Future challenges in speaker diarization

Itshak Lapidot[1], *Hugo Guterman*[2]

[1] Department of Electrical and Electronics Engineering, Sami Shamoon College of Engineering, Beer-Sheva, Israel
[2] Department of Electrical and Computer Engineering, Ben-Gurion University, Beer-Sheva, Israel

itshakl@sce.ac.il, hugo@bgu.ac.il

Abstract

Speaker diarization challenges are usually summarized in answering the question “Who spoke when?”, but in practice there are many questions that either have no answer or have only a partial answer. Several problems are still not under investigation. In this paper we will raise the problems in speaker diarization and give several potential answers and directions for future research.

1. Introduction

Speaker diarization is a crucial issue in speech/speaker recognition technology. The question “Who spoke when?” is of great importance for commercial and forensic applications. In the past, human experts did most of the work to answer the question. The huge size of transmitted speech data makes it impossible to handle and annotate it using human experts.

In recent years, different speaker diarization systems have been developed [1], [2], and [3]. The general task is to separate a conversation according to the speakers’ appearance, when the number of participants and the participants themselves are not known. The state of the art system is based on Joint Factor Analysis (JFA) that requires a huge quantity of data to be trained off-line for channel compensation [3]. The estimation of the number of speakers (validity problem) is mostly done using Bayesian-Information-Criterion (BIC) [4].

Although the results based on JFA are promising, the necessity of channel compensation off-line training is not always possible, e.g., forensic applications, and algorithms without any preliminary training must be applied [5], [6]. These algorithms are still far from the JFA-based system. BIC was initially developed for model selection and not for solving the validity problem. The validity problem solution based on BIC is very unsatisfying, as the tuning parameter of the penalty term is highly dependent on the application and conversation duration. The penalty free version of BIC has problems of model complexity [1]. Therefore, other validity criteria need to be found [7].

Additionally, several other problems are interesting and have to be solved: merging a known speaker in an unsupervised diarization system; diarization with several optional known speakers from a closed set; on-line diarization with a known/unknown number of participants [8] and [3]; coupling between the diarization and speaker verification systems to make it almost without human interference; and automatic diarization quality estimation. All these issues will be the focus of exciting and innovative research in the coming years.

The rest of the paper is as follows: in section 2 we present briefly the advantage of JFA and show the limitations; BIC and its limitations are presented in section 3; the decision about BIC-free systems is given in section 4; section 5 discusses merging known speakers into a diarization process; section 6 deals with automatic quantification of the diarization quality, while the merging of diarization and verification is presented in section 7; section 8 concludes the paper.

2. JFA is good but very expensive

In JFA each super vector consists of a linear combination:

$$s = m + Ux + Vy \qquad (1)$$

Here s is a *speaker*- and a *channel*-dependent supervector; m is a speaker-independent supervector produced by the Universal Background Model (UBM); U is a low rank rectangular matrix where the columns are referred to as eigenchannels; x is a vector with a standard normal distribution; V is a low rank rectangular matrix whose columns are interpreted as eigenvoices; y is a vector with a standard normal distribution. The entries of x are the channel factors while the entries of y are the speaker factors.

For good estimation of those factors, hundreds of hours of speech data are required. A detailed analysis of applying JFA to speaker diarization is given in [3]. The combination of JFA together with a Variational Bayes (VB) system allows performing diarization and the estimation of the number of speakers at the same time.

The JFA-based system is highly dependent on the amount of training data of the factors. If the data do not represent the channels well, it can cause degradation in diarization. The VB-based system is also a very expensive system.

3. BIC cannot solve diarization problems

Initially, the Bayesian Information Criterion was developed for model selection [Shwartz]:

$$\hat{\theta} = \max_{\theta \in \mathcal{M}} \left\{ L(X \mid \theta) - \frac{\#\theta}{2} \log(N) \right\} \qquad (2)$$

where θ is a set of model parameters from a model family $\mathcal{M}$; # is the cardinality, and N is the size of the data. The underlying assumptions are that the parameter θ belongs to a continuous distribution and the data size N is constant. In this case the best set of parameters is the set that gives the maximum log-likelihood, penalized by the model complexity (in this version the complexity is linear with the number of model parameters).

