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,

Source: Advances in Speech Recognition, Book edited by: Noam R. Shabtai,
ISBN 978-953-307-097-1, pp. 164, September 2010, Sciyo, Croatia, downloaded from SCIYO.COM

1965]). 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]

$$a = 1 - |R|^2 \tag{1}$$

and thus the absorption area is

$$A = \bar{a}S \tag{2}$$

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]

$$e\,(t) = 10\log_{10}\int_t^{\infty} h^2(\tau)\,\mathrm{d}\tau - 10\log_{10}\int_0^{\infty} h^2(\tau)\,\mathrm{d}\tau \tag{3}$$

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

$$e\,(\mathrm{RT}) = -60\mathrm{dB}. \tag{4}$$

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]

$$\mathrm{RT} = 0.161\frac{V}{A}. \tag{5}$$

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 ΔMFCC feature vectors, such that

$$\mathbf{c}_t = [c_1^t \ldots c_N^t, \Delta c_1^t \ldots \Delta c_N^t]^{\mathrm{T}} \tag{6}$$

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]

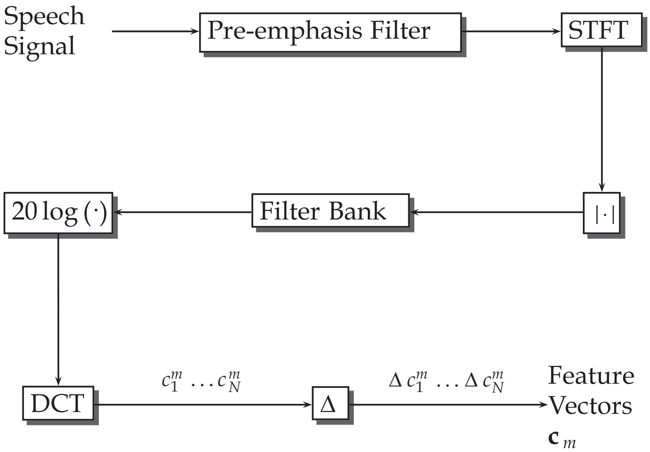


Fig. 1. Extraction of MFCC and ΔMFCC feature vectors from speech signal [Bimbot et al., 2004].

$$\tilde{\mathbf{c}}_t = \mathbf{c}_t - \boldsymbol{\mu} \qquad t = 0 \ldots T-1 \tag{7}$$

where $\boldsymbol{\mu}$ is the sample mean of the series $\mathbf{c}_0 \ldots \mathbf{c}_{T-1}$. The operation of CMS may include variance normalization [Mammone et al., 1996] by dividing the components by the sample *standard deviation* (STD), i.e.,

$$\bar{c}_n^{\,t} = \frac{\tilde{c}_n^{\,t}}{\sigma_n} \qquad \begin{array}{lcl} t & = & 0 \ldots T-1 \\ n & = & 1 \ldots N \end{array} \tag{8}$$

where for every $n = 1 \ldots N$, σ_n is the sample STD of the series $c_n^0 \ldots c_n^{T-1}$.

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

$$\mathbf{X} = [\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{T-1}] \tag{9}$$

be a segment of speech feature vectors $\mathbf{x}_t$ 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 λ_1 is defined to contain the parameters such that a parametric *probability density function* (PDF) $p(\mathbf{X}; \lambda_1)$ would model the conditional PDF $p(\mathbf{X} | H_1)$. In a similar way, λ_0 is defined such that $p(\mathbf{X}; \lambda_0)$ models $p(\mathbf{X} | H_0)$. For example, if the models assume Gaussian distribution, then λ_0 and λ_1 consist of a mean vector and a covariance matrix.

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

$$\Lambda(\mathbf{X}) = \log p(\mathbf{X}; \lambda_1) - \log p(\mathbf{X}; \lambda_0) \begin{cases} \geq \theta & \text{accept hypothetical speaker} \\ < \theta & \text{reject hypothetical speaker} \end{cases} \tag{10}$$

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

$$\log p(\mathbf{X}; \lambda) = \sum_{t=0}^{T-1} \log p(\mathbf{x}_t; \lambda). \tag{11}$$

