IMPROVED BINARY KEY SPEAKER DIARIZATION SYSTEM

Héctor Delgado[1], *Xavier Anguera*[2], *Corinne Fredouille*[3], *Javier Serrano*[1]

[1]CAIAC, Autonomous University of Barcelona, Cerdanyola del Vallès, Spain
[2]Sinkronigo S.L., Barcelona, Spain
[3]CERI/LIA, University of Avignon, Avignon, France

hecdelflo@gmail.com, xanguera@gmail.com
corinne.fredouille@univ-avignon.fr, javier.serrano@uab.cat

ABSTRACT

The recently proposed speaker diarization technique based on binary keys provides a very fast alternative to state-of-the-art systems. However, this speed up has the cost of a little increase in Diarization Error Rate (DER). This paper proposes a series of improvements to the original algorithm with the aim to get closer to state-of-the-art performance. First, several alternative similarity measures between binary key speaker/segment models are introduced. Second, we perform a first attempt at applying Intra-Session and Intra-Speaker Variability (ISISV) compensation within the binary diarization approach through the Nuisance Attribute Projection. Experimental results show the benefits of the newly introduced similarity metrics, as well as the potential of the Nuisance Attribute Projection for ISISV compensation in the binary key speaker diarization framework.

Index Terms— Speaker diarization, binary key, cosine distance, chi-square distance, session variability compensation, nuisance attribute projection

1. INTRODUCTION

Speaker diarization is the task of segmenting an audio file into speaker-homogeneous segments without any prior information about the number of speakers nor their identities. The importance of speaker diarization is well known as a pre-processing tool for many speech-related tasks which take advantage of dealing with speech signals from a single-speaker.

Recently, a speaker diarization approach based on the “binary key” speaker modeling was presented in [1]. Among its advantages, the main one is the system speed, since the technique runs over 10 times faster than real time with little decrease in performance. The approach was further investigated in [2], where it was tested using broadcast data instead of meeting recordings, obtaining similar results. Additionally, [3] addressed the problem of clustering selection and stopping criterion by proposing the use of a global speaker clustering method inspired in [4].

One of the reasons why speaker diarization is such a challenging task is the existence of a high Intra-Session Intra-Speaker Variability (ISISV) in some recordings. Systems usually have to deal with highly varying conditions even within a given audio file. Such variability may lead the diarization system, for example, to model a given speaker by more than one cluster, or to increase speaker errors. This phenomenon has also been shown to have a negative impact in speaker identification and verification systems. To face this problem, several variability compensation techniques have been proposed and successfully applied to speaker recognition. Among those methods, the most popular ones are Nuisance Attribute Projection (NAP) [5], Within Class Covariance Normalization (WCCN) [6], Joint Factor Analysis (JFA) [7], and i-vector [8]. Some of these compensation techniques have been also applied to speaker diarization [4, 9].

In [10], the authors reported the benefits of applying NAP and WCCN to the binary representation of speakers in a speaker recognition task. Therefore, it is reasonable to think that such variability compensation technique should be beneficial in the framework of binary key speaker diarization if a suitable adaptation is achieved.

The main goal of this paper is twofold. First, alternative similarity measures will be evaluated within the binary key speaker diarization system. Secondly, NAP will be adapted to our needs and applied before the final clustering stage. Experiments carried out on the REPERE phase 1 test set of TV broadcast programs show the effectiveness of the introduced similarity measures and the NAP compensation method, outperforming a baseline binary key based speaker diarization system.

The paper is structured as follows: Section 2 describes the baseline binary key speaker diarization system and includes the definition of the similarity measures being evaluated. Section 3 describes the variability compensation technique and its adaptation to the binary key diarization system. Section 4 provides experimental results and discussion. Finally, section 5 concludes and proposes future work.

2. OVERVIEW OF THE BINARY KEY SPEAKER DIARIZATION SYSTEM

A complete description of the binary key diarization system used in this work is given in [2]. Here only a brief overview is done. Essentially, the system is composed of two main blocks. First, the acoustic step transforms the input signal data into a series of binary vectors called Binary Keys (BK), following a supersegment approach. Second, the binary step performs an Agglomerative Hierarchical Clustering (AHC) over the BKs.

