Feature Selection and Model Optimization for Semi-supervised Speaker Spotting

Srikanth Raj Chetupalli[1], Anand Gopalakrishnan[2], and Thippur V. Sreenivas[1]

[1]Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore
[2]Department of Electrical Electronics Engineering, NIT Karnataka, Surathkal

***Abstract*—We explore, experimentally, feature selection and optimization of stochastic model parameters for the problem of speaker spotting. Based on an initially identified segment of speech of a speaker, an iterative model refinement method is developed along with a latent variable mixture model so that segments of the same speaker are identified in a long speech record. It is found that a GMM with moderate number of mixtures is better suited for the task than a large number mixture model as used in speaker identification. Similarly, a PCA based low-dimensional projection of MFCC based feature vector provides better performance. We show that about 6 seconds of initially identified speaker data is sufficient to achieve $> 90\%$ performance of speaker segment identification.**

***Index Terms*—Speaker spotting, Speaker verification, Speaker diarization, Gaussian mixture model (GMM), Mel-Frequency Cepstral Coefficients (MFCCs).**

I. Introduction

We consider the "speaker spotting" problem, in which the goal is to track a given speaker S, i.e., "when did S speak?", in a recorded speech. We consider the specific situation, where the target speaker S is identified manually by a segment in the speech recording by the user, and all other segments of the same speaker have to be identified. This is a semi-supervised approach, because the target speaker model has to be constructed and then used for further spotting. This problem is of interest in indexing speech files or retrieving the portions of speech spoken by a speaker in a conversation. Since the approach requires the user to select a segment for training, the goal is to minimize the effort of the user, which translates to use of limited duration training of target model.

A supervised counterpart to the problem considered is the speaker verification problem [1], [2], such as in forensic applications. In verification, speaker model is estimated from training examples, and the goal is to verify the speaker for each segment in the conversation, and retrieve the speaker segments. In this case, we need to worry about mismatch between the channel and other ambient noise conditions between training and test data. However, in the semi-supervised approach considered here, since the training segment is a part of the same conversation that is to be indexed, the effect of channel and ambient noise mismatch does not arise.

Speaker diarization [3], [4], [5] can be considered as the unsupervised counterpart to the semi-supervised speaker spotting problem. In speaker diarization, the goal is to detect the number of speakers in the conversation, and index the recording according to speaker identities. This is typically achieved by first detecting the speaker change points to identify homogeneous speaker segments, which is followed by clustering of segments corresponding to individual speakers. Model based approaches [6] to speaker diarization estimate the speaker models based on segments obtained after change point detection, and then index the conversation using the estimated speaker models. This problem is different from semi-supervised speaker spotting, because the training segment is identified by the change point detection scheme instead of a user; also multiple target speakers are of interest.

In this paper, we explore semi-supervised speaker spotting in conversations lasting a few minutes. We consider the GMM-UBM (Universal Background Model) approach, in which, the background model is estimated using the entire speech recording, while the speaker model is obtained by adaptation of the background model to the target speaker segment. We propose use of latent variable formulation [7], [8], which allows for the estimation of speaker presence probability. This is a soft-decision approach compared to the hard-decision of a likelihood ratio test [9]. Since the goal is to minimize the amount of training data required, it is necessary to optimize with respect to the variables such as number of mixtures in the model, and dimensionality of the feature vectors. Increase in these variables increases the parameters in the model and hence the amount of training data required. We show that the accuracy of speaker spotting is poor with more number of mixtures. Further, the accuracy decreases with increase in feature dimensions when going from the traditional MFCCs [10] to MFCC+Δ features. We show that reduction of the dimensionality of MFCC+$\Delta\Delta\Delta$ features using PCA (principal component analysis) [11] helps in improving the performance of speaker spotting.

II. Speaker Spotting

Fig. 1, gives an overview of the proposed approach to semi-supervised speaker spotting problem. Given a speech recording and segmental data identified as the target speaker; first we estimate a stochastic model using the feature vectors estimated from the entire speech recording. This model is then adapted to the target speaker segment using maximum a-posteriori (MAP) formulation. The estimated models are then used for spotting the target speaker using a latent variable approach. This formulation provides posterior probability of the target speaker

