INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Realistic multi-microphone data simulation for distant speech recognition

Mirco Ravanelli, Piergiorgio Svaizer, Maurizio Omologo

Fondazione Bruno Kessler (FBK), Trento, ITALY
mravanelli@fbk.eu, svaizer@fbk.eu, omologo@fbk.eu

Abstract

The availability of realistic simulated corpora is of key importance for the future progress of distant speech recognition technology. The reliability, flexibility and low computational cost of a data simulation process may ultimately allow researchers to train, tune and test different techniques in a variety of acoustic scenarios, avoiding the laborious effort of directly recording real data from the targeted environment.

In the last decade, several simulated corpora have been released to the research community, including the data-sets distributed in the context of projects and international challenges, such as CHiME and REVERB. These efforts were extremely useful to derive baselines and common evaluation frameworks for comparison purposes. At the same time, in many cases they highlighted the need of a better coherence between real and simulated conditions.

In this paper, we examine this issue and we describe our approach to the generation of realistic corpora in a domestic context. Experimental validation, conducted in a multi-microphone scenario, shows that a comparable performance trend can be observed with both real and simulated data across different recognition frameworks, acoustic models, as well as multi-microphone processing techniques.

Index Terms: distant speech recognition, simulated data, real data, multi-microphone speech corpora.

1. Introduction

Distant Speech Recognition (DSR) represents a fundamental technology towards natural human-machine interfaces. Despite the recent substantial progress in various related fields, including spatial filtering [1, 2], microphone selection [3], source separation [4], speech dereverberation [5], speaker localization [6], acoustic event detection [7] as well as acoustic modeling [8, 9, 10, 11, 12], DSR still exhibits a lack of robustness, especially when adverse acoustic conditions originated by non-stationary noises and acoustic reverberation are met [13].

The availability of realistic simulated corpora and, more importantly, the definition of common methodologies, algorithms and good practices to generate simulated data plays a crucial role for fostering future research in this field and will eventually help researchers to better migrate laboratory results into real application scenarios. Approaches as contaminated speech training [14, 15, 16], multi-style training [17, 18, 19] and data augmentation [20, 21, 22, 23] have, in fact, been shown very effective in improving the DSR system performance.

During the last decade, some simulated corpora have been made available to the research community under projects or international challenges. Valuable examples are the corpora released under the ChiME [24, 25] and REVERB [26] challenges, which have contributed to define common tasks, baselines and evaluation frameworks across researchers. Other simulated corpora have been released under the CHIL project [27] and, more recently, under the EU DIRHA project [28, 29, 6, 30, 31]. These efforts were extremely important to stimulate research in the DSR field, but in several cases they also pointed out the need of a better coherence between real and simulated data performance. In [25], for instance, the authors state that "*The [CHiME3] challenge has drawn attention to the value of simulated training data, but highlighted the need for better simulation algorithm. It has also demonstrated that caution is needed when interpreting results of challenges that use simulated data evaluation.*". We fully agree with this statement, as our past experience confirms that prudence is needed when using simulated data. This caution is often to be attributed to very subtle differences that may characterize the process of simulation as, for instance, the accuracy and the realism of impulse responses.

The main purpose of this paper is to investigate on the level of agreement in performance trend, that can be obtained with real and simulated signals. A major focus of our work is on reverberation, rather than background noise. Simulations are based on the contamination method described in [15]. Each impulse response (IR) is measured according to the procedure described in [32], while simulated IRs are derived by a modified version of the image method [33], which was experimented in our past works. This modified version differs from the original [33] just for simulating also the directivity pattern of the source, besides sound propagation effects. The experiments, conducted in a new multi-microphone domestic scenario that was developed under DIRHA, demonstrate a good level of agreement in performance, evident with all the investigated acoustic models and processing. We also show the improvement that can be obtained when measured IRs, instead of image-method based ones, are used to train acoustic models.

The paper is organized as follows: Sec. 2 outlines the data simulation approach; Sec. 3 describes the adopted experimental setup, while Sec. 4 reports on the experimental validation of the methodology. Finally, Sec. 5 will draw some conclusions and provide an outlook on future work.

2. Data Simulation

In this work, the data simulation process is achieved according to the following equation:

$$y(t) = x(t) * h(t) + n(t) \quad (1)$$

where $y(t)$ is the simulated distant-talking signal, $x(t)$ is the close-talking speech, $h(t)$ is the impulse response of the acoustic environment for a given source and microphone position, $*$ is the convolution operator, and $n(t)$ is an additive background

http://dx.doi.org/10.21437/Interspeech.2016-731

noise. Several important aspects must be considered for an effective simulation, as discussed in the following.

