New Background Modeling for Speaker Verification

Dat Tran

School of Information Sciences and Engineering
University of Canberra, Australia
dat.tran@ise.canberra.edu.au

Abstract

A new background speaker modelling method is presented in this paper for text-independent speaker verification using Gaussian mixture models. This method does not require speech databases of other speakers to build background speaker models. A background model can be built directly from the same claimed speaker’s database and has a smaller number of Gaussian mixtures compared to the claimed speaker model. Experiments performed on the YOHO database showed a better result for speaker verification using the 64-mixture claimed speaker model and 16-mixture background model compared to current background model set methods using five closest background models.

1. Introduction

In statistical approach, the verification problem is usually formulated as a problem of statistical hypothesis testing. For an input utterance $X$ and a claimed identity $S$, the task is to determine if $X$ was spoken by the claimed speaker $S$. Assuming that $X$ contains speech from only one speaker, the verification task can be regarded as a basic hypothesis test between the null hypothesis $H_0$: $X$ is from the claimed speaker $S$ against the alternative hypothesis $H$: $X$ is not from the claimed speaker $S$. According to Neyman-Pearson Lemma, if the probabilities of both the hypotheses are known exactly, the optimum test to decide between these two hypotheses is a likelihood ratio test given by

$$S(X) = \frac{P(X \mid H_0)}{P(X \mid H)} \begin{cases} > \theta & \text{accept } H_0, \\ \leq \theta & \text{reject } H_0 \end{cases}$$

(1)

where $\theta$ is a predefined decision threshold and $S(X)$ is referred to as the similarity score of the input utterance $X$. This approach provides a good theoretical formulation to speaker verification.

However, in practical verification problems, it is impossible to obtain the exact probability density functions for either the null hypothesis or the alternative hypothesis. A parametric form of the distribution under each hypothesis is assumed to estimate these probability density functions. Let $\lambda_c$ be the claimed speaker model and $\lambda$ be a model representing all other possible speakers, i.e. impostors. Let $P(X \mid \lambda_c)$ and $P(X \mid \lambda)$ be the likelihood functions of the claimed speaker and impostors, respectively. The similarity score is calculated as follows

$$S(X) = \frac{P(X \mid \lambda_c)}{P(X \mid \lambda)} \begin{cases} > \theta & \text{accept } H_0, \\ \leq \theta & \text{reject } H_0 \end{cases}$$

(2)

The denominator $P(X \mid \lambda)$ is called the normalization term. While the model $\lambda_c$ can be estimated using the training data from the claimed speaker, the model for impostors is less well defined. The major difficulty is how to calculate the probability $P(X \mid \lambda)$. Speaker verification performance is strongly affected by variations in signal characteristics, therefore normalization methods have been applied to compensate for these variations [1]. The first approach is to use a subset of impostor models that is representative of the population close (similar) to the claimed speaker [2]. This subset has been called cohort set or background speaker set. Depending on the approximation of $P(X \mid \lambda)$ in (1) by the likelihood functions of $B$ background models $P(X \mid \lambda_i), i = 1, ..., B$, we obtain different normalization methods such as arithmetic mean method [3], posterior probability method [4], geometric mean method [5] and weighted average mean method [6]. The second approach is to use a virtual background speaker model [7]. Assuming all speaker models are Gaussian mixture models (GMMs), the virtual model is regarded as a collection of Gaussian components from impostor models close to the claimed speaker model. The third approach is to pool speech from several speakers to train a single universal background model [8]. The fourth approach is the hybrid cohort-world model [9]. The model $\lambda$ is trained from impostors’ utterances but the number of training utterances is the same as the number of utterances used to train the claimed model.

All of the above-mentioned background modeling methods were discovered by various considerations of the impostors’ models. However, speaker verification systems based on such background models rely on the availability of speaker databases and the acoustic condition. It is known that the speech signal is influenced by the speaking environment, the channel used to transmit the signal, and, when recording it, also by the transducer used to capture the signal. For portable devices such as palm-top computers and wireless handsets, a high demand of computation and memory requirement is not desirable. A background model design for flexible and portable speaker verification systems have been proposed for this purpose [10]. A speaker verification system using left-to-right hidden Markov models consisting of 25 states with 4 Gaussian mixtures per state showed a good performance for this background model. In this approach, background speaker models were directly built from the claimed speaker’s enrolment utterances.

