Efficient combination of parametric spaces, models and metrics for speaker diarization[1]

A Maximum Entropy Approach

Themos Stafylakis, Vassilis Katsouros, George Carayannis
Institute for Language and Speech Processing
Athena - Research and Innovation Center in Information, Communication and Knowledge Technologies
Athens, Greece
{themosst, vsk, gcara}@ilsp.gr

Abstract— **In this paper we present a method of combining several acoustic parametric spaces, statistical models and distance metrics in speaker diarization task. Focusing our interest on the post-segmentation part of the problem, we adopt an incremental feature selection and fusion algorithm based on the Maximum Entropy Principle and Iterative Scaling Algorithm that combines several statistical distance measures on speech–chunk pairs. By this approach, we place the merging–of–chunks clustering process into a probabilistic framework. We also propose a decomposition of the input space according to gender, recording conditions and chunk lengths. The algorithm produced highly competitive results compared to GMM-UBM state-of- the-art methods.**

Keywords— Speaker Diarization; Maximum Entropy; Fusion

Topic area— Single- and Multimedia Indexing

I. INTRODUCTION

Speaker diarization (a.k.a. "Who spoke when?" task) tries to solve the problem of automatically extracting speaker metadata of an audio (or multimedia) document, when no set of specific target – speakers for identification is required or any knowledge about the number of participants is given a priory. The speaker – metadata define the estimated time boundaries of each speaker turn as well as the index of active speaker(–s) during them. They may also carry gender and bandwidth information, as well as speech – music – advertisement characterization.

We consider here only the text–independent Broadcast News (BN) "Who spoke when?" task. Thus, only a single channel audio recording is to be processed and no given or estimated transcript interacts with the diarization system. The structure of the baseline diarization systems is a two–stage algorithm. The first stage labels the audio file to speech-silence – advertisement – etc. areas and estimates the boundaries of speaker turns. Several segmentation methods have been proposed in literature, while most of them adopting the technique of a window that slides over the data and calculates statistical distances. Bayesian Information Criterion (BIC) and Kullback – Leibler (KL) divergence are some common distance metrics used [1]. The second stage is normally the blind clustering of speaker chunks into speaker groups, so that the desired one–to–one mapping between reference speakers and estimated speaker–cluster indices is achieved. The agglomerative (bottom-up) hierarchical clustering is the most common approach, though other methods have been proposed too. Briefly, in bottom-up hierarchical clustering each chunk is initially treated as a discrete cluster and the distances between all cluster–pairs are calculated (given a feature space, a statistical model and a distance metric). At the end of each iteration, the cluster–pair having the minimum distance is merged creating a new cluster and the distance matrix is updated. The stopping criterion will be met when all cluster–pair distances are greater that a predefined threshold.

In this paper we focus on the blind clustering stage of a diarization system and investigate an iterative method of selecting and fusing several feature spaces, statistical models and distance metrics. The final goal is to model the posterior probability of whether a pair of speech chunks belongs to different speakers or not, and utilize it as the final measure in the hierarchical clustering stage. The proposed discriminative modeling uses the minor possible assumptions about the true class–conditional distributions of the intermediate metrics and is based on well defined probabilistic framework (Maximum Entropy Principle). By this approach, one can build an incremental log–linear model where both the significant (salient) features and the corresponding weights are automatically determined in an iterative way [2, 3, 4]. We also propose a further duration-based decomposition of the input space, apart from the gender/bandwidth – baseline categorization. The probabilistic distance measure we propose offers the flexibility to incorporate well-known input space decomposition – statistical integration over each space overcoming heuristic rules and adjustable parameters.

The rest of the paper is organized as follows: In chapter II we refer to some state-of-the-art methods dealing with the diarization task. In chapter III we discuss the motivation behind the probabilistic model we propose, while in chapter IV we describe the several front-end features, models and distance metrics we use. In chapter V we analyze the theory

[1] This work is funded by the Greek General Secretariat of Research and Technology under the program PENED-03/251.

and the training algorithm that leads to the Maximum Entropy model we propose. Finally, in Chapter VI the experimental results are shown, followed by the conclusion and future work directions in Chapter VII.

