Confidence for Speaker Diarization using PCA Spectral Ratio

Orith Toledo-Ronen, Hagai Aronowitz

IBM Research – Haifa, Haifa University Mount Carmel, Haifa 31905, Israel
{oritht,hagaia}@il.ibm.com

Abstract

Confidence scoring is an important component in speaker diarization systems, both for offline speech analytics and for online diarization that are required to produce the speaker segmentation from very little audio. This paper proposes a confidence measure for speaker diarization based on the spectral ratio of the eigenvalues of the Principal Component Analysis (PCA) transformation computed on the pre-segmented audio before diarization is performed on the conversation. We tested our method on two-speaker data and our results show the effectiveness of the PCA's spectral ratio confidence measure for both offline and online diarization. We compare and contrast our proposed confidence measure with other clustering validation methods that provide a quantitative measure of the segmentation quality but are calculated on the segmented data after diarization is performed, and with a related approach that extracts a confidence from the PCA of the pre-segmented audio.

Index Terms: speaker diarization, principle component analysis, confidence measure

1. Introduction

Speaker diarization aims at segmenting a conversation into homogenous segments in which only one speaker presents and then clustering the segments based on speaker identity. In other words, speaker diarization answers the "who spoke when" question for a given audio signal. State-of-the-art algorithms find the speaker turn points and cluster the segments [1]. Speaker diarization is an important component in many speech applications such as two-wire telephony audio analytics, meetings and lectures summarization, and broadcast processing and retrieval. Within the diarization engine, a confidence scoring element that measures the diarization quality, plays a big role in online diarization and is also useful for offline diarization systems. This paper provides a confidence measure for the segmentation quality of the speaker diarization algorithm. The proposed confidence measure does not depend on the specific diarization algorithm in use because it operates on the unsegmented audio data of the conversation.

Other confidence measures for speaker diarization usually use the segmentation output as input to the confidence computation, as described by Vaquero et al. in [2] including: the Bayesian Information Criterion (BIC) measure of the segmentation accuracy, the Kullback-Leibler divergence measure of the distance between the distributions of the two segmented speakers, the convergence rate of the segmentation algorithm. A recent thesis by Vaquero [3] introduces another confidence measure based on a normalized eigenvalue spread by comparing the eigenvalue spread of the whole conversation to the eigenvalue spread of every speaker. As described in the thesis and in [4], the eigenvalue spread ratio is used for determining the correctness of the segmentation of two-speaker conversation datasets. Although related to our technique in terms of extracting information from the eigenvalues, the eigenvalue spread based confidence measure, like other existing clustering quality measures, is computed on the diarization output and not on the input conversation audio like we propose.

A similar approach is presented in context of blind source separation [5] for counting and estimating the audio sources directions based on clustering algorithm using a local confidence measure. The confidence measure proposed in [5] can discriminate the regions with one active source from others. It is defined as the ratio between the largest eigenvalue and the average of all the other eigenvalues computed from the PCA of a local scatter matrix in the time-frequency domain. Although the definition of this confidence measure is similar to ours, the two measures are different and are used for different purposes. Their confidence measure is derived from local information and is used to find regions in the signal where essentially only one source is active, while we compute the confidence measure on all the audio and find if there is a good separation between the two speakers, which gives an indication of the classification quality for speaker diarization.

Other related techniques that are using some variants of the eigengap technique for speaker diarization and operating on the segmented data include: predicting the number of clusters with cluster ensembles [6], spectral clustering [7], for speaker diarization [8], and showing that a drastic drop in the magnitude of the eigenvalues can indicate the number of clusters [6,8], but are not in the focus of our work.

The core idea of this work is using the unsegmented speech data of the conversation to produce a confidence measure for the diarization process. In two-speaker conversations our proposed confidence measure can be used to measure the goodness of the segmentation. One advantage is a potential increase in speaker diarization accuracy. In a call center audio analytics scenario, the confidence measure can be used to filter out some potentially bad segmentation outputs and by that to increase the diarization accuracy on the remaining calls. Moreover, the new confidence measure is useful not only for offline processing but also for online speaker diarization. In online processing, the audio segment from the beginning of the call is used for building the initial diarization models of the speakers in the call. If the initial segment is too short, then it may be that only one speaker is present in the available audio segment. The proposed confidence measure can detect that situation and can indicate that more audio data is needed before performing the segmentation of the call, and by that to increase the overall accuracy of online speaker diarization. On the other hand it can indicate when the initial segment contains a sufficient amount of audio for the diarization, and by that reduce the computation.

