# AN ADAPTIVE INITIALIZATION METHOD FOR SPEAKER DIARIZATION BASED ON PROSODIC FEATURES

*David Imseng*[*]

david.imseng@idiap.ch
Idiap Research Institute, Switzerland
Ecole Polytechnique Fédérale, Switzerland

*Gerald Friedland*[†]

fractor@icsi.berkeley.edu
International Computer Science Institute
Berkeley, CA, 94704

## ABSTRACT

The following article presents a novel, adaptive initialization scheme that can be applied to most state-of-the-art Speaker Diarization algorithms, i.e. algorithms that use agglomerative hierarchical clustering with Bayesian Information Criterion (BIC) and Gaussian Mixture Models (GMMs) of frame-based cepstral features (MFCCs). The initialization method is a combination of the recently proposed "adaptive seconds per Gaussian" (ASPG) method and a new pre-clustering and number of initial clusters estimation method based on prosodic features. The presented initialization method has two important advantages. First, the method requires no manual tuning and is robust against file length and speaker count variations. Second, the method outperforms our previously used initialization methods on all benchmark files that were presented in the 2006, 2007, and 2009 NIST Rich Transcription (RT) evaluations and results in a Diarization Error Rate (DER) improvement of up to 67% (relative).

***Index Terms***— Speaker Diarization, Prosodic features, Gaussian Mixture Models

## 1. INTRODUCTION

The goal of Speaker Diarization is to segment audio into speaker-homogeneous regions trying to answer the question "who spoke when?". Most state-of-the-art systems use a combination of agglomerative hierarchical clustering with Bayesian Information Criterion (BIC) and Gaussian Mixture Models (GMMs) of frame-based cepstral features (MFCCs). Most, if not all, of these approaches ultimately require a certain level of manual tuning of the initialization parameters, including the initial amount of clusters $k$ and the initial number of Gaussians per cluster $g$. As shown in [1], the robustness of these systems can depend heavily on the manual tuning of the above mentioned parameters, which we call *key initialization parameters*.

[*]David Imseng was supported by the Swiss-funded IM2 project.

[†]Gerald Friedland was supported by the Swiss-funded IM2 project and the European-Union-funded AMIDA project

In [1], we recently proposed an "adaptive seconds per Gaussian" (ASPG) approach, which automatically adapts one of the two key initialization parameters. ASPG is based on the relationship between the optimal amount of speech data per Gaussian and the duration of the speech data. It is shown that the system in [1] is robust against meeting length variation and that ASPG generalizes to different datasets.

In this paper, ASPG is combined with a new method based on prosodic and other long-term features. We present how to use prosodic features, that have a good speaker discrimination ability according to the ranking method proposed in [2], to estimate the other key initialization parameter and to perform a non-uniform initialization. A comprehensive evaluation on the 2006, 2007, and 2009 NIST Rich Transcription evaluation sets (also known as RT-06, RT-07, and RT-09, respectively) compares the proposed approach to the ASPG model and to the state-of-the-art baseline system, resulting in a performance gain (from 12% up to 67% relative) compared to the baseline. Further, analysis of the experimental results illustrates that, compared to the baseline, the proposed system is more robust against variable meeting lengths and also promises more robustness against other kinds of data variation such as changes in the number of speakers as well.

The remainder of this paper is organized as follows: Section 2 summarizes some related work. Section 3 presents the baseline system with its key initialization parameters. Section 4 introduces our new approach and experimental results are presented in Section 5. Section 6 concludes the article with thoughts on future work.

## 2. RELATED WORK

Past work in initialization methods for Speaker Diarization has also concentrated on adapting the key initialization parameters and performing non-uniform initialization strategies, however, none of the previously proposed methods that we know of, have proved stable enough across different benchmark sets, recording length variation, and variations in the number of speakers. In [3], the "Cluster Complexity Ratio" (CCR) is used to adapt both key initialization parameters and in [4], the "constant seconds per Gaussian" (CSPG) is

used to adapt one key parameter and to initialize the system non-uniformly. In [5], an approach called "friends and enemies" is presented and a relative performance improvement of 13% on NIST RT-05 data is reported but the method was not robust enough to be used in the NIST RT-07 evaluation. In [6], spatial information, based on Time Delay of Arrival (TDOA) estimation, is used to perform a non-uniform initialization, which yields a relative performance improvement of 4% for multiple channel recordings (NIST RT-07 evaluation set). The method presented in this article is generalizeable also to the single-microphone case where TDOA information is not available. In [7], a uniform initialization is compared to a K-means initialization and it is claimed that the type of initialization does not have significant impact on the result. In [8] a special version of K-means in combination with a maximum likelihood criterion is then used to initialize the described system.