II. Previous work

Many diarization algorithms have been proposed in literature. In [5] the main idea used to confront the speaker clustering task was the statistical modeling of each utterance as a tied-mixture model where the M basis densities (M=128) are estimated from the entire set of speech segments and the weights are estimated for each segment. Advantages of this model are the per-frame likelihoods to the basis densities need to be calculated only once and the weights for merged clusters are computed as a simple averaging of counts. The main drawbacks are the computational cost of on-line training a 128-component Gaussian Mixture Model (GMM) for the entire show. Furthermore, the statistical distance between two speech segments depends on the overall modeling of the show, i.e. it is not an autonomous measure.

In [6], a coupled segmentation-clustering procedure was introduced in order to maximize a global objective function. The latter is decomposed into the overall log-likelihood of the segments to the models as well as the penalty factor, a linear combination of both total number of segments and speakers. Initially, the segmentation process is biased toward over-segmentation and then an iterative segmentation-regrouping process aims to cluster the speech segments. Each segment forms a 8-component diagonal covariance GMM using simple MLE and the Viterbi-based re-segmentation procedure refines the segment boundaries to avoid cutting words. The merging criterion between two GMMs is estimated as the log-likelihood loss for merging the 16 initial Gaussians of both GMMs into a final set of 8 Gaussians. Some drawbacks of this method are the high computational cost (Expectation Maximization algorithm after each merging, lack of closed-form distance measures, etc.) as well as the heuristic complexity penalization rules.

An alternative approach, proposed in [7] was based on state-of-the-art speaker recognition-verification techniques. One Universal Background Model (UBM) with 128 diagonal Gaussians for each of the 4 gender/bandwidth combination was trained. For the initial clustering, the standard Bayesian Information Criterion (BIC) was used, biased towards cluster purity against coverage. Afterwards, for each cluster, maximum a posteriori (MAP) adaptation [8] of the means of the matching Universal Background Model (UBM) is performed. The agglomerative clustering stage was guided by the cross log-likelihood ratio of each pair of segments. The use of Bayesian adaptation as a method to estimate the underlying pdf that generates each cluster is a powerful tool. One can avoid both under (over-) fitting to the data and moreover overcome certain limitations the baseline complexity penalization criteria (BIC, AIC, etc.) suffer from. On the other hand, the method remains computationally heavy, due to the overall GMM-UBM modeling.

III. The benefits from the probabilistic formulation

The main idea behind the algorithm we present is the direct modeling of the posterior probability of whether a pair of speech chunks belongs to different speakers or not. By utilizing this probabilistic distance as the final measure in the hierarchical clustering stage we overcome many heuristic rules and adjustable parameters that are usually posed through the speaker clustering procedure. Due to the statistical formulation of the task, one can create a pool of binary features consisting of arbitrary combinations of front-end features, statistical models, distances and thresholds. The Iterative Scaling algorithm is capable of discovering the most salient features from the pool, resulting in a weighted log-linear modeling of the aforementioned posterior probability.

One major advantage of the probabilistic framework is the well-defined statistical integration over the several input space decompositions that one may choose to apply. One example of such decomposition that we propose is based on the duration of the chunks. Several classes of pair-durations may be defined and trained independently, using the corresponding portion of the available recordings that form the training set. According to the Bayes rule, the evaluation of the overall probabilistic distance of the new examples will then become a trivial statistical integration task (i.e. a combination of the several outputs each duration class produces, weighted by the posterior probabilities of each class given only the pair durations of the new example). By this architecture we overcome the "hard decision" classification of each speech chunk to the predefined categories. The basic architecture is shown in Fig. 1 (only duration – based decomposition is shown).

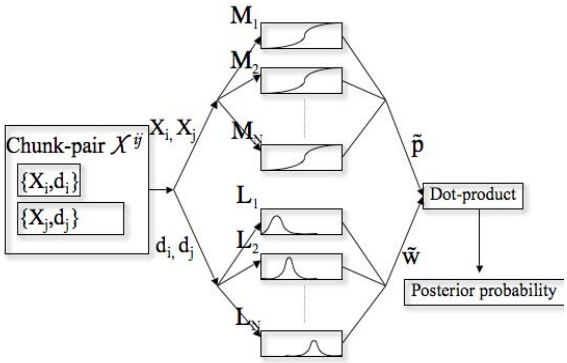

Fig. 1: Architecture of the duration-based decomposition (evaluation stage). M_i correspond to experts that produce their own posteriors, trained with sets of different chunk – pair durations.