In clustering we usually have disagreement with the BIC assumptions: the number of data points is not constant and not all the parameters are continuous. The parameter of greatest interest, i.e., the number of clusters, is discrete.

3.1. BIC as a basic block in diarization systems

BIC can be used twice in the clustering, once for change detection and once during the segment merging process [4]. Commonly, all the models are single multi-dimensional, full covariance, normal distributions.

$$L(X \mid \theta_0) \gtrless L(X \mid \theta_1) + L(X \mid \theta_2) - \lambda \frac{\#\Delta\theta}{2} \log(N) \qquad (3)$$
$$\#\Delta\theta = \#\theta_1 + \#\theta_2 - \#\theta_0$$

The idea is to calculate the likelihood of a segment of length N with parameters θ_0, and the sum of likelihood of the spit segment with lengths N_1 and N_2 data points, such that $N_1 + N_2 = N$, with θ_1 and θ_2 parameters, respectively. The splitting point is chosen to maximize the right side of (3). The penalty term multiplies a hyper-parameter λ, which is found empirically using a development set. If a change point is not found, the segment size is increased and the process repeats.

During the merging phase, the same idea is applied. For each two clusters, the hypothesis H_0 is to merge the clusters while H_1 is not to merge. For each two segments the similarity between the segments is calculated according to (3), and the best candidates are merged if H_0 holds for them. For the merging process a different λ then in the change detection phase should be found empirically.

3.2. BIC and the over clustering problem

As the penalty term is logarithmic with the data size N, and the log-likelihood is approximately linear (for large N), after several cluster merging repetitions, the penalty term can be neglected. As the log-likelihood of two clusters is not lower (and is usually higher) than the log-likelihood of one cluster, the merging process stops.

3.3. BIC cannot solve the validity problem

As can be seen, BIC cannot solve the validity problem as it depends on the penalty term, which can be ignored for large N. It is possible to compensate for it with good accuracy by carefully adjusting the hyper-parameter λ. It can give a quite good solution if all the conversations are approximately of the same length, like in NIST competitions. If the conversations are varied in their durations, as happens in real life scenarios, the adjustment is not possible.

3.4. Several alternatives to standard BIC

Several alternatives are tested in order to deal with the BIC-based validity problems. One approach presented in [9], which penalizes each model according to the data size of the model's data and according the size of all segments. In such a case, instead of the penalty term in (3), the new penalty term will be $\lambda\left(\frac{\#\theta_1}{2}\log(N_1) + \frac{\#\theta_2}{2}\log(N_2) - \frac{\#\theta_0}{2}\log(N)\right)$. Another approach is penalty free [1], which will be discussed in section 4.

Other ideas that should be investigated are: to add a penalty on the discrete variable (which is not penalized in the standard BIC); the number of clusters; the hyper-parameter should be dependent on data size N, and not constant, as it is now.

4. Penalty free diarization

To overcome the penalty problem, a penalty-free BIC was suggested by Ajmera et al. [1]. The basic idea is to estimate the merged cluster with the number of Gaussians, which is the sum of the merged candidate clusters' Gaussians. Such an approach frequently leads to over-clustering.

4.1. Cluster complexity problem

Several problems arise with such an approach:

1. As in GMM, all the mixture weights sum to one, the number of parameters in the two clusters case is smaller by one parameter than in the one cluster case. This can be treated as in regular BIC by penalizing for one parameter.
2. The complexity of two GMMs with M_1 and M_2 mixture components is lower than the complexity of one cluster with $M_1 + M_2$ mixture components, as the dimensionality of the searching space is lower. It is more complex to search in one high-dimensional space than in two low-dimensional sub-spaces. Several possible solutions are given in [9].
3. For large N with a conditional independence assumption, the log-likelihood of the data, given a true *pdf*, is linear with N, while the penalty is logarithmic and can be neglected. It is easy to show for true clusters *pdf*s:
   $$-N \cdot H(X \mid C) \le -N_1 H(X \mid C_1) - N_2 H(X \mid C_2) \quad (4)$$
   when $H(X \mid C)$ is the entropy given the true *pdf* of cluster C. As GMM is a universal *pdf* estimator in the KL sense, when the number of mixture components is high, (4) holds for estimated *pdf*s.

To avoid the problem, the number of cluster parameters must be limited in advance.

5. Sometimes the number of participants is known

In several scenarios, the number of speakers is known in advance. Telephone conversation is the main representative of such a scenario. This information must be supplied to the system.

In bottom-up systems the number of participants defines the stopping criterion. In case of BIC-based systems, the merging is performed as far as the system can find a pair of clusters such that in (3) H_0 holds, i.e., the merged cluster has higher log-likelihood than the log-likelihood of two clusters after penalization. When the number of speakers is known the merging procedure continues even when H_0 does not hold. In such a case, the merged clusters are those that achieve the highest ΔBIC , i.e., subtracting the left side of (3) from the right side.

In top-down usually the diarization begins with one cluster, and the splitting process adds one cluster each time. When the number of participants is known, the system can start working with the correct number of clusters from the beginning [3], [5], and [6].

6. When participants are partially known

In several applications some of the participants are known. For example in TV news the reporters are usually known, while the people being interviewed are not. Most of the diarization systems assume that all the participants are unknown and create new models for all the participants. If we can find a known participant conversation, then we can use an existing “good” model. The problems are first to identify the known speakers and second to use existing models. It is not straight forward to do it. Usually in speaker recognition we use thousands of parameters for a speaker model, e.g., 1024 Gaussian mixtures. In diarization it is well known that the clusters must be much smaller, e.g., 32 Gaussian mixture. One reason is due to (4), as a large model will tend to over clustering.

One solution is to spot the known speaker with a large model and then train a small model from this data and integrate it into diarization system.

7. Initializing the diarization system

In many diarization systems the performance depends on the initial data chosen for each cluster. As the convergence of the system depends on initial models, “wise” initialization can give improvement both in DER and speed of convergence. The common approaches are to equally divide the data between the clusters, either by sequential attribution of the data or random attribution.

In the approach we propose in [5], the data are divided into segments using a Voice Activity Detector (VAD). The simplest assumption is that between two non-speech segments the speech segment has to be attributed to one cluster. We calculate the mean feature vector for each segment and apply a variant of the K-means, which takes into account the segments’ lengths. The algorithm is called Weighted Segmental K-Means (WSKM). Using such initialization gives more than 12% relative improvement in DER and more than 50% reduction in the number of iterations for system convergence.

8. How good is the diarization?

One of the important applications of the diarization system output is speaker verification. If the diarization performance has a high Diarization Error Rate (DER), the verification results will not be reliable. It is important to know for each conversation whether the DER is low and whether the data can be used for verification.

One solution is to have the human expert listen to the diarization outputs. This is not possible when the quantity of data is very large. The solution is to build an automatic system that will grade diarization performances.

Several features can be extracted in order to grade the diarization. Parameters that can give information on the similarity of the clusters can be KL2 distance or any cluster in-between distance. The larger the distance indicated, the better the separation tendency. Expectation of the log-likelihood ratio of the segments also can be an indicator for diarization quality. For each segment, the ratio between the log-likelihood of the chosen cluster to next best cluster is calculated and the average ratio is found. We found that the higher the ratio, the better the diarization. An additional parameter is the average segment length. We observed two-speaker telephone conversations that for low DER the average segment length produced by the system was longer than in case of high DER.

9. Diarization for speaker verification

A general verification system is shown in Figure 1. The common concept is to create two models: Universal Background Model (UBM) and Speaker model. Both models are usually GMMs, where the UBM is trained on a large population, while the speaker model adapted from the UBM is based on a small quantity of data. During the testing, a new utterance is provided together with the UBM and speaker model to the verification system. Different approaches are applied for verification when the state of the art system is based mainly on super-vectors (SV) together with a support vector machine (SVM) [10], or JFA based verification [11].

The common assumption is that the speaker training data has no segments of speech belonging to other speakers or overlapping speech. The same for the test data; it is assumed it all belongs to one speaker only. To achieve it, either the speaker has to volunteer or the full interference of a human expert is required to verify the data. In forensic applications, for example, the speakers usually do not cooperate and the data have to be collected from multi-speaker streams of data, e.g., telephone conversations. A human expert segmentation of data is expensive in terms of time and money, and automatic segmentation is preferred. The straight forward alternative is to use a diarization system.

