On the use of Dot Scoring for Speaker Diarization*

Mireia Diez**, Mikel Penagarikano, Amparo Varona, Luis Javier Rodriguez-Fuentes, German Bordel

GTTS, Department of Electricity and Electronics
University of the Basque Country, Spain
<mireia_diez@ehu.es>

Abstract. In this paper, an alternative dot scoring based agglomerative hierarchical clustering approach for speaker diarization is presented. Dot-scoring is a simple and fast technique used in speaker verification that makes use of a linearized procedure to score test segments against target models. In our speaker diarization approach speech segments are represented by MAP-adapted GMM zero and first order statistics, dot scoring is applied to compute a similarity measure between segments (or clusters) and finally an agglomerative clustering algorithm is applied until no pair of clusters exceeds a similarity threshold. This diarization system was developed for the Albayzin 2010 Speaker Diarization Evaluation on broadcast news. Results show that the lowest error rate that the clustering algorithm could attain for the evaluation set was around 20% and that over-segmentation was the main source of degradation, due to the lack of robustness in the estimation of statistics for short segments.

Index Terms: Speaker Diarization, Dot Scoring, Sufficient Statistics

1 Introduction

Speaker Diarization consists of determining who spoke when in an input audio stream. It involves two main steps: determining the boundaries between speaker turns and clustering segments according to the speaker identity [1]. In recent years, speaker diarization has gained importance as a mean of indexing different types of data such as meetings, broadcast news or telephone conversations.

Most speaker diarization systems apply agglomerative hierarchical clustering with a BIC-based stopping criterion [1]. In this paper, an alternative dot scoring-based agglomerative hierarchical clustering approach is presented. Dot scoring is commonly used as a fast scoring technique in speaker verification. Applying

* This work has been supported by the University of the Basque Country under Grant GIU10/18 and the Government of the Basque Country, under program SAIOTEK (project S-PE10UN87), and the Spanish MICINN, under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds).

** Supported by a research fellowship from the Department of Education, Universities and Research of the Basque Country Government.

dot scoring to diarization has the advantage of a low computational cost for re-training speaker-cluster models.

The dot-scoring speaker diarization system developed for the Albayzin 2010 Speaker Diarization Evaluation (SDE) is based on three subsystems: the audio classifier developed for the Albayzin 2010 Audio Segmentation Evaluation [2], the acoustic change detector module developed for the system submitted to the Albayzin 2006 Speaker Tracking Evaluation [3], and the speaker verification system developed for the NIST 2010 Speaker Recognition Evaluation [4].

The paper is organized as follows: in section 2 all the stages of the proposed diarization system are described: speech/non speech segmentation, acoustic change detection, dot scoring and the clustering algorithm. In section 3 we present the experimental setup used to develop and evaluate the system. Results, as well as the processing time required, are presented in section 4. Finally, conclusions are outlined in section 5.

2 Speaker Diarization System

Speech/non-speech detection, acoustic change detection and clustering are performed separately in this approach.

The speech/non-speech detector developed for this task is based on a 5-class ergodic Continuous Hidden Markov Model including 3 speech sub-classes and 2 non-speech sub-classes. More details can be found in [2]. A simple approach, which uses a XBIC-based measure to detect any change of speaker, background or channel conditions, was applied as defined in [3]. Though it oversegments the audio stream, the set of change points includes almost all the speaker changes. The optimal configuration of the three subsystems was heuristically determined on development data (see section 3.3 for details).

2.1 Dot Scoring

Dot scoring agglomerative hierarchical clustering was performed as follows:

Universal Background Model. A gender independent GMM (Universal Background Model, UBM) was trained. The Sautrela toolkit [5] was used to estimate GMM parameters, applying binary mixture splitting, orphan mixture discarding and variance flooring.