This paper is organized as follows. Section 2 describes our diarization system and gives a brief overview of some existing clustering validation methods. Section 3 describes our proposed confidence method, and Section 4 presents our experimental results. Finally, in Section 5 we provide some conclusions.

2. Speaker Diarization

2.1. Offline diarization

Our offline two-speaker diarization system, which is described in details in [9], is based on supervector parameterization of short audio segments, unsupervised (session based) compensation of intra-speaker within-session variability modeling, followed by clustering based on Principal Component Analysis (PCA). The audio is initially separated into two clusters by the PCA projections, and from there, two speaker's models are trained from the feature vectors of the processed audio and the segmentation is refined by several Viterbi re-segmentation and model training iterations. The PCA plays an important role in our diarization engine as it is used as a mean to separate between the two speakers by looking at the projection of the supervectors on the eigenvector with the largest eigenvalue.

2.2. Online diarization

Our framework for online diarization is based on the offline diarization of a conversation prefix, as described in subsection 2.1, followed by an efficient online processing of the incoming audio of the rest of the conversation, resulting with updates of the diarization with a predefined latency. Our online diarization algorithm is fully described in a recent paper [10]. This online diarization system is useful when the call prefix is short. But for short prefixes, in many case one side of the conversation may be underrepresented and the diarization performance would suffer. To overcome this problem, we proposed to dynamically adjust the prefix size based on confidence measures obtained on the call prefix. In this work we will use the spectral ratio as a confidence measure and compare its performance to other well-know clustering validity measures, described in the next subsection.

2.3. Clustering validity measures

As mentioned above, measuring the diarization quality is important for offline diarization of short calls as well as online diarization with a short prefix. For short calls, when data is sparse, it is important to know if the call is balanced and if both speakers are well-represented in the call. By measuring the clustering quality on the leading prefix of the call in online diarization, we can either reduce the latency of the diarization process if the confidence is high, or prolong the prefix as needed if the confidence is low.

Clustering validity measures that are computed on the segmented audio can be used as confidence measures of the diarization quality. In this work, we explored several existing clustering validation algorithms for guiding the diarization process and compare them to our proposed method that operates on the pre-segmented audio. We applied these clustering validity measures to the segmentation output obtained from diarization of the call. We tested the following four well-known clustering validity measures: the Dunn's validation index [11], the Davies-Bouldin (DB) validation index [12], the C-index [13], and the Silhouette validation method [14], that are briefly described as follows.

The *Dunn's index* aims to maximize inter-cluster distances while minimizing the intra-cluster distances. It is defined as the ratio between the minimal distance between two samples from different clusters and the maximal distance between two samples from the same cluster. Therefore, larger values of Dunn's index indicate better clustering.

The *DB index* is defined as the ratio between the sum of the two standard deviations of the data points' distances in each cluster to their cluster center and the Euclidian distance between the two clusters centers. As a result, good clustering will have a small DB index.

The *C-index* is a normalized sum of the distances of all pairs of samples from the same cluster, which means that good clustering will have a small C-Index.

The *Silhouette* is a measure of the similarity of a data element to the elements in its cluster compared to the elements in other clusters. The silhouette score is computed for every frame f and normalized to the range [0,1], resulting in a frame score s_f. Although the mean of the frame silhouette scores is commonly used as the score of a segment, we define the silhouette confidence score C_{sil} to be the ratio between the number of frames with a silhouette score below a threshold and the total number of frames in the audio segment. For a segment with N frames and threshold T, we define the silhouette confidence measure to be

$$C_{sil} = \frac{1}{N}\sum_{f=1}^{N} H(s_f), \quad \text{where} \quad H(s) = \begin{cases} 1 & s < T, \\ 0 & s \geq T. \end{cases} \qquad (1)$$

