We are IntechOpen, the world’s leading publisher of Open Access books
Built by scientists, for scientists

6,600
Open access books available

177,000
International authors and editors

195M
Downloads

154
Countries delivered to

TOP 1%
Our authors are among the most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™
Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com
The Effect of Reverberation on Optimal GMM Order and CMS Performance in Speaker Verification Systems

Noam R. Shabtai, Boaz Rafaely and Yaniv Zigel
Ben-Gurion University of the Negev, Israel

1. Introduction

In speaker recognition, features are extracted from speech signals to form feature vectors, and statistical pattern recognition methods are applied in order to model the distribution of the feature vectors in the feature space. Speakers are recognized by pattern matching of the statistical distribution of their feature vectors with target models. Speaker verification (SVR) is the task of deciding, upon receiving tested feature vectors, whether to accept or reject a speaker hypothesis, according to the speaker’s model. A popular feature extraction method for speech signal processing is the mel-frequency cepstral coefficients (MFCC) [Davis & Mermelstein, 1980], and Gaussian mixture models (GMM) has become a dominant approach for statistical modeling of speech feature vectors for text-independent SVR [Reynolds et al., 2000].

Speaker verification is widely used in telecommunication or conference room applications, where reverberation is often present due to the surrounding room environment. The presence of reverberation adds distortion to the feature vectors, which results in performance degradation of SVR systems due to mismatched conditions between trained models and test segments.

Feature normalization techniques such as the cepstral mean subtraction (CMS) [Mammone et al., 1996] and variance normalization [Chen & Bilmes, 2007], and score normalization techniques such as the Znorm, Hnorm, Tnorm [Bimbot et al., 2004, Mammone et al., 1996] and Top-norm [Zigel & Wasserblat, 2006], were originally developed to compensate for the effect of a telephone channel [Mammone et al., 1996], or for the effect of slowly varying convolutive noises in general [Reynolds et al., 2000]. For that reason, these techniques may be used to reduce the effect of reverberation, if it is characterized by a short-duration room impulse response (RIR). However, it may be difficult to find research studies in the literature on the effect of CMS on SVR performance under reverberation conditions of long duration RIR, which is often the case in room acoustics.

In cases of long-duration RIR, the target models may be trained using a reverberant speech database, as suggested by Peer et al. [Peer et al., 2008], in order to overcome the mismatched conditions between the models and the reverberant testing speech segments. This method was tested on adaptive-GMM (AGMM) based SVR system, with various values of reverberation time (RT - the time that takes the impulse response to decay by 60dB [Schroeder,
Matching of RT between train and test data was reported to reduce the equal error rate (EER) from 16.44% to 9.9% on average, when using both Znorm and Tnorm score normalizations.

The methods that were described in the previous paragraph used fixed GMM order, and were automatically performing feature normalization. This chapter shows that the effect of reverberation on the feature vectors might decrease the optimal GMM order, for Bayesian and Kullback information criteria (BIC and KIC, respectively). As a feasibility study, the relatively simple case of GMM without adaptation was used, as currently AGMM systems are designed for using constant model order. However, the study in this chapter might suit a future adjustment of AGMM systems.

The investigation of the effect of GMM order is based on a study performed by the authors [Shabtai et al., 2008a], where only simulated RIRs were used. This chapter also investigates the effect of reverberation on the performance of CMS applied to MFCC feature vectors in SVR. In that sense, it serves as an extension of an early study of the authors [Shabtai et al., 2008b], where only simulated RIRs were used to form reverberant speech. Here both simulated and measured RIRs are employed.

2. Room parameters

Room parameters can either have a direct relation to the physical characteristics of the room, or some relation to the RIR. Associated with the physical characteristics of the room we have the geometrical characteristics, which are the volume $V$ and the surface area $S$, and the reflection coefficient of the room boundaries, $R$. The absorption coefficient of the room boundaries $a$ is defined as [Kuttruff, 2000]

$$ a = 1 - |R|^2 $$

and thus the absorption area is

$$ A = \bar{a} S $$

where $\bar{a}$ is the average absorption coefficient along the room boundaries.