Sufficient statistics. Let $\lambda \equiv \{\omega_k, \mu_k, \Sigma_k | k = 1..K\}$ be a GMM composed by K Gaussians of dimension F with diagonal covariance matrices Σ_k. Let f_t be the feature vector at time t. Let $\gamma_k(t)$ be the posterior probability of Gaussian k at time t. We define:

$$n_k = \sum_t \gamma_k(t) \tag{1}$$

$$x_k = \sum_t \gamma_k(t) \, \Sigma_k^{-\frac{1}{2}} (f_t - \mu_k) \tag{2}$$

The sets of parameters vectors $\nu = [\nu_{ij}]$, where $\nu_{ij} = n_i$, $i \in [1..K]$, $j \in [1..F]$, and $x = [x_1, \ldots, x_K]$ (each x_i being a F-dimensional vector) are known as the zero and first order sufficient statistics, respectively. Given a dataset c, the one-iteration relevance-MAP adapted and normalized mean vectors $m = \Sigma^{-\frac{1}{2}}(\mu_c - \mu_{ubm})$ can be computed according to the following expression[1] [6,4]:

$$m = (\tau \mathbf{I} + diag(\nu))^{-1} \cdot x \tag{3}$$

Dot scoring similarity measure. Dot-scoring is a simple and fast technique used in speaker verification that makes use of a linearized procedure to score test segments against target models [6]. Given a feature stream f (the target signal) and a speaker model λ_s, the first-order Taylor-series approximation to the GMM log-likelihood is:

$$\log P(f|\lambda_s) \approx \log P(f|\lambda_{ubm}) + m_s^t \cdot \nabla P(f|\lambda_{ubm}) \tag{4}$$

where m_s denotes the vector of normalized means corresponding to speaker s, ∇ denotes the gradient vector w.r.t the standard-deviation-normalized means of the UBM, and $\nabla P(f|\lambda_{ubm}) = x_f$ is the vector of first order statistics corresponding to the target signal f. The log-likelihood ratio between the target model and the UBM used for scoring can be approximated as follows:

$$score(f, s) = \log \frac{P(f|\lambda_s)}{P(f|\lambda_{ubm})} \approx m_s^t \cdot x_f \tag{5}$$

For the diarization task, the similarity $sim(a, b)$ between two segments a and b was defined as:

$$sim(a, b) = min\{score(f_a, b), score(f_b, a)\} = min\{m_b^t \cdot x_a, m_a^t \cdot x_b\} \tag{6}$$

Score normalization. TZ normalization was applied to dot-scores. Two independent sets of development data were used for the estimation of T-norm (normalization w.r.t. the test utterance) and Z-norm (normalization w.r.t. the speaker cluster) parameters. Taking into account score normalization, the similarity measure was redefined as:

$$sim(a, b) = min\{score_{TZ}(f_a, b), score_{TZ}(f_b, a)\} \tag{7}$$

2.2 The clustering algorithm

The similarity measure defined above was used to perform agglomerative hierarchical clustering. Given two segments (or two clusters of segments), if they are clustered together, computation of sufficient statistics for the joint cluster is straightforward:

$$\begin{aligned} x_{a+b} &= x_a + x_b \\ n_{a+b} &= n_a + n_b \end{aligned} \tag{8}$$

[1] $diag(\nu)$ stands for a square matrix with the elements of ν in the diagonal

This leads to a very simple clustering algorithm:

1. **Find** $s_{max} = \max_{\forall(a,b)} \{sim\,(a,b)\}$
   $(a*, b*) = \operatorname{argmax}_{\forall(a,b)} \{sim\,(a,b)\}$
2. **If** $s_{max} < \Theta$ **then** STOP
3. **Set** $x_{a*} = x_{a*} + x_{b*}$
   $n_{a*} = n_{a*} + n_{b*}$
4. **Remove** cluster $b*$
5. **Jump to** 1

3 Experimental setup

3.1 Databases

We decided to keep independence between training and development data, therefore the Albayzin 2010 SDE database was used for development and the KALAKA database [7] for training the GMMs.