3. PCA and Spectral Ratio

PCA is a mathematical transformation of a multivariate observation dataset that finds its uncorrelated principal components called eigenvectors. It is a widely-used technique for dimension reduction in unsupervised learning of neural computation, data mining, feature extraction in signal processing, and source separation. For more details, see [15]. PCA-based diarization is using the projection of the first eigenvector (the one with the largest eigenvalue) as the criterion for separating the two speakers [9]. In addition to using the PCA largest eigenvector for diarization, we are now looking also at the eigenvalues of the PCA transform, and in particular, at the first and second largest eigenvalues. The spectral ratio is the ratio between the largest eigenvalue and next largest eigenvalue. Hence, if the eigenvalues of the covariance matrix are ordered in descending order $[\lambda_1, \lambda_2, \ldots]$, then the spectral ratio is:

$$R = \lambda_1 / \lambda_2. \qquad (2)$$

We propose to use R as a confidence measure for speaker diarization. The spectral ratio is an efficient measure that can predict the performance of the diarization. Generally speaking, if the spectral ratio for a given conversation is large, the separation between the two speakers will be good.

The clustering validation measures described in subsection 2 are all computed on the clustering result, while our proposed method is based on computing the confidence on the unsegmented audio. Besides comparing our method to the post-segmentation clustering validity measures, we also compare it with another pre-segmented confidence measure which is described in [5]. The empirical confidence measure proposed in [5], compares the largest eigenvalue to the mean of all the other eigenvalues, while we compare the largest eigenvalue to the second largest eigenvalue. This confidence measure behaves differently from our proposed one. It is less robust and does not perform as well as our proposed confidence, as we show next in the experimental results section.

4. Experimental Setup and Results

4.1. Setup

A subset of the NIST-2005 SRE core dataset was used as a development set (131 sessions), and a disjoint subset was used as an evaluation set (929 sessions). We artificially convert the stereo datasets to mono by summing both channels. The ground truth was derived from the automatically produced transcripts provided by NIST. We use the standard speaker error rate (SER) measure and do not include speech/non-speech errors. SER is computed according to the standard protocol for evaluation of a two-speaker segmentation task, which is available from NIST [16]. In our experimental setup, we perform the diarization on several prefix lengths of the five-minute calls, ranging from 15sec up to 240sec. Hence, the total number of test segments is fixed in all our experiments but they vary in audio length depending on the prefix size. Our diarization engine uses a 16-Gaussians UBM model for generating the supervectors for diarization. The UBM is trained offline on the development data and not on the audio of the test call. Similarly, a single NAP transformation is trained on the development set and used for compensation of all the test calls. The front-end of our system is based on Mel-frequency cepstrum coefficients (MFCC). The feature vector consists of 13 cepstral coefficients extracted every 10 msec using a 25 msec window. We remove non-speech frames using an adaptive energy-based voice activity detector (VAD) with Viterbi smoothing.

4.2. Improved diarization accuracy

In this experiment we show that the proposed confidence measure *R*, defined in Equation 2, can be used to filter out potentially bad segmentation. Based on the value of confidence measure, we select a subset of the test calls with the highest confidence and filter out the rest of the calls. By removing some percentage of the calls with low confidence we can increase the overall accuracy of the rest of the data. This result is presented in Figure 1 by showing the SER performance as a function of the total amount of segmented calls for several prefix lengths. For example, the result for 90% segmented calls means that 10% of the calls with the lowest spectral ratio confidence value were filtered out. The performance of the confidence filtering on the entire conversations (total of 300sec) is almost identical to that of the 240sec prefix, and, therefore, is omitted from the figure.

We can see the effect of filtering the data with the spectral ratio confidence. In all cases the overall performance improves on the filtered portion of segmented calls obtained by the confidence measure. From analysis of these experiment results we find that, for example, by filtering out 20% of the calls with the lowest confidence values we can decrease the overall SER by up to 10%-45% relative to the prefix baseline performance depending on the prefix length.

If we translate these experimental results to an operational diarization system, each point in the graph corresponds to some threshold on the confidence measure that can be used for filtering out the calls for achieving overall better accuracy at the price of ignoring or requesting for more input audio on the filtered subset.

