Generic modeling applied to speaker count

Ananth N Iyer
Uchechukwu O Ofoegbu
Robert E Yantorno, Temple University
Brett Y Smolenski
Generic Modeling Applied to Speaker Count

A. N. Iyer*, U. O. Ofòegb hu*, R. E. Yantorno* and B. Y. Smolenski†

*Speech Processing Laboratory, Temple University, Philadelphia, PA 19122, USA
E-mail: {aniyer,uche1,byantorn}@temple.edu
†Airforce Research Laboratory/IFEC, Rome, NY 13441-4505, USA
E-mail: Brett.Smolenski@rl.af.mil

Abstract—The problem of determining the number of speakers participating in a conversation and building their models in short conversations, within an unknown group of speakers, is addressed in this paper. The lack of information about the number of speakers and the unavailability of sufficient data present a challenging task of efficiently estimating the speaker model parameters. The proposed method uses a novel generic speaker identification (GSID) system as a guide in the model building process. The GSID system is designed performing speaker identification where the speaker associated with the test data may not be enrolled. The models in the GSID system are employed as initial speaker models, representing the persons participating in the conversation, and are subjected to a classification-adaptation procedure. The classification is performed based on the Bhattacharyya distance between the model database and the test data being analyzed. The model database of the system is designed to consist of simple and well separated models. A technique to generate such generic models is introduced. The proposed method was applied to the speaker count problem and has produced an overall accuracy of 75.3% in determining if there were 1, 2 or 3 speakers in a conversation.

I. INTRODUCTION

The problem of building speaker models from conversational speech consists of three parts: i) determining the number of speakers participating in the conversation, ii) learning the speaker models and iii) indexing the temporal activity of each speaker. The method presented in this paper deals mainly with the first part. The need for a method to determine the speaker count in a conversation arises mainly in telephone monitoring systems. The developed method is designed for monitoring telephone lines in a prison, where placing three-way telephone calls are forbidden.

In this research, the lack of information about the speakers is overcome by the use of a generic model set, which refers to a set of models that need not represent the speakers in the conversation, but can be used to perform speaker identification (SID), i.e. open-set speaker identification. The general framework adopted will be referred to as the generic speaker identification (GSID) system and will be used to perform speaker count by identifying the generic model set, and then mapping that set to speaker specific models representing the speakers in the conversation. The approach is summarized in the form of a block diagram and is shown in Figure 1.

One of the first concerns in the use of generic speaker models is the choice of a speaker model. The Gaussian Mixture Model (GMM) has been a favorite choice for modeling, as it delivers excellent results in speaker identification when large amounts of data are available for testing [1]. The concept of generic speaker models using GMMs has been successfully applied in speaker indexing on broadcast news data [2]. For such data, however, single speaker utterances usually have long durations [3] as compared to telephone conversations where speaker contiguous utterances may last for about 1 second. This data limitation degrades the performance of the GMM approach. A method to choose the optimal model (GMM or VQ) using the Bayesian Information Criterion has been previously introduced [4]; however the technique is extremely complex. To obtain a balance between performance and complexity, a single Gaussian model is chosen to represent speakers. Furthermore, the Gaussian model is preferred as its parameters can be easily estimated and adjusted, which is a desirable feature of the generic model set. Also, in a recent study it was shown that a single Gaussian can sufficiently model speakers for speaker identification [5].

Prior studies have shown that techniques based on generic models are dependent on their initialization [6]. Various methods based on the Universal background model [7], Monte Carlo Markov Chain [2], and Speaker Model Quantization methods [6] have been proposed. An initial model set is desired to span the entire model space and the models chosen to be equidistant. A new method which ensures that the models in the initialization will be maximally separated is proposed in this paper. In Section II, results from closed-set SID experiments are presented for the purpose of determining the best speech features, and that is followed by; an algorithmic description to identify the generic model set and a discussion of the open-set speaker identification method. Experimental setup and results are discussed in Section III. Finally, conclusions and further studies are discussed in Section IV.

