ON THE USE OF EARLY-TO-LATE REVERBERATION RATIO FOR ASR IN REVERBERANT ENVIRONMENTS

Alessio Brutti and Marco Matassoni

Fondazione Bruno Kessler, Center for Information and Communication Technology, Trento, Italy

{brutti,matasso}@fbk.eu

ABSTRACT

This work presents an analysis of distant-talking speech recognition in a variety of reverberant conditions, correlating ASR performance to the acoustic characteristics of a given propagation channel. In particular we show how, for a digit recognition task, the ASR accuracy is directly related to the Early-to-Late Reverberation ratio of the room impulse response, capturing in a single parameter the reverberation properties of a given channel independently of the setup. Consequently, this measure can be successfully considered for acoustic model training either selecting the most suitable model for a given spatial configuration, or defining the subset of RIRs to be used for the creation of multi-condition models. Experimental results on simulated data as well as on data generated with real impulse responses support our claims.

Index Terms— distant ASR, reverberation, room impulse response, direct-to-reverberant ratio, multi-condition training

1. INTRODUCTION

Distant speech recognition is progressively gaining major attention since the usage of close-talking microphones is inconvenient or difficult in specific scenarios [1]. As a result, distant speech is usually acquired by means of microphone arrays, which allow the implementation of selective spatial acquisition or other speech enhancement techniques [2], but have a restricted view of the space of interest. Alternately, networks of distributed microphones could be employed guaranteeing a uniform acoustic coverage of the monitored area independently of the source position and orientation. This scenario is being investigated under the EU project DIRHA where a vocal system for the control of home devices is targeted.

In such configurations the adoption of array-processing techniques is often neither possible nor effective [3] and alternative approaches can be successfully applied, as for example channel selection [4] or source separation [5, 6]. It is well known that reverberation and background noise degrade speech recognition performance, but few studies have investigated the relation between acoustic conditions and recognition rates in a multi-microphone scenario, where the speaker position is not constrained or known in advance. Indeed, in this distributed setup, various spatial factors impact on the signal quality: the distance and the orientation of the pair source-microphone, the consequent different SNR, the acoustic propagation in the enclosure. Hence it is of interest to correlate ASR performance with some purely acoustic measurements [7]. In [8], the authors proposed a temporal-domain method for predicting recognition performance in unseen noisy environments. This estimate can be usefully exploited during setup to increase the robustness of the resulting system, for example selecting or training a suitable acoustic model. Authors in [9] studied the harming parts of room impulse responses, discussing the contribution of early and late reflections to ASR performance while in [10] the inter-frame correlation of reverberant feature vectors is analyzed. In a related work the adjustments of dereverberation algorithms to ASR systems are evaluated [11]. More recently the reverberation problem has been addressed from different perspectives [12, 13, 14, 15] showing the need for effective solutions to cope with the related masking effects.

This work presents an analysis of the correlation between the accuracy of distant-talking speech recognition and the properties of the acoustic channel. The main assumption is that the ASR performance is directly related to the Early-to-Late Reverberation Ratio (ELR) of the Room Impulse Response (RIR), whose correlation with speech and music clarity has been already investigated [16]. In this way, we can characterize the behavior of the ASR with a single parameter in spite of the variety of acoustic conditions typical of a domestic scenario, due to different room layouts, source positions, orientations and directivity patterns. As a consequence we can introduce a criterion for clustering the channels and designing suitable data contamination strategies [17]. In [4, 18] similar metrics are adopted but their correlation with the recognition performance is not investigated.

The paper is organized as follows: Section 2 introduces the problem of ASR in reverberation and presents the proposed RIR characterization. The experimental framework is introduced in Section 3, describing the data and the recognition task. Results are discussed in Section 4 while Section 5 draws some conclusions.

2. ROOM ACOUSTIC AND ASR PERFORMANCE

In enclosures, acoustic waves propagate through multiple paths due to the presence of reflecting surfaces (e.g. walls, furniture). This results in the so-called reverberation that consists in multiple replicas of the emitted signal reaching the microphone. The effects of the enclosure acoustics are usually described through the convolution between the RIR h and the clean speech signal $s(t)$:

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

where the RIR is assumed time-invariant, $*$ denotes convolution, $y(t)$ is the reverberated signal captured by the microphone and $\eta(t)$ is the environmental noise. Although it usually plays an equally crucial role on ASR performance, addressing the environmental noise is beyond the scope of this work and the term $\eta(t)$ is neglected hereafter.