An important room parameter that can be measured from the RIR is RT, which is the time that takes the energy in a room to decay by 60 dB once the source is turned off. By assuming that until the source was turned off it had been producing a stationary white noise, RT can be calculated from the RIR by using Schroeder’s energy decay curve [Schroeder, 1965]

$$ e(t) = 10 \log_{10} \int_{t_1}^{t_2} h^2(\tau) \, d\tau - 10 \log_{10} \int_{t_0}^{t_1} h^2(\tau) \, d\tau $$

where $h(t)$ is the RIR, and numerically solving

$$ e(\text{RT}) = -60 \text{dB}. $$

In the ISO 3382 standard [ISO 3382:1997, 1997], RT is calculated from a least squares based linear fitting of Schroeder’s energy decay curve in order to compensate for the non-linearity and for the noise-floor effect.

Room response from a source to a receiver can be given in the frequency domain by the room transfer function (RTF). In rectangular rooms, the RTF is known to be a combination of
natural or eigen modes. At frequencies where the density of the eigenmodes is more than three eigenmodes for a 3dB bandwidth of a given eigenmode, the sound field is usually considered to sufficiently satisfy the assumptions of diffuse field theory. In diffuse fields, RT is related to the volume by Sabine formula [Kinsler et al., 2000]

\[ RT = 0.161 \frac{V}{A}. \]

3. Feature extraction and normalization

A commonly used procedure of MFCC feature extraction is shown in Fig. 1 [Bimbot et al., 2004]. The pre-emphasis filter is applied to enhance the high frequencies of the spectrum, which are generally reduced by the speech production process. The STFT block splits the signal in the time domain into overlapping frames where the signal is considered to be stationary, and calculates the fast Fourier transform (FFT) of each frame. Then, filter banking is applied by integrating the magnitude FFT of the signal frames with triangular windows in the mel-frequency domain. Afterwards, the dB level is calculated. This results in a series of energy scalars for every frame. Discrete cosine transform (DCT) is calculated, from which coefficients are selected to form MFCC feature vectors. Applying a discrete-time derivative results in \( \Delta \)MFCC feature vectors, such that

\[ c_i = [c_1^i, \ldots c_N^i, \Delta c_1^i, \ldots \Delta c_N^i]^T \]

is the feature vector of the \( t \)th frame (\( t \) here is a discrete time index), where \( N \) is the number of MFCC coefficients.

Transmission channels may add a convolutive effect to the speech signal prior to the process of feature extraction. This may result in feature vectors distortion. For that reason feature normalization may be used. In this chapter we discuss the CMS technique, which is the operation of subtracting the sample mean [Bimbot et al., 2004]

![Fig. 1. Extraction of MFCC and \( \Delta \)MFCC feature vectors from speech signal [Bimbot et al., 2004].](image-url)
\[
\tilde{c}_t = c_t - \mu \quad t = 0 \ldots T - 1
\]

where \( \mu \) is the sample mean of the series \( c_0 \ldots c_{T-1} \). The operation of CMS may include variance normalization [Mammone et al., 1996] by dividing the components by the sample standard deviation (STD), i.e.,

\[
\tilde{e}_n = \frac{e_n}{\sigma_n} \quad t = 0 \ldots T - 1 \quad n = 1 \ldots N
\]

where for every \( n = 1 \ldots N \), \( \sigma_n \) is the sample STD of the series \( c_0 \ldots c_{T-1} \).

4. Speaker verification with GMM approach

In this section we represent a brief description on SVR with GMM approach [Bimbot et al., 2004, Mammone et al., 1996]. Speaker verification is the task of accepting or rejecting a tested speaker as a hypothetical speaker. Let

\[
X = [x_0, x_1, \ldots, x_{T-1}]
\]

be a segment of speech feature vectors \( x_t \) of discrete time \( t \in \{0, 1, \ldots, T - 1\} \). Let \( H_1 \) represent the event that the tested speaker is the hypothetical speaker, and let \( H_0 \) represent the opposite event.

