INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

On the importance of efficient transition modeling for speaker diarization

Itshak Lapidot[1], *Jean-François Bonastre*[2]

[1]Afeka Tel-Aviv College of Engineering, ACLP, Israel
[2]University of Avignon, LIA, France
itshakl@afeka.ac.il, jean-francois.bonastre@univ-avignon.fr

Abstract

In recent years speaker diarization becomes an important issue. In previous works, we presented the Hidden Distorsion Model (HDM) approach, in order to overcome the limitations of traditional HMMs in terms of emission and transition modeling. In this work, we show that HDM allows to build more efficient speaker diarization systems both in terms of diarization error rated and in terms of memory footprint. The best diarization performance is obtained using smaller than usual emission models which constitutes potentially a key advantage for embedded applications with limited memory resources and computational power. A significant memory size reduction was observed using LDC CALLHOME (American) for both SOM- and GMM-based emission probability models.

Index Terms: speaker diarization, hidden Markov model (HMM), hidden distortion model (HDM), Gaussian mixture model (GMM), self organizing map (SOM)

1. Introduction

Speaker diarization in general and two-speakers diarization more specifically is useful for many application fields like forensics, homeland security or call center. Many different approaches were applied in order to solve this problem [1]. These approaches can be divided into two groups, depending on the availability of reliable data for off-line training processes. The first group corresponds to the situation where this kind of data is available. I-vectors-based diarization systems [2–5] and *extended hidden Markov model* (e-HMM)-based diarization system [6] are good representative of such approaches. On the other side, we find approaches which do not rely on off-line training. For these approaches, the diarization process is entirely based on the audio recording to be processed. The *Bayesian information criterion* (BIC)-based diarization systems with and without penalty terms [7, 8], HMM-based systems [9,10] and its generalization to *hidden distortion model* (HDM)-based system [11] are good candidates as these approaches could be applied when there is no reliable data for the off-line training or when the conditions are frequently varying.

The work presented in this article is a continuation of [11]. In [11] we presented hidden distortion model (HDM), a new approach relative to the second group which could be seen as a generalization of the Viterbi-based HMM approach. HDM presents two advantages. The first is to rely not only on the probabilistic framework but also on any additive distortion emission models, e.g., sum of Euclidian distances or sum of the negation of the log-likelihood, and non-probabilistic transition costs. The second advantage is that it allows to scale the ratio between the transitions and the emissions, similarly to [12]. The main difference to [12] is the fact that HDM allows to optimize the model parameters together with the scaling factor, which was not allowed in the quoted previous work.

In [11] we showed that different transition cost matrices can lead to different results for *self organizing map* (SOM) [13] emission models, with adjusted scaling factors. In the present work we explore the performance of HDM using *Gaussian mixture model* (GMM) as emission models. We will also see if adjusting the scaling factor has an impact on the optimal emission model sizes, using SOM or GMM emission models. As both training process and Viterbi decoding complexity are almost linearly dependent on the model size, the diarization speed is also directly linked to this meta parameter.

The rest of this article is organized as follows: Section 2 presents a quick overview of the HDM approach. A specific highlight is set on the theoretical constraints and the scaled log-likelihood transition costs. The diarization system is described in 3. Experimental results on speaker diarization are given in section 4. Finally, in section 5, we conclude with a discussion about the benefits of HDM in the context of a diarization system.

2. Hidden distortion model - HDM

Let us assume a system with K states. Each state k is defined by a distortion model DM_k, and let $C_{qk} = cost\left(s_n = q | s_{n-1} = k\right)$ be the transition cost of being in state q at any discrete time n, given being in state k at time $n-1$. We define $\mathcal{C} = \left[C_{qk}\right]_{|q,k \in \{1,\ldots,K\}}$ to be the cost transition matrix. $d_k\left(x_n\right)$ is a distortion of the data vector $x_n \in X = \{x_1, \ldots, x_N\}$, given a model DM_k, where X is the sequence of data vectors. The distortion must be additive, i.e., $D\left(X|DM\right) = \sum_{x_n \in X} d\left(x_n\right)$. A GMM is an example of such a model, with $d\left(x_n\right) = -\log\left(l\left(x_n\right)\right)$, where $l\left(x_n\right)$ is the likelihood of the model given the observation vector x_n. In addition, there is a vector of initial costs, for being in state k at time $n = 1$: $\Pi = \left[\pi_1, \ldots, \pi_K\right]^T$. Our model can be defined as a triplet $\mathcal{M} = \{\{DM_k\}, \mathcal{C}, \Pi\}$. The minimal cost of the data and the states path S, given the model is given in eq. (1):