2.1. Close-Talking Recordings

Our experience in data simulation suggests that the availability of a high-quality close-talking data set is crucial for generating realistic distant-talking simulated data. Particular attention should be directed to ensuring dry and noiseless recordings, since the possible presence of noise sources, saturation, reverberation effects due to the room acoustic as well as distance between speaker and microphone can produce artifacts in the later simulation process. The quality and the characteristics of the microphone can also influence the realism of the simulations.

In the context of the DIRHA project, high quality close-talking speech signals have been acquired under extremely quiet conditions (with a SNR of at least 50-60 dB for each sentence), in an acoustically treated recording room, using a high-quality microphone (Neumann TLM 103) and a professional audio card (RME Octamic II).

2.2. Impulse Response

The impulse response is the most representative feature characterizing an acoustic space. In the assumption of linear time-invariant reverberant rooms, IRs provide a complete description of the changes a sound signal undergoes when traveling from one point in space to another [34]. The impulse response can be either measured in the targeted environment or geometrically inferred by simulations.

Several techniques have been proposed in the last decade for measuring the IR of an acoustic enclosure, including solutions based on Maximum Length Sequence (MLS) [35], Linear Chirps [36], or Exponential Sine Sweeps (ESS) [36]. In [32], a comparison between these different methods has been proposed for distant speech recognition purposes, showing that ESS outperforms the other methods, especially when long (1 minute) and high dynamic excitation signals can be emitted in the acoustic environment. This result is due to a better management of the harmonic distortions introduced by the loudspeaker as well as to a more favorable SNR. That study also revealed that using a professional loudspeaker for exciting the acoustic environment (e.g., a professional Genelec 8030) leads to a much more realistic impulse response measurement, if compared to what obtained with a cheaper loudspeaker. Following these guidelines, an IR measurement campaign has been conducted in the context of the DIRHA project to acoustically characterize a real apartment equipped with a network of microphones. As discussed in [28], about 9000 IRs were estimated.

Synthetic room impulse responses can be generated by the well-known Image-source Method (IM) [33], based on a geometric model accounting for room size, source and microphone positions, and ideal propagation/reflection paths within the enclosure. The baseline method only considers attenuation and (approximated) time instants of arrival of reflections, which results in quite unrealistic IRs. Several improvements have been proposed in order to achieve IRs with characteristics that better match with those measured in real environments [37, 38]. In this work, for instance, a modified version of the standard algorithm allowing us to simulate directive sources is considered. This version has shown to be effective to generate IRs better reflecting real-world conditions. However, such simplified propagation models, assuming an empty shoebox geometry, cannot reproduce the complex patterns of sound propagation in real rooms, as will be shown in Sec.4.5.

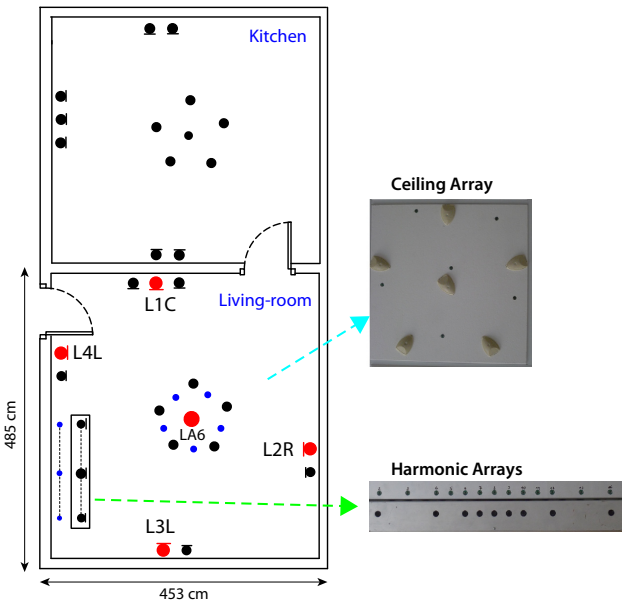

Figure 1: An outline of two rooms of the DIRHA apartment considered for this study. Small blue dots represent digital MEMS microphones, red ones refer to the channels considered for the following experimental activity, while black ones represent the other available microphones. The right pictures show the ceiling array and the two linear harmonic arrays installed in the living-room.