The model \( \lambda_1 \) is defined to contain the parameters such that a parametric probability density function (PDF) \( p(X; \lambda_1) \) would model the conditional PDF \( p(X|H_1) \). In a similar way, \( \lambda_0 \) is defined such that \( p(X; \lambda_0) \) models \( p(X|H_0) \). For example, if the models assume Gaussian distribution, then \( \lambda_0 \) and \( \lambda_1 \) consist of a mean vector and a covariance matrix.

The decision is then made according to the log-likelihood ratio test (LLRT)

\[
\Lambda(X) = \log p(X; \lambda_1) - \log p(X; \lambda_0) \begin{cases} \geq \theta & \text{accept hypothetical speaker} \\ < \theta & \text{reject hypothetical speaker} \end{cases}
\]

where \( \Lambda(X) \) is referred to as the score function, and \( \theta \) is the LLRT threshold. If the feature vectors in \( X \) are assumed independent, then for each model, \( \log p(X; \lambda) \) may be calculated by

\[
\log p(X; \lambda) = \sum_{t=0}^{T-1} \log p(x_t; \lambda).
\]

In applications where different speakers have a different number of feature vectors, the score function may be normalized by \( T \) to form

\[
\tilde{\Lambda}(X) = \frac{1}{T} \Lambda(X)
\]

in order not to bias the score in favor of speakers with more feature vectors.

According to the GMM approach, if \( x \) is a feature vector, and \( \lambda \) is a set of parameters, then

\[
p(x; \lambda) = \sum_{i=1}^{M} \alpha_i p_i \left( x ; \lambda^{(i)} \right)
\]
where $M$ is the number of Gaussians in the GMM, or, the model order, the weights $\omega_i$ apply

$$
\sum_{i=1}^{M} \omega_i = 1,
$$

and $p_i(x; \lambda^{(i)})$ is a parametric normal PDF. Hence, the sub-model $\lambda^{(i)}$ consists of a mean vector $\mu_i$ and a covariance matrix $\Sigma_i$ parameters of a single Gaussian. Hence,

$$
p_i(x) = p(x; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} e^{-\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)}
$$

where $d$ is the dimension of $x$. According to (15), the model $\lambda$ in (13) can be denoted as [Reynolds et al., 2000]:

$$
\lambda = [\omega_i, \mu_i, \Sigma_i]_{i=1...M}
$$

The parameters $\omega_i$, $\Sigma_i$, and $\mu_i$ are estimated using the expectation maximization (EM) algorithm [Dempster et al., 1977]. The covariance matrix $\Sigma$ can be selected as either diagonal or a full matrix. The interpretation of a diagonal covariance matrix is that the feature vector coordinates are independent of one another. The computation of the parametric PDFs is much simpler in this case. The advantage of the full covariance matrix, however, is the enhanced generalization of the parametric PDFs in modeling the conditional PDFs. In practice, GMM is used with diagonal covariance matrices to approximate the case of one Gaussian with a full covariance matrix with less computational effort.

Speakers that are known to a certain hypothesis are referred to as target speakers of that hypothesis, and impostor speakers to other hypotheses. Performance analysis of SVR is measured with miss probability, $P_{miss}$, which is the probability that a target model was rejected.

$$
P_{miss} = P(\lambda(X) < \theta \mid \text{target}),
$$

and with the probability of false alarm, $P_{fa}$, which is the probability that an impostor speaker was accepted

$$
P_{fa} = P(\lambda(X) > \theta \mid \text{impostor}).
$$

Both $P_{miss}$ and $P_{fa}$ are functions of the threshold $\theta$, and they each come at the expense of the other. The threshold $\theta$ is used as a parameter to yield the detection error trade-off (DET) curve, which plots $P_{miss}$ as a function of $P_{fa}$. The point on the DET curve where $P_{miss}$ equals $P_{fa}$ is the EER. The EER is usually used as a scalar measure of the performance of SVR systems.

5. The effect of reverberation on the feature vectors in GMM

For reverberant speech, if the RT is larger than the short time Fourier transform (STFT) frame size, there will be time-smearing of the feature vectors. An increase in RT increases this time-smearing. This effect may cause the Gaussian means of the GMM to come closer together. In order to examine this, the weighted average distance between the Gaussians in the GMM and the overall mean feature vector can be calculated in the following form:
where $M$ is the GMM order, $\omega_i$ is the weight of the $i$th Gaussian, $\mu_i$ is the mean vector of the $i$th Gaussian, and $\mu$ is the overall mean feature vector. It is assumed that if an increase of RT results in closer Gaussians, then $D$ should decrease.