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

$$\tilde{\Lambda}(\mathbf{X}) = \frac{1}{T} \Lambda(\mathbf{X}), \tag{12}$$

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

According to the GMM approach, if $\mathbf{x}$ is a feature vector, and λ is a set of parameters, then

$$p(\mathbf{x}; \lambda) = \sum_{i=1}^{M} \omega_i p_i \left(\mathbf{x}; \lambda^{(i)} \right) \tag{13}$$

where M is the number of Gaussians in the GMM, or, the model order, the weights ω_i apply

$$\sum_{i=1}^{M} \omega_i = 1, \tag{14}$$

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

$$p_i(\mathbf{x}) = p(\mathbf{x}\ ;\boldsymbol{\mu}_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{d}{2}} \left|\Sigma_i\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}} \Sigma_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)} \tag{15}$$

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

$$\lambda = \{\omega_i, \boldsymbol{\mu}_i, \Sigma_i\}_{i=1\ldots M} \tag{16}$$

The parameters ω_i, $\boldsymbol{\Sigma}_i$, and $\boldsymbol{\mu}_i$ are estimated using the *expectation maximization* (EM) algorithm [Dempster et al., 1977]. The covariance matrix $\boldsymbol{\Sigma}_i$ 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*, $\mathrm{P_{MISS}}$, which is the probability that a target model was rejected.

$$\mathrm{P_{MISS}} = \ \mathrm{P}(\Lambda(\mathbf{X}) < \theta\ \mid \text{target}), \tag{17}$$

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

$$\mathrm{P_{FA}} = \mathrm{P}(\Lambda(\mathbf{X}) > \theta\ \mid \text{impostor}). \tag{18}$$

Both $\mathrm{P_{MISS}}$ and $\mathrm{P_{FA}}$ are functions of the threshold θ, and they each come at the expense of the other. The threshold θ is used as a parameter to yield the *detection error trade-off* (DET) curve, which plots $\mathrm{P_{MISS}}$ as a function of $\mathrm{P_{FA}}$. The point on the DET curve where $\mathrm{P_{MISS}}$ equals $\mathrm{P_{FA}}$ 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:

$$D = \sum_{i=1}^{M} \omega_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})^{\mathrm{T}} (\boldsymbol{\mu}_i - \boldsymbol{\mu}), \tag{19}$$

where M is the GMM order, ω_i is the weight of the i'th Gaussian, $\boldsymbol{\mu}_i$ is the mean vector of the i'th Gaussian, and $\boldsymbol{\mu}$ 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