IV. Front-end, models and distance metrics

In this chapter we analyze the several front-end features, statistical models that describe the pdf of each speech chunks and metrics/criteria used to form a distance measure between chunks. For the rest of the paper we will refer to the three categories as triplets, that is every valid combination of each category forms a triplet. We emphasize that the

probabilistic model we form is not restricted to the choice of triplets we propose.

A. Front-end spectral parameters

One of the most common parametric spaces used in speech processing is Mel–Frequency Cepstral Coefficients (MFCC), i.e. the DCT of the Mel-frequency warped log–spectrum of a Hamming windowed frame. In our experiments we use 32ms window with 0.5 overlap (16KHz sampling frequency) to form a 24–dimensional MFCC static feature space denoted by $MFCC_s$. We also discard the 9 higher DCT components and appended 1st order time derivatives forming a 26–dimensional feature spaces (static and differential coefficients, 13–dimensional each) denoted by $MFCC_d$ and 2nd order time derivatives forming a 39–dimensional feature space (static, differential and acceleration coefficients, 13–dimensional each) denoted by $MFCC_a$. We do not apply Cepstral Mean Subtraction (CMS) or Relative Spectral (RASTA) filtering to MFCC, since we use the (common in the diarization problem) assumption that the recording conditions for each speaker remains invariant through the entire broadcast.

An alternative feature space we use in our experiment is Line–Spectrum Pairs (LSP) [9]. This is a fully parametric space, derived from Linear Prediction theory. LSP are widely used in sound class discrimination as well as in speaker diarization applications with great success [10]. In our experiments, we use LSP static (18–dimensional) denoted by LSP_s and augmented versions by appending their 1st and 2nd time derivatives, denoted by LSP_d (36–dimensional) and LSP_a (54–dimensional) respectively.

We also refer to other feature spaces (e.g. PLP, RASTA–PLP, MVDR) as well as pitch that one may include in the feature selection algorithm [11].

B. Statictical modeling of speech chunks

A common way of modeling the density of a speech chunk in speaker diarization is a multivariate single Gaussian Model with full covariance matrix of the underline parametric space (denoted by GM_f). Obviously, this cannot model accurately the wide range of phonemes. However, is capable of producing satisfactory results especially for short duration speech chunks. Moreover, a Gaussian Model with diagonal covariance matrix – denoted by GM_d – might be parsimonious, when the durations are even shorter (<5 sec), assuming a feature space having relatively low correlation between its coefficients (MFCC have this property, LSP don't).

A more sophisticated modeling of speech chunks is by fitting a Gaussian Mixture Model (GMM) using either Maximum–Likelihood method (i.e. Expectation Maximization algorithm) or MAP–adaptation techniques [8]. As we already noticed, the main drawback of these models is their high computational demands. However, one may incorporate GMM models, too, in the proposed framework.

In our experiments, we consider only single Gaussian Models, both GM_f and GM_d for $MFCC_k$ and only GM_f for LSP_k, $k \in \{s,d,a\}$.

C. Statistical distance metrics

Several distance metrics have been proposed in the literature. We will consider here only d–dimensional GM_f and GM_d. One of the most extensively used metrics is ΔBIC [1] defined as

$$\Delta\mathrm{BIC}_{ij} = \mathrm{GLR}_{ij} - \mathrm{PF}_{ij}, \quad (1)$$

which is decomposed as follows,

$$\mathrm{GLR}_{ij} = \frac{m_i}{2}\log\left|\Sigma_i\right| + \frac{m_j}{2}\log\left|\Sigma_j\right| - \frac{m_{i\cup j}}{2}\log\left|\Sigma_{i\cup j}\right| \quad (2)$$

and

$$\mathrm{PF}_{ij} = \frac{1}{2}\theta n_p \log(m_{i\cup j}). \quad (3)$$

In our notation, GLR denotes generalized–likelihood ratio, PF penalty factor, i (j), m_i (m_j) and Σ_i (Σ_j) the i–th (j–th) chunk, its length and its covariance matrix respectively, $|\cdot|$ the determinant, $i \cup j$ the union of i–th and j–th chunks, n_p the number of free parameters in the model, i.e.

$$n_p = \begin{cases} d + d(d+1)/2, & \text{for } GM_f \\ 2d, & \text{for } GM_d \end{cases}, \quad (4)$$

and θ the penalization coefficient, usually in the range from 1 to 7.