The conversion of a set of acoustic features into a BK is carried out thanks to a UBM-like model called Binary Key Background Model (KBM), which is trained using the test data itself (refer to [2] for details). The number of Gaussian components of this model can be of several hundreds and will determine the dimension of the BKs.

Once the KBM is trained, any set or sequence of input feature vectors can be converted into a Binary Key. A BK $\mathrm{v}_f = \{v_f[1], ..., v_f[N]\}, v_f[i] = \{0, 1\}$ is a binary vector whose dimension N is the number of components in the KBM. Setting a position $v_f[i]$ to 1 (TRUE) indicates that the i-th Gaussian of the KBM coexists in the same area of the acoustic space as the majority of the acoustic data being modeled. The BK can be obtained in two steps. First, for each feature vector, the best N_G matching Gaussians in the KBM are selected (e.g. the N_G components which provide the highest likelihood for the current feature vector). The IDs of all selected Gaussians are stored in a $n \times N_G$ matrix, where the rows represent the n feature vectors. Second, the count of the occurrences of each Gaussian ID in the previous matrix is calculated and stored in a Cumulative Vector (CV) of dimension N, where each position i represents the i-th Gaussian of the KBM. Then, the final BK is obtained by setting to 1 the M top positions of the CV. Note that this method can be applied to any set of features, either a sequence of features from a short speech segment, or a feature set corresponding to a whole speaker cluster.

The last step before switching to the binary process step is the clustering initialization. In this paper we opt for a simple uniform cluster initialization by splitting the input data into N_{init} equal-sized chunks.

The binary step implements an AHC clustering approach. However, all operations are done with binary data, which makes the process much faster than using classic GMM-based approaches (please refer to [2] to see execution time figures). First, BKs for the initial clusters are calculated using the method explained above. Then, the input data, previously converted into a sequence of BKs, is reassigned to the current clusters, by comparing input BKs to all current cluster BKs by using some similarity measure (section 2.1). Once data have been redistributed, BKs are trained for the new clusters. Next, similarities between all cluster pairs are calculated, and the cluster pair with the highest score is merged, reducing the number of clusters by one. The iterative process is repeated until a single cluster containing all the input BKs is obtained. Finally the output clustering must be selected from all partial clusterings obtained in all the iterations. This is done by calculating the student T-test T_s metric as explained in [1] to all clustering solutions. Then, the clustering which maximizes T_s is returned.

2.1. Similarity measures for binary keys and cumulative vectors

In this subsection we propose two similarity measures that exploit the CV counts.

The originally proposed similarity metric used in binary key speaker diarization [1] is defined as

$$S(\mathrm{a}, \mathrm{b}) = \frac{\sum_{i=1}^{N}(a_i \wedge b_i)}{\sum_{i=1}^{N}(a_i \vee b_i)} \tag{1}$$

which involves bit wise-operations between two BKs. The effectiveness of this metric for comparing BKs was assessed and demonstrated to be effective in previous work [1, 2].

As explained above, positions equal to one within a BK indicate that the Gaussians of the KBM associated to those positions are the ones that best fit the sequence of feature vectors being converted. The selected components are the ones most frequently chosen as top-scoring components for the segment/cluster, i.e. the M highest positions in the Cumulative Vector (CV, see section 2). A CV is a vector of positive integers in which each position i stores the frequency of activation of the i-th Gaussian component in the KBM. This could be interpreted as a set of weights specifying the importance of each component in the given feature set. But these weights are lost in the process of conversion from a CV to a BK. However, it seems reasonable to think that this removed information could also be beneficial for discriminating between speakers, thus CVs could be used as speaker models in place of BKs. Using CVs instead of BKs has already been addressed for speaker verification with success in [10]. Speaker recognition and speaker diarization are tasks closely related, thus we think that using CVs as speaker models could also provide benefits in the speaker diarization task.