3. Experimental Setup

This section describes the microphone setup, the task, the corpora as well as the speech recognition framework considered in this work.

3.1. Multi-microphone Setup

The apartment used in the DIRHA project is equipped with high-quality omnidirectional microphones (Shure MX391/O), connected to multichannel clocked pre-amp and A/D boards (RME Octamic II), which provide a synchronous sampling at 48 kHz, with 24 bit resolution. The living-room and the kitchen comprise the largest concentration of sensors and devices. As shown in Fig. 1, the living-room includes three microphone pairs, a microphone triplet, two 6-microphone ceiling arrays (one consisting of MEMS digital microphones), two harmonic arrays (consisting of 15 electret microphones and 15 MEMS digital microphones, respectively). The experiments in this work refer to the use of the five microphones depicted in red in Fig.1. The reverberation time T_{60} of the considered room is about 0.75 seconds, which indicates that the acoustic characteristics are quite challenging for DSR studies.

3.2. Task and corpora

The task considered in this work is the Wall Street Journal (WSJ-5k), in agreement with the task addressed in the CHiME 3 challenge. While CHiME 3 was pretty focused on robustness against noise, in this work the main source of disturbance is reverberation. For testing purposes we employed both real and simulated data, which are derived from recordings in the DIRHA apartment. Real data were collected from four native

Data Type / *A.M.*	*Single Distant Microphone*		*Delay-and-Sum Beamforming*		*Oracle Microphone Selection*	
	Real Data	Sim Data	Real Data	Sim Data	Real Data	Sim Data
Mono	62.2	64.7	56.8	58.8	49.6	51.9
Tri1	39.8	41.1	33.9	34.9	28.0	29.2
Tri2	33.0	33.6	28.4	29.1	22.6	23.2
Tri3	21.5	22.3	18.0	19.1	13.6	14.9
Tri4	19.9	21.4	17.5	17.4	12.6	13.8
DNN	12.0	13.2	10.7	11.6	7.2	7.6

Table 1: WER(%) obtained in a distant-talking scenario with real and simulated data across different acoustic models and microphone processing.

US English speakers (two females and two males) uttering a total of 272 WSJ sentences in different positions of the apartment. In particular, each subject was positioned in the living-room and read the material from a tablet, standing still or sitting on a chair, in a given position. After a set of 10-12 sentences, she/he was asked to move to a different position and take a different orientation. In order to allow a fair comparison between real and simulated data, we asked the same speakers to utter the same sentences in our recording studio, using the acquisition set-up described in Sec.2.1. Moreover, for each position/orientation of the speaker in the real recording, a corresponding IR was measured, allowing us to derive a simulated corpus well-matching with the speaker positions used for the real data. The training phase is based on the WSJ0 database (LDC catalog number LDC93S6A), which was contaminated with an impulse response measured in a position different from those used for testing purposes.

3.3. ASR framework

The experimental part of this work is based on the Kaldi toolkit [39]. The recipe considered for training and testing the DSR system is similar to the s5 recipe proposed in the Kaldi release for WSJ data. In short, the speech recognizer is based on standard MFCCs and acoustic models of increasing complexity. "*Mono*" is the simplest system based on 48 context-independent phones of the English language, each modeled by a three state left-to-right HMM (overall using 1000 gaussians). A set of context-dependent models are then derived. In "*tri1*" 2.5k tied states with 15k gaussians are trained by exploiting a binary regression tree."*Tri2*"is an evolution of the standard context-dependent model in which a Linear Discriminant Analysis (LDA) is applied. In both "*tri3*" and "*tri4*" models Speaker Adaptive Training (SAT) is also performed. The difference is that "*tri4*" is bootstrapped by the previously computed "*tri3*" model. The considered "*DNN*", based on the Karel's recipe, is composed of 6 hidden layers of 2048 neurons, with a context window of 19 consecutive frames (9 before and 9 after the current frame) and an initial learning rate of 0.008. The weights are initialized via RBM pre-training, while the fine tuning is performed with stochastic gradient descent optimizing cross-entropy loss function.

4. Results

This section provides some speech recognition results, with the purpose of validating the proposed data simulation approach. In the following sub-section, a close-talking baseline is provided, while in subsections 4.2, 4.3 and 4.4 distant-talking experiments with single microphone, beamforming on the ceiling array, and oracle microphone selection are respectively presented.