It should be noted that ΔBIC takes values in $\mathbb{R}$ rather that in $\mathbb{R}^+$. GLR is a distance term, while PF penalizes the complexity of the model. Thus, ΔBIC is treated differently when we apply thresholds (see Section A in Chapter V).

We consider also the Kullback-Leibler (KL) divergence, defined by

$$\mathrm{KL}_{ij} = \frac{1}{2}\left(\log\left(\frac{\left|\Sigma_j\right|}{\left|\Sigma_i\right|}\right) + tr\left(\Sigma_i\Sigma_j^{-1}\right) + \left(\mu_i - \mu_j\right)^T \Sigma_j^{-1}\left(\mu_i - \mu_j\right) - d \right), \quad (5)$$

where μ_i and μ_j denote the means of the i-th and j-th chunk respectively (column vectors), $tr(\cdot)$ the trace and the rest as above. KL is not a real distance since it is asymmetric. However, we can symmetrize it by using either the arithmetic or the harmonic mean. The symmetrized distances are denoted by KL_a and KL_h respectively [12].

Another statistical distance is the Arithmetic–Harmonic Sphericity (AHS), defined by

$$\mathrm{AHS}_{ij} = \log\left(tr\left(\mathrm{C}_j\mathrm{C}_i^{-1}\right) tr\left(\mathrm{C}_i\mathrm{C}_j^{-1}\right)\right) - 2\log(d). \quad (6)$$

In diarization task, one may use correlation matrices C_i and C_j, instead of covariance matrices, in order to keep information about the means, following the invariant recording conditions assumption as mentioned above [13].

We finally refer to T_2 – Hotelling distance, defined by

$$\mathrm{H}_{ij} = \frac{m_i m_j}{m_i + m_j}\left(\mu_i - \mu_j\right)^T \Sigma_{i\cup j}^{-1}\left(\mu_i - \mu_j\right). \quad (7)$$

T_2 – Hotelling distances are useful in cases where chunks are too short to estimate reliably covariance matrices and use them in discrimination.

In our experiments, we consider ΔBIC, KL_a, KL_h and AHS for all parametric spaces and models and H only for $MFCC_s$ and LSP_s, i.e. static parameters, since T_2 –Hotelling distances are zero for distributions having identical means,

as in the case when we consider differential and acceleration parameters.

Therefore, the multiple metrizable parametric spaces are defined as a triplet $t \in \mathcal{T}$, $\mathcal{T} \equiv (\mathcal{S}, \mathfrak{M}, \mathcal{D})$ of all the valid combinations of parametric spaces, models and distance metrics. For example, t = ($MFCC_s$, GM_f, KL_h) refers to the triplet where the parametric space results from static MFCC, modeled as a multivariate single Gaussian with full covariance while the distance metric is the KL diversity symmetrized by the harmonic mean. In our experiments, we have used $|\mathcal{T}|$ = 39 such triplets as explained above.

V. The Proposed Algorithm

In order to train the exponential model used in the Maximum Entropy (ME) algorithm, we need a labeled training set $\tilde{D} = \{(x_j, b_j)\}$, $j = 1, \ldots, N$ (or multiple sets $\tilde{D}$ for each category of gender/bandwidth/chunk lengths) as well as a set of triplets $\mathcal{T}$. The variable x corresponds to a pair of raw audio files (the front-end feature space is defined by the triplet) while the binary random variable b takes the value of 0 when the speech chunks belong to the same speaker and 1 otherwise.

A. Binarization of the triplets

Our approach to the problem of fusing triplets is based on the statistical framework of feature selection for random fields. We choose to binarize the set $\mathcal{T}$ using multiple thresholds $\mathcal{U}_t^k$ on the triplets, so that we obtain a set of binary valued features:

$$f_t^k(x,b) \in \{0,1\},\ k \in \{1,2,\ldots,K\},\ b \in \{0,1\},\ t \in \mathcal{T}, \quad (8)$$

where K is total number of thresholds for each triplet.

Each binary feature indicates whether the prediction of the class that the pair x belongs to is correct or wrong:

$$f_t^k(x,b) = \mathbf{1}_{\{g_t^k = b\}} \in \{0,1\}, \quad (9)$$

where $g_t^k(x) \in \{0,1\}$ is the corresponding predictor of the class of x and the notation $\mathbf{1}_{\{\cdot\}} \in \{0,1\}$, stands for the indication function.

In our experiments, the number of thresholds for each t is K=9, resulting in L = 351 total binary features. The multiple thresholds (apart from the triplets having ΔBIC as metric) are determined as follows.

For each triplet, we calculate the conditional mean and variance of the two classes assuming that they follow normal distributions. To increase gaussianity we transform the distances to the log – domain. The 5th threshold (i.e. the middle one) is equal to the equiprobable point. The remaining thresholds are determined such that the 1st and 9th thresholds are equal to the means of the "same speaker" class (b=0) and "different speaker" class (b=1) respectively. We use linearly spaced thresholds for each of the 2 directions (see Fig. 2).

Notice that the threshold determination is the only step we use generative statistics for the class-conditional pdf of each triplet. For the rest of the analysis the goal is to estimate discriminative statistics, i.e. to directly model the posterior probabilities of each class. For the triplets having ΔBIC as distance metric we use discrete values of θ in the range 1 to 5.

Thus, the binary valued features are given by:

$$f_t^k(x,1) = \begin{cases} 1, \text{ if } t(x) > \mathcal{U}_t^k \\ 0, \text{ otherwise} \end{cases} \quad (10)$$

and

$$f_t^k(x,0) = \overline{f_t^k(x,1)}. \quad (11)$$

For simplicity of notation, we form a unique index $i = 1,2,\ldots,L$ to refer to all the binary features independently from the triplet and the threshold indices. The set of candidate features will be denoted as $F = \{f_i\}$, $i = 1,2,\ldots,L$.

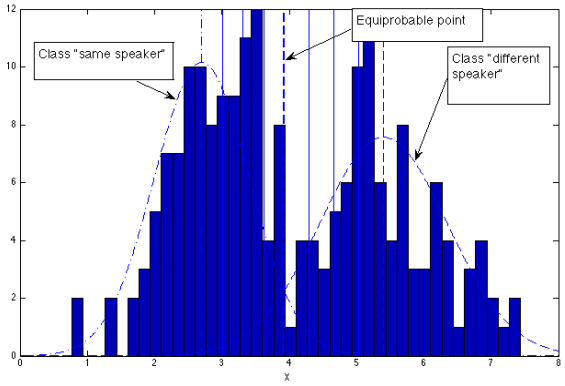

Fig. 2: The determination of the thresholds for a triplet. The histogram and the pdf of each class approximated with a single gaussian. The middle threshold corresponds to the equiprobable point.

B. Feature selection and training for Exponential Models

The algorithm we describe here is based on ME principle [2, 3, 4]. The objective is to learn the optimal exponential weights and the corresponding salient features from the training data by using the following linear exponential family of distributions

$$q_\lambda(b \mid x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_i \lambda_i f_i(x,b) \right), \quad (12)$$

where $Z_\lambda(x)$ is a normalizing factor, such that

$$q_\lambda(0 \mid x) + q_\lambda(1 \mid x) = 1. \quad (13)$$

The parameters are estimated by minimizing KL divergence between the estimated model q_λ and the empirical distribution

$$\tilde{p}(x,b) = \begin{cases} \frac{1}{|\tilde{D}|}, \text{ if } (x,b) \in \tilde{D} \\ 0, \text{ otherwise} \end{cases}. \quad (14)$$