$$D_{\text{norm}} = \frac{D}{D_{\text{RT}=0}} \tag{20}$$

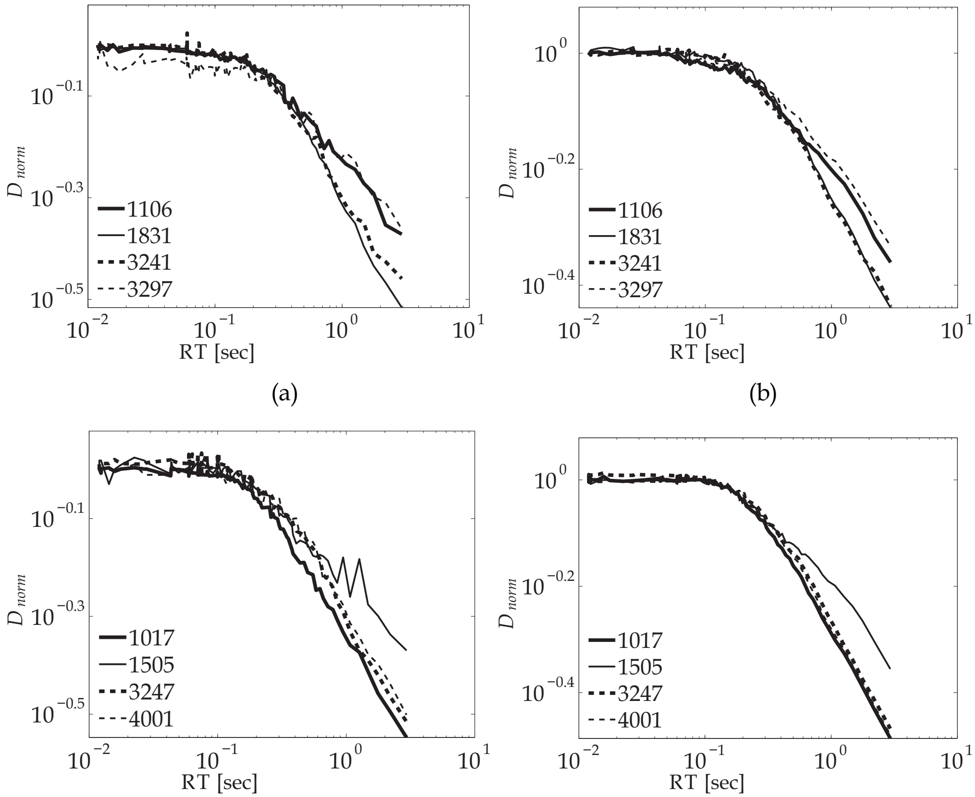


Fig. 2. Normalized distance between Gaussians and overall mean in a GMM of different speakers. Numbers in the legend indicate speaker index in NIST-SRE database. (a) 10 Gaussians, male speakers; (b) 50 Gaussians, male speakers; (c) 10 Gaussians, female speakers; and (d) 50 Gaussians, female speakers.

was used, where $D_{\mathrm{RT}=0}$ 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

$$\mathrm{BIC}_{\lambda_M,\mathbf{X}} = -\log p(\mathbf{X}|\lambda_M) + \frac{1}{2}M(2d+1)\log N \tag{21}$$

$$\mathrm{AIC}_{\lambda_M,\mathbf{X}} = -\log p(\mathbf{X}|\lambda_M) + M(2d+1) \tag{22}$$

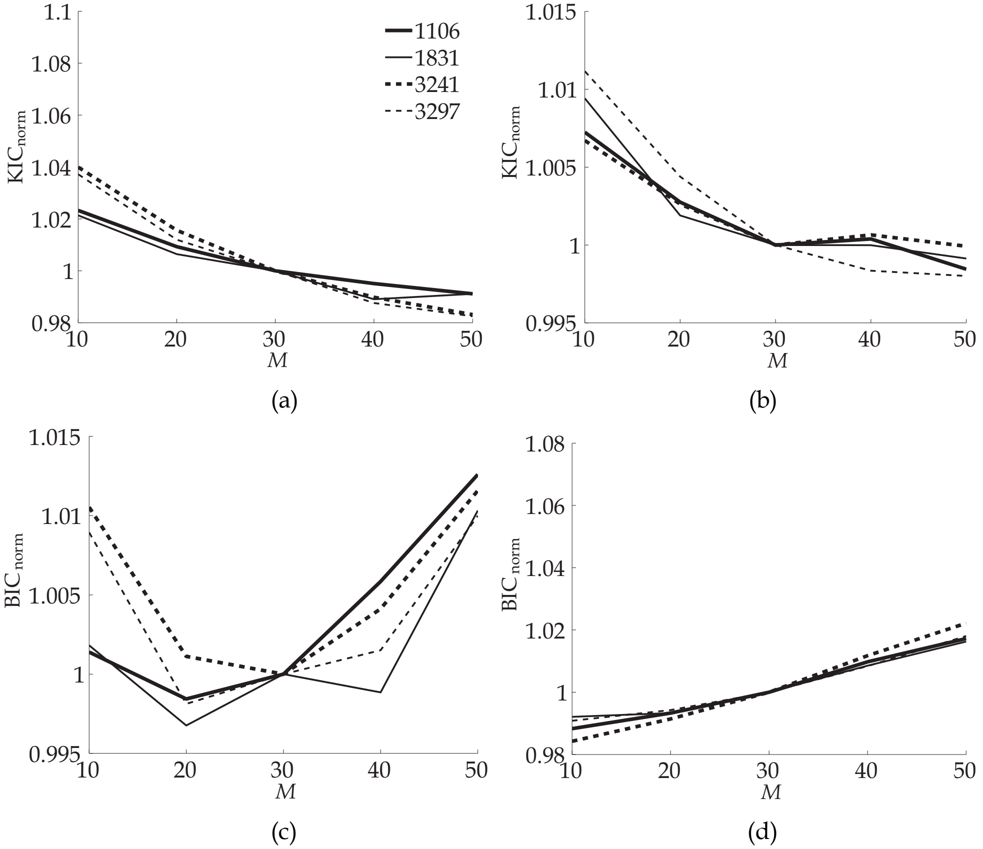
$$\mathrm{KIC}_{\lambda_M,\mathbf{X}} = -\log p(\mathbf{X}|\lambda_M) + \frac{3}{2}M(2d+1) \tag{23}$$

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

$$M^\star_{\mathrm{IC}} = \arg \min_{M\in\{10,20,30,40,50\}} \mathrm{IC}_{\lambda_M,\mathbf{X}} \tag{24}$$