$$C\left(X, S^* | M\right) = \min_S \left\{ \pi_{s_1} + d_{s_1}\left(x_1\right) + \sum_{n=2}^{N} \left(d_{s_n}\left(x_n\right) + C_{s_n s_{n-1}} \right) \right\} \quad (1)$$

A parameter estimation problem is solved in the Viterbi sense: Given Q observation vector sequences $\mathbf{X} = \{X_q\}_{q=1}^{Q}$, each of length M_q, $X_q = \left\{x_{q1}, \ldots, x_{qM_q}\right\}$ and Q sequences of states $\mathbf{S} = \{S_q\}_{q=1}^{Q}, S_q = \left\{s_{q1}, \ldots, s_{qM_q}\right\}$, find new model parameters $\mathcal{M}$ that will minimize the total cost. First

http://dx.doi.org/10.21437/Interspeech.2016-503

let us find the total cost of the data and state sequences given the model:

$$C(\mathbf{X},\mathbf{S}|\mathcal{M}) = \sum_{q=1}^{Q}\left\{\pi_{qs_1} + d_{qs_1}(x_{q1}) + \sum_{n=2}^{M_q}\left(d_{s_{qn}}(x_{qn}) + C_{s_{qn}s_{q(n-1)}}\right)\right\} =$$
$$= \sum_{q=1}^{Q}\left(\pi_{qs_1} + \sum_{n=2}^{M_q} C_{qs_n q(s_{n-1})}\right) + \sum_{q=1}^{Q}\sum_{n=1}^{M_q} d_{qs_n}(x_{qn}) =$$
$$= C(\mathbf{S}|\mathcal{M}) + D(\mathbf{X}|\mathbf{S},\mathcal{M}) \quad (2)$$

Where $C(\mathbf{S}|\mathcal{M})$ is the total cost of the sequences given the model and $D(\mathbf{X}|\mathbf{S},\mathcal{M})$ is the total distortion of the data given the state sequences and the model.

As one can see, the distortion component and the cost component are disjoint and can be minimized separately. Although the two terms are disjoint, the data associated with each state are highly dependent on the choice of the distortion measure, which means that two different measures will lead to different clustering (partition of the data). Different partitions inherently affect the emission distortion models and the transition costs. Due to the fact that we are able to use different scaling hyper-parameters (fudge factor) in the system and to optimize the parameters together with the scaling hyper-parameter, this approach enables to balance the emission distortions and the transition costs. When only one sequence is available (as it happens in the targeted class of diarization problems), it is not possible to estimate the initial costs. As there is no prior information, all the initial costs are set to zero.

2.1. The scaled log likelihood constraint

In order to estimate the emission model parameters and the transition costs, Viterbi decoding is performed. In this work we apply only the scaled log-likelihood constraint for the transition costs. A more detailed discussion about the constraints can be found at [11].

There is a close relation between HDM with a scaled log-likelihood constraint and the HMM with a "fudge factor" [12].

The scaled log-likelihood criterion is:

$$\sum_{q=1}^{K} e^{-\alpha C_{qk}} = \sum_{q=1}^{K} a_{qk} = 1 \quad (3)$$

When C_{qk} is the cost of transition to be at state q at time n coming from state k at time $n-1$. For $\alpha = 1.0$ this constraint is identical to the one in HMM when the transition probability is defined as $a_{qk} = e^{-C_{qk}}$. In this case, we use a constraint similar to HMM's one but instead of putting only the cost in the exponent, we use a scaled cost (in the case of Viterbi decoding it is equivalent to use a_{qk}^{α} instead of a_{qk}). Thus, the costs become:

$$C_{qk} = -\frac{\ln(a_{qk})}{\alpha} = \frac{1}{\alpha}\ln\left(\frac{\sum_{p=1}^{K} N_{pk}}{N_{qk}}\right) \quad (4)$$

When N_{qk} is the number of times in the path (usually according to the Viterbi decoding) a transition from state k to state q was observed.