The λ_i parameters correspond to the Lagrange coefficients of the entropy maximization problem, that is, "maximize the entropy of the model given the following constrains":

$$\sum_{x,b} \tilde{p}(b,x) f_i(x,b) = \sum_{x,b} \tilde{p}(x) q_\lambda(b \mid x) f_i(x,b),\ \forall f_i \in F. \quad (15)$$

The maximization with constrains including only 1st order statistics leads to the log-linear model, while the estimation of the Lagrange coefficients (i.e. the training of the algorithm) to the maximization of the likelihood function of the model with respect to the data set. From (15) one may notice that we constrain the expected values of features f_i with respect to the model to be equal to the expected values with respect to the empirical distribution.

The maximization problem is concave with respect to the λ_i parameters. However, no closed – form solution can be reached. Furthermore, the pool of candidate features might be too large to handle. Thus, an iterative scaling algorithm is adopted, which is capable of both extracting the most salient features while training the current model to reach the optimal λ_i too. Each iteration may be decomposed into two steps. During the feature induction step, we select the feature that improves the current model the most, as follows:

$$f^*_{iter} = \arg\max_{f_{iter} \in C_{iter}} \left\{ \sup_{\alpha} \left\{ D\left(\tilde{p} \,\|\, q\right) - D\left(\tilde{p} \,\|\, q_{a, f_{iter}}\right) \right\} \right\}. \quad (16)$$

In the above notation, the subscript "$_{iter}$" denotes the iteration count and C_{iter} the candidate pool of the remaining features. The model

$$q_{a, f_{iter}} = \frac{q_\lambda(b \mid x) \exp\left(a \cdot f_{iter}\right)}{Z_a(x)}, \quad (17)$$

is augmented by the f_{iter} feature. The optimal α is calculated using Newton Method. We emphasize here that the function is concave with respect to α, thus Newton Method will attain the global maximum [2].

During the second step we estimate the weights λ_i that minimize the KL divergence defined by:

$$D\left(\tilde{p} \,\|\, q_\lambda\right) = \sum_x \tilde{p}(x) \sum_{b \in \{0,1\}} \tilde{p}(b \mid x) \log \frac{\tilde{p}(b \mid x)}{q_\lambda(b \mid x)}. \quad (18)$$

By decomposing the logarithm into numerator and denominator and observing that first term contains only the empirical distribution, the minimization of (18) is equivalent to the maximization of the log–likelihood of the estimated model to the training set:

$$L_{\tilde{p}}\left(q_\lambda\right) = \sum_x \sum_{b \in \{0,1\}} \tilde{p}(x,b) \log q_\lambda(b \mid x). \quad (19)$$

The calculation of λ_i can be done iteratively using the update rule

$$\lambda_i^{'} = \lambda_i + \Delta\lambda_i, \quad (20)$$

where

$$\Delta\lambda_i \propto \log\left(\frac{\sum_{x,b} \tilde{p}(b,x) f_i\ (x,b)}{\sum_{x,b} \tilde{p}(x) q_\lambda(b \mid x) f_i\ (x,b)} \right). \quad (21)$$

Again, the global minimum will reached since the function is convex with respect to the coefficients vector [2]. The update formula (21) is called Generalized Iterative Scaling (GIS). For a comparison between several update formulas we refer to [14].

The algorithm terminates either when the number of desired active features is reached, or when the KL reduction drops below a predefined threshold (i.e. no further significant information can be gained from the feature pool).

When modeling posterior probabilities, one should carefully determine the priors of each class. Since the cost of a false merging is much higher that a missed one (the later can be smoothed in hierarchical clustering), one may use $N_0 << N_1$. An alternative approach is to use $N_0 = N_1 = N/2$, and adjust the threshold of merging in hierarchical clustering to be less than 0.5. In our algorithm we used the second approach, because it offers the flexibility to decouple the prior probabilities from the posteriors.

C. The input space decomposition

Apart from the baseline gender/bandwidth categorization, we further decompose the input space according to the duration of the chunk pairs. The main idea behind this decomposition is to train different models (experts) for each duration category and integrate over the models to obtain the final posterior probability for each new example x. We form 5 chunk duration categories, thus N_d=15 (i.e. 5(5+1)/2) chunk pair categories. The duration statistics for each category are calculated in the log – domain. The mean values of each class are 3.0, 5.1, 8.7, 14.7 and 25.6 seconds. In order to create the training set for each category we first form the desired means and variances for each class (assuming normal class conditional distributions in log – domain) and then we force each set to have the desired duration statistics by chopping the audio file and obtaining the desired length. Each category has the same number of training examples.

The evaluation of the posterior probability is achieved by firstly evaluating each of the N_d experts' outputs and then integrating over them. The weights we use at the integration are the posterior probability that the pair belongs to the i-th duration class, given the pair durations:

$$q_\Lambda\left(b \mid x\right) = \sum_{i=1}^{N_d} p\left(c_i^d \mid d_x, \Theta^d\right) q_{\lambda_i}\left(b \mid x\right), \quad (22)$$

$$p\left(c_i^d \mid d_x, \Theta^d\right) = \frac{p\left(c_i^d\right) p\left(d_x \mid \theta_i^d\right)}{\sum_{j=1}^{N_d} p\left(c_j^d\right) p\left(d_x \mid \theta_j^d\right)}, \quad (23)$$

$$p\left(c_i^d\right) = \frac{1}{N_d},\ i \in \left\{1, 2, \ldots, N_d\right\} \quad (24)$$

In the above notation, $\Theta^d = \left\{\theta_i^d\right\},\ i \in \left\{1, 2, \ldots, N_b\right\}$ corresponds to the duration parameters of the N_d experts and c_i^d to the i-th duration class.

VI. Experiments

In order to train the models we used a subset of the WSJCAM0 British English Speech Corpus for Large Vocabulary Speech Recognition. The training set consists of 39 female and 53 male speakers. Each speaker read about 90 training sentences. We divide it into two subsets, 21 female and 26 male for training and the rest to form the evaluation set. We trained each of the N_d=15 duration categories separately for female and male genders. The average classification error is shown in Fig. 3. We compared them with a baseline 8-component GMM classifier ($MFCC_d$) as well as with the binary feature that produced best results in training set for each duration category. Keep in mind that

classification means only to identify whether a pair of chunks belongs to the same speaker or not.

For optical purposed we created a unique index to number the pair-duration categories. The numbering increases as follows: {(1,1), (2,1), (3,1), (4,1), (5,1), (2,2), (3,2), (4,2), (5,2), (3,3), (4,3), (5,3), (4,4), (5,4), (5,5)}, where each entry corresponds to the single chunk duration category explained in Chapter V.

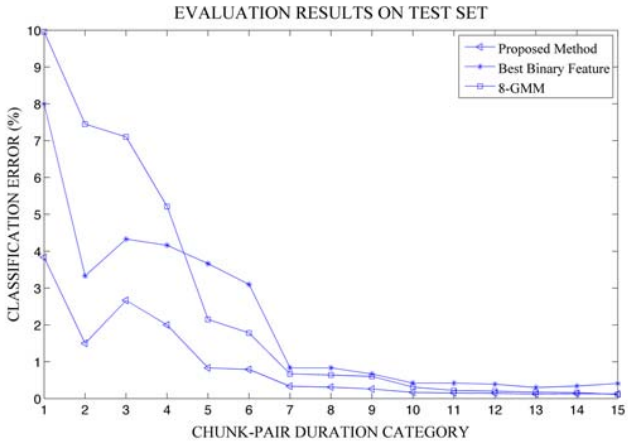

Fig. 3: Comparison of the classification error by the proposed algorithm. The results are averaged between female and male test sets.