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

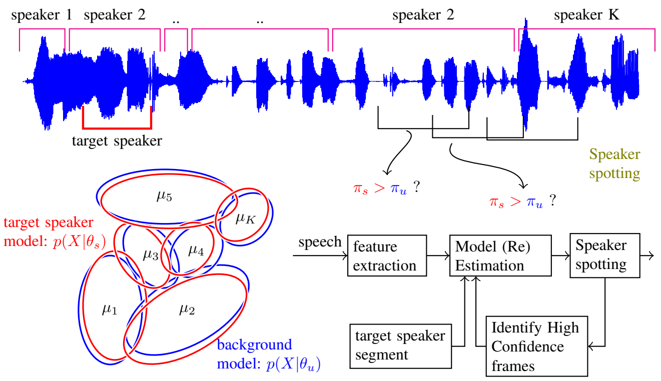


Fig. 1. Overview of the proposed approach to speaker spotting.

for each segment; which is then used to select high confidence frames (of posterior probability > 0.9). The selected segments are then grouped with the initial training segment to improve upon the speaker model. The updated speaker model is then used for speaker spotting in an iterative manner. The individual blocks of the proposed approach are discussed below.

A. Speaker spotting using latent variable model

Consider the model for the generation of a speech conversation as shown in Fig. 2. The feature vector $\mathbf{X}_n$ at frame index n is modeled as coming from one of the two states (indexed by z_n): (i) target speaker, or (ii) non-target speaker (referred to as background), and silence. Assuming a normal quality recording of $> 30\ dB$ (signal to ambient noise ratio), we can easily suppress the silence frames from any modeling consideration by thresholding short-time energy of the signal. Hence, silence is not taken as a separate state in the conversation model. The probability density function (pdf) of the observation is modeled as a mixture of the observation densities for the two states of conversation, using the latent variable approach as

$$\mathbb{P}(\mathbf{X}|\{\boldsymbol{\theta}_s, \boldsymbol{\theta}_u, \pi\}) = \pi_s \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}_s) + \pi_u \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}_u), \qquad (1)$$
$$\text{and } \pi_s + \pi_u = 1,$$

where $\mathbb{P}(\mathbf{X}|\boldsymbol{\theta}_s)$ is the pdf (parameterized by $\boldsymbol{\theta}_s$) of the target speaker, and $\mathbb{P}(\mathbf{X}|\boldsymbol{\theta}_u)$ is the pdf of the background. The variables π_s, and π_u denote the mixture weights of the constituent densities as latent variables. We formulate the speaker activity detection as that of estimating π_s, and π_u given a set of observation vectors and the pdfs of the target speaker and background. The estimates for π_s, and π_u can be interpreted as the a-posteriori probabilities of the target or non-target speaker presence.

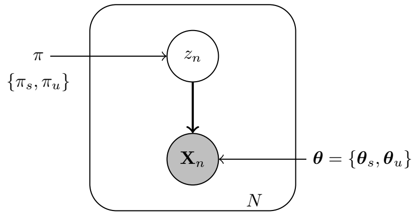

π
$\{\pi_s, \pi_u\}$
z_n
$\mathbf{X}_n$
$\boldsymbol{\theta} = \{\boldsymbol{\theta}_s, \boldsymbol{\theta}_u\}$
N

Fig. 2. A generative model for speech conversation.

Let $\boldsymbol{\mathcal{X}} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$ be a segment of N consecutive feature vectors. Given $\mathbb{P}(\mathbf{X}|\boldsymbol{\theta}_s)$, $\mathbb{P}(\mathbf{X}|\boldsymbol{\theta}_u)$, and $\boldsymbol{\mathcal{X}}$, we consider the following optimization problem,

$$\underset{\pi_s, \pi_u}{\text{maximize}}\ \mathbb{P}(\boldsymbol{\mathcal{X}}) \equiv \prod_{n=1}^{N} \left[\pi_s \mathbb{P}(\mathbf{X}_n|\boldsymbol{\theta}_s) + \pi_u \mathbb{P}(\mathbf{X}_n|\boldsymbol{\theta}_u)\right], \qquad (2)$$
$$\text{subject to } \pi_s + \pi_u = 1.$$