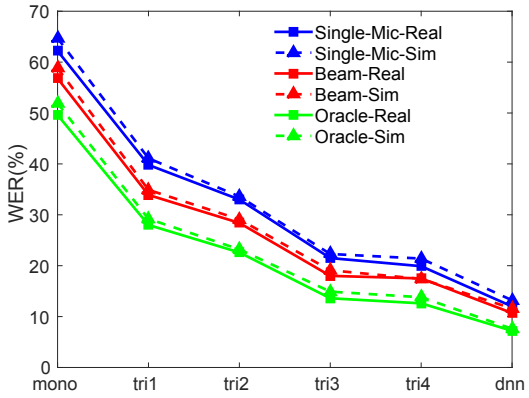

Figure 2: Graphical representation of the performance trends reported in Table 1.

4.1. Close-talking baseline

The Word Error Rate (WER%) obtained by decoding the close-talking WSJ sentences recorded in the recording studio is 3.7% (using DNN models trained with the original clean WSJ data set). It is worth nothing that, under such favorable acoustic conditions, the DNN model leads to a very accurate sentence transcription. For reference purposes, the average WER with close-talking signals recorded in the DIRHA apartment is about 5%.

4.2. Single distant-microphone performance

The results reported in the first column of Table 1 show the performance obtained when a single distant microphone (i.e., the "*LA6*" ceiling microphone depicted in Fig. 1) is considered. Results clearly highlight that in the case of distant-speech input the ASR performance is dramatically reduced, if compared to a close-talking case. The use of robust DNN models trained with contaminated speech material leads, as expected, to a substantial improvement of the WER when compared to other GMM-based systems. The most interesting result, however, is that a similar performance trend is obtained for both real and simulated data over different acoustic models. This trend can also be appreciated by comparing the continuous (real data) and dashed (sim data) blue lines of Fig. 2. In particular, the average relative WER distance between such data-sets computed over the considered acoustic models is about 6%. We believe that this is a significant result, especially if one considers that part of this variability can be attributed, despite our best efforts for aligning simulated and real data, to the fact that in the two recording sessions speakers inevitably uttered the same sentence in a dif-

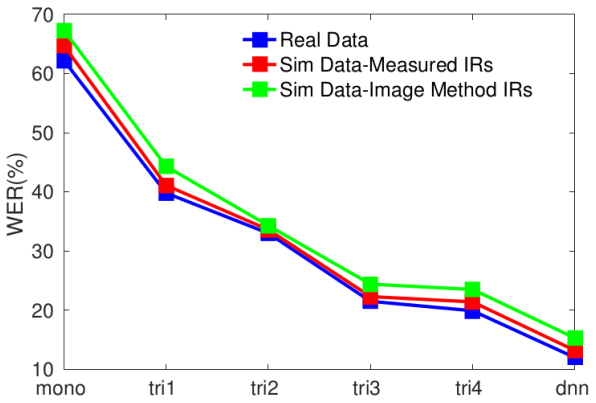

Figure 3: Comparison between real and simulated data with contaminated training performed with a measured IR.

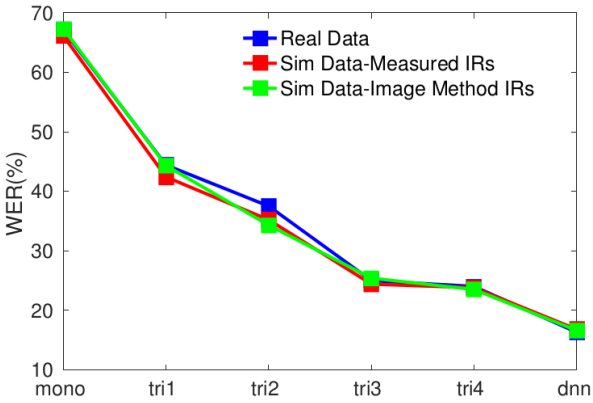

Figure 4: Comparison between real and simulated data with contaminated training performed with an image method IR.

ferent way.