## 3. BASELINE SYSTEM

For the experiments presented in this article, we used the ICSI Speaker Diarization engine. The baseline system is presented in [9] and the novel initialization method is implemented in the current version that was used for the NIST RT-09 evaluation. This study investigates the behavior of the agglomerative clustering algorithm which is described in [1], also including an overview over all tunable parameters. In this work, only a short overview over the tunable key parameters (the number of initial clusters $k$ and the amount of Gaussians per initial cluster $g$) is given.

The algorithm is initialized using $k$ clusters, where $k$ is larger than the number of speakers that are assumed to appear in the recording. Every cluster is modeled with a Gaussian Mixture Model containing $g$ Gaussians. In order to train initial GMMs for the $k$ speaker clusters, an initial segmentation is generated by uniformly partitioning the audio into $k$ segments of the same length.

NIST distinguishes between recordings with multiple distant microphones (MDM) and recordings with one single distant microphone (SDM). In the case of MDM, beamforming is typically performed to produce a single channel out of all available ones and often the delay between different channels is used as a feature and combined with MFCCs as in [9]. In this article we present results for both, SDM and MDM recordings. In the case of MDM we are using the beamformed channel but we do not use the delays between channels as an additional feature stream.

The output of a Speaker Diarization system consists of meta-data describing speech segments in terms of starting time, ending time, and speaker cluster name and is usually evaluated against manually annotated ground truth segments. A dynamic programming procedure is used to find the optimal one-to-one mapping between the hypothesis and the ground truth segments so that the total overlap between the reference speaker and the corresponding mapped hypothesized speaker cluster is maximized. The difference is expressed as Diarization Error Rate (DER), which is defined by NIST[1]. The DER can be decomposed into two main components: Speech/non-speech Errors (misses and false alarms) and Speaker Errors (mapped reference is not the same as hypothesized speaker). For this study we focus on Speaker Errors and disregard parameter tuning for speech/non-speech detection as this is usually seen as a separate task.

| Category | Feature ID | Short description |
|---|---|---|
| pitch | f0_median | median of the pitch |
| pitch | f0_min | min of the pitch |
| pitch | f0_mean_curve | mean of the pitch tier |
| formants | f4_stddev | std dev of the 4th formant |
| formants | f4_min | min of the 4th formant |
| formants | f4_mean | mean of the 4th formant |
| formants | f5_stddev | std dev of the 5th formant |
| formants | f5_min | min of the 5th formant |
| formants | f5_mean | mean of the 5th formant |
| harmonic | harm_mean | mean of the harmonics-to-noise ratio |
| formant | form_disp_mean | mean of the formant dispersion |
| pitch | pp_period_mean | mean of the pointprocess of the periodicity contour |

**Table 1**. The 12 prosodic features used in the proposed initialization method (see also [2]).

## 4. AUTOMATIC PARAMETER ESTIMATION AND NON-UNIFORM INITIALIZATION

The analysis presented in [1], showed that it is possible to achieve a lower Diarization Error Rate by tuning the two key initialization parameters, namely the number of initial clusters $k$ and the number of Gaussians per initial cluster $g$. These two parameters can be summarized into a seconds per Gaussian parameter *sec per gauss* $= \frac{\text{speech duration in seconds}}{g \cdot k}$. It was found that by fixing $g = 4$ and estimating $k$ with a simple linear regression based on *sec per gauss*, the system behaves more robust, especially for shorter meetings [1]. In this section, we present another method to estimate $k$ (see Figure 1) and propose to use the aforementioned linear regression to adapt $g$ accordingly. The presented method estimates the number of initial clusters and also provides a non-uniform initialization for the agglomerative clustering procedure based on the long-term feature study and ranking presented in [2]. Derived from the ranking in [2], the 12 top-ranked prosodic features (listed in Table 1) are extracted on all the speech regions (speech/non-speech detector, see [9]) in the recording. The features are extracted with the help of

[1] http://www.itl.nist.gov/iad/mig/tests/rt/2009/index.html

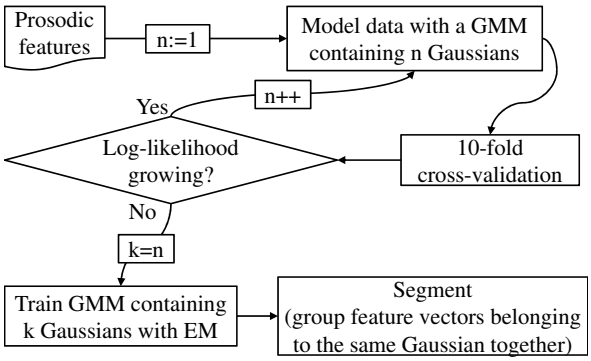


**Fig. 1**. A schematic view of the new method. Prosodic features are used to estimate the number of initial clusters $k$ and the clustering results in a non-uniform initialization.