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 $\mathrm{IC}_{\mathrm{norm}}$. 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 ($M^\star_{\mathrm{KIC}}$ 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 ($M^\star_{\mathrm{BIC}}$ in Fig. 3(c)), whereas for reverberant speech with RT=0.85 sec (Fig. 3(d)) it reduces to 10. The general effect of



(a)

(b)

(c)

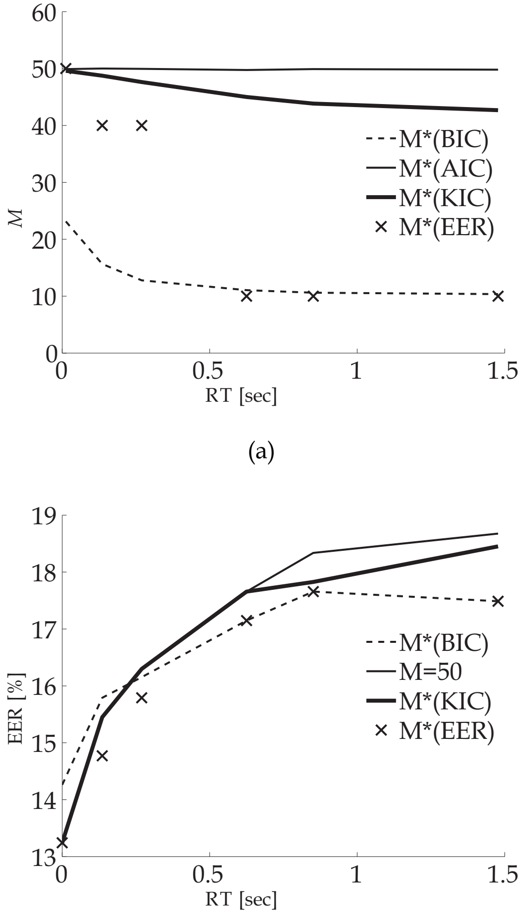
(d)

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.



(a)

(b)

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

NIST-98 SRE database, using reverberant speech with the same RT of the test speech, and a constant model order of 256 Gaussians. No channel compensation was used.

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, $M^{\star}_{\mathrm{KIC}}$ is similar to the optimal model order in terms of minimum EER. For RT > 0.5 sec, $M^{\star}_{\mathrm{BIC}}$ is similar to the optimal model order in terms of minimum EER. $M^{\star}_{\mathrm{AIC}}$ 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 $M^{\star}_{\mathrm{KIC}}$ for RT < 0.5 sec, but worse than using $M^{\star}_{\mathrm{BIC}}$ for RT > 0.5 sec. Since $M^{\star}_{\mathrm{BIC}}$ 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 & Berkley, 1979]. It was shown that for high RTs, the efficiency of 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 Tab. 1. Measured RIRs 1 ÷ 10 were measured with Brüel & Kjær 4295 Omni-Source loudspeaker and Brüel & Kjær 4942 $\frac{1}{2}$-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

RIR	Environment	RT [sec]	V_i[m^3]
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

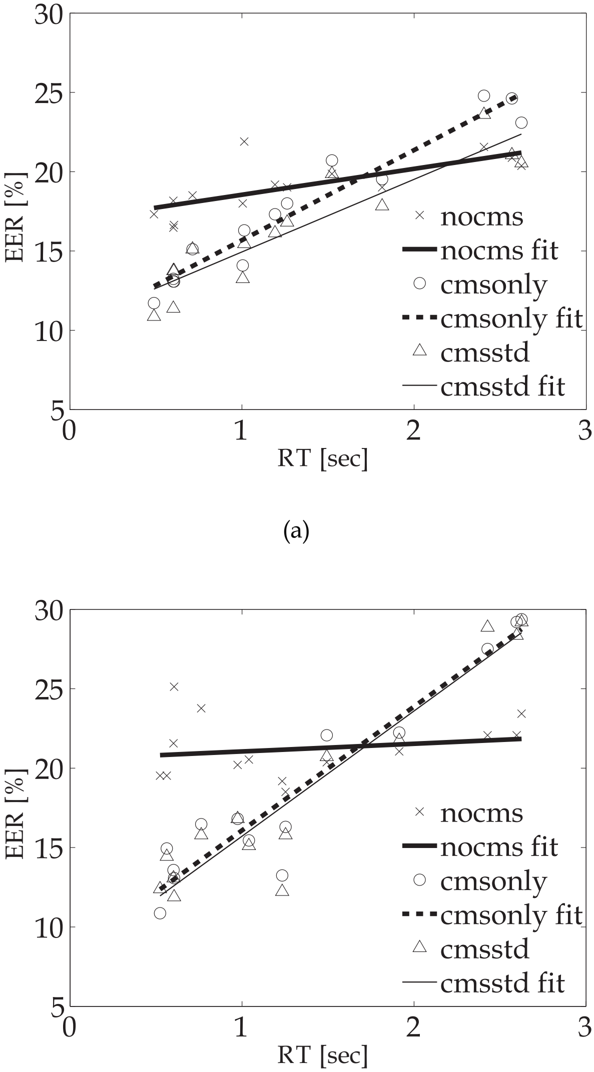
Table 1. Rooms in which RIRs were measured.

