Speaker Diarization in Commercial Calls

Itshak Lapidot, Lidiya Aminov, Tal Furmanov, and Ami Moyal

ACLP - Afeka Center for Language Processing, *Afeka Tel-Aviv Academic College of Engineering, Israel*

itshakl@afeka.ac.il, lidiyaa@afeka.ac.il, talf@afeka.ac.il, amim@afeka.ac.il

Abstract

Speaker diarization is becoming increasingly important component of speech and speaker recognition technologies, particularly those utilized by commercial and forensic applications. The output of a diarization system is a time-stamped conversational "diary" of participating speakers. This output can in turn be used for training additional applications such as speaker verification and automatic speaker adaptation.

In the current work we f focuses on speaker diarization for commercial applications. In comparison to "friendly" telephone conversations, conversations recorded from live commercial call-centers, introduce a new set of challenges. Various events, such as, music segments, totally zero signal intervals or DTMF and DTMF-like signals, all complicate the diarization task and inject additional errors into the system.

The tools and algorithms used to handle these complications and errors will be described in detail. System performance on real-life calls will be presented and analyzed, and performance tuning solutions will be presented. It will be shown that the baseline system which achieved on LDC America CallHome about 12.71% *diarization error rate* (DER) performed initially on real call center data with DER of about 37.5% and after additional special events treatment DER has been reduced to 23.02%.

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.

Mostly, the research works are done on databases of relatively high quality speech. It can be either a telephone conversation [1]-[3], meeting [4] or a show [5]. A lot of work was done in the last year on speaker diarization with variety of approaches [1]-[8]. A good review can be found at [9]. In most cases the preprocessing includes a *voice activity detector* (VAD) which detects non-speech events such as silence (which is basically the background noise) and frequently also a block which detects music. When applying the diarization system on real call center data, new challenges arise. In the current work we use the diarization system for telephone conversations which described in [2] and applied it to the real call center data. The degradation in the performances was dramatic. In the next sections we will present the problems and the solutions which were added to the baseline system, in order to overcome the new challenges.

The rest of the paper is as follows: In section 2 the baseline system is described, which is based on *hidden-distortion—model* (HDM); in section 3 the real call-center conversation problems are discussed and the solutions are provided; the experiment and the result are given in in section 4, while section 5 concludes the paper.

2. HDM-based diarization system

An HDM-based diarization system is shown in Fig. 1.

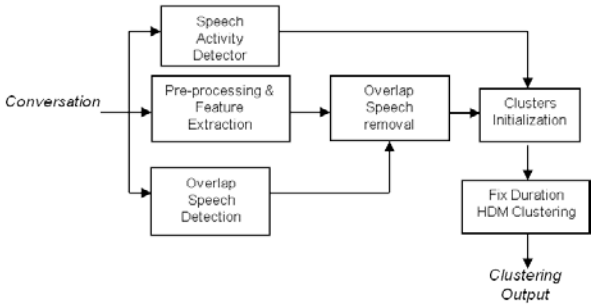

Fig. 1: HDM-based diarization system.

The system is very similar to the one which is based on *hidden-Markov-model* (HMM) [1]. First, a standard pre-processing following by extraction of 12 *mel-frequency-cepstral-coefficients* (MFCC) is performed. In parallel, energy-based speech activity detection is applied and overlapping speech detection in the time domain is performed [10]. The overlapping speech removed from the conversation, and the remaining speech segments are clustered using *weighted-segmental-K-means* (WSKM) [11] in order to initiate the speaker models. Each cluster is used to train an initial speaker model, which is a *self-organizing-map* (SOM) of size 6×10. Additional model of the same size is trained for the non-speech model with the segments which were detected as non-speech.

We use HDM instead of HMM as HMM suffers from several limitations:

1. The HMM trained using Viterbi statistics, and as such, the training of the transition matrix and the emission probabilities are disjoint. It may lead to an imbalance between the emission likelihoods and the transition probabilities. It might happen that the

global likelihood is depends mostly on the transition probability, and is almost independent from the input samples. It might be the cases when the state changes are rare. In such a case the self-loop transition probability is very high (close to one), while other transition probabilities are very small. A regularization parameter can help to improve the performance. However, in the probabilistic framework (HMM), there is no such regularization. To emphasize this point, HMM training process increasing the global likelihood, involving both, transitions and emissions parts, while in diarization, the goal is to decrease the diarization error. We hope that the tow criteria are highly correlated (the likelihood maximization and the diarization error minimization), however it is not necessarily so.

2. HMM approach is based on the probabilistic paradigm, i.e., the state models have to be statistical models (GMM for example). In several tasks like Damerau–Levenshtein distance calculation in strings comparison (DNA protein sequences), it is a limitation. Such representation of distances/distortions, in the probabilistic framework is not natural.

To overcome these two limitations, an extension of Viterbi statistics based HMM, was proposed as the *hidden-distortion-model* (HDM) [2], which combines the advantages of HMM-based approach and at the same time allows to use distortion-based approaches together with regularization parameter encapsulation. The HDM limit only to a family of additive distortions, $Distortion(x_1,\ldots,x_N)=\sum_{n=1}^{N} distortion(x_n)$, i.e. the distortion of a N vectors sequence is the sum of N individual distortions applied each on one vector. The negation of the log-likelihood is an example of such additive distortion. Following this, Viterbi-based HMM can be viewed as a particular case of HDM. Instead of the emission probabilities, emission distortions are calculated; while the transition probability matrix and initial probability vector replaced by the transition cost matrix and the initial cost vector. In this framework, a regularization of transition costs becomes a natural part of the model. The regularization parameters have to be determined based on some development data. A more detailed analysis of HDM can be found at [2].

3. Problems with real call-center data

The HDM-based diarization system performed well on telephone calls which collected by LDC, such as America Callhome [12], and achieved *diarization-error-rate* (DER) of 12.71%. However, the results on call-centers data achieved much higher DER of 37.33%. In the next sub-sections the real call-centers data problems are discussed together with the actions which were taken to solve the problem.

3.1. Music detection

In general case, a telephone conversation usually consists of four types of events:

1. Speaker 1 is speaking.
2. Speaker 2 is speaking.
3. Overlapping speech – both speakers are speaking at the same time.
4. Non-speech event – only background noise as working computer and/or air-condition, chair movement, etc.

In call-center calls quite frequently there are also music segments. Music segments are not detected by a non-speech detector and assigned to speaker clusters. As such, these segments are part of the diarization error, and as they used to train the speaker models, the models are not correct and the segmentation is not precise as it could be. As a solution, we replace our energy-based non-speech detector by another non-speech detector which also trained to detect music segments. The music segments were eliminated from the segmentation.

3.2. Complete silence removal

Non-speech events are usually an environmental noise. In call-center calls when the recording system detects a non-speech event, zero values are assigned instead of the noise values. Such segments cause to a problem in both, non-speech detection (in case of the energy-based non-speech detector as it place the lower threshold to zero) and training the non-speech model. After DC removal, the zero values have some finite small value. An MFCC of a constant segment is a Kronecker delta multiplied by a constant, i.e., first filter energu is non-zero, while all the other filters outputs are zero. Consequently, the MFCC vector is a vector of same values (constant vector). If there are several segments of zeroes, it cause to badly trained non-speech model., as several code-words drift toward this vector.

To overcome this problem, the zero segments were detected in advance, and removed from the conversation.

3.3. DTMF removal

Additional event that frequently appears at the call-center telephone conversations and does not appear in "friendly" calls is the *dual-tone multi-frequency* (DTMF) event. These tones are combinations of two sinusoidal signals with different frequencies. These tones are caused by pressing the telephone buttons. To overcome the problem, a DTMF detector was added to the system and all segments which detected as DTMF are removed from the conversation before the models initialization stage.

After adding all the new detected events, the overall system is as shown in Fig. 2,

4. Experiments and results

We apply our HDM approach on a two-speaker telephone speaker diarization task. The features we used are the classical Mel-Frequency Cepstral Coefficients (MFCC) which are extracted on 20ms signal window with 50% of overlap (12 MFCC coefficients). The speaker diarization system has 3 hyper-states (non-

speech, speaker A, speaker B). A fixed duration constraint of 20 tied states (200ms) is used during the first 5 iterations, and one additional iteration to increase the resolution used with only 10 tied states (giving a total of 6 iterations). Each cluster model is a Self-Organizing Map (SOM), with size of 6×10. In all HDM experiments, the model distortion measure is the square Euclidian distance. The non-speech model is initialized using the non-speech segments provided by the SAD while the two other models are initialized according to a weighted segmental K-means [11] which applied only on the speech segments.

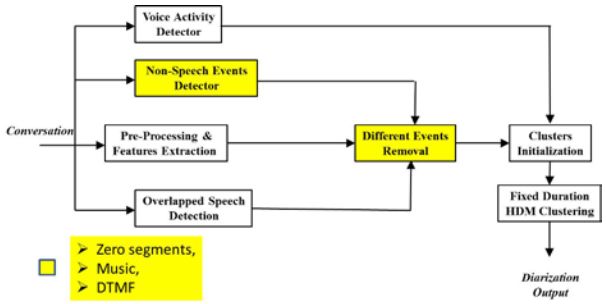

Fig. 2: The new system.

4.1. Database

Two databases are used for the experiments: LDC America CallHome [12]. 108 conversations CallHome conversations are used for LDC of about 30 minutes duration each, but with only about 10 minutes with human transcription. Only this transcribed part is used here. The other database is telephone conversations from several call-centers. The data are sampled at 8kHz in a 2 channel μ-law format and the two channels are summed in order to have one channel conversations.

4.2. Diarization Error Rate (DER)

The performance is evaluated using frame-based diarization error rate (DER), as defined by NIST in [13]. The DER defined as following:

$$DER = 100\frac{\sum_{s=1}^{S}\left\{dur(s)\cdot\left(\max\left(N_{Ref}(s),N_{Sys}(s)\right)-N_{Correct}(s)\right)\right\}}{\sum_{s=1}^{S}\left\{dur(s)\cdot N_{Ref}(s)\right\}}$$

Given S speech segments in the conversation, the DER calculated according to the following notation:

- $dur(s)$ – Duration of the segment s.
- $N_{Ref}(s)$ – The number of speakers assigned to segment s.
- $N_{Sys}(s)$ – The number of speakers assigned by the system to segment s.
- $N_{Correct}(s)$ – The number of speakers assigned by the system to segment s which actually takes part in s.

As diarization task is an unsupervised task, the indexing of the speakers by the system do not use any prior knowledge about their identity. As such, the DER should be calculated for all possible permutations of the speaker indices, and the minimal DER is taken.
The DER calculation is performed excluding a 0.5 seconds time-window around the changing points (i.e., the errors inside 0.25 second on each side of the changing points are not taken into account).

4.3. Experiments with LDC America CallHome

Table 1 presents the DER for the baseline following by the improvement achieved by adding an additional block to the system. Before starting analyze the results, it is important to remind that on LDC America Callhome database [12] the DER was 12.71%.

Table 1: Results of the basic system and with additional events detection blocks.