In this paper, we present a background modelling method for text-independent speaker verification systems using Gaussian mixture models (GMMs) based on the above-mentioned background model design. We use the same training data to build the claimed speaker model and background model. The difference between these two models is the
number of Gaussian mixtures. The background model should have a smaller number of Gaussian mixtures compared to the claimed speaker model. Experiments were performed on the YOHO speech database with different number of Gaussian mixtures. Experimental results showed that a low verification error rate is obtained if the number of Gaussian mixtures in the background model is less than half of those in the claimed speaker model. Compared to current background model set methods, the proposed method using 64-mixture GMMs for claimed speaker models and 16-mixture GMMs for background speaker models showed a better performance.

2. Proposed Background Speaker Model

Consider the model space of GMM in which each model can be viewed as a point. According to a new approach to utterance verification [11], given a model \( \lambda_c \), nested neighborhoods in the model space can be defined as follows:

- Zero neighborhood: this consists of the model \( \lambda_c \) only.
- Tight neighborhood: a very small neighborhood surrounding the model \( \lambda_c \) which indicates a robust representation of the model \( \lambda_c \).
- Medium neighborhood: a medium size and significantly larger than the tight neighborhood. This neighborhood possibly includes all competing models, which are by definition close to the model \( \lambda_c \) in model space.
- Large neighborhood: this is larger in size and should cover all impostors’ models.
- Infinity neighborhood: this has an infinity size and cover the entire model space, which represents non-speech events.

According to this approach, we can say that all of the normalization methods [2] – [9] mentioned in the previous section are to find background speaker models existing in the medium neighbourhood. In other words, any model found in this medium neighbourhood can be used as a background speaker model. If we can find a model transformation method which can transform the model in the zero or tight neighborhood to a model in the medium neighborhood, then the transformed model can be used as a background model. The approach in [10] can be regarded as a model transformation. Other transformation methods proposed are as follows:

- Using a subset of the claimed speaker’s training set to train the background speaker model;
- Using a “weaker” (lower acoustic resolution) claimed speaker model as the background speaker model. For example, if the claimed speaker model is a 64-mixture GMM, then the background speaker model might be a 16-mixture GMM or a 4-mixture GMM; and
- Swapping values in a model parameter, e.g., mixture values between GMM components in the claimed model to obtain the background model.

In this paper we use the second method to obtain background speaker models for text-independent speaker verification systems. We used the same training data to build the claimed speaker model and background speaker models. We trained GMMs for each speaker in the YOHO database with 2, 4, 8, 16, 32 and 64 Gaussian mixtures. The 16-mixture, 32-mixture and 64-mixture GMMs were used as claimed speaker models and background speaker models were GMMs having smaller number of mixtures. We also compared the proposed method to the arithmetic mean method \( S_a(X) \) [3] and geometric mean method \( S_g(X) \) [5] using background model set

\[
S_a(X) = \log P(X | \lambda_c) - \log \left( \frac{1}{B \sum_{i=1}^B P(X | \lambda_i)} \right)
\]

\[
S_g(X) = \log P(X | \lambda_c) - \frac{1}{B \sum_{i=1}^B \log P(X | \lambda_i)}
\]

3. Experimental Results

3.1. Database description

The YOHO corpus was designed for speaker verification systems in office environments with limited vocabulary. There are 138 speakers, 108 males and 30 females. The vocabulary consists of 56 two-digit numbers ranging from 21 to 97 pronounced as “twenty-one”, “ninety-seven”, and spoken continuously in sets of three, for example “36-45-89”, in each utterance. There are four enrolment sessions per speaker, numbered 1 through 4, and each session contains 24 utterances. There are also ten verification sessions, numbered 1 through 10, and each session contains 4 utterances. All waveforms are low-pass filtered at 3.8 kHz and sampled at 8 kHz.