Figure 2 shows an example of the weighted average distance between the Gaussians in GMMs that are trained from reverberant speech signals, which are the result of a convolution with simulated RIRs. A normalized form of this distance

$$D_{\text{norm}} = \frac{D}{D_{RT=0}}$$

Fig. 2. Normalized distance between Gaussians and overall mean in a GMM of different speakers. Numbers in the legend indicate speaker index in NIST-SRE database. (a) 10 Gaussians, male speakers; (b) 50 Gaussians, male speakers; (c) 10 Gaussians, female speakers; and (d) 50 Gaussians, female speakers.
was used, where $D_{RT=0}$ indicates the weighted average distance between the Gaussians and the overall mean feature vector in the case of clean (non-reverberant) speech. The feature space of the GMMs in Fig. 2 is of 24 dimensions, and the feature vectors consist of 12 MFCC and 12ΔMFCC coefficients. The normalized distance is displayed as a function of RT in both logarithmic axes. The STFT frame size is 30 ms. The numbers in the legend of Fig. 2 represent indices of speakers from the NIST-99 SRE database (see Sec. 7). The GMMs were trained using 10 and 50 Gaussians both for male and female speakers. The value of $D_{norm}$ seems to decrease with the increase of RT. Hence, the Gaussian means of the GMM come closer together. As a result, the GMM might need fewer Gaussians. Also seen from Fig. 2 is a knee between 100 and 200 msec, where RT is considerably larger than the STFT frame size. It should be pointed out that this knee applies to all speakers at similar RTs.

6. The effect of reverberation on the optimal GMM order

We aim to find an optimal GMM order for reverberant speech. The Bayesian, Akaike, and Kullback information criteria (BIC, AIC, and KIC, respectively) [Chen & Huang, 2005] were used to estimate the unknown order of the target models with the training observation. The criteria are defined as follows

$$BIC_{\lambda,M,X} = -\log p(X|\lambda_M) + \frac{1}{2}M(2d+1)\log N \quad (21)$$

$$AIC_{\lambda,M,X} = -\log p(X|\lambda_M) + M(2d+1) \quad (22)$$

$$KIC_{\lambda,M,X} = -\log p(X|\lambda_M) + \frac{3}{2}M(2d+1) \quad (23)$$

where $M$ is the order of the model $\lambda_M$, $N$ is the number of feature vectors in the realization $X$, and $d$ is the feature vector dimension. The optimal model order $M^*$ was selected for an information criterion $IC$, as the one whose model $\lambda_M$ amongst $M \in \{10, 20, 30, 40, 50\}$ yields the minimum criterion value for $X$, or,

$$M_{IC}^* = \arg \min_{M \in \{10, 20, 30, 40, 50\}} IC_{\lambda,M,X} \quad (24)$$

where IC is one of the information criteria defined above.

Figure 3 shows an example of the KIC and BIC values of GMMs that are trained from both non-reverberant speech signal and reverberant speech signal, which is the result of a convolution with simulated RIR. The IC values with 10, 20, 30, 40, and 50 Gaussians were normalized for each speaker with the IC value of 30 Gaussians to yield $IC_{norm}$. The numbers in the legend of Fig. 3(a) represent indices of speakers from the NIST-99 SRE database (see Sec. 7), and apply to all sub-figures in Fig. 3. It can be seen that optimal model order in terms of minimum KIC for clean speech is 50 ($M_{KIC}^*$ in Fig. 3(a)), whereas for reverberant speech with RT=0.85 sec (Fig. 3(b)) it reduces to some value in the range of 30 ÷ 50. Optimal model order in terms of minimum BIC is in the range of 20 ÷ 40 ($M_{BIC}^*$ in Fig. 3(c)), whereas for reverberant speech with RT=0.85 sec (Fig. 3(d)) it reduces to 10. The general effect of
Fig. 3. KIC and BIC values as a function of GMM order (normalized with KIC and BIC values in case of 30 Gaussians), using clean and reverberant speech of male speakers. (a) KIC without reverberation, (b) KIC with RT=0.85 sec, (c) BIC without reverberation, and (d) BIC with RT=0.85 sec.

reverberation is therefore to reduce the optimal model order. It should be noted, however, that the results have a large variance of model order, but a low variance of IC values. Therefore, the significance of model order should be examined in terms of minimum EER of a SVR system.

