Geometric contamination for GMM/UBM speaker verification in reverberant environments

Alessio Brutti, Maurizio Omologo

Fondazione Bruno Kessler-CIT Irst, via Sommarive 18 Trento, Italy

brutti@fbk.eu, omologo@fbk.eu

Abstract

Reverberation generated by multi-path acoustic propagation in enclosures is one of the most critical issues for distant-speech speaker verification systems. While late arrivals can be treated as additive noise, early reflections critically affects the speech spectral properties that allow differentiating among speakers. Considering a standard GMM/UBM speaker verification system based on MFCC, a geometric data contamination scheme relying on a rough modeling of the sound propagation through the image method is presented to attack the reverberation. The contribution of this paper is twofold: first, an analysis of the amount of knowledge of the impulse response needed for an effective contamination is carried out on simulated data; second, the geometric contamination is applied to real impulse responses to validate its use in real application contexts. Preliminary experimental results are provided to support the theoretical study and show the effectiveness of the approach.

Index Terms: speaker verification, distant speech, data contamination

1. Introduction

The goal of text-independent speaker verification is to validate the claimed identity of a speaker without knowing the spoken utterance [1, 2]. This technology is crucial in many speech enabled applications, in particular when access permission and user profiling are concerned. To this regard, an emerging application field is related to home automation where a natural voice interaction with the house appliances is allowed, as envisaged in the EC funded DIRHA project[1]. In this scenario, verification is performed in distant-talking mode; under these conditions non-stationary environmental noise and reverberation represent crucial issues to address.

So far the problem of speaker verification has been mainly addressed considering telephone or close-talking recordings, typically using Gaussian Mixture Model (GMM) in combination with Mel Frequency Cepstral Coefficient (MFCC) [3]. Only recently, solutions have been investigated to tackle the effects of reverberation. A first set of approaches focus on modifying the test signals in the feature domain. Examples are: Cepstral Mean Subtraction (CMS) [4], RASTA processing [5], Feature Warping [6] and long-term feature normalization [7]. Other works have addressed the reverberation problem by means of multichannel processing [8, 9, 10]. Some research activities instead focus on different feature sets or classifiers which may result less sensitive to noise and reverberation than the traditional GMM based on MFCC: Support Vector Machines (SVM) [11], Joint Factor Analysis (JFA) [12], Gaussian SuperVectors (GSV) [13] and Total Variability (TV) [14].

[1] See http://dirha.fbk.eu for details.

As an alternative to the methods listed above, noise and reverberation can also be tackled through model adaptation by using specific training material matching the test conditions [15]. Similarly, data contamination [16] is a technique that exploits some information about the real environment, such as the impulse responses and the background noise level of the room, in order to reduce the mismatch between the training and test conditions. The main drawback of these approaches is the need of specific material. This solution is not feasible in a domestic scenario since it would require acquiring training material, or measuring impulse responses, for each possible condition and position of speaker and microphone.

Following the strategy adopted in [17], we try to compensate the effects of reverberation through a data contamination approach. But instead of measuring the real Room Impulse Response (RIR) between the source and the microphone, we estimate the RIR through the method of image sources [18], which describes the acoustic propagation with a simple geometric model. With respect to traditional data contamination strategies, this way we avoid the time consuming activity of acquiring specific material or measuring real impulse responses.

Similar approaches have been recently investigated in [19, 20, 21] where multiple speaker models are trained from some previously reverberated material (either real or simulated), representing a variety of environmental conditions and source positions. In particular, in [20] contaminated models based on artificial RIRs created with the image method are used in real rooms. With respect to these works, the approach described here presents several differences. First of all, we only focus on early arrivals which affect directly the speaker related spectral characteristics, without considering the reverberation tail: this reduces the mismatch between the real recordings and the ideal conditions reproduced by the application of the image method. Secondly, we employ an enhanced version of the image method which accounts for source orientation and directivity. Finally, we do not consider multiple models, and the related selection.