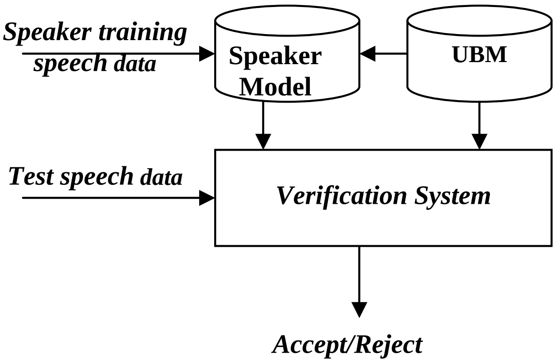

Figure 1: General speaker verification system

In the training phase human expert interference is required. Several strategies can be used:

1. After the diarization the human expert will listen to the output file and decide the channel that will be used for the training.
   Most of diarization errors occur at the boundaries of the segments. It causes the human expert to not be cooperative. Thus, it is important to minimize his involvement.
2. Let the human expert mark several seconds of the target speakers. After the diarization, the marked speech data is used to find the cluster, whose data should be used for speaker model training.
3. Let the human expert mark several seconds of both speakers (in the case of two-speaker telephone conversation, for example) and use these data to initiate the models for diarization. In this way it will be partially supervised and let us know in advance which cluster data should be used for speaker model training.

In the verification case, the data of each cluster have to be verified. If one of the verifications gave a positive result, then it is assumed that the target speaker participated in the conversation.

The training and verification should be also changed, as the assumption that all the data belong to one speaker only is not valid. One of the possible solutions might be to give a score to each segment, and then to use only the high scored segments. Each segment that contributes to the training/verification should be weighted according to its score.

For this application, the problem of the diarization quality, described in section 8, is very important. Conversations that are marked as "bad" should be rejected or treated suspiciously.

10. Conclusions

In this paper we described the challenges that were not under investigation at all or have not been investigated in depth. We did not give the solutions for the issues we raise, but only made a start on it. For several issues we presented possible directions that can give solutions.

There is a lot of work that has to be done, and there are other issues that were out of the scope of this paper, such as overlapping speech detection and inter-conversation speaker diarization, i.e., gathering the same speaker from different conversations under the same label.

References

[1] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, "Unknown-multiple speaker clustering using HMM," *Proc. International Conference on Spoken Language Processing*, pp. 573-576, September 16-20, 2002, Denver, Colorado, USA.

[2] I. Lapidot (Voitovetsky), H. Guterman, and A. Cohen, "Unsupervised speaker recognition based on competition between self-organizing maps," *IEEE Trans. on Neural Networks*, vol. 13, no. 4, pp. 877-887, July 2002.

[3] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of Telephone Conversations using Factor Analysis'" *IEEE Journal of Selected Topics in Signal Processing*, December 2010.

[4] S. S. Chen and P. S. Gapalakrishnan, "Clustering via the Bayesian criterion with applications to speech recognition," *in Proceedings of ICASSP'98*, vol. 2, 1998, pp. 645-648.

[5] O. Ben-Harush, I. Lapidot, and H. Guterman, "Weighted segmental K-means initialization for SOM-based speaker clustering," *in Proceedings of Interspeech'08, 2008*, Brisbane, Austalia.

[6] O. Ben-Harush, I. Lapidot, and H. Guterman, "Entropy based overlapped speech detection as a pre-processing stage for speaker diarization," *in Proceedings of Interspeech'09*, 2009, Brighton, UK.

[7] T. Stafylakis, V. Katsouros, and G. Carayannis, "Redefining the Bayesian Information Criterion for speaker diarisation," *in Proceedings of Interspeech'09*, 2009, Brighton, UK.

[8] O. Ben-Harush, I. Lapidot, and H. Guterman, "Incremental diarization of telephone conversations," *in Proceedings of Interspeech'10*, 2010, September 26-30, 2010, Kaihin Makuhari, Japan.

[9] T. Stafylakis, X. Anguera, "Improvements to the equal-parameter BIC for Speaker Diarization," *in Proceedings of Interspeech'10*, 2010, September 26-30, 2010, Kaihin Makuhari, Japan.

[10] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

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