7. Experimental study of the effect of GMM order on SVR

In this section, an experimental study of the effect of GMM order on the EER of SVR is presented. Reverberant speech training data were generated for several values of RT. The image method of Allen and Barkley [Allen & Berkley, 1979] was used to generate a simulated impulse response of a room. RT is measured on the impulse response according to [Schroeder, 1965]. Speech segments were taken from the national institute of standards and technology (NIST) – 1999 speaker recognition evaluation (SRE) database [Martin and Przybocki, 2000] for training target GMMs.
The Effect of Reverberation on Optimal GMM Order and CMS Performance in Speaker Verification Systems

Fig. 4. Comparing results of optimal GMM order using BIC, AIC, and KIC, to the optimal GMM order in terms of minimal EER. (a) Optimal model order, (b) EER values.

Figure 4(a) shows the optimal order for an average of 198 male speakers with one-minute long speech segment each. Figure 4(a) also compares the optimal order of BIC, AIC, and KIC to the optimal order in terms of minimum EER (the model order among $M \in \{10, 20, 30, 40, 50\}$ that yields the minimum EER). EER results were generated by a loglikelihood based SVR experiment. This experiment included 686 half-minute long reverberant test speech segments, generated from the NIST-99 SRE database. The test speech was introduced to the reverberant target GMMs and to a background GMM. The Background GMM was generated from 50 speakers, each with a one-minute long speech segment, taken from the...
It can be seen that the optimal model order is reduced with the increase of RT for model selection according to BIC and KIC. For $RT < 0.5$ sec, $M_{KIC}^*$ is similar to the optimal model order in terms of minimum EER. For $RT > 0.5$ sec, $M_{BIC}^*$ is similar to the optimal model order in terms of minimum EER. $M_{AIC}^*$ is constant 50.

Fig. 4(b) shows the EER results for the optimal model order of KIC and BIC, compared with EER of a constant model order 50, and to the minimum EER. It can be seen that in terms of EER values, using a constant model order 50 is similar to using $M_{KIC}^*$ for $RT < 0.5$ sec, but worse than using $M_{BIC}^*$ for $RT > 0.5$ sec. Since $M_{BIC}^*$ decreases with the increase of RT, reducing model order can reduce the EER of SVR in a highly reverberant environment.

8. Experimental study on the performance of CMS applied in SVR under reverberation

An early study of the authors [Shabtai et al., 2008b] has investigated the effect of reverberation on the efficiency of CMS in improving the performance of SVR. The performance of an SVR system was measured by calculating the EER in rooms with different RTs and volumes. Test speech segments were made reverberant with RIRs that were simulated using the image method of Allen and Barkley [Allen & Barkley, 1979]. It was shown that for high RTs, CMS decreases.

In this section we extend the research to reverberant speech generated by convolution with measured RIRs. The environments in which the RIRs were measured are tabulated in Table 1. Measured RIRs 1 ÷ 10 were measured with Brüel & Kjær 4295 Omni-Source loudspeaker and Brüel & Kjær 4942 1/2-inch diffuse-field microphone, at selected rooms in Ben-Gurion University of the Negev, Israel (BGU). Measured RIRs 11 ÷ 14 were taken from the Concert Hall Research Group (CHRG) project [CHR, 2004]. In order to compare the results with simulated RIRs, the image method was used to simulate RIRs of rooms with similar dimensions and RTs to the rooms in Tab. 1.