The Albayzin 2010 SDE consists of 24 sessions of TV broadcast news in Catalan, most sessions being 4 hours long (some of them being shorter). The database, recorded from the 3/24 TV channel, includes around 87 hours of audio, split into 2 sets: train/development (16 sessions, 2/3 of the total amount of data) and test (8 sessions, the remaining 1/3). Even though 3/24 TV mostly contains speech in Catalan, around 1/6 of the speech segments are spoken in Spanish. The database contains male, female and overlapped speech and the number of speakers per recording varies from 30 to 250. The distribution of background conditions within the database is the following: Clean speech: 37%; Music: 5%; Speech with music in background: 15%; Speech with noise in background: 40%; Other: 3% [8].

KALAKA materials were also extracted from (wide-band) TV shows. The database, which was designed to build language recognition systems, contains speech in four target languages: Basque, Catalan, Galician and Spanish, all of them official languages in Spain. KALAKA contains more than 12 hours of speech per target language.

3.2 Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCC) were used as acoustic features. The MFCC set, comprising 13 coefficients, including the zero (energy) coefficient, was computed in frames of 32 ms at intervals of 10 ms for the two first modules (speech/non-speech detection and acoustic change detection). In the clustering approach, the MFCC set was computed in frames of 20 ms at intervals of 10 ms and augmented with dynamic coefficients (13 first-order and 13 second-order deltas), resulting in a 39-dimensional feature vector. Also, an energy based voice activity detector (VAD) was applied to remove those fragments (short silences) with an energy level 30 dB (or more) under the maximum. All the speech processing computations were done by means of the Sautrela toolkit [5].

3.3 Parameter optimization

Speech/non-speech detector. A 5 state ergodic Continuous Hidden Markov Model was estimated using the Sautrela toolkit [5], under the Layered Markov Models framework. Preliminary experiments on a subset of the development data revealed that best audio segmentation performance was achieved when the number of mixtures was 512. The emission distributions were independently estimated for each state, applying the Baum-Welch algorithm on the corresponding sets of segments extracted from the reference segmentations of 12 development sessions. The number of mixtures per state and the transition probabilities (auto-transitions fixed to 0.999999, transitions between states and final state transitions fixed to $2 \cdot 10^{-7}$) were optimized on audio segmentation experiments over the remaining 4 development sessions. Considering a 2-class speech/non-speech classification setup, the false alarm and the miss error rates were around 1% for the speech class (including the three sub-classes mentioned in 3.1). Note that, since we are mistaking around 2% of the speech frames, our speaker diarization error will be, at best, of that order.

Gaussian Mixture Models. Although the evaluation was limited to Catalan TV speech, in order to increase the speaker variability, TV broadcast speech in Spanish, Catalan, Galician and Basque, taken from the Kalaka database [7], was used to train gender independent GMMs (Universal Background Model, UBM) consisting of 256, 512 and 1024 mixture components. Again, the Sautrela toolkit was used to estimate GMM parameters, applying binary mixture splitting, orphan mixture discarding and variance flooring.

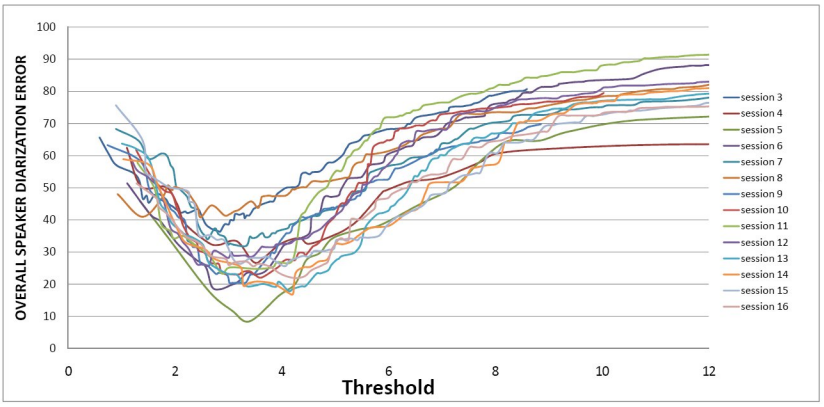

Fig. 1. Overall Speaker Diarization Error as a function of the similarity threshold applied as stopping criterion in the clustering algorithm, for sessions 3-16 of the development set using a 256-mixture GMM system.