3. The diarization system

We apply the HDM approach to two-speaker telephone speaker diarization task. The system used for this experimental evaluation is mostly the same as the one presented in [9] and [11]. Figure 1 presents the system's block diagram. It is composed of a set of pre-processing steps, feature extraction, speech/non-speech detection, and overlapped speech detection, followed by the diarization step.

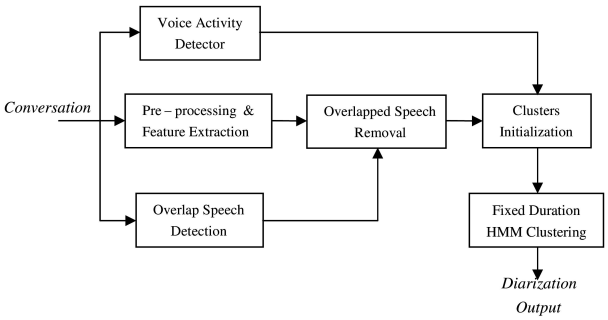

Figure 1: *Speaker diarization system.*

First, classical 12 *mel-frequency cepstral coefficients* (MFCC) are extracted every 20 ms with 50% of overlap. Speech activity detection is performed using a simple energy threshold. Overlapped speech detection is performed as described in [9]. The detected overlapped speech segments are taken out before to apply the diarization process.

The speaker diarization system has three hyper-states for non-speech, speaker *A*, and speaker *B*. A fixed-duration constraint of 20 tied states (200 ms) is used during the first 5 iterations of Viterbi decoding and parameters estimation. Inside the tied states, the transition to other hyper-state is forbidden. It can be viewed as 20 states with shared parameters when only one transition is available. An example of two state HDM with fixed-duration constraint is shown in 2. In order to increase the resolution, only 10 tied states are used for the last iteration (giving a total of 6 iterations). Two kind of state emission models are examined: First *self-organizing map* (SOM) is used (it is an Euclidean distance viewed as a distortion model); the second is the *Gaussian mixture model* (GMM) (the negative log-likelihood is viewed as a distortion model). The non-speech segments provided by the speech activity detector are used to initialize the non-speech state emission model, while the other two models are initialized using weighted segmental K-means [10, 14] which is applied only on the speech segments. As each conversation is processed separately, no initial costs are used.

4. Experiments and results

In this section we present the experiments we conduct on a subset of LDC CALLHOME database [15]. We are using 108 two-speakers conversations of about 30 minutes each from the English/American subset. Only about 10 minutes of speech per conversation are labeled and are used for the diarization task. The average duration of the speech segments is about 2.07 sec.

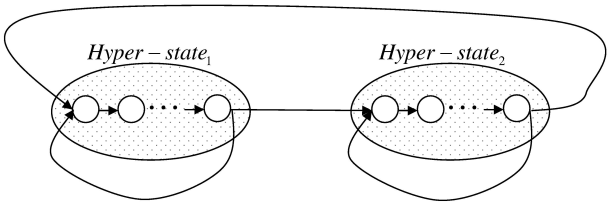

Figure 2: *Two-state fixed-duration HDM system.*

4.1. Evaluation Criterion

Performance is evaluated using the frame-based *diarization error rate* (DER), as defined in [16]. The DER is calculated with a $0.5\,\text{sec}$ window around each changing point (i.e., errors that take place within the $0.25\,\text{sec}$ on either sides of a given changing point are not taken into account).

$$DER = \frac{\sum_{s=1}^{S} dur(s) \cdot \Big(max(N_{Ref}(s), N_{Sys}(s)) - N_{Cor}(s)\Big)}{\sum_{s=1}^{S} dur(s) \cdot N_{Ref}(s)}. \tag{5}$$

Where s is a speech segment and $dur(s)$ is its duration; $N_{Ref}(s)$, the number of assigned speakers; $N_{Sys}(s)$, the number of speakers assigned by the system and $N_{Cor}(s)$, the number of speakers correctly assigned by the system. This equation allows the DER criterion to take into account incorrect classifications as well as classifications with too many or too few detected speakers.