Here, we assume that the feature vectors are independent and identically distributed. For the case of $N = 1$, solving the above optimization problem leads to the solution,

$$\pi_s = \begin{cases} 1, & \text{if } \mathbb{P}(\mathbf{X}_n|\boldsymbol{\theta}_s) > \mathbb{P}(\mathbf{X}_n|\boldsymbol{\theta}_u) \\ 0, & \text{otherwise,} \end{cases} \qquad (3)$$

which is the same as the criterion for likelihood ratio test, assuming equal priors for the two states of the recording. For $N > 1$, the solution needs to be estimated using an iterative scheme such as the Expectation-Maximization (EM) algorithm [12], shown in Algorithm 1. The solution for π_s, can be interpreted as the probability of the target speaker state given the observations. This is a useful measure for segments containing overlapped speakers which is a common case in real recordings [4], since it gives the probability with which the observations are explained by the target speaker model [8]. π_s also gives the expected number of frames for which the likelihood under target speaker state is higher than the likelihood under the background state. The value of π_s can be thresholded to make a decision on the presence of target speaker. Note that, the solution obtained after thresholding using this approach may be the same as that of the likelihood ratio test with equal prior assumption, but the value of π_s gives a measure of the confidence on the decision, which can be used for refining the speaker model.

Algorithm 1 Speaker Spotting Algorithm

```
Initialize $\pi_s^0 = \pi_u^0 = \frac{1}{2}$, $i = 0$.
Compute the likelihoods $\mathbb{P}(\mathbf{X}_n|\boldsymbol{\theta}_s)$, and $\mathbb{P}(\mathbf{X}_n|\boldsymbol{\theta}_u)$, $\forall n$.
repeat
    $i \leftarrow i + 1$
    Compute:
        $\gamma(z_n, s) = \frac{\pi_s^{i-1}\mathbb{P}(\mathbf{X}_n|\boldsymbol{\theta}_s)}{\pi_s^{i-1}\mathbb{P}(\mathbf{X}_n|\boldsymbol{\theta}_s) + \pi_u^{i-1}\mathbb{P}(\mathbf{X}_n|\boldsymbol{\theta}_u)}$.
    Update:
        $\pi_s^i = \frac{1}{N}\sum_{n=1}^{N}\gamma(z_n, s)$, and $\pi_u^i = 1 - \pi_s^i$.
until Convergence, i.e., $|\pi_s^i - \pi_s^{i-1}| \leq 0.01$.
```

B. Target speaker and Background models

The formulation above assumes the knowledge of the pdfs of the target speaker and the background. We used GMMs to represent the pdfs. Background model is trained first using the feature vectors of the entire speech recording. The parameters of the model are estimated with maximum-likelihood criterion using the EM algorithm. Now, the segment marked as target speaker is used to estimate the speaker model through adaptation, i.e., through the MAP adaptation technique [2], [13]. The idea here is to have the background model capture speaker-independent distribution of the feature vectors, while the adaptation personalizes the model to the target speaker. Through this the difference between the speakers is captured better by the model differences. Also, the training data for the background model and for the target speaker adaptation are taken from the same speech recording.

C. Iterative refinement

Performance of the proposed approach depends very much on the target speaker model adapted from the background model, which is data dependent. The accuracy and reliability with which it represents the speaker is directly related to the number of "reliable" feature vectors from initial training segment, based on which the model is adapted. The performance can be improved by updating the model with high-confidence segments obtained after speaker spotting and this can be iterated along with speaker spotting until convergence, i.e., negligible change in the estimated speaker segments. We define the high-confidence segments as those segments where the posterior speaker probability (π_s) obtained through the speaker spotting algorithm, to be greater than 0.9. These segments of feature vectors are appended to the initial training segment and speaker model is updated using the MAP adaptation similar to the previous section. The updated model is then used to refine the speaker positions in the speech recording.

D. Feature Extraction

MFCCs, which describe the perceptually relevant timbre features, are the most widely used feature vectors for speech recognition tasks. The MFCC vectors are often augmented with their first and second derivatives to account for the temporal dynamics of the speech signal. The addition of the derivative vectors increases the dimensionality of the total feature vectors, which in turn increases the parameters of the underlying distribution. For a given number of feature vectors, increasing the feature vector dimension results in poor modeling of the underlying distribution (curse of dimensionality). To overcome this problem, we consider dimensionality reduction of the MFCC vectors appended with the derivative features. The first three derivatives of the 13-dimensional MFCC vectors are appended with MFCCs to create a 51 $(12 + 3 \times 13)$-dimensional feature vector (by omitting energy component of only the MFCCs and including the energies of all the derivatives). PCA is then used to compute the 12 largest principal components of the 51-dimensional feature vector, which is then used as the feature vector for modeling. Feature vectors from the entire speech recording are used for estimating the principal components. Feature vectors are computed every 10 ms using a Hann window of duration 25 msec.