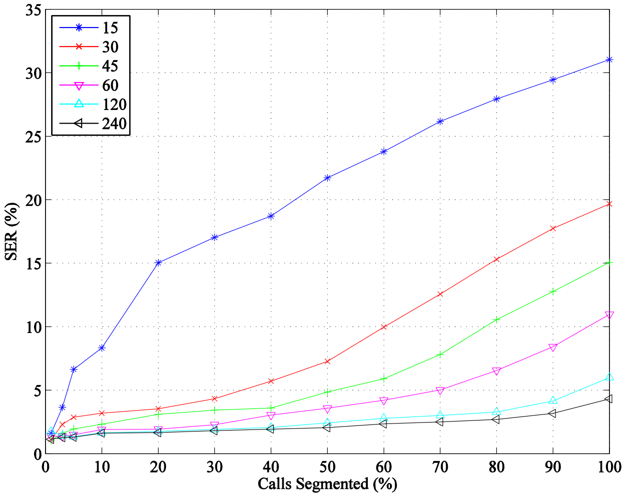

Figure 1: SER after filtering with the spectral ratio confidence measure as a function of the percentage of segmented calls for several prefix sizes.

4.3. Comparison with other confidence measures

We now compare the performance of the confidence measure *R* to the four post-segmentation confidence measures described in subsection 2.3. For the silhoutte confidence as defined in Equation 1, a threshold of 0.5 is used. In addition, we compare the spectral ratio confidence to the empirical eigenvalue-based confidence proposed in [5]. The results are shown in Figure 2, with the spectral ratio confidence ('Eig SR') and the other pre-segmentation confidence ('Eig Spread'). Both pre-segmentation methods are shown in solid lines while all the post-segmentation validation methods are shown in dotted lines. The results in the graph are for processing the entire conversations with 300 seconds of audio. We can see that both the pre-segmentation confidence measures are significantly better than the post-segmentation clustering validation measures. Moreover, our proposed confidence measure performs better than the confidence measure proposed in [5], and achieves the best result in all filtering conditions. As for the comparison between the different post-segmentation measures, we can see that the DB Index, the C-Index, and the silhoutte all perform roughly the same. All three measures give some gain in filtering out the segmentations with low confidence, but the performance on the 10% of the segmentations with highest confidence is worse than the baseline. The Dunn's Index is the only post-segmentation measure that gives a consistent gain in performance for all the filtering range. However, its results are not as good as those of the pre-segmentation confidence measures.

We performed the same experiment of comparing the six confidence measure on shorter prefix lengths. The results show that on average, the performance on the spectral ratio confidence is better than the 'Eig Spread' confidence for all prefix lengths, and that it is more beneficial as the audio segment gets longer. The spectral ratio confidence is also always better than the four post-segmentation measures except for the shortest prefix of 15 seconds, for which the DB-Index and the C-Index performed better.

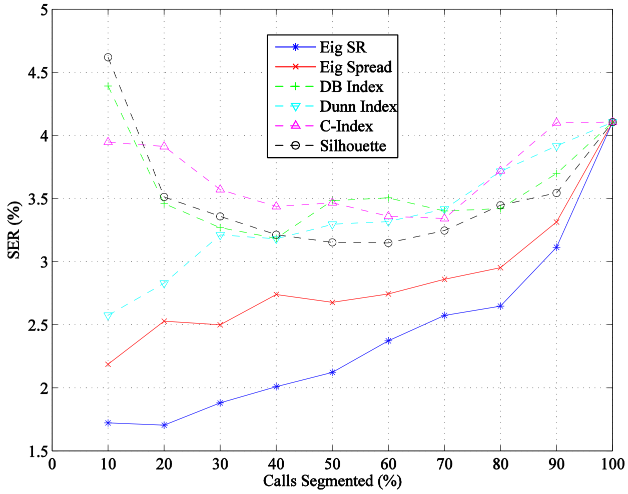

Figure 2: SER after filtering with several confidence measures computed on the entire 300-second calls as a function of the percentage of segmented calls.

4.4. Using other diarization algorithms

Next we show that the spectral ratio confidence is a general measure that can be used not only with our PCA-based diarization but also in conjunction with other diarization algorithms. To demonstrate that, we performed the filtering experiment with the spectral ratio, as in subsection 4.2, and measured the overall diarization performance of a BIC-based system described in [9]. The results are summarized in Table 1 as a function of the filtered calls (0% is the baseline with no filtering). We can see that the spectral ratio confidence improves the overall diarization accuracy with both diarization algorithms.

Table 1: SER after filtering with the spectral ratio confidence comparing between the PCA-based diarization and BIC diarization.