II. METHODS

A speaker model is represented by a single Gaussian $N \sim N(\mu, \Sigma)$ with parameters $\mu$ and $\Sigma$ being the mean vector and the covariance matrix. A set of speaker models...
is created with speech data from a standard database, and is represented as $\Omega = \{N_1, N_2, \ldots, N_Q\}$, with $Q$ being the total number of speakers in the database. The methods described in this section are based on constructing a metric space with each speaker model representing an element in that space (referred to as the model space). Based on studies on speaker discrimination [5], the Bhattacharyya distance was found to yield a reasonable SID accuracy level when the number of models was small. Hence the Bhattacharyya distance was chosen as an appropriate metric in this work and is defined as:

$$d(N_p, N_q) = \frac{1}{4}(\mu_p - \mu_q)^T(\Sigma_p + \Sigma_q)^{-1}(\mu_p - \mu_q) + \frac{1}{2} \log \left( \frac{\Sigma_p + \Sigma_q}{2\sqrt{\Sigma_p \Sigma_q}} \right). \quad (1)$$

A. Feature Selection

The task of selecting a good speech feature for the GSID system was performed using closed-set speaker identification (SID) experiments. The SID system consisted of 50 enrolled speakers from the HTIMIT database, and each speaker was represented as a Gaussian model. The minimum Bhattacharyya distance criterion was applied to determine the SID accuracy. The SID accuracy performance with four popular speech features: Linear Predictive Cepstral Coefficients (LPCC), Line Spectral Pairs (LSP), Log-Area Ratio (LAR) and Mel-Frequency Cepstral Coefficients (MFCC), is shown in Figure 2. The values presented are the average of 100 trials of the SID experiment and the errorbars in each case shows the standard deviation. Note that the LSP features achieved the highest average SID performance as well as the lowest standard deviation. Hence the LSP speech features were used in all experiments.

B. Generic Model Set

A desired feature of the generic model set is that the models be maximally separated, i.e., they span a large area in the model space and have small (and preferably no) overlap regions. One can imagine that such a set would be able to represent a large population of speakers. One possible way to obtain such a set is to apply a clustering method to group similar models and represent them using a centroid model [6]. The disadvantage in such an approach is that the centroid model does not represent one speaker, but a combination of many speakers. A more appropriate way to obtain the maximally separated subset $S$ is by formulating an optimization problem to maximize a cost function $J$:

$$J = \sum_{N_p, N_q \in S} d(N_p, N_q), \quad (2)$$

where the summation is performed on all the models in the subset $S$. The solution to this optimization problem can be obtained by a brute-force combinatorial search for the optimal subset, and in general, such a searching procedure is computationally expensive. For example, consider a set $\Omega$ with $Q = 50$ and the number of models in the subset $P = 10$ is required. It can be determined that cost function $J$ has to be evaluated $C_P \simeq 10^{10}$ times. To reduce the computational magnitude, a suboptimal strategy is proposed and is described in Algorithm 1 below.

**Algorithm 1** To determine maximally separated subset.

**Require:** Number of models in subset $K$; Initialize set $S^{(2)} = \{N_p, N_q\}$, such that $N_p$ and $N_q$ are the two farthest models; $k = 2$;

1: \textbf{while} $k < K$ \textbf{do}
2: \hspace{1em} Choose model $N_i$ which has the largest distance to set $S^{(k)}$. The distance between a point and a set is defined as the minimum of all the distances from the model to all the models in the set. This step can be mathematically written as
3: \hspace{1em} $N_i = \arg\max_{N_i \in \Omega} \min_{N_j \in S^{(k)}} d(N_i, N_j)$

4: \hspace{1em} $S^{(k+1)} = S^{(k)} \cup N_i$
5: \hspace{1em} $k = k + 1$.
6: \textbf{end while}
7: \textbf{return} $S = S^{(k+1)}$

The validity of the algorithm was determined by evaluating it on a random set of points sampled from a uniform distribution, and the result is shown in Figure 3. The gray circles represent the points in the set $\Omega$ and the black asterisks represent the points in the optimally chosen subset $S$.