III. Experiments and Results

We first study the performance of the proposed algorithm using a synthetic (concatenated) database with known ground truth. We used GMMs with diagonal covariance matrices for the background and target speaker models. Only mean adaptation is performed for the estimation of speaker model. Evaluation is performed on segments of duration 2 sec (200 vectors) with 400 msec (40 vectors) overlap between successive segments. Fig. 3 illustrates a typical result of the speaker spotting algorithm using a recording of $K = 3$ speakers. A 5 sec sub-segment from the first segment (shown in red) is chosen as the training data. We observe that, (i) estimated speaker posterior probability π_s is > 0.9 for the training segment, (ii) for non-target speaker segments (blue and cyan), π_s is consistently smaller except for a few segments, (iii) for target speaker segments, π_s is more than 0.5 for most of the segments except for the segment around 100 sec (false negative), (iv) at the transition segments from target to non-target speaker, π_s value decreases gradually due to the overlapped evaluation of the segments, and increases gradually at the transitions from non-target to target speaker, and (v) for the segment (after 20 sec) with overlapped speakers, we see that π_s is significant during the overlapped portions, this can be used to identify target speaker in such segments also.

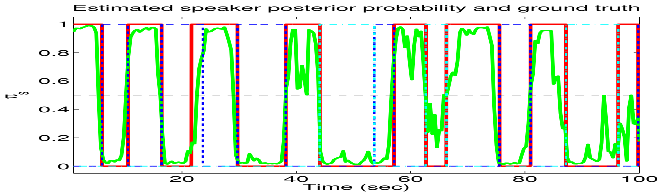


Fig. 3. Estimated speaker posterior (green) along with the ground truth for target speaker (red), and the other two speakers (blue and cyan).

A. Concatenated Records

We generated a conversation example using the speech recordings from Starkey database [14]. The database consists of recordings from 8 male speakers and 8 female speakers reading a single passage. The recordings are down-sampled to 8 KHz sampling rate from the original F_s of 44.1 KHz. A conversation is generated by concatenating speech segments taken from different speakers. For a conversation with K speakers, we picked K speakers randomly from the total of 16; and for each segment of the conversation, one of the selected K speakers is randomly chosen, and a segment is taken from a random position in the recording, corresponding to that speaker. Length of each segment is random with a uniform distribution between 3 and 10 sec. Number of segments in each conversation is chosen to be 40 (so a typical record generated is of length 120 to 400 sec). Experiments are carried out on 100 test conversations each, for different number of speakers K in the record.

B. Performance Measures

The performance of speaker spotting is obtained as the fraction of true positives, false positives, and false negatives, as defined below (in set notation):

$$\text{True Positive (TP)} = \frac{|A \cap B|}{|B|}$$

$$\text{False Positive (FP)} = \frac{|A \cap B^c|}{|B^c|}$$

$$\text{False Negative (FN)} = \frac{|A^c \cap B|}{|B|}$$

where, $|.|$ is the cardinality of the set argument, A is the set of frame indices where speaker is noted as present by the algorithm, and B is the set of frame indices from the ground truth. The overall performance is studied using the *F-measure*, defined as,

$$\textit{F-measure} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{4}$$

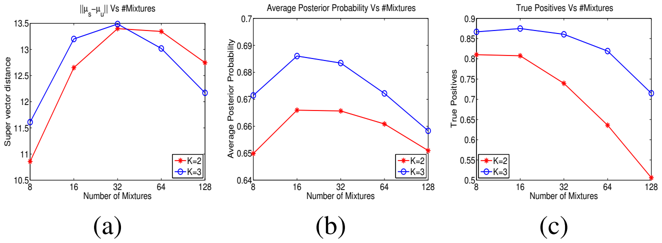
In the experiments below, we used $M = 16$ mixtures in GMM, 6 sec of training segment for adaptation and PCA reduced features, unless specified explicitly.