Filtered (%)	0	10	20	30	40	50
PCA	4.1	3.1	2.6	2.6	2.4	2.1
BIC	6.5	5.6	4.8	4.6	4.1	3.8

4.5. Timing Results

The time reduction using our confidence scoring algorithm relative to the full diarization algorithm followed by a clustering quality evaluation depends on the specific diarization algorithm and the specific clustering validation algorithm used. In particular, if using our speaker diarization engine, which is based on the PCA and requires all the processing for computing the confidence plus some extra steps for producing the diarization, the time saving is not high (~25%), given that we apply a low-complexity clustering validation algorithm such as the DB Index. But if it will be used with other diarization algorithms that may be of higher complexity and potentially more accurate the time reduction could potentially be more significant.

5. Conclusions

In this paper we introduce a confidence measure for speaker diarization, based on the PCA eigenvalue spectral ratio. The spectral ratio confidence measure is computed on the incoming audio before diarization is performed. We show that the proposed confidence can be used for filtering out calls which will potentially produce bad segmentation and by that to increase the overall accuracy of the system on the remaining calls during offline processing. The confidence-based approach is also important for online diarization, where the system runs on a short prefix of the call and a confidence measure is necessary for determining if the quality of the segmentation is sufficient for a given prefix length.

The proposed confidence measure performs better than other existing pre-segmentation and post-segmentation clustering validation methods that we have tested. We show the usefulness of the spectral ratio confidence measure for both filtering out the bad segmentations and for selecting the good segmentations on short calls. Moreover, this confidence measure, which operates on the pre-segmented audio, could be integrated with other diarization engines.

6. Acknowledgements

This work is supported by the Dem@Care FP7 project, partially funded by the EC under contract number 288199.

7. References

[1] S. Tranter, D. Reynolds, “An Overview of Automatic Speaker Diarisation Systems”, *IEEE Transactions on Audio, Speech and Language Processing*, 2006, pp. 1557-1565.

[2] C. Vaquero, A. Ortega, J.A. Villalba, A. Miguel, and E. Lleida, "Confidence measures for speaker segmentation and their relation to speaker verification", in Proc. *Interspeech*, 2010, pp.2310-2313.

[3] C. Vaquero "Robust Diarization for Speaker Characterization", *Ph.D. Thesis*, 2011.

[4] C. Vaquero, A. Ortega, E. Lleida, “Partitioning of Two-Speaker Conversation Datasets”, in Proc. *Interspeech*, 2011, 385-388.

[5] S. Arberet, R. Gribonval, F. Bimbot, “A Robust Method to Count and Locate Audio Sources in a Multichannel Underdetermined Mixture”, *IEEE Transactions on Signal Processing*, Vol. 58, No. 1, 2010, pp. 121-133.

[6] N. Bassiou, V. Moschou, C. Kotropoulos, “Speaker Diarization Exploiting the Eigengap Criterion and Cluster Ensembles”, In Proc. of *IEEE Transactions on Audio, Speech & Language Processing*. 2010, pp. 2134-2144.

[7] U. von Luxburg, “A tutorial on spectral clustering,” *Statistics and Computing*, vol. 17, no. 4, pp. 395-416, 2007.

[8] H. Ning, M. Liu, H. Tang, and T. S. Huang, “A spectral clustering approach to speaker diarization”, in Proc. *ICSLP*, 2006.

[9] H. Aronowitz, "Unsupervised Compensation of Intra-Session Intra-Speaker Variability for Speaker Diarization", in Proc. *Speaker Odyssey*, 2010.

[10] H. Aronowitz, Y. Solewicz, O. Toledo-Ronen, "Online Two-Speaker Diarization", in Proc. *Speaker Odyssey*, 2012.

[11] J. Dunn, “Well separated clusters and optimal fuzzy partitions”, *Journal of Cybernetics*, Vol. 4, 1974, pp. 95-104.

[12] D.L. Davies, D.W. Bouldin, “A cluster separation measure”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 1, No. 2, 1979, pp. 224-227.

[13] L. Hubert, J. Schultz, “Quadratic assignment as a general data-analysis strategy”, *British Journal of Mathematical and Statistical Psychology*, Vol. 29, No. 2, 1976, pp. 190-241.

[14] P.J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”, *Journal of Computational and Applied Mathematics*, Vol. 20, 1987, pp. 53-65.

[15] I. Jolliffe, “Principal component analysis”, Springer, 2nd ed., 2002.

[16] NIST segmentation scoring script (2002), Available online: http://www.itl.nist.gov/iad/mig/tests/sre/2002/SpkrSegEval-v07.pl.