Threshold selection. Threshold optimization was performed on the development set. Figure 1 shows the performance measured on development sessions 3-16 for each threshold value using a 256-mixture GMM system. Based on the average system performance, similarity thresholds were set to the values shown in Table 1.

Table 1. Threshold value selection for each GMM system based on average system performance.

	#mixtures		
	256	**512**	**1024**
Threshold	3.80	3.74	3.98

3.4 Performance criteria

The diarization error rate (DER) defined by NIST [9] was the primary metric used in the evaluation, applying a scoring "forgiveness collar" of 250 ms around each reference segment boundary [8].

4 Results

Experiments carried out showed almost no difference in performance among the GMM systems using 256, 512 and 1024 mixtures. Therefore, the 256-mixture GMM system was selected for further analyses, due to the lower cost of sufficient statistics and similarity matrix computations. Table 2 shows the performance of the clustering algorithm described above on the evaluation set, using four different segmentations:

- Seg1: Reference Speaker Segmentation.
- Seg2: Reference Speaker Segmentation + GTTS Acoustic Change Detection.
- Seg3: Reference Speech/Non-Speech Segmentation + GTTS Acoustic Change Detection.
- Seg4: GTTS Speech/Non-Speech Detection + GTTS Acoustic Change Detection.

Table 2. Overall Speaker Diarization Error obtained by applying the clustering algorithm on four different segmentations of the evaluation set (see text for details).

DER %	**Seg1**	**Seg2**	**Seg3**	**Seg4**
256-m GMM	20.48	26.14	29.61	33.16

The Overall Speaker Diarization Error obtained with the Reference Speaker Segmentation (Seg1, 20.48%) would be the best performance that our clustering system could reach for the evaluation set. The difference between this result and the result obtained with the fully automated system (Seg4, 33.16%) may be explained as follows:

- Difference between Seg3 and Seg4: 3.55%. Seg3 starts from a perfect Speech/Non-Speech classification, whereas Seg4 applies the GTTS Speech/Non-

Speech detection system. So, the difference can be explained by the Speech/Non-Speech classification error.

- Difference between Seg1 and Seg2: 5.66%. Since both systems take the reference speaker segmentation as a starting point, the difference in performance can only be due to over-segmentation introduced by the GTTS acoustic change detector. Applying the acoustic change detector on the optimal speaker segmentation does not remove speaker boundaries but produces many short segments whose statistics strongly depend on local variabilities. This explains why the performance of the clustering algorithm, which is based on those statistics, degrades for short segments.
- Difference between Seg2 and Seg3: 3.47%. Seg2 includes all the speaker boundaries (plus a number of acoustic changes inside speaker turns), whereas Seg3 may be missing some of them. This explains the difference.

4.1 Processing time

Table 3 shows the CPU time (expressed as real-time factor, $\times$RT) employed in six separate operations: (1) feature extraction for segmentation; (2) speech/non-speech segmentation; (3) acoustic change detection; (4) feature extraction for clustering; (5) computation of sufficient statistics; and (6) hierarchical clustering of speech segments, for both the reference speaker segmentation and the automatic segmentation. Note that the CPU time employed in clustering is almost four times higher for the automatic segmentation than for the reference segmentation, because of the different number of speech segments: 7.24 and 3.62 segments/minute, respectively. The total CPU time of the speaker diarization system is 0.2932$\times$RT.