4.3. Delay-and-sum beamforming performance

The simulation methodology described in Sec. 2, can be extended in a very straightforward way to a multi-microphone scenario. It would be thus of crucial importance to ensure that the similar trend between real and simulated data achieved with a single microphone is preserved even when multi-microphone processing is applied to the data. Here a standard delay-and-sum beamforming, based on source-microphone delays computed with the GCC-PHAT algorithm [40], is applied to the six microphones of the ceiling array of Fig. 1. Table 1 and Fig. 2 show that beamforming is helpful in improving the system performance. One can also note that, as hoped, a similar performance trend between the data-sets is reached when applying beamforming. For instance, in the case of real data coupled with DNN acoustic models, delay-and-sum beamforming leads to a relative improvement of about 12% over the single microphone case, which is similar to the improvement of 13% obtained with the simulated data.

4.4. Oracle microphone selection performance

To further confirm the result achieved in the previous sections, an oracle microphone selection is applied to both real and simulated data. An oracle microphone selection is performed by selecting, for each sentence uttered by the speaker, the best WER from the five signals acquired by the red microphones in Fig. 1. Table 1 and Fig. 2, confirm that the consistency between real and simulated data is largely preserved. The experimental results also show that an optimal microphone selection would be particularly helpful for improving the DSR performance. A proper channel selection has a great potential even when compared with a microphone combination based on delay-and-sum beamforming. For instance, in the case of real data with DNN acoustic models, a WER of 7.2% is obtained with an oracle channel selection against a WER of 10.7% achieved with beamforming.

4.5. Measured vs Geometric Modeling of IRs

In this section we compare the simulations based on measured IRs, so far considered, with simulations derived by image method-based IRs. For the latter case, the geometry of the targeted living room, the spatial coordinates of microphones and speakers, as well as the reverberation time T_{60} of 0.75s are imposed to the IM algorithm. As outlined in Sec. 2.2, a certain source spatial directivity similar to that exhibited by a real speaker, is considered. Fig. 3 and Fig. 4 show the performance observed using two different training strategies. In particular, Fig. 3 reports the trend obtained when the training set is contaminated with an impulse response measured in the target environment, while Fig. 4 presents the results obtained when using an image method-based IR. Results confirm that, in both matching and mismatching conditions, simulated data obtained with measured IRs exhibit a trend very similar to that observed with real data. For instance, in the case of DNN, performance with Real, Sim-Measured IRs, and Image Method, are 12.0%, 13.2%, and 15.3%, respectively. On the other hand, despite our best efforts for increasing the realism of image method-based IRs, the performance with such simulation approach is still unsatisfactory. In particular, in the case of DNN the relative performance loss using image-method based IRs, instead of measured IRs, for contaminated training is 36% (i.e., from 12% to 16.3%).

5. Conclusion

In this paper we discussed our best practices to generate realistic multi-microphone data for training and testing distant-speech recognition systems. Our approach has been validated by comparing real data with simulated data obtained by convolving close-talking dry speech sequences with impulse responses measured in a domestic environment. The experimental results show that a very similar performance trend can be obtained between real and simulated data over different experimental conditions, involving different acoustic models and multi-microphone processing techniques. This study also revealed that data simulation based on IRs measured in the targeted environment ensures much better results than those obtained with an IR simulator based on Image method. However, in the perspective of a real application, measuring every time the IRs can be unpractical. The results reported in this paper are thus just a starting point towards a future work, which will study more in depth how the gap between measured and synthetic IRs can be reduced. An ideal solution would be to automatically analyze the recorded speech and to drive an unsupervised adaptation of initial IRs possibly generated by simulation.

6. References

[1] M. Brandstein and D. Ward, *Microphone arrays*. Springer, Berlin, 2000.

[2] W. Kellermann, *Beamforming for Speech and Audio Signals*. in HandBook of Signal Processing in Acoustics, Springer, 2008.

[3] M. Wolf and C. Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Communication*, vol. 57, pp. 170–180, Feb. 2014.

[4] S. Makino, T. Lee, and H. Sawada, *Blind Speech Separation*. Springer, 2010.

[5] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation.* Springer, 2010.

[6] A. Brutti, M. Ravanelli, P. Svaizer, and M. Omologo, "A speech event detection/localization task for multiroom environments," in *Proc. of HSCMA 2014*, pp. 157–161.

[7] A. Temko, C. Nadeu, D. Macho, R. Malkin, C. Zieger, and M. Omologo, "Acoustic event detection and classification," in *Computers in the Human Interaction Loop*. Springer London, 2009, pp. 61–73.

[8] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *Proc. of ASRU 2013*, pp. 285–290.

[9] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *Proc. of ICASSP 2014*, pp. 5542–5546.