As said before, a CV contains frequencies of component activation. This counts are calculated in a per-frame basis: for each feature vector, the positions corresponding to the N_G top-scoring Gaussians are incremented by one, implying that the final absolute values depend on the duration of the segment/cluster being converted. However, we are more interested in the relative variations among the CV positions than in the vector's magnitude. Therefore, it seems a case where the cosine similarity may be suitable for comparing them, as it is a measure related to the angle between the two vectors, which depends on the vectors' directions. The cosine similarity between vectors a and b is defined as

$$S_{cos}(a, b) = \frac{a \cdot b}{\|a\| \, \|b\|} \tag{2}$$

Finally, as a CV counts how many times each Gaussian component in the KBM has been selected as a top-scoring Gaussian for the feature set being converted, it could be considered as some sort of histogram. A well-know distance for comparing histograms is the chi-square χ^2 distance, defined by equation 3. We propose its use to measure similarity between

CVs.

$$D_{\chi^2}(a,b) = \frac{1}{2}\sum_{i=1}^{N}\frac{(a_i - b_i)^2}{a_i + b_i} \tag{3}$$

In order to avoid by-zero divisions in the denominator, a constant value (equal to the minimum increment representable by double-precision numbers in Matlab, which is 2.2×10^{-16}) is summed to all CV positions. Furthermore, the CVs are normalized before computing similarity.

3. SESSION VARIABILITY COMPENSATION IN BINARY KEY SPEAKER DIARIZATION

Dealing with session variability has become a must for any modern speaker recognition system. The binary key speaker modeling is not an exception and popular compensation methods such as NAP and WCCN have been successfully applied to it in a speaker verification task [10]. Speaker diarization systems also have to deal with varying conditions within the audio signal to be processed. Thus, compensation of this variability will presumably be beneficial and will result in improvements in accuracy. In this work we propose to apply NAP for binary key speaker diarization by adapting it to the particular needs that speaker diarization entails. Given its nature, NAP can be applied to any speech representation in form of high-dimensional vectors called supervectors (e.g. Gaussian supervectors, speaker factors, i-vectors, and CVs in this work).

3.1. Nuisance Attribute Projection

NAP [5] assumes that the within-class variability is restricted to a low dimensional subspace. In order to remove this variability, the supervectors are projected onto an orthogonal complementary subspace. First, the within-speaker scatter matrix is calculated on appropriate labeled data as

$$\mathrm{W} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(z_i^s - \bar{z}^s)(z_i^s - \bar{z}^s)^t \tag{4}$$

where S is the number of speakers, n_s is the number of utterances by speaker s, z_i^s is the supervector representing the i-th utterance of speaker s, and $\bar{z}^s$ is the mean of the supervectors of speaker s. Then, the projection matrix is obtained according to $P = (I - UU^t)$, where U is the rectangular matrix of the k eigenvectors associated with the k largest eigenvalues (obtained after solving the eigenvalue problem $Wu = \lambda u$). Finally, the transformation of supervector x is the result of applying the projection as $y = Px$.

Given its nature, NAP can be applied to any sort of supervector representation of an utterance/cluster. In our system, we first convert the input data into a sequence of CVs. Then, all the CVs are compensated by applying the projection P. Next, in the AHC stage, CVs for the new clusters are trained as usual, and then compensated in the same way as the segment CVs. The rest of the process remains the same as the baseline system.

With regard to the estimation of the within-class scatter matrix W, a development set is used. For each audio file in the development set, all the segments of each participating speaker are pooled together by speaker, and divided into segments of one second. Those segments are used as speaker utterances in the computation of W. As our system estimates the KBM on the test audio file, the CVs of the development set have to be calculated for each test audio file using its own KBM.

Once matrix W is estimated, the projection P is calculated. As said above, U is the matrix formed by the k eigenvectors associated to the top k eigenvalues. Instead of using a fixed value of k for all input audio files, we estimate k as a function of the proportion p of the total eigenvalue mass as follows:

$$\min_k \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{j=1}^{D}\lambda_j} \geq p \tag{5}$$

being D the dimension of matrix W.