![Fig. 2. Speaker Identification performance with the Bhattacharyya distance for various speech features. The bar chart represents the average performance value and the errorbars show the standard deviations.](image)

![Fig. 3. Result of Algorithm 1 to determine the maximally separated subset (asterisks) on a random set sampled from uniform distribution (circles).](image)
C. Generic Speaker Identification

The initial generic model set $S$, determined in the previous step, is used as the database to perform speaker identification on the test data being analyzed. Speaker identity is obtained at an utterance level, where each utterance is expected to consist of data from only one speaker. Statistics obtained from the SWITCHBOARD conversational speech database [8] suggests that most utterances in telephone conversations are at least 1 second in duration (see Figure 4). The database consists of 2435 conversations between two speakers recorded over a telephone line lasting approximately 6 minutes each. The length of utterances, i.e., the length of speaker homogeneous segments were determined and a probability density function (pdf) of the results is shown in Figure 4. All the conversations present in the database were used to generate the statistics. Segmentation of a conversation was performed by selecting non-overlapping sections (utterances) of the conversations of length 1 second. Speaker identification is performed by determining the closest model, i.e. the minimum Bhattacharyya distance criterion, to each of the utterances independently, and therefore, associating each utterance with a generic speaker model. One can imagine that it is highly probable that same speaker utterances would be associated with a particular generic model. The chosen models are sequentially adapted to fit the incoming utterances one at a time. The adaptation is performed by updating the mean vector $\mu$ and covariance matrix $\Sigma$ using the following formulas:

$$
\hat{\mu} = (1 - \alpha)\mu + \alpha\mu_{\text{utt}} \\
\hat{\Sigma} = (1 - \alpha)[\Sigma + \mu\mu^T] + \alpha[\Sigma_{\text{utt}} + \mu_{\text{utt}}\mu_{\text{utt}}^T] - \hat{\mu}\hat{\mu}^T
$$

(4)

(5)

where $\mu_{\text{utt}}$ and $\Sigma_{\text{utt}}$ are the mean vector and the covariance matrix estimated from the utterance and $\alpha$, with $0 < \alpha < 1$, is a control parameter to adjust the rate of adaptation. The control parameter enables the adaptation to be performed in a constrained manner. A value of $\alpha$ close to unity suggests that the model adapts rapidly to the observed utterance and is not desirable as the model tends to become specific to the utterance. Similarly, $\alpha$ value close to zero suggests slow model adaptation and is again undesirable as the resulting model might not represent the speaker sufficiently. Based on experimentation a value of $\alpha = 0.8$ was found to produce the best performance.

D. Model Analysis

The next step in finding the speaker count and the models themselves is to analyze the result of the GSID operation. One simple way to determine the number of speakers in the conversation is to identify the models in the generic model set which were altered by the data being analyzed. However, such an approach is error-prone as small utterances, which are not significant enough to represent a speaker might be considered in the speaker counting process. It was found during experimentation that models could be identified effectively based on two factors. The first factor is how well the model can represent the data ($P_f$), and the second factor is the amount of data it represents ($P_d$). An overall quality measure $q$ is computed for each model in the set as:

$$
q = \lambda P_f + (1 - \lambda) P_d
$$

(6)

where $\lambda$ is a tuning parameter determined based on the performance of the GSID system. The use of $P_f$ in the quality measure is obvious, while $P_d$ plays an important role in removing models that are chosen spuriously. It was demonstrated that the data metric itself was sufficient for determining the presence of a third speaker in conversations [9].

The quantity $P_f$ is inversely proportional to the average distance between the model and its associated speech frames. In the case of the Gaussian model, $P_f$ represents the likelihood of the data being generated by the selected model. The data metric $P_d$ is determined as the ratio of the data length represented by the model to the overall length of the data being analyzed. The quality measure is computed for all the models in the generic model set, and the number of models which have a quality measure above a specific level of quality ($q_{th}$) represents the speaker count. The quality measure threshold is left as a free parameter and is chosen based on experimental evaluation. Note that the metrics $P_f$ and $P_d$ are normalized in the range $[0,1]$ to avoid scaling errors while using a preset threshold.