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.

It is often convenient to split the RIR into three parts [16], each of them impacting on the emitted signal in different ways:

$$h(\tau) = h_d(\tau) + h_e(\tau) + h_r(\tau), \quad (2)$$

where $h_d(\tau)$ is the direct propagation path, $h_e(\tau)$ describes the early arrivals up to some tens of ms and $h_r(\tau)$ represents the late diffuse reverberation typical of the RIR tail.

Early arrivals are in general not so harmful in speech recognition since ASR systems, similarly to the human auditory system, typically benefit from the energy boost produced by replicas of the same signal arriving at the microphone in a very limited time [9]. Conversely, the reverberation tail critically affects the ASR behavior [10]: due to the time smearing, phonemes are mixed up with the preceding sound, strongly degrading the decoding stage.

In the past, ASR performance in reverberant environments has been mainly associated to the reverberation time T_{60} or to the distance between the source and the microphones [19, 20], keeping all the other factors affecting the RIR fixed (source directivity and orientation, room dimensions and wall absorption coefficients). Recently, the Direct-to-Reverberant Ratio (DRR) has become a popular way to measure the amount of distortion introduced by a given RIR, independently of the specific environment and experimental setup. The DRR measures the ratio between the energy propagating along the direct path (i.e. without reflections) and the reverberant energy [21]:

$$\mathrm{DRR} = \frac{\int_\tau h_d(\tau)^2 d\tau}{\int_\tau \left(h_e(\tau) + h_r(\tau)\right)^2 d\tau} \quad (3)$$

The metric is mostly used in dereverberation or speech enhancement, either to measure the performance or to characterize the experimental conditions. However, since the decoding step typically gets benefit from early arrivals, we consider a different metric, as follows:

$$\mathrm{ELR}_T = 10 \log_{10} \frac{\int_{\tau=0}^{T} h(\tau)^2 d\tau}{\int_{\tau=T}^{\infty} h(\tau)^2 d\tau} \quad (4)$$

where T determines the time instant when we split between early and late arrival. Basically, it is a generalization of the *clarity* C_{80} used to characterize the music transparency in concert halls [16]. Assuming that the component due to the reverberation tail is uncorrelated with the contribution due to the early arrivals and can be modeled as additive noise, the proposed ELR_T measure can be interpreted as a sort of SNR. An equivalent metric, the *definition* D50 [16], was used in [11] for a similar investigation (setting $T = 50$ms).

3. EXPERIMENTAL SETUP

The hypothesis that ELR_T is an effective feature for predicting the complexity of a reverberant speech recognition task is validated using two well-known corpora (TIDIGITS and TIMIT). The reverberant material was created using both artificially generated and real (measured) RIRs.

3.1. Real and Simulated Impulse Responses

Given a real apartment equipped with a large number of microphones on walls and ceilings, whose map is sketched in Figure 1, a set of RIRs was measured by reproducing exponential chirp signals by means of a loudspeaker and considering different positions and orientations in different rooms [22, 23]. In Figure 1 circles indicate the source positions while arrows represent the orientation. All RIRs were measured at 48kHz. Through the use of the image method [24, 25] the same set of RIRs was simulated at 16kHz in the living room, in the kitchen and in the bathroom, varying the reverberation time from 0.2 to 0.8 seconds and considering three source emission patterns (from omnidirectional to very directional). A variation of the original formulation of the image method was employed to account for source directivity and orientation. Overall, 1155 different conditions were generated in the living room, 324 in the kitchen and 18 in the bathroom.