4. EXPERIMENTS AND RESULTS

This section describes the experimental setup and results for two different experiments. First, the proposed similarity measures are evaluated on the baseline diarization system (without variability compensation). Second, NAP is applied by estimating the within-class-scatter-matrix on development data. Note that in NAP experiments we only use the cosine similarity since after applying the projection on the CVs space, none of the S and S_{χ^2} are suitable for measuring similarities between the vectors projected to the new subspace.

All the experiments are performed using "oracle" speech/non-speech labels, with the aim to asses the proposed improvements without the impact of impurities introduced by false alarm errors. With regard to overlapping speech, although our system is not prepared to handle such overlapped speech, regions with more than one active speaker are included in both the diarization process and DER computation.

All tests are performed on the REPERE phase 1 test dataset of TV data. This database was developed in the context of the REPERE Challenge [11]. It consists of a set of TV shows from several French TV channels. For NAP projection estimation, the REPERE phase 1 development dataset is used.

4.1. Experimental setup

Parameters and settings of the various modules of the binary key speaker diarization system are described here. Feature extraction is performed using standard 19-order MFCCs, computed using a 25ms window every 10ms.

For training the KBM, single Gaussian components are obtained using a 2s window over the signal. Window rate is set according to the input audio length, in order to obtain an initial pool of 2000 Gaussians. Then, the final number of components is reduced to N components by following the Gaussian component selection algorithm explained in [1].

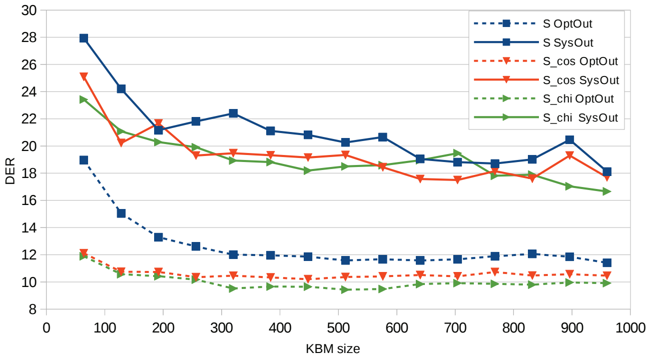

Fig. 1. Diarization performance measured in DER for the proposed similarity measures, in function of the KBM size.

However, in this work we use the cosine distance between the means of the Gaussian components as a distance measure in the component selection algorithm, instead of using KL2 distance. We have observed that comparing the means using cosine distance is discriminative enough and also much faster to compute than KL2 [12].

With regard to binary key estimate parameters, the top 5 Gaussian components are taken in a per frame basis, and the top 20% components at segment level.

Finally, in the AHC stage, BKs are computed for each 1s segment, augmenting it 1s before and after, totaling 3s.

For performance evaluation, the output labels are compared with the reference ones to compute the DER. As said before, overlapping speech regions are included in both diarization processing and calculation of DER. In such regions with more than one speaker simultaneously, our system assigns only one speaker label.

4.2. Similarity measure experiments

Evaluation results of the proposed similarity measures are collected in Figure 1. Continuous lines show system output (SysOut) performance returned by the final clustering selection algorithm, while dashed lines show performance of the optimum clusterings selected manually (OptOut). We include these results in order to set a performance ceiling. Let's first put the focus on the optimum clusterings selected manually (dashed lines). It can be appreciated how both S_{cos} and S_{χ^2} outperform the original similarity measure S (considered as the baseline similarity metric), providing a decrease of DER of around 1% absolute in the case of the cosine similarity, and of around 2% absolute with the chi-square similarity (with some overlap in the curves for some KBM sizes). A second fact observed is that performance does not improve after a certain KBM size of around 512 components. Regarding the output returned by the final clustering selection algorithm (SysOut), the new similarity measures also provide gains in performance of around 2-3% absolute. Contrarily to the case of optimum clusterings, the automatic clustering selection benefits from larger KBMs. However the gap between performance of system output and optimum clusterings is still too large, exhibiting performance differences of around 8% absolute DER. This shows that the clustering selection algorithm is not very accurate at detecting the optimum clustering. In any case, it is confirmed that the information contained in the CVs contribute to the discrimination between speakers and that the proposed similarity metrics provide good discrimination.