III. EXPERIMENTAL EVALUATION

A. Experimental Specifications

The initial generic model set consisted of 20 speaker models. The models were generated using data from the HTIMIT speech corpus [10] to form the standard database set. The corpus consists of 384 speakers and all the speakers were used to form the set $\Omega$ as described in Section II. Each set of speaker data consists of ten files, of which five files were used for training purposes. The remaining files were used to generate the test data. Due to the unavailability of multi-speaker telephone

---

1The temporal speaker activity was obtained using transcriptions provided by the Mississippi State University (http://www.cavs.msstate.edu/hse/ies/projects/switchboard/index.html).
conversations, which are expected to typically occur as three-way or conference calls, the conversations were simulated by abutting utterances from different speakers. One thousand trials of the speaker count experiment were performed on 20 second conversations with 1, 2, 3 and 4 participants. In each of the trials, the test data was generated by using a set of speakers that does not overlap with the speakers in the generic model set, hence preserving the intended notion of open-set speaker identification.

An illustration of the GSID system on a three speaker simulated conversations is shown in Figure 5. The data associated with each of the models are shown in different colors. The vertical dotted lines represent the actual speaker boundaries.

![Illustration of speaker modelling on conversational data: the data associated with each model is represented by a different shade of gray in a three-speaker conversation. The vertical dotted lines show the actual speaker boundaries.](image)

The accuracy [%] of the GSID system in determining the speaker count (SC) is tabulated in Table 1. The values across columns represents the determined speaker count (presented as Determined SC) and the values across the rows represent the true speaker count (True SC).

<table>
<thead>
<tr>
<th>True SC</th>
<th>One</th>
<th>Two</th>
<th>Three</th>
<th>Four</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>88.0</td>
<td>11.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Two</td>
<td>18.4</td>
<td>71.2</td>
<td>9.6</td>
<td>0.8</td>
</tr>
<tr>
<td>Three</td>
<td>6.0</td>
<td>27.2</td>
<td>37.6</td>
<td>9.2</td>
</tr>
<tr>
<td>Four</td>
<td>2.8</td>
<td>2.6</td>
<td>37.2</td>
<td>33.2</td>
</tr>
</tbody>
</table>

The overall performance of the system was evaluated as follows: let $sc_e$ represent the estimated speaker count and $sc_t$ be the true speaker count. The probability of correct identification of the speaker count is defined as:

$$P = \sum_{k=1}^{4} p(sc_e = k | sc_t = k)p(sc_t = k).$$

The operational parameters were chosen as $\lambda = 0.1$ and $q_{th} = 0.4$ and was found to result in the highest accuracy $P$ defined above. The GSID system was found to be correct 62.5% of the times in determining the speaker count as 1, 2, 3 or 4. The probabilities, $p(sc_t = k) = \frac{1}{4}$ was used, and one can imagine that the number of speakers in the conversations is not equally probable in reality. The occurrence of 4 speaker conversations is significantly smaller in number as compared to the 2 and 3 speaker conversations. In counting the presence of 1, 2 or 3 speakers in a conversation, the GSID system was able to attain an accuracy level of 75.3%.

IV. CONCLUSIONS

A method to determine the speaker count and build speaker models from conversational speech data is proposed. The main focus of this research was restricted to determining the number of speakers participating in a conversation, which is cast as the problem of determining the number of speaker models needed to best represent the data being analyzed. The major challenge has been that no a priori information about the speakers is available. This problem is more difficult with telephone conversations as only short utterances are available for analysis. To overcome these problems, a method to map a set of generic models into speaker specific models was introduced. Experimental evaluation has shown reasonable results in determining the speaker count in a conversation. Further enhancements of the proposed approach could include initializing models based on the dialect, gender, age or accent of the speakers. Allowances could also be made to increase the complexity of the models during each iteration, if required.

ACKNOWLEDGMENT

This effort was sponsored by the Air Force Research Laboratory, Air Force Material Command, and USAF, under agreement number FA8750-04-1-0146. We also wish to thank Michael A. Carlin for his contributions.

REFERENCES