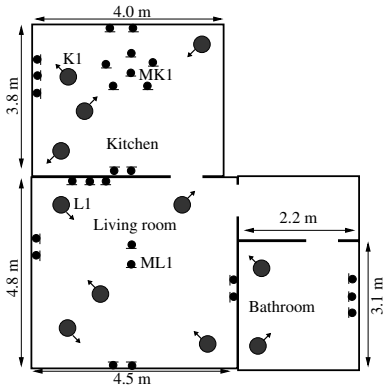


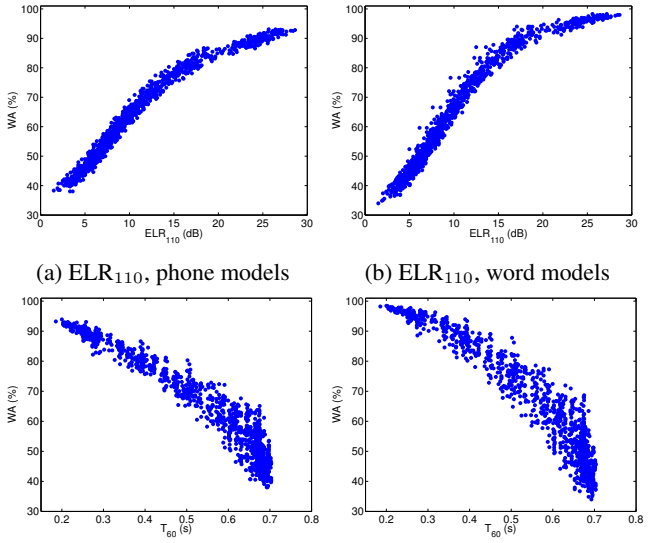
Fig. 1: Map of the real apartment (comprising a living room, a kitchen, and a bathroom) used for the experimental setup. Large circles indicate the source positions while small circles represent the 27 microphones.

3.2. ASR task

To study ASR performance in the reverberant conditions described above, a connected-digit recognition task has been selected, based on a popular HTK architecture. Two alternative systems have been considered in order to analyze performance trends according to different recognition complexities. The first task is based on word-models and the related acoustic model is obtained from the TIDIGITS corpus (about 8600 sentences for 12 whole-word HMMS). A parallel set of experiments is based on phone models, in this case trained on the likewise well-known TIMIT database (about 5000 sentences for 40 monophones). The recognition experiments are performed using MFCC features: speech is segmented into frames of 25ms with a frame shift of 10ms using the Hamming window, and MFCC are obtained from the log mel-spectrum by applying DCT. The feature vector is augmented with the zeroth cepstral coefficient and the dynamic coefficients (Δ and $\Delta\Delta$); Cepstral Mean Subtraction is then used. Since the number of experiments is quite large, our test set is represented by a subset of the standard test portion of TIDIGITS: we have selected 870 sentences sampling uniformly the original set, assuming a consistent performance trend.

4. RESULTS

As a first analysis we verify if the proposed parameter ELR_T is actually correlated with the ASR Word Accuracy (WA) and to derive the best value of T. Figure 2 plots the WA for the 1155 conditions in the living room against the corresponding ELR_{110} for phone and word models using the clean material for training. For the same conditions, Figures 2c and 2d plot the accuracy against the T_{60}, measured with respect to the maximum of the RIRs as there may be no line-of-sight. The Figure confirms that the proposed parameter



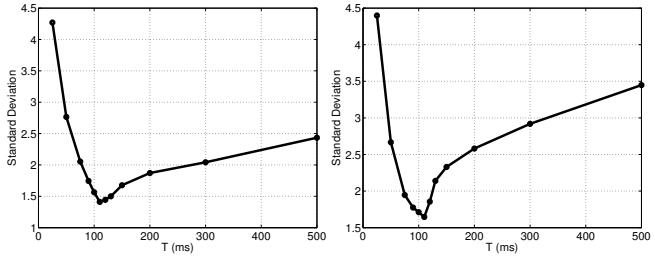
(a) ELR_{110}, phone models (b) ELR_{110}, word models

(c) T_{60}, phone models (d) T_{60}, word models

Fig. 2: ASR accuracy against ELR_{110} and T_{60} for phone and word models trained on clean material. Each dot corresponds to a given condition in the living room.

is strongly correlated to the ASR performance, although some variance is present mainly due to the probabilistic nature of the ASR and to the non-linear influence of dictionary and language model (i.e. a digit-loop). Conversely, the T_{60} presents a lower correlation with the ASR behaviour, in particular for high reverberation times.

Figure 3 plots the standard deviation of the WAs (computed on the 1155 replicas of the test set) as function of T: T=110ms is the value that minimizes this curve and is therefore adopted in the rest of the experimental analysis. In the forthcoming experiments we present results using the word models only. Not reported experiments confirm a similar trend for the phone models.