The SVR system was using 20 msec speech frames in which MFCC and ΔMFCC were calculated to form 24-dimensional feature vectors, for which CMS was either applied or not. Target models were trained using the AGMM approach [Reynolds et al., 2000]. A background GMM (BGM) of 1024 Gaussians was generated from one-minute long non-reverberant speech segments of 50 speakers, taken from the NIST-1998 SRE database. This BGM was used to train target AGMMs for 198 male speakers, with one-minute long non-reverberant speech segments, taken from the NIST-1999 SRE [Martin and Przybocki, 2000] database. Test speech segments were taken from NIST-1999 SRE, for 686 male speakers with half-minute long speech segment each. The test speech segments were made reverberant by convolving them with simulated and measured RIRs. The EER results were calculated by introducing the reverberant test speech segments to the target AGMMs and BGM of non-reverberant speech.

Figure 5 shows a scatter plot of EER values as a function of RT. The cross, circle, and triangle marks on Fig. 5 represent EER values when either feature normalization was not used, or CMS was applied, or CMS was applied along with variance normalization, respectively. Linear fitting to the EER values is shown in Fig. 5. Thick solid curves denote using no
The Effect of Reverberation on Optimal GMM Order and CMS Performance in Speaker Verification Systems

RIR | Environment | RT [sec] | $V_\text{[m}^3\text{]}$
--- | --- | --- | ---
1 | Building 33 Office 126 | 0.8 | 37
2 | Building 33 Office 427 | 0.6 | 42
3 | Building 34 Classroom 103 | 0.6 | 120
4 | Building 33 Lecture room 102 | 0.5 | 147
5 | Building 34 Classroom 202 | 1 | 301
6 | Building 33 Teaching lab 204 | 0.6 | 339
7 | Building 26 Auditorium 4 | 1.5 | 793
8 | Building 26 Auditorium 5 | 1.2 | 1142
9 | Building 26 Auditorium 6 | 1.3 | 1142
10 | Sonnenfeld lecture room | 1 | 2529
11 | Mechanics Hall (Worchester, MA) | 2.4 | 8367
12 | Troy Music Hall (Troy, NY) | 2.6 | 11320
13 | Boston Symphony Hall (Boston, MA) | 2.6 | 16611
14 | Kleinhans Music Hall (Buffalo, NY) | 1.9 | 18241

Table 1. Rooms in which RIRs were measured.

Feature normalization, dashed curves denote using CMS, and with thin solid curves denote using CMS along with variance normalization. Figures 5(a) and 5(b) refer to simulated and measured RIRs, respectively. In the case of simulated RIRs as well as in the case of measured RIRs, it can be seen that CMS is improving the performance of SVR in a reduced manner with the increase of RT. Moreover, it can be seen that for some high values of RT, CMS may increase the EER rather than decrease it. These results support previous results [Shabtai et al., 2008b] in which it was shown that CMS is improving the performance of SVR in a reduced manner with the increase of RT, and validate them with measured RIRs.

9. Conclusion

The effect of GMM order on SVR with reverberant speech was investigated. Time-smearing of the feature vectors due to reverberation reduces the optimal GMM order in terms of minimum BIC and KIC. When tested on a GMM-based SVR system, reducing model order improves system performance for highly reverberant speech. A future adjustment to AGMM may be proposed in this direction.

The effect of room volume and RT on the performance of CMS applied to MFCC feature vectors in SVR was investigated. It was shown that the performance of CMS may degrade with the increase of RT. In some cases of high RT, CMS may increase the EER of SVR rather than decrease it. Hence, in these cases, CMS should not automatically be used. As a future work, we purpose combining a CMS decision block in SVR.
Fig. 5. EER values of SVR with reverberant speech as a function of RT. Cross marks ("x") denote no feature normalization (linear fitting with thick solid line), circles ("o") denote CMS (linear fitting with thick dashed line), and triangles ("Δ") denote using CMS with variance normalization (linear fitting with thin solid line). Test speech segments were made reverberant by convolution with (a) simulated, and (b) measured RIRs.
10. References


www.intechopen.com
In the last decade, further applications of speech processing were developed, such as speaker recognition, human-machine interaction, non-English speech recognition, and non-native English speech recognition. This book addresses a few of these applications. Furthermore, major challenges that were typically ignored in previous speech recognition research, such as noise and reverberation, appear repeatedly in recent papers. I would like to sincerely thank the contributing authors, for their effort to bring their insights and perspectives on current open questions in speech recognition research.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:
