A MULTI-CHANNEL CORPUS FOR DISTANT-SPEECH INTERACTION IN PRESENCE OF KNOWN INTERFERENCES

Erich Zwyssig, Mirco Ravanelli, Piergiorgio Svaizer, Maurizio Omologo

Fondazione Bruno Kessler, Trento, Italy

{zwyssig,mravanelli,svaizer,omologo}@fbk.eu

ABSTRACT

This paper describes a new corpus of multi-channel audio data designed to study and develop distant-speech recognition systems able to cope with known interfering sounds propagating in an environment. The corpus consists of both real and simulated signals and of a corresponding detailed annotation. An extensive set of speech recognition experiments was conducted using three different Acoustic Echo Cancellation (AEC) techniques to establish baseline results for future reference. The AEC techniques were applied both to single distant microphone input signals and beamformed signals generated using two state-of-the-art beamforming techniques. We show that the speech recognition performance using the different techniques is comparable for both the simulated and real data, demonstrating the usefulness of this corpus for speech research. We also show that a significant improvement in speech recognition performance can be obtained by combining state-of-the-art AEC and beamforming techniques, compared to using a single distant microphone input.

Index Terms— microphone array, ASR, barge-in, acoustic echo cancellation

1. INTRODUCTION

The European DIRHA project (Distant-speech Interaction for Robust Home Applications) investigates voice-enabled automated control of services and devices in an apartment. Hands-free speech interaction with an automatic system using distant microphones distributed in the environment requires multi-channel front-end processing including optional spatial filtering for selective acquisition of the desired speaker and a technique for suppression of disturbing contributions from unwanted acoustic sources [1]. Some of these interferences may be directly acquired at their source (e.g. audio from radio/TV or the acoustic messages generated by the spoken dialogue management system) and used as references to suppress their effect.

A key requirement for developing these front-end components and the entire speech recognition chain is suitable multi-microphone speech corpora. During the past decade, several projects and collaborative programmes have addressed the development of corpora and challenges for studying distant-speech interaction scenarios, such as CHIL [2], AMI [3], REVERB [4], PASCAL-CHIME [5], GRID [6]. Most of these corpora were developed to investigate distant-speech recognition, speech separation and enhancement in noisy environments, based on single or multiple distant microphones. It is also worth noting that some of these data sets were developed using simulations. The convenience of using realistic simulated data to train distant-speech recognition systems was explored previously [7], most recently in the context of the DIRHA project which has created a new set of multi-microphone speech corpora [8, 9].

The aim of this paper is to introduce a corpus designed to develop voice interaction solutions that are also effective when known audio sources are active, and in particular when the user tries to interrupt a voice message generated by the system, corresponding to what is referred to as the barge-in condition [10, 11].

To the best of our knowledge, this is the first public corpus designed to investigate the problem of suppressing known audio interferences prior to the application of a speech recognition system. The corpus comprises both a very realistic simulated data set and a large set of real audio signals, consisting of high-quality sample-synchronous recordings.

The simulated and real data sets were developed to address three main scenarios: a) when the user interrupts the dialog manager prompt while no other interferers are active; b) when the user speaks while the TV is on, and c) the combination of a) and b). To address each of these situations, three state-of-the-art AEC techniques, i.e. Subband-based Acoustic Echo Cancellation (SAEC) [12, 13], Frequency Domain Adaptive Filtering (FDAF) [14] and Semi-Blind Source Separation (SBSS) [15] were applied in order to establish the corresponding ASR performance baselines. The aim of acoustic echo cancellation is to remove the contribution of the overlapping sources. Although in the literature AEC performance is generally reported in terms of objective measurements such as the Echo Return Loss Enhancement (ERLE) or other misalignment metrics [16], the task proposed here is addressed based on evaluating each technique through its impact on ASR performance, i.e., on the resulting Word Error Rate (WER). Two well-known beamforming tools, BeamformIt [17] and mdm [18] were also used to investigate the case of a multi-microphone input combined with AEC and the speech recognition chain.