4.2. Experiments with GMM as an emission model

In this subsection we use GMM as emission model for each state. At the beginning $\alpha = 1.0$ is used, i.e., in this case, HDM is equivalent to a standard HMM. Different numbers of mixture components are examined in order to find out the optimal HMM configuration. The DER results are presented in Table 1.

Table 1: *HDM results depending on the number of mixture components for* $\alpha = 1.0$.

#Components	**6**	**10**	**16**	**21**	**24**
DER [%]	26.07	23.84	22.45	21.54	21.35

As can be seen from table 1, 21 and 24 mixture components give approximately the same DER. So, we set the number of mixture components to 21 for the rest of the experiment and we change the value of the scaling hyper-parameter. The results are presented at Table 2.

We can see that when the scaling hyper-parameter is tuned in a proper way, a significant improvement can be achieved (more than 37%). The best achieved results are for $\alpha = 0.05$. The question we asked ourselves is whether the number of mixture components we found at table 1 is also optimal after α optimization. Table 3 presents the results of an experiment where we use the optimal α and vary the model sizes. We observe that the number of mixture components can be dramatically reduced! The best results are obtained with 16 mixture components. With this configuration, we also observe a small DER reduction of about 7% of relative improvement. The results for 10 mixture components are very close to the ones obtained with 16 components emission models. Even for 4 mixture components emission models, the DER is almost equivalent. These results mean that when the scaling hyper-parameter is well tuned the emission models can be significantly simplified.

Table 2: *HDM results for* 21 *mixture components with different values of* α.

α	**0.02**	**0.05**	**0.1**	**0.5**	**1.0**
DER [%]	18.51	13.55	14.48	19.55	21.54

Table 3: *HDM results for* $\alpha = 0.05$ *with different number of mixture components.*

#Components	**4**	**6**	**10**	**16**	**21**
DER [%]	13.78	13.90	12.59	12.80	13.55

4.3. Experiments with SOM as an emission model

In this subsection we present experiments similar than in the previous section but using SOM emission model at each state. SOM can be view as a log-likelihood estimator [17] or just as an Euclidean distance distortion measure. As for GMMs, we start by setting $\alpha = 1.0$ and different SOM sizes (number of neurons or code-words) are examined to determine the optimal configuration. The results in terms of DER are presented in Table 4.

Table 4: *HDM results as the number of SOM size for* $\alpha = 1.0$.

SOM size	6×4	6×6	6×8	6×9	6×10
DER [%]	20.91	18.50	17.79	18.04	17.58

It can be viewed from table 4 that from SOM size of about 6×6 to size of 6×10, the differences in terms of DER are quite small, even if the best results are obtained with the largest models. It is consistent with the conclusions in [10].

As for GMM emission models, we evaluate the SOM performance for different values of the scaling hyper-parameter α. The results are presented at Table 5.

As in the GMM case the optimally tuned scaling hyper-parameter improves significantly the diarization performance (almost 26% of relative improvement). The best achieved results are for $\alpha = 0.2$. Again we aim to figure out whether the codebook size found at table 5 is the optimal size for the optimized value of α. The results are summarized at Table 6. It can be seen that the SOM size can be reduced from 6×10 to 6×6 or even 6×5, i.e., by $40 - 50\%$, from 60 code-words to only 30. For SOM size of 6×4, the DER climbs to 14.56% but still remains relatively low.

Table 5: *HDM results for* 6×10 *SOM size with different values of* α.

α	**0.1**	**0.2**	**0.3**	**0.5**	**1.0**
DER [%]	13.76	13.03	13.38	15.08	17.58

Table 6: *HDM results for* $\alpha = 0.2$ *with different number of code-words.*

SOM size	6×5	6×6	6×7	6×8	6×10
DER [%]	12.93	12.90	12.91	12.97	13.03

5. Conclusions

In this work we presented the HDM and evaluated its performance using GMM and SOM as emission models. The transition cost matrix was defined according to the scaled log-likelihood constraint. It was shown that an appropriate scaling hyper-parameter can reduce dramatically the DER. Furthermore, to tune optimally the scaling hyper-parameter allows to reduce significantly the size of the emission models and such to have much less parameters to estimate. The DER using the simplified models are not degraded but even slightly better compared to large models DER. This result can be explained by the fact that for large DER, the frames associated with each state are not belonging only to one specific speaker. It makes the *pdf* much more complicated and require many parameters (Gaussian components or neurons, code-words) for the emission model to describe it. When the number of errors is small, the frames associated with one state belong mainly to only one speaker. The variability is smaller and the *pdf* could be estimated with a smaller number of parameters, giving a more robust estimation (due to a better number of data / number of parameters to estimate ratio).

It is interesting to understand the meaning of the costs we obtained. With HMM trained via Viterbi statistics, the transition probabilities are $a_{qk} = N_{qk} \Big/ \sum_{p=1}^{K} N_{pk}$, then the transition cost is $C_{qk} = -\ln(a_{qk})$. When using the scaling hyper-parameter the transition cost is also divided by it and becomes $C_{qk} = -\frac{\ln(a_{qk})}{\alpha}$. As the optimal hyper-parameter is $\alpha = 0.05$ and $\alpha = 0.2$ for GMM and for SOM (as emission models) respectively, it means that we increased the transition cost by 20 in the GMM case and by 5 in the SOM case. This fact leads to the conclusion that the transition costs are much more important in order to obtain a good diarization than what can be achieved using a standard HMM. Of course it is possible that other transition costs criteria can be even better (see [11] for additional examples) and also other emission models. The choice of the transition cost criterion and the emission model has to be done depending of the diarization problem to solve and the evaluation criterion to optimized.

As we were able to achieve better performance in terms of DER using much smaller emission models when the scaling factor is optimal, it becomes easier to apply diarization on embedded systems. This is due to the fact that the memory requirements as well as the time for model training and for Viterbi decoding are all approximately linear with the number of mixture components (or the number of code-words).

6. References

[1] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356 –370, feb. 2012.

[2] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059 –1070, dec. 2010.

[3] S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, and J. Glass, "Exploiting intra-conversation variability for speaker diarization," in *Proceedings of Interspeech 2011*, 2011.

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

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

[6] C. Fredouille, S. Bozonnet, and N. W. D. Evans, "The LIA-EURECOM RT'09 Speaker Diarization System," in *RT 2009, NIST Rich Transcription Workshop, May 28-29, 2009*, Melbourne, USA, month = 05,

[7] J. Ajmera, H. B. I. Lapidot, and I. McCowan, "Unknown-multiple speaker clustering using HMM," in *In Proceedings of ICSLP-2002*, 2002, pp. 573–576.

[8] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Interspeech*, Aug. 2013.

[9] 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 2009*, 2009.

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

[11] I. Lapidot and J.-F. Bonastre, "Generalized viterbi-based models for time-series segmentation applied to speaker diarization," in *ODYSSEY 2012 -The Speaker and Language Recognition Workshop*, 2012.

[12] B. Lecouteux, G. Linares, Y. Esteve, and G. Gravier, "Generalized driven decoding for speech recognition system combination," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-april 4 2008, pp. 1549 –1552.

[13] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464 –1480, sep 1990.

[14] O. Ben-Harush, I. Lapidot, and H. Guterman, "Weighted Segmental K-Means Initialization for SOM-Based Speaker Clustering," in *Proceedings of Interspeech 2008*, 2008.

[15] "Linguistic data consortium," LDC97S42, Catalog, 1997, available: http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC97S42.

[16] "Spring 2005 (rt-05s) rich transcription meeting recognition evaluation plan," available: http://www.itl.nist.gov/iad/mig/tests/rt/2005-spring/rt05s-meeting-eval-plan-V1.pdf.

[17] I. Lapidot, "SOM as Likelihood Estimator for Speaker Clustering," in *Proc. Eurospeech'03*. Geneva, Switzerland: ISCA, September 1-4 2003, pp. 3001–3004.