3.2. Speech processing and algorithmic issues

Speech processing was performed using HTK V2.0, a toolkit [12] for building hidden Markov models (HMMs). The data were processed in 32 ms frames at a frame rate of 10 ms. Frames were Hamming windowed and pre-emphasized. The basic feature set consisted of 12th-order mel-frequency cepstral coefficients (MFCCs) and the normalized short-time energy, augmented by the corresponding delta MFCCs to form a final set of feature vector with a dimension of 26 for individual frames.

GMMs are initialized as follows. Mixture weights, mean vectors, and covariance matrices were initialized with essentially random choices. Covariance matrices are diagonal, i.e., \( \sigma_{ik} = \sigma_{ik}^2 \) and \( \sigma_{ik} = 0 \) if \( i \neq j \), where \( \sigma_{ik}^2 \), \( 1 \leq k \leq K \) are variances. A variance limiting constraint was applied to all GMMs using \( \sigma_{ik} = \sigma_{ik}^2 \) on elements of all variance vectors in the GMM in our experiments.

Each speaker was modelled by using 96 training utterances in four enrolment sessions without end-point detection. Error rates therefore were not too low to allow meaningful comparisons between the different background speaker modelling methods for speaker verification. GMMs were trained in text-independent mode.

3.3. Speaker Verification Results

Experiments were performed on 138 speakers using each speaker as a claimed speaker and rotating through all speakers. The total number of claimed test utterances and
Impostor test utterances are 5520 (138 claimed speakers x 40 test utterances) and 756240 ((138 x 137) impostors x 40 test utterances), respectively.

Figure 1 shows a comparison of DET (Detection Error Tradeoff) curves for claimed speaker models consisting of 16 Gaussian mixtures and background speaker models consisting of 2, 4 and 8 mixtures using speaker-independent threshold. Larger numbers of mixtures are considered in Figures 2 and 3. It can be seen from the three figures that the highest performance is obtained if the ratio of the number of Gaussian mixtures in the claimed speaker model and background model should be greater than 2.

**Figure 1:** Comparison of DET curves for the 16-mixture claimed speaker model and background speaker models consisting of 2, 4 and 8 Gaussian mixtures. The same enrollment utterances were used to build all of these GMMs.

**Figure 2:** Comparison of DET curves for the 32-mixture claimed speaker model and background speaker models consisting of 2, 4, 8, 16 and 32 Gaussian mixtures. The same enrollment utterances were used to build all of these GMMs.

Figures 4, 5 and 6 compare the verification error rates of the proposed method with the arithmetic mean method in (3) and the geometric mean method in (4) where background speaker sets (closest speakers) with \( B = 5 \) were used. The proposed method shows a worst performance in Figures 4 and 5. However, in Figure 5, a better performance is obtained for the proposed method using 64-mixture claimed speaker models and 16-mixture background speaker models.

**Figure 3:** Comparison of DET curves for the 64-mixture claimed speaker model and background speaker models consisting of 2, 4, 8, 16 and 32 Gaussian mixtures. The same enrollment utterances were used to build all of these GMMs.

**Figure 4:** Comparison of DET curves for the 16-mixture claimed speaker model and: 1) Proposed method using the 4-mixture background model, 2) Geometric mean method using 5 closest background speaker models, and 3) Arithmetic mean method using 5 closest background speaker models.

4. **Discussion**

The proposed background speaker modeling method has shown a comparable performance to current normalization methods. We are investigating other model transformation methods to improve this approach. The proposed method can be applied when only the enrollment data is available. The method is useful for users who want to select a password in any language or use any microphone under any acoustic condition. The method is also useful for practical and portable devices such as palm-top computers or wireless handsets since a higher demand of computation and memory requirement may not be desirable for such applications.
5. Conclusion

We have presented a background modeling method for speaker verification. The same training data set was used to train the background speaker model and the claimed speaker model. The background speaker model had a lower acoustic resolution. This was done by using a smaller number of Gaussian mixtures compared to the claimed speaker model. Experiments performed on the YOHO database showed an acceptable verification error rate. A good performance was obtained if the ratio of the number of Gaussian mixtures in the claimed speaker model and background model is greater than 2. Using the 64-mixture GMM for the claimed model and the 16-mixture GMM for the background model, the lowest error rate was achieved.

6. Acknowledgement

The author would like to acknowledge the support of the Divisional Research Institute Grant, University of Canberra, Australia.

7. References