(a) phone models (b) word models

Fig. 3: Standard deviation of the WA using clean models for different values of T.

It is now interesting to investigate if the correlation between WA and ELR_{110} still holds when different layouts (i.e. rooms) are considered. Figure 4 reports the results related to the three rooms: the points of the RIRs in the kitchen and in the bathroom fit the distribution of RIRs in the living room, here represented through the average WA with steps of 2dB. This confirms that the proposed RIR classification is effective and almost independent of the room under analysis.

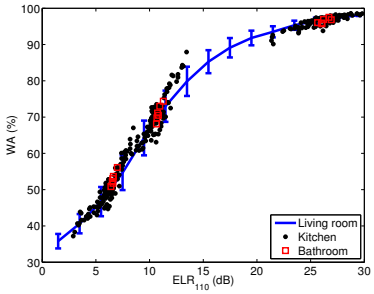
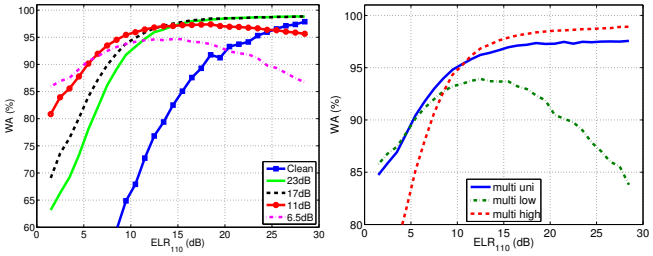


Fig. 4: ASR accuracy against the ELR_{110} for clean word models considering the three rooms. The points of the livingroom are here represented through a continuous line.

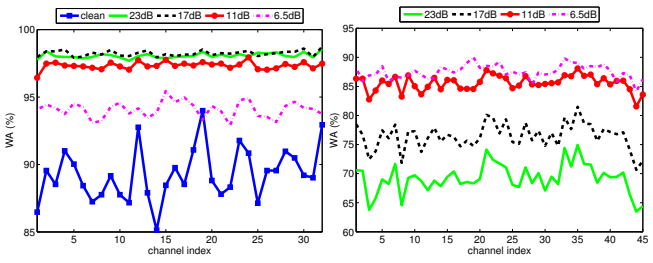
The scope of this work is not only to introduce an efficient characterization of the RIRs based on ELR_{110} measure but also to prove that RIRs with similar ELR_{110} have very similar reverberation properties with respect to ASR performance, despite other influencing factors. This means that, given a set of similar RIRs, we can use just one of them to generate contaminated training material and derive an acoustic model that partly compensates the acoustic mismatch [22]. For this purpose we selected 4 RIRs, among the 1155 available in the living room, with ELR_{110} respectively equal to 23, 17, 11 and 6.5dB and, via the contamination procedure [22], we created the corresponding acoustic models. Clustering the RIRs with a step of 1dB, Figure 5a shows the average WA per bin when different acoustic models are employed. "Clean" indicates acoustic models trained on the dry anechoic speech signals. The impact of the contamination



(a) single-condition models (b) multi-condition models

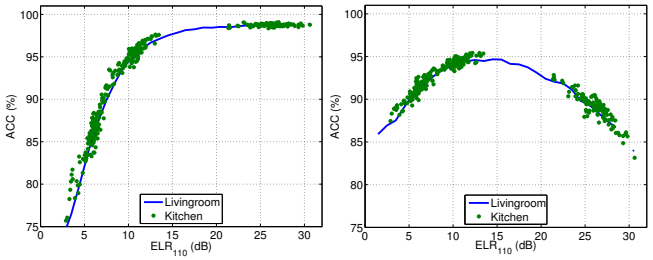
Fig. 5: Average WA for models trained on: (a) single-condition material using a single channel; (b) multi-condition with uniform sampling of the space of ELR_{110} and using only low and high ELR_{110}.

process is clear and improves WAs in a consistent way across the different classes, confirming that the RIRs in a given cluster have very similar properties. We can also state that, for a given channel, the best acoustic model is the one generated with the closest ELR_{110}, independently of the layout and the environment in the test and training data. For a better analysis, Figure 6a reports the WA for the 32 RIRs with ELR_{110} around 17dB, while Figure 6b focuses on the 45 RIRs with ELR_{110}=3.5dB. The Figures prove that the acoustic model trained on a single RIR is effective also for all the other RIRs with similar ELR_T. Interestingly, this also holds if we consider the RIRs of another room. Figure 7 plots the ASR accuracies obtained in the kitchen when using the acoustic models of the living room for ELR_{110} equal to 17dB (Fig. 7a) and 6.5dB (Fig 7b). The points match the line related to the average accuracy in the living room, supporting our hypothesis.