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.



(a)

(b)

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

[CHR, 2004] (2004). *Concert Hall Research Group CD v.3*. Concert Hall Research Group, 327F Boston Post Road. Sudbury, MA 01776. Attention: Timothy J. Foulkes, email: chrg@cavtocci.com, phone: 978.443.7871, fax: 978.443.7873.

[Allen and Berkley, 1979] Allen, J. B. and Berkley, D. A. (1979). Image method for efficiently simulating small room acoustics. *J. Acoust. Soc. Am.*, 65(4):943–950.

[Bimbot et al., 2004] Bimbot, F., Bonastre, J. F., Fredouille, C., Gravier, G., Meignier, S., Merlin, T., Garcia, J. O., Chagnolleau, I. M., Delacretaz, D. P., and Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *EURASIP J. Appl. Signal Processing*, 2004(4):430–451.

[Chen and Bilmes, 2007] Chen, C. P. and Bilmes, J. A. (2007). MVA processing of speech features. *IEEE Trans. Speech Audio Process.*, 15(1):257–270.

[Chen and Huang, 2005] Chen, H. and Huang, S. (2005). A comperative study on model selection and multiple model fusion. In *Proc. FUSION*, volume 1, pages 820–826.

[Davis and Mermelstein, 1980] Davis, S. B. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-28(4):357–366.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.*, 39(1):1–38.

[ISO 3382:1997, 1997] ISO 3382:1997 (1997). Acoustics – measurement of the reverberation time of rooms with reference to other acoustical parameters.

[Kinsler et al., 2000] Kinsler, L. E., Frey, A. R., Coppens, A. B., and Sanders, J. V. (2000). *Fundamentals of Acoustics*. John Wiley, New York.

[Kuttruff, 2000] Kuttruff, H. (2000). *Room Acoustics*. Spon Press, New York.

[Mammone et al., 1996] Mammone, R. J., Zhang, X., and Ramachandran, R. P. (1996). Robust speaker recognition: a feature-based approach. *IEEE Signal Process. Mag.*, 13(5):58–71.

[Martin and Przybocki, 2000] Martin, A. and Przybocki, M. (2000). The NIST 1999 speaker recognition evaluation - an overview. *Digital Signal Processing*, 10(1–3):1–18.

[Peer et al., 2008] Peer, I., Rafaely, B., and Zigel, Y. (2008). Reverberation matching for speaker recognition. In *Proc. ICASSP*, pages 4829– 4832.

[Reynolds et al., 2000] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1):19–41.

[Schroeder, 1965] Schroeder, M. R. (1965). New method of measuring the reverberation time. *J. Acoust. Soc. Am.*, 37(3):409–412.

[Shabtai et al., 2008a] Shabtai, N. R., Rafaely, B., and Zigel, Y. (2008a). The effect of GMM order and CMS on speaker recognition with reverberant speech. In *Proc. HSCMA*, pages 144–147.

[Shabtai et al., 2008b] Shabtai, N. R., Rafaely, B., and Zigel, Y. (2008b). The effect of room parameters on speaker verification using reverberant speech. In *Proc. IEEEI*, pages 231–235.

[Zigel and Wasserblat, 2006] Zigel, Y. and Wasserblat, M. (2006). How to deal with multiple targets in speaker identification systems? In *Proc. IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, pages 1–7.

Advances in Speech Recognition
Edited by Noam Shabtai

ISBN 978-953-307-097-1
Hard cover, 164 pages
Publisher Sciyo
Published online 16, August, 2010
Published in print edition August, 2010

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.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821