2. Multi-path Acoustic Propagation

In enclosures, acoustic waves propagate through multiple paths due to the presence of reflecting surfaces (e.g. walls). This results in multiple replica of the emitted signal reaching the microphone, each one with a given delay and attenuation depending on the specific propagation path. The multi-path propagation, which results in the so-called reverberation, is detrimental for distant-speech related applications. The effects of the enclosure acoustics are described through the RIR h:

$$y(t) = h * s(t) + \eta(t), \tag{1}$$

where the RIR is assumed time-invariant, $*$ denotes convolution, $s(t)$ is the emitted speech signal, $y(t)$ is the reverberated

signal captured by the microphone and $\eta(t)$ is the environmental noise. In this paper we are not interested in addressing the environmental noise and the term $\eta(t)$ will be discarded hereafter.

It is a common approach to model the RIR as a sequence of pulses representing all the possible propagation paths connecting the source to the microphone:

$$h(\tau) = \sum_k \beta_k \delta\left(\tau - \frac{d_k}{c}\right), \tag{2}$$

where c is the speed of sound and d_k is the length of the k-th path. The attenuation β_k includes the propagation loss, the directivity gains of the microphone and the source and the energy absorption due to reflections [22]. This representation is particularly effective in modeling the direct path and early arrivals, which are in general strong and sparse and hence can be described with a linear system. Hence, it is often convenient to split the RIR into two parts [23]:

$$h(\tau) = \sum_{k=0}^{K-1} \beta_k \delta\left(\tau - \frac{d_k}{c}\right) + h_r(\tau), \tag{3}$$

where K early arrivals are considered and $h_r(\tau)$ represents the late diffuse reverberation in the RIR tail. In this work we are mainly interested in modeling the first part related to early arrivals, which introduce a reshaping of the signal spectrum depending on the inter-path delay and relative attenuation. In the simple case where only one reflected path is present together with the direct path, the RIR behaves as a comb filter [22]. When multiple reflected paths reach the microphone, with varying attenuation, the complex combination of each comb filter results in an irregular sequence of peaks and deeps in the amplitude of the RIR frequency response. These fluctuations affect the speaker specific information, e.g. the formants and pitch, reducing the performance of verification algorithms. Conversely, the reverberation tail produces a time smearing of the spectrum which is not so crucial for the verification task, since the likelihoods are typically evaluated on the full input sentence.

3. Geometric Data Contamination

For an empty room, assuming that geometry, microphone location, source positions and orientation as well as wall reflection coefficients are known, realistic RIRs can be artificially generated using the image method [18] by mirroring the microphone (or the source) with respect to the 6 surfaces of a parallelepipedic enclosure up to the images of order I. With image order we denote the number of reflections in the corresponding propagation path. If we limit our study to the main reflections, the image method can be used to approximate the real RIR as:

$$\hat{h}(\tau) = \sum_{i=0}^{N_I-1} \hat{\beta}_i \delta\left(\tau - \frac{\hat{d}_i}{c}\right), \tag{4}$$

where i indexes the image sources while $\hat{\beta}_i$ and $\hat{d}_i$ are obtained by the image method technique. In our implementation, the path attenuation $\hat{\beta}_i$ is derived using a modified version of the original formulation, which accounts for source directivity and orientation through a parameterized acoustic radiation model [24]. N_I in eq. 4 is the number of propagation paths with order lower or equal than I: in our work we consider relatively small values for I (smaller than 20) to address only early arrivals and neglect aspects related to diffuse reverberation.

Similarly to what done in [17], we use the approximated RIR to contaminate the clean signals and try to compensate the mismatch between the models and the test signals. For each training and enrollment sentence $s_T(t)$ we create a contaminated version through $\hat{h}$:

$$s_T^C\left(t, \hat{h}\right) = \hat{h} * s_T(t), \tag{5}$$

which will be used in model training. Note that with respect to [17], environmental noise is not included in the contamination procedure, i.e. only reverberation effects are here addressed.