(a) Livingroom channels at 17dB (b) Livingroom channels at 3.5dB

Fig. 6: WA of each channel in the 17dB class (a) and in the 3.5dB class (b) for different acoustic models.



(a) model at ELR_{110}=17dB (b) model at ELR_{110}=6.5dB

Fig. 7: ASR accuracy of each channel of the kitchen when using the living room models trained at ELR_{110}=17dB and ELR_{110}=6.5dB.

4.1. RIR-based multi-condition training

Multi-condition training is a strategy for acoustic model training that attempt to cope with variable acoustic conditions by creating a multi-condition traing dataset. Since we have observed that the ELR_T of both the test channel and the training channel are tightly related to the WA, it can be used to select a representative set of RIRs for the multi-condition dataset. To validate this hypothesis, we created 3 multi-condition models based on the use of: 3 RIRs with low ELR_{110} ("multi low"), 3 RIRs with high ELR_{110} ("multi high") and 3 RIRs spanning in a uniform way the ELR_{110} of all RIRs (4.5dB, 8dB and 17dB). The average WA of the three models are reported in Figure 5b. As expected the acoustic model resulting from the uniform multi-condition training shows improved robustness with respect not only to the models based on a single reverberant condition (i.e. a single ELR_{110}) but also with respect to the non-uniform multi-condition models. It is worth noting also how the "multi low" and "multi high" models perform very similarly to the related single models, 6.5dB and 17dB respectively, since they comprise a limited range of ELR_{110}.

4.2. Real RIRs

Interestingly, the ELR_{110} seems to be an absolute measure, for a given recognition task, and preserves its significance also on real data. Figure 8 plots together the ASR results for data coming from both simulated and real IRs; for each ELR_{110} the cloud of points of the simulated data is substituted by mean and standard deviation.

The points associated to the real RIRs, whose ELR_{110} ranges between 7 and 14dB, present a very good match with the simulated data. Concerning the contaminated acoustic models, Table 1 reports the average WA of the real data using various acoustic models trained

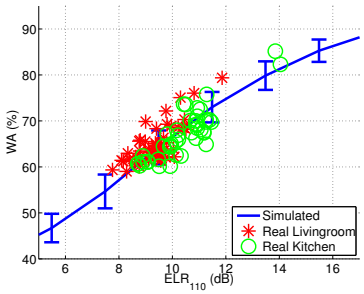


Fig. 8: ASR accuracy against the ELR_{110} for clean word models for simulated and real data.

Model	ELR_{110}	Real RIRs	Livingroom	Kitchen
L1-ML1	10.9	95.3	95.4	95.2
K1-MK1	10.8	95.5	95.7	95.4
sim1	17	91.9	91.4	92.5
sim2	11	93.6	93.5	93.8
sim3	6.5	92.5	91.8	92.1

Table 1: Average WA on the real-RIRs data using various acoustic models, trained with real (L1-ML1, K1-MK1) and simulated (sim1, sim2, sim3) data. 55 channels are considered in the Livingroom and 52 in the Kitchen.

on single independent RIRs: "L1-ML1" and "K1-MK1" refer to the two real source-microphone pairs reported in Figure 1, the labels "sim1", "sim2" and "sim3" refer to models based on RIRs generated through the image method. Note that again the best model is the one with ELR_{110} closest to the average ELR_{110} of the real data which is 9.9dB. Also, the average performance using models trained on the real data is similar as the range of ELR_{110} is limited. The minor gap between the real and synthetic models may be related to the different sampling rates.

5. CONCLUSIONS

In this work a study of distant-talking speech recognition in reverberant conditions is presented: exploiting a large variety of both simulated and real impulse responses, we show that the ASR performance correlates with the Early-to-Late Reverberation Ratio. Although the ASR errors distribution is influenced also by the active dictionary, the analysis proves that it is possible to approximately predict recognition accuracies from acoustic parameters derived from the room impulse responses (i.e. ELR_{110}).