C. Effect of Number of Mixtures

Consider the distance between the mean super-vectors of the background model and the speaker model adapted from the same, i.e., $\|\bar{\boldsymbol{\mu}}_u - \bar{\boldsymbol{\mu}}_s\|_2$, where $\bar{\boldsymbol{\mu}}_{()}$ is the concatenation of the mean vectors of all the mixture components. Fig. 4(a), shows the distance averaged over all the ground truth segments from 100 conversations, as a function of the number of mixtures. Interestingly, we see that the distance increases with the number of mixtures, initially up to 32 mixtures, but then decreases as the model size becomes large. A similar behavior is observed in the average posterior probability of the target speaker evaluated on the training segment, as shown in Fig. 4(b). We can attribute this to the limited data adaptation. In the case when number of mixture components is high, the number of feature vectors per parameter of the model will be low, and hence the adapted speaker models will be only slightly different from the background model, resulting in similar likelihoods for the background and target speaker models. Fig. 4(c) shows the speaker spotting performance (in terms of avg. true positive values) as a function of the number of mixtures, for 100 simulated conversations comprising different speakers. We see that the performance is poorer at higher number of mixtures, which is again attributed to the limited adaptation data as described above. Interestingly, better performance is obtained for $K = 3$ speaker data compared to $K = 2$, which is studied further in sec. III-F.

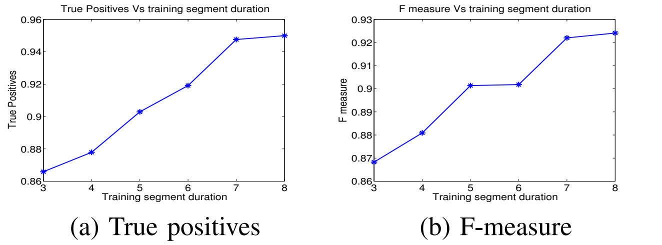
D. Feature set selection

The number of training examples required to train a model depends on the number of parameters in the model. Increasing the feature dimension, such as by appending the derivative features to MFCCs in a typical speech application, results in an increase in the number of model parameters for the same number of mixture components in the GMM. Hence, if the size of training set is fixed, features with higher dimensionality learn poor models compared to features with smaller dimensionality. Table I shows the spotting performance with different feature sets. The experiment is performed on conversations with $K = 3$ speakers. We see that the performance decreases as expected with increase in the dimension of the feature vectors. Best performance is obtained for a 12-dimension PCA reduced features. This shows that the derivative features contain complementary information, which is useful for the task at hand, and this is captured through PCA reduced features, resulting in improved performance.



(a) (b) (c)

Fig. 4. (a) Background and target speaker model separation, (b) average posterior probability of target speaker on the training segment and (c) True positives as a function of the number of mixtures.



(a) True positives

(b) F-measure

Fig. 5. Performance as a function of the initial training data duration.

TABLE I
TRUE POSITIVES FOR DIFFERENT FEATURES

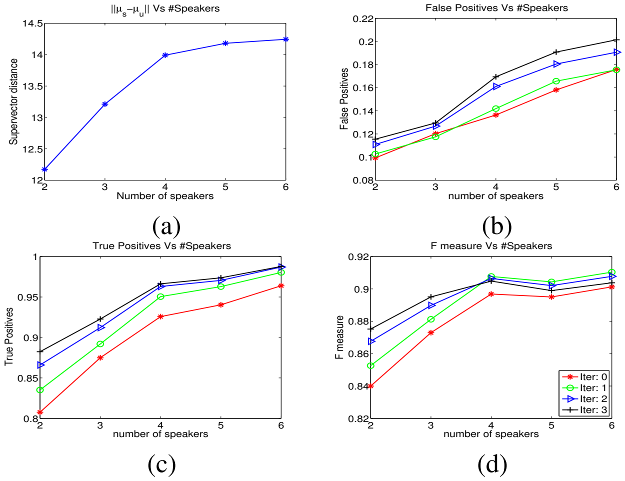
Feature (#dimensions)	TP
MFCC (12)	0.8303
MFCC+Δ (25)	0.7887
MFCC+$\Delta\Delta$ (38)	0.7307
MFCC+$\Delta\Delta\Delta$ (51)	0.6791
PCA reduced (12)	**0.8749**

E. Effect of Duration of training segment

Estimation of the parameters of a model depends on the amount of training data available. Fig. 5 shows the performance of the proposed speaker spotting as a function of the duration of the training segment marked by the user. The conversations used for the experimentation have $K = 3$ speakers. We used PCA reduced features as feature vectors in this experiment. We see that the performance increases as more training data is made available for training. The *F-measure* is more than 0.9 for training segment durations above 5 seconds.