2. THE DIRHA_AEC CORPUS

This section provides an introduction to the DIRHA_AEC corpus. In Sec. 2.1 we describe the multi-microphone experimental setup adopted to develop the corpus, while Sec. 2.2 and 2.3 detail the simulated and real data sets.

2.1. The Experimental Setup

In the DIRHA project, a microphone-equipped apartment is available for experiments as well as for the development of a real-time prototype. The flat comprises five rooms which are equipped with a network of several microphones. This study focuses on the microphone network in the living room (shown in Fig. 1) which includes three pairs of sensors and one triplet installed on the walls and a ceiling array with six microphones. All these sensors are high-quality omnidirectional condenser microphones (SHURE MX-391). An additional harmonic array, developed under the DICIT project

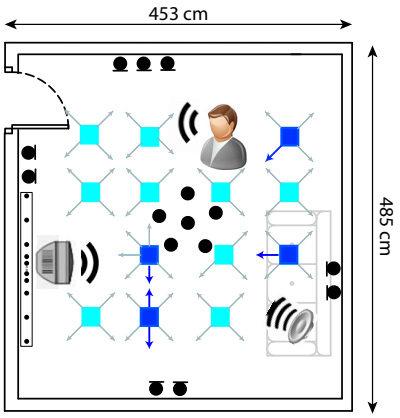

Fig. 1. The multi-microphone experimental setup adopted for simulated and real data in the DIRHA living room environment. Squares and arrows show impulse response positions and directions available for simulation purposes, dark blue ones indicate positions and directions used for real recordings.

Fig. 2. A picture of the DIRHA living room, showing a pair of wall microphones and the ceiling and DICIT arrays.

[19] and composed of 13 omnidirectional electret microphones, is located above the television.

Within the microphone pairs and triplets, individual sensors are at a distance of 30 cm. In the ceiling array five microphones are arranged in a star configuration at a distance of 30 cm to the central one. The linear harmonic DICIT array is based on a total distance between the first and last microphones of 192 cm with the spacing between the 13 sensors progressively halved towards the centre. To allow the study of cross-room propagation effects, 13 microphones were installed in the adjacent kitchen, featuring the same geometry as in the living room. Finally, a close-talking microphone was worn by each speaker in order to compute standard AEC metrics such as ERLE. Therefore, the close-talking microphone and clean versions of all the known interferences (i.e. the TV and prompt) are included in both the real and simulated data sets. A total of 42 channels of audio data are available for single and multi-microphone AEC experiments. Each room in the DIRHA appartment is equipped with a ceiling loudspeaker which enables the dialog with the user.

The mean reverberation time (T_{60}) of the living room is 0.77 s, indicating that the acoustic characteristics are quite challenging for

ID	Sources	Overlap [%]	Commands
Sim-S0	Cmds	0	450
Sim-S1	Cmds + Prompt	20	450
Sim-S2	Cmds + TV	100	450
Sim-S3	Cmds + Prompt + TV	20 / 100	450
Real-S0	Cmds	0-20	650
Real-S1	Cmds + Prompt	50-100	650
Real-S2	Cmds + TV	100	650
Real-S3	Cmds + Prompt + TV	50-100 / 100	650

Table 1. Main features of the adopted simulated and real data sets. The column *Sources* indicates possible additional known interferences, *Overlap* reports the percentage of overlap between the speaker and the interferer activities, while *Commands* refers to the overall number of speech commands available in the data set.

distant-talking speech processing. Table 1 summarises the main features of the simulated and real corpora, characterised by a sampling frequency of 48 kHz and a 16 bit accuracy. The SNR of the speech commands to the background noise is approx. 20 dB and 0 dB to the interferences. Note that all the microphone signals are sample-synchronised, based on a common clock transmitted to a set of professional audio cards (OCTAMIC RME II).

2.2. Simulated Data

The DIRHA_AEC corpus presented here is an extension of the existing DIRHA simulated corpora [8, 9] which are multi-microphone, multi-room and multi-language databases consisting of several domestic acoustic scenarios. In contrast to the other simulated databases, the DIRHA_AEC corpus is specifically designed for acoustic echo cancellation and is currently available in Italian. The simulated corpus has been generated using a technique capable of reconstructing multi-microphone signals of typical acoustic scenarios in a very realistic manner. As in [8], a set of high-quality multi-microphone impulse responses measured in a target environment [20, 21]; a collection of clean speech and non-speech signals recorded in a professional studio; and different background noises recorded in the DIRHA living room have been combined to generate the simulated data set.

A simple acoustic scene involving a speech signal $s(t)$ overlapped with a TV signal $x(t)$ can be simulated as

$$d(t) = s(t) * h_s(t) + x(t) * h_x(t) + w(t). \quad (1)$$

The simulated signal $d(t)$ is obtained by convolving $s(t)$ and $x(t)$ with the impulse responses corresponding to the speaker position $h_s(t)$ and TV position $h_x(t)$, with $w(t)$ being a background noise recorded in the target environment.

The complete multi-microphone simulated data set consists of 100 acoustic simulations (called scenes) of 60 seconds each, involving a total of 30 speakers (15 males and 15 females). Each acoustic simulation is composed of a variable number (ranging from 3 to 6) of short speech commands, uttered in one of 74 predefined positions/orientations (see Fig. 1). Each simulated acoustic scene is replicated in four different scenarios of increasing complexity. In the simpler scenario (Sim-S0), each acoustic scene consists of speech commands without any overlapping interference, while the other three more challenging scenarios are based on the same commands progressively overlapped with a prompt speech signal emitted by a loudspeaker (Sim-S1), a TV signal (Sim-S2) and both of these (Sim-S3). A comprehensive annotation and documentation is provided for each simulated scene, i.e. an XML annotation file is

available for each microphone detailing all sequences and their simulation conditions. The annotation format is compliant with the previous multi-microphone data sets generated in the DIRHA project and is described in detail in [8].

2.3. Real Data

In addition to the simulated data, a corpus of real data has also been recorded. This corpus was acquired under similar acoustic conditions and in the same living room environment used for the simulated data set. The real recordings involved asking 13 speakers (6 males and 7 females) to read a list of 50 commands in five different positions/orientations in the room (indicated by the blue squares in Fig. 1). As in the simulated sequences, all the commands were repeated with no overlapping sources (Real-S0), an overlap with a prompt (Real-S1), an overlap with the television (Real-S2) and with both overlaps (Real-S3). All the real sequences were manually annotated by an expert.

3. AEC AND BEAMFORMING

We performed baseline front-end processing for automatic recognition of the commands given by the speaker in the above scenarios. This comprised both acoustic echo cancellation (see Fig. 3) to suppress the known interferences and delay-sum beamforming (BF) to enhance the speech captured by distant microphones. We did not implement an optimised combination of AEC and BF [22, 23, 24, 25] but only considered a simple connection of both components which we called "AEC first" and "BF first". In the former case, an AEC is needed for every microphone channel, yielding good performance but at a high computational cost. In the latter case, the computational cost is lower, but the time-variability of the adaptive beamformer reduces the effectiveness of the downstream AEC [26] (in practice, this solution is generally adopted only in the case of fixed beamformers).

Before applying front-end processing all the signals were downsampled from 48 kHz to 16 kHz sampling rate.

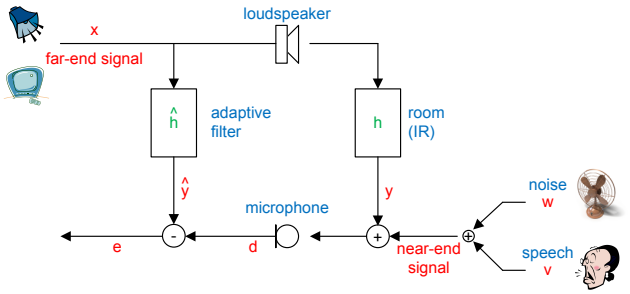

Fig. 3. Principle of acoustic echo cancellation. x is the known signal emitted by a loudspeaker (prompts reproduced by the DIRHA system, or TV audio), v is the speech component captured by the microphone and w is the environmental noise.