The accordance between the simulation and the real measurements confirms the validity of the proposed analysis, envisaging a practical method for training and selecting proper acoustic models for a given acoustic environment. Indeed, it is possible to synthetically generate a number of RIRs of the targeted environment and, according to the ELR_T ranking, select a small subset of them for acoustic training.

Future work will investigate the influence of other important factors, namely the possible background noise, the complexity of the recognition task or the language itself. Another interesting issue is the capability of blindly estimating the ELR_T, or equivalent metrics, directly from the audio signals [21, 26], without requiring a full knowledge of the RIRs.

6. REFERENCES

[1] M. Wölfel and J. McDonough, *Distant speech recognition*, Wiley, 2009.

[2] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Digital Signal Processing - Springer-Verlag. Springer, 2001.

[3] K. Kumatani, J. McDonough, J.F. Lehman, and B. Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *HSCMA*, 2011.

[4] M. Wolf and C. Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Communication*, 2013.

[5] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: An overview of challenge systems and outcomes," in *ICASSP*, 2013, pp. 126–130.

[6] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S. Hahm, and A. Nakamura, "Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds," *Computer Speech & Language*, vol. 27, no. 3, pp. 851–873, 2013.

[7] K. Takeda, M. Kondo, and F. Itakura, "An acoustic measure for predicting recognition performance degradation," in *ICASSP*, 2000, vol. 3, pp. 1739–1742 vol.3.

[8] H. Hermansky, E. Variani, and V. Peddinti, "Mean temporal distance: predicting ASR error from temporal properties of speech signal," in *ICASSP*. IEEE, 2013.

[9] R. Petrick, K. Lohde, M. Wolff, and R. Hoffmann, "The harming part of room acoustics in automatic speech recognition.," in *INTERSPEECH*, 2007.

[10] A. Sehr and W. Kellermann, "On the statistical properties of reverberant speech feature vector sequences," in *IWAENC*, 2010.

[11] A. Sehr, E. Habets, R. Maas, and W. Kellermann, "Towards a better understanding of the effect of reverberation on speech recognition performance," in *IWAENC*, 2010.

[12] A. Krueger and R. Haeb-Umbach, "Model-Based Feature Enhancement for Reverberant Speech Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1692–1707, 2010.

[13] O. S. Sadjadi, H. Bořil, and J. H. L. Hansen, "A comparison of front-end compensation strategies for robust LVCSR under room reverberation and increased vocal effort," in *ICASSP*, 2012, pp. 4701–4704.

[14] Randy Gomez, Keisuke Nakamura, and Kazuhiro Nakadai, "Robustness to speaker position in distant-talking automatic speech recognition," in *ICASSP*, 2013, pp. 7034–7038.

[15] L. Wang, Z. Zhang, A. Kai, and Y. Kishi, "Distant-talking speaker identification using a reverberation model with various artificial room impulse responses," in *Signal Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, 2012, pp. 1–4.

[16] H. Kuttruff, *Room Acoustics*, Elsevier Applied Science, fifth edition edition, 1991.

[17] M. Matassoni, M. Omologo, D. Giuliani, and P. Svaizer, "Hidden Markov Model training with contaminated speech material for distant-talking speech recognition," *Computer Speech & Language*, vol. 16, no. 2, 2002.

[18] L. Couvreur and C. Couvreur, "On the use of artificial reverberation for ASR in highly reverberant environments," in *Benelux Signal Processing Symposium*, 2000.

[19] B. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *ICASSP*, 1997.

[20] M. Seltzer, *Microphone Array Processing for Robust Speech Recognition*, Ph.D. thesis, Brown University, 2003.

[21] P. Naylor, N. Gaubitch, and E. Habets, "Signal-based performance evaluation of dereverberation algorithms," *Journal of Electrical and Computer Engineering*, vol. 2010, 2010.

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

[23] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," in *Language Resources and Evaluation Conference*, 2014.

[24] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics,," *Journal of the Acoustical Society of America*, vol. 65, no. 4, April 1979.

[25] P. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1527–1529, April 1986.

[26] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *EUSIPCO*, 2011.