praatlib, a library that uses Praat[2], on all the speech regions of the recordings (one-second-Hamming-window). The 12-dimensional feature vectors are clustered using one GMM with diagonal covariance. This clustering serves as an initialization for an agglomerative clustering algorithm, therefore it is desired that the model selection tends to over-estimate the number of speakers. The agglomerative clustering algorithm will merge redundant clusters; however, it is not able to split clusters. To determine the number of Gaussians, a 10-fold cross-validation is used to calculate the log-likelihood of GMMs with different numbers of Gaussians (iteration loop in Figure 1). Then, Expectation Maximization is used to train the GMM (consisting of the previously determined number of Gaussians) on all the feature vectors extracted from the speech segments of the entire recording. Finally, every feature vector is assigned to one of the Gaussians in the GMM. We can group all of the feature vectors belonging to the same Gaussian into the same initial segment. The clustering thus results in a non-uniform initialization where the number of initial clusters is automatically determined (the number of initial clusters $k$ is set to the number of clusters in the Expectation Maximization algorithm). The extraction of the long-term features and the clustering procedure to estimate the number of initial clusters and to perform a non-uniform initialization adds about realtime to the processing time.

## 5. EXPERIMENTS

To show the performance of the proposed method, experiments on all previous NIST RT evaluation datasets since 2006 were performed. All evaluation sets (RT-06, RT-07 and RT-09) were only used for testing, not for training or manual tuning. The development set from the NIST RT-06 was used to determine the linear regression that was used to estimate $g$ and to tune the baseline system parameters. The complete meetings and also shorter segments (100-, 300- and 500-second segments), split as described in [1], were processed. The baseline system uses the static initialization parameters $k = 16$ and $g = 5$ (manually tuned) and ASPG estimates $k$, based on the linear regression, and uses a fixed $g = 4$ ([1]). The new system, proposed in this paper, uses ASPG to estimate $g$, and prosodic features to estimate $k$ and to perform a non-uniform initialization. Our baseline system performed better than the related approaches (mentioned in Section 2) on the RT-06 and RT-07 evaluation sets[3]. For the RT-09 evaluation set, no results are available for the related work. Therefore we compare the new approach only to the baseline and ASPG.

In Table 2, the results of the experiments on the RT-06 and RT-07 evaluation datasets are presented (about 5.4 hours of data). The new approach outperforms the baseline by relative improvements ranging from 18% to 67%. The new method yields a better performance than ASPG and the baseline system, for all different meeting lengths and for both recording conditions (MDM as well as SDM). The results of the experiments on the RT-09 evaluation set are also presented in Table 2 (about 3 hours of data). RT-09 was considered much more difficult than the previous ones because it contains more speakers (up to 11) and more overlap (up to 37% per meeting). Nevertheless, the proposed approach behaves robustly on that dataset as well, resulting in relative improvements from 12% up to 52%. Interestingly, in the case of complete MDM recordings (all evaluation sets), the new approach leads to a high performance gain compared to ASPG and is able to lower the DER compared to the baseline system.

It can be seen that the new approach is more robust to meeting length variations. The correlation between the difference in Speaker Error (between baseline and our proposed approach) and the number of speakers in a meeting was calculated for the NIST RT-09 evaluation set (RT-09 contains up to 11 speakers per meeting). The positive correlation value of 0.45 indicates that the new approach yields more improvement if there are more speakers and is thus also more robust to the variability of the number of speakers in a recording than the baseline.

## 6. CONCLUSION AND FUTURE WORK

In this paper we proposed a new method to automatically initialize a typical state-of-the-art Speaker Diarization system, i.e. a system that uses agglomerative hierarchical clustering of Gaussian Mixture Models representing cepstral features. The novel initialization method requires no manual tuning and therefore contributes to the robustness of Speaker Diarization yielding a relative improvement of up to 67 % compared to a state-of-the-art baseline system. The new system is more robust to meeting length variation and variations in the number of speakers. In addition, these conclusions generalize to all NIST RT evaluation sets since 2006.