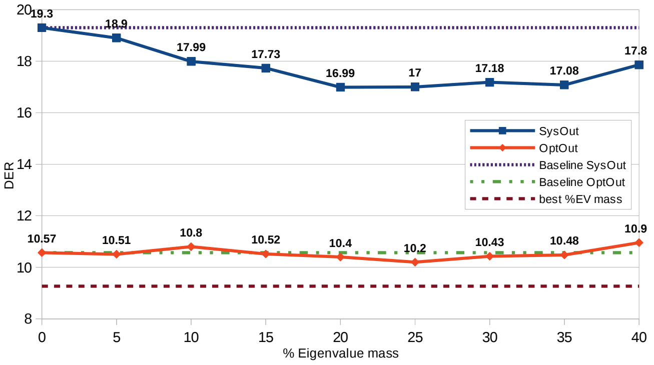

Fig. 2. Diarization performance with NAP variability compensation, measured in DER, in function of the KBM size.

4.3. Session compensation experiments

Figure 2 shows results obtained after applying NAP compensation within the diarization process, using a KBM size of 896, in function of the proportion p of eigenvalue mass used for selecting the k of eigenvectors for estimating the NAP projection. Like in Figure 1, results for the system output (SysOut) and for the optimum clusterings selected manually (OptOut) are shown. Concerning the optimum clusterings selected manually (orange line), we observe performance improvements just slightly above the baseline system performance (dashed green line). The best result is 10.20% DER with $p = 25$, versus 10.57% DER of the baseline. However, in the case of the returned output by the clustering selection algorithm, it seems that the NAP compensation contributes to improve such final clustering selection. Greater performance improvements are achieved, obtaining 16.99% DER with $p = 20$, versus 19.3% DER of the baseline, getting an absolute improvement of 2.3% DER. Additionally, the result by tuning the value of p for each audio file is plotted by the dashed red line. This remarks the importance of selecting the right number k of eigenvectors for each audio file in the computation of NAP. Although the selection of k as a function of a given proportion of the eigenvalue mass results beneficial, there is still room for improvement, and a more accurate method for deciding k may result in further performance gains.

In order to illustrate the hypothetical improvement of a more suitable selection method of k, performance for each show (OptOut) is shown before and after applying NAP by setting the optimum value of p manually. It can be observed that the use of NAP is beneficial in almost all audio files. The overall, time-weighted DER is 1.3 points lower after applying NAP (10.54% DER versus 9.27% DER).

Show ID	Baseline	NAP	
	DER	p	DER
BFMTV_BFMStory_1	8.3	25	5.93
BFMTV_BFMStory_2	16.28	25	15.74
BFMTV_BFMStory_3	5.75	40	4.6
BFMTV_BFMStory_4	4.35	40	4.21
BFMTV_CultureEtVous_1	26.54	10	22.55
BFMTV_CultureEtVous_2	29.61	25	30.65
BFMTV_CultureEtVous_3	20.7	25	17.97
BFMTV_CultureEtVous_4	3.88	35	5.74
BFMTV_CultureEtVous_5	11.34	40	10.01
BFMTV_CultureEtVous_6	18.65	25	18.37
BFMTV_CultureEtVous_7	24.8	5	25.26
LCP_CaVousRegarde_1	8.02	25	8.02
LCP_CaVousRegarde_2	9.11	25	9.1
LCP_CaVousRegarde_3	20.79	30	15.04
LCP_EntreLesLignes_1	34.23	15	31.97
LCP_EntreLesLignes_2	11.49	20	9.29
LCP_EntreLesLignes_3	6.92	10	5.49
LCP_LCPInfo13h30_1	8.54	25	6.91
LCP_LCPInfo13h30_2	2.8	20	0.57
LCP_LCPInfo13h30_3	11.98	35	8.96
LCP_PileEtFace_1	16.29	20	15.89
LCP_PileEtFace_2	15.35	10	13.31
LCP_PileEtFace_3	26.51	20	32.88
LCP_PileEtFace_4	14.23	35	10.64
LCP_PileEtFace_5	7.97	10	5.26
LCP_TopQuestions_1	3.83	30	2.87
LCP_TopQuestions_2	1.54	15	0.57
LCP_TopQuestions_3	3.48	30	2.76
Overall	10.57	-	**9.27**

Table 1. DER of optimum clusterings in a per-show basis of the baseline system, and by applying NAP. p specifies the proportion of eigenvalue mass used to decide the number of eigenvectors for NAP.