F. Effect of Number of speakers in the conversation, and Model Re-estimation

The background model for a conversation with more number of speakers will capture more diverse features than that for a conversation with say only two speakers. Due to this, the speaker model adapted from such a diverse background model will represent better the target speaker information compared



(a) (b) (c) (d)

Fig. 6. (a) Background and target speaker model separation, (b) False positives, (c) True Positives, and (d) F-measure, as a function of the number of speakers K (Iter=0 corresponds to speaker spotting with training segment chosen by user).

to when the number of speakers is less. Fig. 6 (a) shows a plot of the average of the distance between the super-vectors of the background model and the adapted speaker models as a function of the number of speakers in the conversation. We see that, the distance increases with increase in the number of speakers as predicted from the above observations. Now, we expect that the performance of speaker activity detection algorithm should improve with increase in the number of speakers. Fig. 6 (b, c) show the false positives and true positives as a function of the number of speakers. We see that the true positives increase with increase in the number of speakers in the conversation. However, we notice that the false positives also increase with increase in the number of speakers. Fig. 6(d) shows the *F-measure*, which is increasing with the number of speakers in the conversation showing better overall performance obtained for conversation with more speakers. Fig. 6 also shows the performance as a function of the iterations for model re-estimation using ”high confidence” data as discussed in earlier sections. We see that the performance improves significantly in the poor performance regions, i.e., when the number of speakers is less.

G. Real recordings

The experiments so far considered concatenated records. In this section, we experiment with a real recording of 5 minutes duration (sampling rate $Fs = 8$ KHz). The recording contains $K = 3$ speakers, one male (speaker-1) and two females (speakers$-2, 3$), reading different paragraphs from a novel. The recording was made in a lab environment with fan noise in the background, and also have page turns in a few places. Speaker 1 is closer to the microphone compared to the other two speakers, and hence louder in the recorded conversation. We used PCA reduced features with $M = 16$ mixtures in GMM for the speaker spotting algorithm. Training segment duration is taken to be 6 sec. Table II shows the performance for the three individual speakers. The true positive rate is more than 0.95 for all the speakers, however there are significant number of false positives for speaker-2. Almost perfect detection of speaker-1 is due to the proximity to microphone (signal-to-ambient noise ratio is higher) and also due to the energy level difference with respect to the other two speakers.

TABLE II
PERFORMANCE ON A REAL RECORDING

Target	True Positives	False Positives	F-measure
Speaker 1	0.9960	0.0103	0.9919
Speaker 2	0.9559	0.4270	0.80251
Speaker 3	0.9792	0.0392	0.9702

IV. Conclusion

We studied the speaker spotting problem in speech conversations using latent variable formulation and GMM-UBM framework. Experimentally, we see that for conversations lasting a 3 to 5 minutes, using PCA reduced features along with 8 to 32 mixture GMM models and 6 to 8 sec of training data for adaptation give F-measure > 0.9. Also, the proposed approach gives a soft-decision on the speaker presence which is useful in conversations with over-lapped speakers.

V. Acknowledgements

This research is also supported by the DRDO-IISc Frontiers Research Program.

References

[1] J. P. Campbell Jr, “Speaker recognition: a tutorial,” *Proc. of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[3] S. E. Tranter and D. A. Reynolds, “An overview of automatic speaker diarization systems,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, 2006.

[4] M. H. Moattar and M. M. Homayounpour, “A review on speaker diarization systems and approaches,” *Speech Communication*, vol. 54, no. 10, pp. 1065–1103, 2012.

[5] X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, “Speaker diarization: A review of recent research,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, 2012.

[6] P. Woodland, T. Hain, S. Johnson, T. Niesler, A. Tuerk, and S. Young, “Experiments in broadcast news transcription,” in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, May 1998, pp. 909–912 vol.2.

[7] C. Bishop, *Pattern Recognition and Machine Learning*, ser. Information science and statistics. Springer, 2013.

[8] H. Sundar, T. Sreenivas, and W. Kellermann, “Identification of active sources in single-channel convolutive mixtures using known source models,” *Signal Processing Letters, IEEE*, vol. 20, no. 2, pp. 153–156, Feb 2013.

[9] H. Poor, *An Introduction to Signal Detection and Estimation*, ser. Springer Texts in Electrical Engineering. Springer New York, 2013.

[10] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[11] H. Hotelling, “Analysis of a complex of statistical variables into principal components.” *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journ. of the royal stat. soc. Series B (methodological)*, pp. 1–38, 1977.

[13] J.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains,” *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[14] “Starkey Hearing Technologies. (2013). Open access stimuli for the creation of multi-talker maskers.” http://www.starkeyevidence.com, 2013.