3.1. Acoustic Echo Cancellation

The three AEC techniques explored in this work are characterised as follows:

- **SAEC**: Subband-based Acoustic Echo Cancellation (SAEC) was implemented using analysis/synthesis filter banks and the well-known time-domain NLMS (normalised least mean square) adaptation in each subband. Non-critically sampled filter banks were employed in order to avoid aliasing effects. High computational efficiency suitable for real-time implementation – also in the case of long impulse responses – was achieved by using uniform DFT filter banks based on a lowpass prototype, modulation through FFT and polyphase decomposition [27]. In each subband the NLMS adaptation step was controlled using the "delay coefficient" method [12, 13].
- **FDAF**: Frequency Domain Adaptive Filtering (FDAF) is the standard frequency-domain FIR adaptive filter algorithm with frequency-bin step size normalisation [14] characterised by computationally efficient block processing and fast and uniform convergence across frequencies.
- **SBSS**: Semi-Blind Source Separation (SBSS) is an extension of the BSS paradigm in order to include a priori knowledge as a constraint in its demixing adaptation. We tested an SBSS algorithm implementing frequency-domain BSS based on Independent Component Analysis (ICA), with a constraint on the reference signals, i.e. the known echo source signals to be cancelled [15]. A robust fast convergence behaviour was obtained by applying scaling normalisation to the constrained natural gradient. This method is particularly interesting as it can tackle the combined problems of multi-channel AEC and source separation.

3.2. Acoustic Beamforming

In order to enhance the speech of the user and obtain an improved SNR we adopted standard delay-sum techniques, namely the open source BeamformIt (bfi) toolkit[1] [17] and the mdm tools developed in the AMI project [18]. Both use GCC-PHAT for TDOA estimation [28] and employ these values for delay-sum beamforming. BeamformIt applies sophisticated TDOA smoothing, resulting in improved recognition, as previously shown in [29] and [30].

4. EXPERIMENTS

In the experiments described in what follows, front-end processing was applied to the six microphones of the ceiling array. Note also that oracle voice activity detection (VAD) was used to assess the performance of our system, thus avoiding a further source of variability in the ASR performance.

4.1. ASR system and Task

The speech recognition system adopted in this work is a standard HMM-GMM system based on the HTK toolkit [31]. The speech signal is blocked into frames of 25 ms with 10 ms time shift after which 12 Mel-frequency Cepstral Coefficients (MFCCs) plus the log-energy are extracted.

For the acoustic model, similarly to [32], a set of 26 phone-like units of the Italian language was chosen. Each unit is modelled with a three-state left-to-right continuous density HMM, with mixtures of 128 Gaussian components for each state. The acoustic model was trained using the phonetically-rich APASCI database [33] which was contaminated using a single impulse response measured in the DIRHA living room, as proposed in [32].

The decoding was performed based on a small command recognition task (390 words) using a bigram language model trained on a

[1] http://www.xavieranguera.com/beamformit/

Corpus	Close-Mic	Single Far-Mic	Beam-bfi	Beam-mdm
Sim-S0	9.7	33.1	26.2	31.3
Sim-S1	9.7	47.8	39.5	39.5
Sim-S2	9.7	61.3	57.6	60.4
Sim-S3	9.7	70.0	67.2	67.1
Real-S0	12.6	39.4	36.1	38.0
Real-S1	11.6	68.0	56.3	64.3
Real-S2	18.0	72.1	48.6	52.9
Real-S3	13.2	77.0	65.6	67.2

Table 2. WER [%] results without acoustic echo cancellation.

small text corpus of typical commands given in a domestic environment.

4.2. Results and Discussion

In this section we report experimental results from using the different AEC and beamforming techniques mentioned in Sec. 3 and from different types of known interferences that overlap with the targeted speech input.