5. CONCLUSIONS

This work proposes a series of improvements to binary key speaker diarization. First, several similarity measures between BKs and/or CVs are introduced in the diarization framework. Second, first attempts of applying ISISV compensation are done within the binary key speaker diarization framework through the Nuisance Attribute Projection. The inclusion of the new similarity measures was proven to be beneficial, whereas our approach to NAP for binary key diarization also provided performance gains. Although the method for selecting the right value of k has been proved to be effective, the per-audio-file analysis done suggests that there is still room for further improvement if a more effective method is found. One more negative aspect of our approach to NAP is the need to convert the development set into CVs using the KBM estimated on the current test audio file, which introduce an extra computational cost. Since one of the strongest points of the binary key speaker diarization technique is its speed, a method which reduced the extra computational cost would be highly desirable. For instance, our baseline system using the cosine distance and KBM size of 896 presents a Real-Time factor (xRT) of 0.07, while the system using NAP compensation presents a much higher computation time of 0.5 xRT using the same KBM size. For example, this problem could be solved by using a global KBM for all test audio streams, which would also be used to compute the CVs of the development set. In this way, the development CVs are compute only once and could be reused for all tests without the need of re-estimating then over an over again. Finally we remark the weakness of the final clustering selection algorithm, as the returned solutions are far from the performance ceiling. This issue is addressed in [12] by proposing a new clustering selection method.

6. ACKNOWLEDGMENTS

This work is part of the project "Linguistic and sensorial accessibility: technologies for voiceover and audio description", funded by the Spanish Ministerio de Economía y Competitividad (FFI2012-31024). This work was partially done within the French Research program ANR Project PERCOL (ANR 2010-CORD-102). This article is supported by the Catalan Government Grant Agency Ref. 2014SGR027.

REFERENCES

[1] Xavier Anguera and Jean-François Bonastre, "Fast speaker diarization based on binary keys," in *ICASSP*, May 2011, pp. 4428–4431.

[2] Héctor Delgado, Corinne Fredouille, and Javier Serrano, "Towards a complete binary key system for the speaker diarization task," in *INTERSPEECH*, 2014, pp. 572–576.

[3] Héctor Delgado, Xavier Anguera, Corinne Fredouille, and Javier Serrano, "Global speaker clustering towards optimal stopping criterion in binary key speaker diarization," in *Proc. IberSPEECH*, 2014, pp. 59–68.

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

[5] Alex Solomonoff, Carl Quillen, and William M. Campbell, "Channel compensation for svm speaker recognition," in *ODYSSEY*, 2004, pp. 57–62.

[6] Andrew O. Hatch, Sachin Kajarekar, and Andreas Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Proc. of ICSLP*, 2006, p. 14711474.

[7] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, July 2008.

[8] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, May 2011.

[9] Hagai Aronowitz, "Unsupervised compensation of intra-session intra-speaker variability for speaker diarization," in *ODYSSEY*, 2010.

[10] Gabriel Hernandez-Sierra, Jose R. Calvo, Jean-François Bonastre, and Pierre-Michel Bousquet, "Session compensation using binary speech representation for speaker recognition," *Pattern Recognition Letters*, vol. 49, no. 0, pp. 17 – 23, 2014.

[11] J. Kahn, O. Galibert, L. Quintard, M. Carre, A. Giraudel, and P. Joly, "A presentation of the REPERE challenge," in *CBMI*, June 2012, pp. 1–6.

[12] Héctor Delgado, Xavier Anguera, Corinne Fredouille, and Javier Serrano, "Novel stopping criterion for binary key speaker diarization," Submitted.