System Improvements	DER [%]
Basic system	37.33
VAD + Music removal	26.43
Complete silence removal	22.51
DTNF removal	22.46

As can be seen from Table 1, music segments have a crucial impact on the diarization system performance. Music removal block improve the DER by more than 10%, which is about 29.2% relative improvement. The impact of this removal is not only in reducing the errors caused by a wrong labeling of these segments (if they attributed to one of the speaker), but also in the fact that the models which have music segments are wrongly trained, as the subspace of the music MFCCs is not the same as speech or non-speech MFCCs.
Complete silence removal is the removal of segments which have zero value. It was found to be very useful to remove these segments out, before the HDM start to iterate (about 3.9% additional absolute reduction in DER). For our understanding, the presence of complete silence segments can affect the SAD decision as it influence on thresholds which are data adaptive. At the same time, the non-speech SOM model is not trained well as part of the code-words places are influenced by these segments MFCCs.
The last block is DTMF removal. Although the improvement in DER is minor, it still has to be done. First of all, it is the right thing to do. DTMF signal is not one of the speakers signal and should be taken out. Moreover, it has positive influence on the accuracy of the diarization. If the diarization is performed for the further process of the data, such as speaker verification or automatic speaker adaptation, we would prefer to lose some data, and to have more pure clusters. It is more important than having small miss with bigger false alarm. Additional wrong data in the cluster may damage the further systems performance.

5. Conclusions

In this study we implement the HDM-based speaker diarization system to real call-center conversations. It is clear that these conversations are much harder to diarize than the conversations in the LDC databases. The reason is not only in the channel quality which is frequently not high, but also due to additional effects which are not speech, but also not typical non-speech. In this work we focused on three main non-typical events: music, DTMF and complete science. The removal of these events reduces the DER by 40% approximately. It clearly shows the importance in learning the specificities of the data, in order to deal the specific problems which occur.
After removing the problematic events, the DER is still much higher than the one on LDC America Calhome. For our understanding, the difference in DED is according to the following reasons:

1. The music and DTMF removal us not perfect,
2. The speech quality is not very high,
3. Some of the conversations are very short (less than a minute) and there is not enough data to obtain a good statistic and as a consequence, the diarization system does not perform well.

From the comparison between the performances of the diarization system on LDC America Callhome and the real call-center conversation it is evident that a large improvement can be done. To do so, more data investigation should be done in order to characterize the bottlenecks which must be released.
This work is in progress and more events should be still analyzed and additional actions should be taken. As a continuation of this work we extend the diarization of call-center calls with more than two speakers and we expect to find new events which might damage the diarization performance and will have to be treated.

Acknowledgements

This work was supported by Grant #49084 provided by the Chief Scientist of the Israeli Ministry of Economy for developing Robust Speaker Diarization for the multi-speakers telephony environment. The research was carried out as part of the Magneton program, which encourages the transfer of knowledge from academic institutions to industrial companies, in this case the Afeka Center for Language Processing (ACLP) and Nice Systems Ltd.

References

[1] O. Ben-Harush, O. Ben-Harush, I. Lapidot, and H. Guterman, “Initialization of iterative-based speaker diarization for telephone conversations,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 414-425, February 2012.

[2] I. Lapidot and J.-F. Bonastre, “Generalized Viterbi-based models for time-series segmentation applied to speaker diarization,” *Odyssey2012*, June 25-28, 2012, Singapore.

[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. Shum, N. Dehak, R. Dehak, and J. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015-2028, 2013.

[5] M. Huijbregts, Marijn and D. van Leeuwen, "Diarization-based speaker retrieval for broadcast television archives," In *Interspeech-2011*, pp.1037-1040, Florence Italy.

[6] 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.

[7] I. Lapidot, J.-F. Bonastre, and S. Bengio, “Telephone Conversation Speaker Diarization Using Mealy-HMMs,” *Speaker Odyssey 2014*, June 16-19, Finland.

[8] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *Transactions on Audio, Speech, and Language Processing, IEEE/ACM*, vol. 22, no. 1, pp. 217-227, 2014.

[9] A. X. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Transactions on Audio, Speech, and Language*, vol.20, no.2, pp.356-370, Feb. 2012

[10] 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.

[11] 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.

[12] Liguistic data consortium. LDC97S42, Catalog, 1997. Available: http://www.ldc.upenn.edu/Catalog.

[13] “Nist diarization criterion,” available: http://www.itl.nist.gov/iad/mig/tools/.