Table 2 which gives baseline results obtained without applying any AEC technique, shows a dramatic decrease between the performance of a close-talking microphone compared to a single distant microphone. A general trend of decreasing performance from Sim-S0 to Sim-S3 can also be observed due to the increased difficulty of the task, as described in Sec. 2. The results with the real data show a similar trend, although with a general reduction in performance, except for the Real-S2 data set. This is due to the SNR of the TV audio output being 6 dB, and not 0 dB as defined earlier and used in the simulations (Sim-S2). Note also that beamforming techniques without AEC do not produce a significant improvement for various reasons, such as the relatively limited number of microphones and the frequent switching of the beam from the speech source to the interferer and back. In addition, Beam-bfi generally outperforms Beam-mdm. For this reason only Beam-bfi results will be reported in what follows.

Table 3 shows the results obtained using each AEC technique on a single input channel (i.e. the central channel of the ceiling array). The real data again show a lower performance than the simulated data due to the smaller overlap of the commands and prompts in the latter data set. More importantly, the AEC processed output performs significantly better than that obtained from a single distant microphone input (compare Table 3 with the second column of Table 2).

Note that the performance of SBSS is lower compared to the other AEC methods for the datasets Real-S1 and Real-S3. Each AEC method has different timing requirements for its internal filter adaptation. The nature and length of speech overlap affects AEC performance. SBSS requires larger blocking for adaptation and therefore fails to match the other AEC methods for short adaptation times.

Corpus	SAEC	FDAF	SBSS
Sim-S1	34.9	35.6	34.5
Sim-S2	40.6	36.8	36.7
Sim-S3	45.0	40.4	40.5
Real-S1	36.9	42.1	55.9
Real-S2	42.0	43.5	41.8
Real-S3	47.0	48.3	59.5

Table 3. WER [%] performance for single-channel acoustic echo cancellation.

	BF first			AEC first		
Corpus	SAEC	FDAF	SBSS	SAEC	FDAF	SBSS
Sim-S1	31.1	32.9	30.7	26.0	26.0	26.1
Sim-S2	42.4	40.3	39.6	30.8	27.2	29.7
Sim-S3	51.6	50.2	46.7	34.6	30.1	31.2
Real-S1	34.7	39.5	48.4	25.9	30.0	41.3
Real-S2	36.1	38.7	34.9	31.5	31.4	30.8
Real-S3	49.5	48.1	50.2	34.9	34.9	45.0

Table 4. WER [%] performance for multi-channel acoustic echo cancellation.

Table 4 shows results from the joint use of beamforming and AEC in the two configurations referred to as "BF first" and "AEC first". The latter always gives significantly better performance, in line with what is reported in the literature and was reviewed in Sec. 3 [22, 23, 24, 25, 26].

As expected, this combined processing (in particular the "AEC first" option) outperforms that of single processing alone as shown in Tables 2 and 3. Overall our proposed "AEC first" scheme achieves less than 30% WER on both the simulated and real data.

5. CORPUS RELEASE

The ASR system scripts used for evaluation purposes and a portion of the corpus will be made available on the DIRHA website. Some short examples can be already downloaded from `http://dirha.fbk.eu/DIRHA_AEC`. The full simulation data set and a portion of the real data will then be the object of a further public distribution to allow other researchers to reproduce baseline results and report on new techniques.

6. CONCLUSIONS AND FUTURE WORK

In this paper we presented a new corpus of multi-channel audio data to study distant-speech recognition systems when known interfering sounds are propagating in the environment. The corpus has been used to produce a set of ASR baseline results and analyse the combination of a speech recognition engine, beamforming and AEC techniques. Recognition performance achieved using a distant-talking command task in the real application scenario of the DIRHA project, shows the benefit which can be obtained with a suitable front-end processing. This needs to be further investigated in a joint optimisation perspective.

The corpus is being extended to other languages, including English, and part will be made available at public level along with the other DIRHA corpora.

7. ACKNOWLEDGEMENTS

The research presented here has been partially funded by the European Union's 7th Framework Programme (FP7/2007-2013) under grant agreement no. 288121 DIRHA (see `http://dirha.fbk.eu`).

We would also like to thank Luca Cristoforetti and Alessandro Sosi for their help in recording the real AEC corpus and Marco Pellin for his help in the data annotation.

8. REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer-Verlag, 2008.

[2] D. Mostefa et al., "The CHIL Audiovisual Corpus for Lecture and Meeting Analysis inside Smart Rooms," *Language, Resources and Evaluation*, vol. 41, no. 3, pp. 389–407, 2007.

[3] J. Carletta et al., "The AMI Meeting Corpus: A Pre-announcement," in *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, 2006, pp. 28–39.

[4] K. Kinoshita et al., "Proceedings of ieee reverb challenge," in *Proceedings of IEEE WASPAA*, 2013, pp. 1–4.

[5] 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.

[6] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, November 2006.

[7] 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, no. 2, pp. 205–223, 2002.

[8] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmüller, and P. Maragos, "The DIRHA simulated corpus," in *Proceedings of LREC*, 2014.

[9] 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 *Proceedings of INTERSPEECH*, 2014, pp. 1613–1617.

[10] A. Neustein, *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, Springer Publishing Company, Incorporated, 2010.

[11] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, 2001.

[12] C. Breining et al., "Acoustic echo control. An application of very-high-order adaptive filters," *Signal Processing Magazine, IEEE*, vol. 16, no. 4, pp. 42–69, 1999.

[13] M. Matassoni, M. Omologo, and C. Zieger, "In-car audio compensation based on NLMS for hands-free speech recognition," in *Proceedings of IEEE ICASSP*, 2004, pp. 2591–4594.

[14] J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Magazine*, vol. 9, no. 1, pp. 14–37, 1992.

[15] F. Nesta, T. Wada, and B. Juang, "Batch-online semi-blind source separation applied to multi-channel acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 583–599, 2011.

[16] C. Paleologu, J. Benesty, and S. Ciochina, *Sparse Adaptive Filters for Echo Cancellation*, Morgan & Claypool Publishers, Synthesis Lectures on Speech and Audio Processing, 2010.

[17] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[18] G. Lathoud, I. McCowan, and D. Moore, "Segmenting multiple concurrent speakers using microphone arrays," in *Proceedings of Eurospeech*, 2003, pp. 2889–2892.

[19] A. Brutti et al., "WOZ acoustic data collection for interactive TV," in *Proceedings of LREC*, 2010.

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

[21] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proceedings of 108th AES Convention*, 2000.

[22] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *Proceedings of IEEE ICASSP*, 1997, pp. 219–222.

[23] W. Herbordt, S. Nakamura, and W. Kellermann, "Joint optimization of LCMV beamforming and acoustic echo cancellation for automatic speech recognition," in *Proceedings of IEEE ICASSP*, 2005, pp. 77–80.

[24] G. Reuven, S. Gannot, and I. Cohen, "Joint noise reduction and acoustic echo cancellation using the transfer-function generalized sidelob canceller," *Speech Communication*, vol. 49, no. 7–8, pp. 623–635, 2007.

[25] K. Kammeyer, M. Kallinger, and A. Mertins, "New aspects of combining echo cancellers with beamformers," in *Proceedings of IEEE ICASSP*, 2005, pp. 137–140.

[26] W. Kellermann, "Acoustic echo cancellation for beamforming microphone arrays," in *Microphone Arrays*, D. Ward M. Brandstein, Ed., pp. 281–306. Springer, 2001.

[27] M. Omologo and C. Zieger, "Comparison between Subband and fullband NLMS for in-Car audio compensation and hands-free speech recognition," in *Proceedings of IWAENC*, 2005, pp. 189–192.

[28] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[29] E. Zwyssig, S. Renals, and M. Lincoln, "On the effect of snr and superdirective beamforming in speaker diarisation in meetings," in *Proceedings of IEEE ICASSP*. IEEE, 2012, pp. 4177–4180.

[30] E. Zwyssig, *Speech processing using digital MEMS microphones*, Ph.D. thesis, The University of Edinburgh, 2013.

[31] S. Young et al., *The HTK Book*, Entropic Cambridge Research Laboratory, Cambridge, United Kingdom, 2002.

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

[33] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, "Speaker independent continuous speech recognition using an acoustic-phonetic Italian corpus," in *Proceedings of ICSLP*, 1994, pp. 1391–1394.