To evaluate the proposed method in the diarization task we used the English Broadcast News RT–02 Rich Transcription corpus. The corpus consists of 6 shows of one hour duration each. A fraction of 30 minutes per show is used for evaluation of the diarization algorithm. The training set for the telephone bandwidth category was collected from the 2002 NIST Speaker Recognition Evaluation Corpus. We used 24 female and 31 male speakers. The merging threshold was set to 0.41 (i.e. we merge speech chunks i and j iff $p(1| X^{ij}) < 0.41$). We also evaluated it using three other methods: The most significant triplet (denoted by *MST*) derived from the algorithm, an 8-component, diagonal covariance GMM with $MFCC_d$ and finally a 128-component GMM (denoted by GMM_a), MAP-like adapted from the corresponding UBM (denoted by GMM_b). Only the means were adapted as in [8]. The total speaker error for all the 4 methods is shown below:

TABLE I.TOTAL SPEAKER ERROR (%) ON ENGLISH BROADCAST NEWS RT-02 RICH TRANSCRIPTION CORPUS

Show	*MNB*	*PRI*	*NBC*	*CNN*	*VOA*	*ABC*	*Total*
ME	3.61	4.42	10.27	9.61	2.44	15.78	7.69
MST	7.30	11.8	15.95	17.44	6.35	24.91	14.13
GMM_a	6.13	13.4	15.19	16.95	9.41	21.70	13.80
GMM_b	1.73	3.51	11.56	7.13	2.78	11.24	6.32

Only speaker error is considered, i.e. the portion of the speech that is mapped to a different from the reference speaker, after the optimal one-to-one mapping is performed. Areas of overlapping speakers are excluded from the evaluation. Moreover, perfect segmentation has been applied to all shows to focus on the speaker clustering task.

VII. CONCLUSION AND FUTURE WORK

In this paper we proposed a framework of inducing features in speaker diarization problem. By this approach, one may obtain highly competitive results compared to more complex models using only single Gaussian models. Moreover, the decomposition of the input space according to the duration of speech chunk pairs offers the flexibility of training several models independently and increasing the overall system performance.

As a future work, we propose the incorporation of prior distribution of the λ_i, in order to control the coefficients range of values and avoid overfitting. An exponential distribution would be a natural choice for the particular problem we face, since all λ_i should be non-negative. Moreover, the optimization function remains convex with respect to λ_i. The method can be extended to the segmentation step, in a straightforward way.

REFERENCES

[1] S.S. Chen and P.S. Gopalakrishnan, "Clustering via the Bayesian Information Criterion with Applications in Speech Recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seattle, WA*, May 1998, vol. 2, pp. 645–648.

[2] S. D. Pietra, V. D. Pietra, and J. Lafferty, "Inducing features of random fields," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 380-393, April 1997.

[3] A. Berger, S. D. Pietra, and V. Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, (22-1), March 1996.

[4] W. H.-M. Hsu and S.-F. Chang, "A statistical framework for fusing mid-level perceptual features in news story segmentation," in *IEEE International Conference on Multimedia and Expo*, 2003.

[5] S. E. Tranter, M. J. F. Gales, R. Sinha, S. Umesh, P. C. Woodland, "The development of the Cambridge University RT-04 diarization system," Cambridge University Engineering Department, Trumpington Street, Cambridge.

[6] S. E. Tranter and D. A. Reynolds, "Speaker Diarisation for Broadcast News," in *Proc. Odyssey Speaker and Language Recognition Workshop*, June 2004, pp. 337–344.

[7] Barras, C., Zhu, X., Meignier, S. and Gauvain, J.-L.: 2004, "Improving speaker diarization," *Fall 2004 Rich Transcription Workshop* (RT04), Palisades, NY.

[8] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[9] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Amer*. 57 (Suppl. 1) (1975) 35.

[10] L. Lu, H. Jiang, and H. J. Zhang, "A robust audio classification and segmentation method," in *Proc. ACM 2001, Ottawa, ON, Canada*, 2001, pp. 203–211.

[11] Gu, L. and Rose, K., "Perceptual Harmonic Cepstral Coefficients as the Front-end for Speech Recognition," in *Proc. ICSLP'00*.

[12] D. H. Johnson and S. Sinanovic, "Symmetrizing the Kullback-Leibler distance," *Technical report, Rice University*, 2001.

[13] F. Bimbot and L. Mathan, "Text-Free Speaker Recognition using an Arithmetic Harmonic Sphericity Measure," in *Proc. European Conference on Speech Communication and Technology (Eurospeech), Berlin, Germany*, September 1993, vol. 1, pp. 169–172.

[14] Malouf, R. "A comparison of algorithms for maximum entropy parameter estimation," in *Proceedings of CoNLL*, 2002.