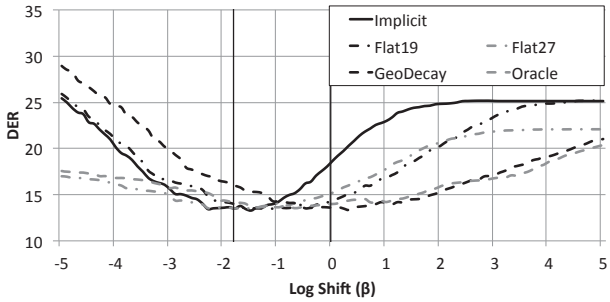


(a) Acoustic

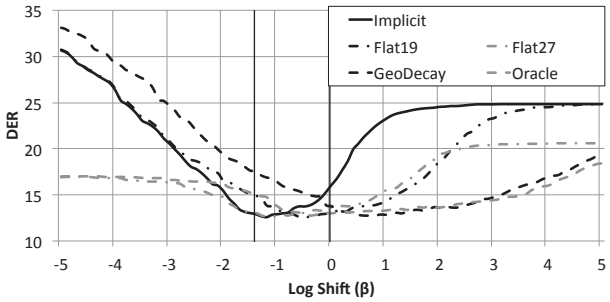


(b) DNN

Figure 3: *RMS of the error for predicted number of speakers as a function of calibration shift for the acoustic and DNN diarization systems with several priors. Priors in gray utilize oracle information and the thresholds from unsupervised calibration are marked along with no calibration (0).*



(a) Acoustic



(b) DNN

Figure 4: *DER as a function of calibration shift for the acoustic and DNN diarization systems with several priors. Systems in gray utilize oracle information, and the thresholds from unsupervised calibration are marked along with no calibration (0).*

iment, a reasonably selected geometric decay is not obviously worse than the true oracle information, with the exception of the small error difference for extremely negative shifts.

4.3. Diarization with Calibration Error

We also examined the relationship of priors and calibration for the diarization itself, with results shown in terms of diarization error rate (DER), which combines miss, false alarm, and speaker confusion errors, in Fig. 4. However, in these experiments, we observed the typical practice of using oracle speech activity marks, and so the DER in Fig. 4 only corresponds to speaker confusion error.

In many ways, these results follow similar patterns to the speaker counting results in the previous section, especially for positive log shifts. The implicit prior increases in error for the lowest shift, followed by the flat priors, while the geometric decay and oracle priors are most resistant to increases in error for positive log shifts.

However, for negative shifts, the cost of bad priors is more clearly shown here than for speaker counting. In this case, the geometric decay performs at a roughly 5% absolute DER worse than all other priors for most negative log shifts. This is because the geometric decay prior gives the greatest mass to estimating a single speaker, and so diarization with that prior is most prone to merging all segments into a single speaker. The alternate version of this effect can also be seen for the flat27 and oracle priors, which hardly increase in error at all, even for the largest negative shifts, because merging all segments to a single speaker is prohibited by those priors.

So, in this case, it is clearly seen that, while good information in the prior probabilities can significantly widen the threshold range for optimal performance, bad information in the prior probabilities can hurt performance.

It is also worth noting that Fig. 4 is also strikingly similar to Fig. 2, indicating that the diarization results are indeed related to similarity between the oracle and realized prior.

The thresholds from unsupervised calibration (-1.88 for acoustic and -1.40 for DNN) are also reasonable choices for all priors except the geometric decay, but, as the priors improve, their optimal bowl widens towards and beyond a shift of 0. And so, it appears that improved knowledge of the speaker distribution can help mitigate a poor threshold, but it also appears that the value of improved priors is largely neutralized if an ideal threshold is selected.

5. Conclusion

This work derived the role of speaker priors in speaker diarization with AHC, and, given this understanding, showed that the absence of a prior is in practice a prior that grows geometrically with number of speakers. It was subsequently shown that incorporating more reasonable priors increases stability for speaker counting and speaker diarization to errors from inaccurate calibration. This development is especially important if in-domain calibration, either supervised or unsupervised, is not accurate (or possible). It is also possible that partial information may be known in advance, and this approach allows the incorporation of that knowledge as well. The work presented here suggests that limited knowledge of the speaker distribution can significantly improve performance in the face of suboptimal calibration, even if that limited knowledge is something as simple as a flat prior.

6. References

[1] 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*, vol. 19, no. 4, pp. 788–98, May 2011.

[2] N. Brümmer and E. de Villiers, "The speaker paritioning problem," in *Proceedings of Odyssey*, 2010.

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

[4] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "A Sticky HDP-HMM with Application to Speaker Diarization," *Annals of Applied Statistics*, 2010.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] S. J. D. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[11] G. Sell, D. Garcia-Romero, and A. McCree, "Speaker Diarization with I-Vectors from DNN Senone Posteriors," in *Proceedings of Interspeech*, 2015.

[12] N. Brümmer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, University of Stellenbosch, December 2010.