Computations were made in two servers. The first one, devoted to speech/non-speech segmentation and acoustic change detection, was a Dell PowerEdge 1950, equipped with two Xeon Quad Core E5335 microprocessors at 2.0GHz (allowing 8 simultaneous threads) and 4GB of RAM. The second one, devoted to clustering, was a Dell PowerEdge R610, equipped with 2 Xeon 5550 (each featuring 4 cores) at 2.66GHz and 32GB of RAM.

Table 3. CPU time (real-time factor, $\times$RT) employed by the different modules of the speaker diarization system.

	Ref. segm.	**GTTS segm.**
Features (segmentation)	-	0.0033
Speech/non-speech segmentation	-	0.0375
Acoustic change detection	-	0.1058
Features (clustering)	0.0026	
Statistics	0.0050	
Clustering	0.038	0.139

5 Conclusions

In this paper a new speaker diarization approach, which applies agglomerative hierarchical clustering based on dot scoring, has been described.The system consists on a chain of four uncoupled modules: speech/non-speech segmentation, acoustic change detection, computation of sufficient statistics and hierarchical clustering of speech segments. Despite its simplicity, the proposed system attained competitive results in the Albayzin 2010 Speaker Diarization Evaluation.

Experiments carried out on different segmentations showed: (1) that the best performance that the clustering algorithm could attain for the evaluation set was around 20%; and (2) that over-segmentation introduced by the acoustic change detector was the main source of degradation, because of the lack of robustness in the estimation of statistics for short segments. Future work will involve trying to improve the robustness of the clustering algorithm to short segments, or alternatively, to avoid over-segmentation while keeping the detection rate of speaker boundaries. Besides, alternative feature parameterizations will be studied.

References

1. S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1979–1986, 2006.
2. L. J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, "GTTS Systems for the Albayzin 2010 Audio Segmentation Evaluation," in *FALA 2010 "VI Jornadas en Tecnología del Habla" and II Iberian SLTech Workshop*, (Vigo, Spain), November 2010.
3. L. J. Rodriguez, M. Penagarikano, and G. Bordel, *A Simple but Effective Approach to Speaker Tracking in Broadcast News*, vol. LCNS 4478 of *Lecture Notes in Computer Science*, pp. 48–55. Springer Verlag, Berlin Heidelberg: Pattern Recognition and Image Analysis (IbPRIA 2007), Joan Martí, José Miguel Benedí, Ana Maria Mendonça and Joan Serrat (Eds.), 2007.
4. M. Penagarikano, A. Varona, M. Diez., L. J. Rodriguez-Fuentes, and G. Bordel, "University of the Basque Country System for NIST 2010 Speaker Recognition Evaluation," in *Proceedings of the II Iberian SLTech Workshop*, (Vigo, Spain), 2010.
5. M. Penagarikano and G. Bordel, "Sautrela: A Highly Modular Open Source Speech Recognition Framework," in *Proceedings of the ASRU Workshop*, (San Juan, Puerto Rico), pp. 386–391, December 2005.
6. A. Strasheim and N. Brümmer, "SUNSDV system description: NIST SRE 2008," in *NIST Speaker Recognition Evaluation Workshop Booklet*, 2008.
7. L. J. Rodriguez-Fuentes, M. Penagarikano, G. Bordel, A. Varona, and M. Diez, "KALAKA: A TV Broadcast Speech Database for the Evaluation of Language Recognition Systems," in *7th International Conference on Language Resources and Evaluation*, (Valleta, Malta), 17-23 May 2010.
8. M. Zelenak, H. Schulz, and J. Hernando, "Albayzin 2010 Evaluation Campaign: Speaker Diarization," in *FALA 2010 "VI Jornadas en Tecnología del Habla" and II Iberian SLTech Workshop*, (Vigo, Spain), November 2010.
9. *The 2009 NIST Rich Transcription Evaluation.* http://www.itl.nist.gov/iad/mig/tests/rt/.