[10] F. Weninger, S. Watanabe, J. Le Roux, J. Hershey, Y. Tachioka, J. Geiger, B. Schuller, and G. Rigoll, "The MERL/MELCO/TUM System for the REVERB Challenge Using Deep Recurrent Neural Network Feature Enhancement," in *IEEE REVERB Workshop*, 2014.

[11] S. S. Masato Mimura and T. Kawahara, "Reverberant speech recognition combining deep neural networks and deep autoencoders," in *IEEE REVERB Workshop*, 2014.

[12] A. Schwarz, C. Huemmer, R. Maas, and W. Kellermann, "Spatial Diffuseness Features for DNN-Based Speech Recognition in Noisy and Reverberant Environments," in *Proc. of ICASSP 2015*.

[13] E. Hänsler and G. Schmidt, *Speech and Audio Processing in Adverse Environments*. Springer, 2008.

[14] M. Ravanelli and M. Omologo, "Contaminated speech training methods for robust DNN-HMM distant speech recognition," in *Proc. of INTERSPEECH 2015*, pp. 756–760.

[15] 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, pp. 205–223, 2002.

[16] M. Ravanelli and M. Omologo, "On the selection of the impulse responses for distant-speech recognition based on contaminated speech training," in *Proc. of INTERSPEECH 2014*, pp. 1028–1032.

[17] A. Sehr, C. Hofmann, R. Maas, and W. Kellermann, "Multi-style training of HMMS with stereo data for reverberation-robust speech recognition," in *Proc. of HSCMA 2011*, pp. 196–200.

[18] L. Couvreur, C. Couvreur, and C. Ris, "A corpus-based approach for robust ASR in reverberant environments." in *Proc. of INTERSPEECH 2000*, pp. 397–400.

[19] T. Haderlein, E. Nöth, W. Herbordt, W. Kellermann, and H. Niemann, "Using Artificially Reverberated Training Data in Distant-Talking ASR." ser. Lecture Notes in Computer Science, vol. 3658. Springer, 2005, pp. 226–233.

[20] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *Proc. of ICASSP 2014*, pp. 5582–5586.

[21] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. ASRU 2015*, pp. 436–443.

[22] S. R. A. Ragni, K. Knill and M. Gales, "Data augmentation for low resource languages," in *Proc. of INTERSPEECH 2014*, pp. 5582–5586.

[23] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. of INTERSPEECH 2015*, pp. 3586–3589.

[24] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge." *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.

[25] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Proc. of ASRU 2015*.

[26] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The reverb challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," in *Proc. of WASPAA 2013*, pp. 1–4.

[27] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. Chu, A. Tyagi, J. Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet, "The CHIL Audiovisual Corpus for Lecture and Meeting Analysis inside Smart Rooms," *Language resources and evaluation*, vol. 41, no. 3, pp. 389–407, 01/2008 2007.

[28] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. of LREC 2014*, pp. 2629–2634.

[29] M. Matassoni, R. Astudillo, A. Katsamanis, and M. Ravanelli, "The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones," in *Proc. of INTERSPEECH 2014*, pp. 1616–1617.

[30] E. Zwyssig, M. Ravanelli, P. Svaizer, and M. Omologo, "A multi-channel corpus for distant-speech interaction in presence of known interferences," in *Proc. of ICASSP 2015*, pp. 4480–4484.

[31] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The DIRHA-ENGLISH corpus and related tasks for distant-speech recognition in domestic environments," in *Proc. of ASRU 2015*, pp. 275–282.

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

[33] J. Allen and D. Berkley, "Image method for efficiently simulating smallroom acoustics," in *J. Acoust. Soc. Am*, 1979, pp. 2425–2428.

[34] H. Kuttruff, *Room acoustic*, 5th ed. Spon Press, 2009.

[35] M. Schroeder, "Diffuse sound reflection by maximum-length sequences," in *J. Acoust. Soc. Am*, vol. 57(1), 1975, pp. 149–150.

[36] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. of the 108th AES Convention*, 2000, pp. 18–22.

[37] P. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," in *J. Acoust. Soc. Am*, vol. 80(5), 1986, pp. 1527–1529.

[38] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," in *J. Acoust. Soc. Am*, vol. 124(1), 2008, pp. 269–277.

[39] D. Povey at all, "The Kaldi Speech Recognition Toolkit," in *Proc. of ASRU 2011*.

[40] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," vol. 24, no. 4, pp. 320–327, 1976.