4. GMM-UBM speaker verification

To evaluate the effectiveness of the proposed data contamination approach, a traditional GMM-UBM [3] system is adopted. Note that the contamination scheme can be couple to any other, possibly more effective, verification system. The GMM-UBM approach was adopted as a standard benchmark to evaluate our contamination strategy. The feature distribution of each speaker is modeled by means of a GMM trained on speaker specific material; the distribution of any other speaker is modeled by a Universal Background Model (UBM). A GMM is completely identified by the weights w_n, the mean vector μ_n, the covariance matrix Σ_n and the number of Gaussian components G ($n = 1, \cdots, G$). In the following, the GMM of the l-th speaker and the UBM are denoted as $\Gamma_l = (w_{l,n}, \mu_{l,n}, \Sigma_{l,n})$ and $\Gamma_U = (w_{U,n}, \mu_{U,n}, \Sigma_{U,n})$ respectively. In the GMM-UBM approach a set of sentences that are representative of the population is used to train a UBM through the iterative Expectation Maximization (EM) algorithm. Then, speaker dependent models are adapted using the Maximum A Posteriori (MAP) algorithm applied only to the UBM mean vector $\mu_{U,n}$.

The adopted front-end samples the incoming speech signals at 16 kHz and codes them into a 30 dimensional feature vector sequence composed of 15 MFCC with their first order derivatives. The analysis window length is 20 ms and the analysis step is 10 ms. Pre-emphasis is applied with a first order FIR filter: $H_{\text{pe}}(z) = 1 - 0.97z^{-1}$.

For a feature vector $\mathbf{x}$, the corresponding likelihood given the model Γ_l is defined as:

$$p(\mathbf{x}|\Gamma_l) = \sum_{n=1}^{G} w_{l,n} p_{l,n}(\mathbf{x}), \tag{6}$$

where $p_{l,n}(\mathbf{x})$ is the n-th multidimensional Gaussian probability density function with mean vector $\mu_{l,n}$ and covariance matrix $\Sigma_{l,n}$. In general, the speaker identity is verified on the full spoken sentence. Denoting with $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$ the feature vector sequence, where T is the number of feature vectors extracted from the utterance, the average log-likelihood given the model Γ_l is:

$$\mathcal{L}(\mathbf{X}|\Gamma_l) = \frac{1}{T} \sum_{t=1}^{T} \log p\left(\mathbf{x}_t|\Gamma_l\right). \tag{7}$$

The resulting log-likelihood is then compared with the UBM log-likelihood, through a likelihood ratio:

$$\lambda(\mathbf{X}|\Gamma_l) = \mathcal{L}(\mathbf{X}|\Gamma_l) - \mathcal{L}(\mathbf{X}|\Gamma_U). \tag{8}$$

If $\lambda(\mathbf{X}|\Gamma_l)$ is above a given threshold the speaker identity l is accepted. Figure 1 sketches a block diagram of the proposed contamination and verification system.

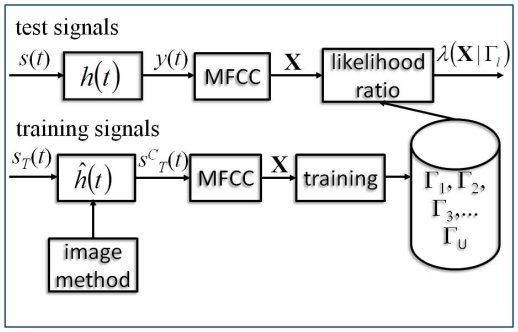

Figure 1: Simplified block diagram of the proposed speaker verification system.

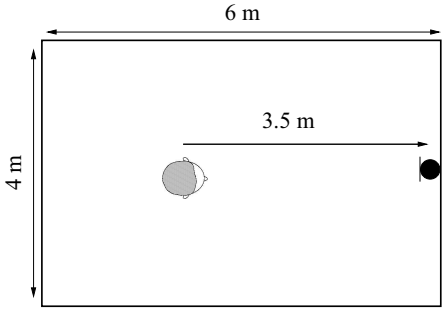

Figure 2: Map of the set up adopted in the generation of the simulated database.

5. Experiments and Results

An extended version of the Acoustic Phonetic And spontaneous Speech Corpus of Irst (APASCI) database was used to evaluate the impact on speaker verification performance of the proposed geometric contamination procedure. The database includes 164 speakers, each of them uttering about 20 italian phonetically rich sentences. The average utterance length is about 4.8 seconds. 80 speakers were selected for training of the UBM, resulting in 1710 utterances, while the remaining speakers were chosen for individual model adaptation and testing purposes. For each speaker, 10 utterances were selected for the enrollment stage to adapt models from the UBM. The remaining utterances were used for testing, resulting in 928 correct user trials and 77024 impostor trials. Following a best practice in speaker verification, performance is evaluated in terms of Detection Error Tradeoff (DET) curves and Equal Error Rate (EER). In the next sections, we report experimental results on artificial as well as real data.

5.1. Simulated Data

As a first study, we consider artificially generated data. The goal of this experimental analysis is to understand if modeling the early reflections only is actually effective and what is a good choice for I. The APASCI sentences were reverberated with RIRs generated with the image method. Various reverberation times T_{60} were considered for a room whose dimensions are $6\times4\times4$ meters. The source was placed in a central position and oriented toward the microphone which was mounted on one of the walls at 3.5m from the source, as depicted in Figure 2. Note that given the same set up, different reverberation times are obtained by properly changing the wall absorption coefficients (we assume identical coefficients for all walls). Experiments varying the position of the source led to a similar trend of results, and hence are not reported in this work.

The training material was created by reverberating the dry signals with shorter RIRs, obtained with the same image method procedure, but limiting the maximum reflection order I (basically the RIR is truncated up to a maximum length).

Figure 3 reports the EER obtained when G=256 for 4 reverberation times and various I. In the figure, "mismatch" refers to model training on clean material and test on the reverberated speech while "match" indicate experiments in perfectly matched conditions where the same full RIR is used for reverberating the training and test material. First of all note the regular decreasing trend of the EER as I increases. When $T_{60} = 0.2$s, considering only 6-th order images is enough to get almost the same performance as for the matched case ($I \simeq 75$ in the full RIR). For the mid-low reverberation time $T_{60} = 0.4$s, the RIR truncated at $I = 8$ almost achieves the minimum EER. For higher T_{60} larger I are needed. However, also in these challenging cases a considerable improvement with respect to the mismatch case is obtained for values of I around 8. Note that, in this setup, when $I = 8$ the longest path takes approximately 157ms to reach the microphone. Finally, it is worth noting the gap in performance for the 4 reverberation times in the EER with matched models. It clearly shows the effects of the diffuse reverberation which just introduces entropy in the feature vectors and cannot be modeled even with full knowledge of the RIR (in this case an inverse filter dereverberation strategy would be more successful).

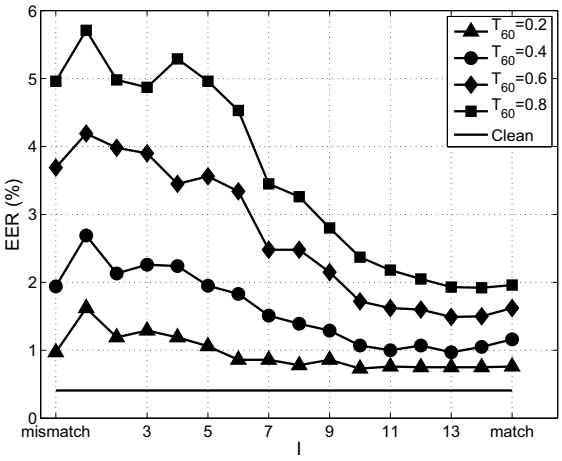

Figure 3: EER on the simulated data varying I and T_{60} for G=256. The continuous line at the bottom is the lower bound obtained on the clean material. "Mismatch" refers to models trained on clean material and test performed on reverberated speech. "Match" refers to experiments in matched conditions.

5.2. Experiments with real RIRs

Although interesting from a theoretical point of view, the previous analysis requires a perfect knowledge of the RIR up to a given reflection order I. In this section we investigate the applicability of the proposed approach to real RIRs. Given knowledge of the room dimensions and of source and microphone positions, an artificial RIR is generated by the image method procedure. With respect to the simulated data, we introduce here a further degree of uncertainty since the artificial RIR is an estimation based on a simplified propagation model. In this experiment the nominal values for the source and microphone positions as well as for the room dimensions are available. In a more general framework, the position and orientation of the speaker could be estimated automatically [25].

To conduct this analysis the APASCI database was reverberated using a RIR measured in the living room of a real apartment through the use of specific exponential sweep excitation signals [26]. The dimensions of the room are similar to those adopted in the simulations and the reverberation time ranges between 0.6 and 0.8s, depending on the source position. Real environmental noise was then added creating a Signal-to-Noise Ratio (SNR) of about 30 dB. The geometrically based RIRs for data contamination were created for various image orders and two reverberation times: 0.3s and 0.6s.

Figure 4 shows the DET curves for T_{60}=0.6s and for I equal to 4, 18 and for the full RIR, in comparison with the match and mismatch cases, when the GMM models include G=256 gaussians.

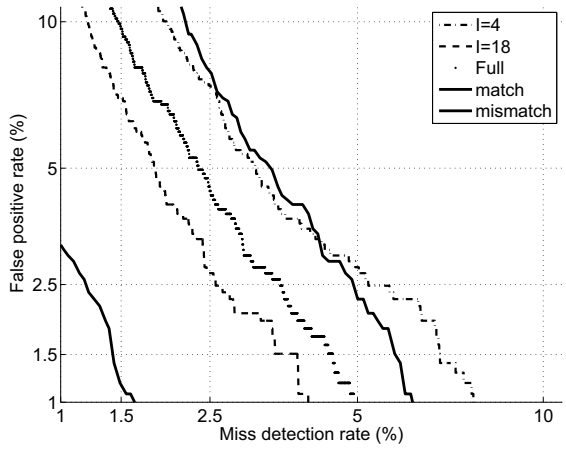

Figure 4: DET curves on the real data for T_{60}=0.6s. To make easier the interpretation of this figure, only curves for I=4, I=18 and for the full RIR are shown together with the match and mismatch cases.

Figure 5 reports the EER for the two reverberation times. The two lines (continuous and dashed dotted) represent the lower and upper bounds obtained under matched and mismatched conditions respectively. Note that a gain with respect to the mismatch case is achieved for I=10, while using larger image orders does not provide any significant improvement. The best performance is obtained using 0.6s as reverberation time and I=18: the EER is reduced from 4.01% to 2.57%. Interestingly, the quality of the contaminated material seems to be not so sensitive to the selected reverberation time in the artificial RIR generation, with models for T_{60}=0.3s performing slightly worse than those obtained with T_{60}=0.6s. This confirms the

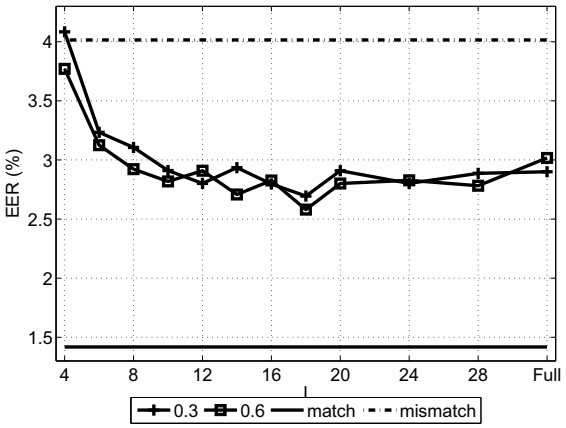

Figure 5: EER on the real RIR data. The two lines (continuous and dashed dotted) represent the lower and upper bounds in matched and mismatched conditions. RIRs are estimated using two T_{60} and various image orders.

fact that the effect of early arrivals is dominant in the speaker verification task. As a matter of facts, the amplitude of low order reflections is mostly influenced by the room geometry rather than by the reflection coefficients, which are instead crucial for the reverberation tail.

6. Conclusions

This paper addressed the effects of early arrivals on a GMM-UBM approach for speaker verification in reverberant speech. A data contamination scheme based on artificially generated RIRs through the image method is presented to account for the spectral reshaping due to the early reflections. Experimental results on simulated data justify the idea of focusing on early arrivals only, rather than including the reverberation tail. Further experiments on real RIRs confirm the feasibility of the approach even in real application contexts.

One of the main drawbacks of the proposed approach is the time consuming re-training on the contaminated data which is required as soon as the speaker position or the room geometry change. This limitation could be addressed by considering a spatial grid of possible source positions for which pre-computed models are available or by using adaptation, starting from some general models, instead of training from scratch.

A further aspect to address in future works is the robustness against the accuracy of the source and microphone positions as well as of the room geometry. This is of particular interest because in real applications the source position and orientation are not known (they can be obtained employing a source localization algorithm) and the room layout is typically available with nominal and not accurate measures.

Finally, the image method could be replaced by an automatic estimation of the main propagation paths (i.e., dominant reflections) in the RIR.

7. Acknowledgments

The research leading to these results has partially received funding from the European Union's 7th Framework Programme (FP7/2007-2013) under grant agreement n. 288121 - DIRHA.

8. References

[1] S. Furui, *An overview of speaker recognition technology*, ser. In C.-H. Lee, F. K. Soong and K.K. Paliwal (Eds.), Automatic speech and speaker technology. Boston: Kluwer Academic, 1996.

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4, 2004.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, pp. 19–41, 2000.

[4] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254 – 272, apr 1981.

[5] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578 –589, oct 1994.

[6] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proceedings of Speaker Odyssey 2001 conference*, 2001, pp. 213–218.

[7] S. Ganapathy, J. Pelecanos, and M. Omar, "Feature normalization for speaker verification in room reverberation," in *ICASSP*, 2011, pp. 4836 –4839.

[8] J. Qin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2023–2032, September 2007.

[9] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *ICSLP*, 1996.

[10] I. McCowan, J. Pelecanos, and S. Sridharan, "Robust speaker recognition using microphone arrays," in *Proceedings of Speaker Odyssey 2001 conference*, 2001, pp. 101–106.

[11] V. Wan and W. Campbell, "Support vector machines for speaker verification and identification," in *Proceedings of the 2000 IEEE Signal Processing Society Workshop*, vol. 2, 2000, pp. 775–784.

[12] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1435 –1447, may 2007.

[13] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

[14] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Interspeech*, 2009, pp. 1559–1562.

[15] K. K. Yiu, M. W. Mak, and S. Y. Kung, "Environment adaptation for robust speaker verification," in *Eurospeech*, 2003, pp. 2973–2976.

[16] M. Matassoni, M. Omologo, D. Giuliani, and P. Svaizer, "HMM-training with contaminated speech material for distant-talking speech recognition," *Computer Speech and Language*, vol. 16, pp. 205–223, 2002.

[17] C. Zieger and M. Omologo, "Combination of clean and contaminated GMM/SVM for far-field text-independent speaker verification," in *Interspeech*, 2008, pp. 1949–1952.

[18] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of Acoustic Society of America*, vol. 65, no. 4, pp. 943–950, April 1979.

[19] J. Gammal and R. Goubran, "Combating reverberation in speaker verification," in *IEEE Proceedings of Instrumentation and Measurement Technology Conference*, vol. 1, May 2005, pp. 687–690.

[20] P. De Leon and A. Trevizo, "Speaker identification in the presence of room reverberation," in *Biometrics Symposium, 2007*, sept. 2007, pp. 1 –6.

[21] A. Akula, V. Apsingekar, and P. De Leon, "Speaker identification in room reverberation using GMM-UBM," in *Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009*, 2009, pp. 37 –41.

[22] H. Kuttruff, *Room Acoustics*. Elsevier Applied Science, 1991.

[23] T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 791 – 803, November 2003.

[24] A. Brutti, M. Omologo, and P. Svaizer, "An environment aware ML estimation of acoustic radiation pattern with distributed microphone pairs," *Signal Processing*, vol. 93, no. 4, pp. 784–796, apr 2013.

[25] A. Brutti, M. Omologo, and P. Svaizer, "Speaker localization based on oriented global coherence field," in *Interspeech*, 2006.

[26] M. Ravanelli, A. Sosi, P. Svaizer, and M. Omologo, "Impulse response estimation for robust speech recognition in a reverberant environment," in *EUSIPCO*, 2012, pp. 1668 –1672.