[2]http://www.fon.hum.uva.nl/praat/

[3]http://www.itl.nist.gov/iad/mig/tests/rt/2007/index.html

| | | NIST RT-06/RT-07 evaluation sets | | | | NIST RT-09 evaluation set | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | MDM | | SDM | | MDM | | SDM | |
| Duration | Configuration | Spkr Err | Rel. +/- | Spkr Err | Rel. +/- | Spkr Err | Rel. +/- | Spkr Err | Rel. +/- |
| Entire Meeting | baseline | 12.80% | - | 16.40% | - | 18.20% | - | 24.80% | - |
| | ASPG | 14.50% | +13.28% | 13.00% | -20.73% | 19.50% | +7.14% | 19.30% | -22.18% |
| | ASPG & prosodic | 10.50% | -17.97% | 12.80% | -21.95% | 16.10% | -11.54% | 19.00% | -23.39% |
| 500 | baseline | 16.40% | - | 20.40% | - | 18.30% | - | 23.80% | - |
| | ASPG | 14.20% | -13.41% | 16.90% | -17.16% | 16.70% | -8.74% | 19.60% | -17.65% |
| | ASPG & prosodic | 11.80% | -28.05% | 11.80% | -42.16% | 15.50% | -15.30% | 19.40% | -18.49% |
| 300 | baseline | 23.80% | - | 27.40% | - | 23.60% | - | 27.30% | - |
| | ASPG | 15.40% | -35.29% | 17.10% | -37.59% | 17.90% | -24.15% | 20.60% | -24.54% |
| | ASPG & prosodic | 14.00% | -41.18% | 14.80% | -45.99% | 17.20% | -27.12% | 18.90% | -30.77% |
| 100 | baseline | 44.00% | - | 50.10% | - | 41.10% | - | 41.40% | - |
| | ASPG | 22.10% | -49.77% | 21.70% | -56.69% | 19.50% | -52.55% | 20.70% | -50.00% |
| | ASPG & prosodic | 18.00% | -59.09% | 16.60% | -66.87% | 19.70% | -52.07% | 19.80% | -52.17% |

**Table 2**. Comparison of the new approach to ASPG and the baseline on the NIST RT-06, RT-07 and RT-09 evaluation sets (single distant microphone and multiple distant microphone case). Entire meetings and shorter segments are compared (see [1]). The changes are calculated relative to the baseline, only the Speaker Error is shown.

Based on the results shown in the discussion in Section 5 we will further investigate the extraction of prosodic features, and study the effect of other external influence such as overlap to the prosodic feature extraction to further improve the robustness. Other future work includes performing experiments with longer meetings (rather than short ones) as well as the generalization to other audio domains, such as broadcast news.

## 7. ACKNOWLEDGMENT

We thank Mary Knox from ICSI and Mathew Magimai Doss from Idiap Research Institute for helpful comments on our work.

## 8. REFERENCES

[1] David Imseng and Gerald Friedland, "Robust speaker diarization for short speech recordings," in *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, December 2009, pp. 432–437.

[2] G. Friedland, O. Vinyals, Y. Huang, and C. Muller, "Prosodic and Other Long-term Features for Speaker Diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 985–993, July 2009.

[3] Xavier Anguera Miró, *Robust Speaker Diarization for Meetings*, Ph.D. thesis, Universitat Politecnica de Catalonia, 2006.

[4] David A. Leeuwen and Matej Konečný, "Progress in the AMIDA Speaker Diarization System for Meeting Data," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*. 2008, vol. 4625/2008 of *LNCS*, pp. 475–483, Springer-Verlag.

[5] X. Anguera, C. Wooters, and J. Hernando, "Friends and enemies: a novel initialization for speaker diarization," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP'06)*, September 2006, pp. 689–672.

[6] J. Luque, C. Segura, and J. Hernando, "Clustering initialization based on spatial information for speaker diarization of meetings," in *Proceedings of Interspeech*, September 2008, pp. 383–386.

[7] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *In Proceedings of IEEE Workshop on Automatic Speech Recognition Understanding*, 2003, pp. 411–416.

[8] J. Huang, E. Marcheret, K. Visweswariah, and G. Potamianos, "The IBM RT07 Evaluation Systems for Speaker Diarization on Lecture Meetings ," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*. 2008, vol. 4625/2008 of *LNCS*, pp. 497–508, Springer-Verlag.

[9] Chuck Wooters and Marijn Huijbregts, "The ICSI RT07s Speaker Diarization System," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007*. 2008, vol. 4625/2008 of *LNCS*, pp. 509–519, Springer-Verlag.