Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 47 (2018) 314–332

www.elsevier.com/locate/csl

Cepstral distance based channel selection for distant speech recognition☆

Cristina Guerrero Flores[a,b], Georgina Tryfou[a,b], Maurizio Omologo[b,*]

[a] *University of Trento Via Sommarive, 14, 38123 Trento, Italy*
[b] *Fondazione Bruno Kessler-irst Via Sommarive 18, 38123 Trento, Italy*

Received 9 November 2016; received in revised form 9 August 2017; accepted 18 August 2017
Available online 25 August 2017

Abstract

Shifting from a single to a multi-microphone setting, distant speech recognition can be benefited from the multiple instances of the same utterance in many ways. An effective approach, especially when microphones are not organized in an array fashion, is given by channel selection (CS), which assumes that for each utterance there is at least one channel that can improve the recognition results when compared to the decoding of the remaining channels. In order to identify this most favourable channel, a possible approach is to estimate the degree of distortion that characterizes each microphone signal. In a reverberant environment, this distortion can vary significantly across microphones, for instance due to the orientation of the speaker's head. In this work, we investigate on the application of cepstral distance as a distortion measure that turns out to be closely related to properties of the room acoustics, such as reverberation time and direct-to-reverberant ratio. From this measure, a blind CS method is derived, which relies on a reference computed by averaging log magnitude spectra of all the microphone signals. Another aim of our study is to propose a novel methodology to analyze CS under a wide set of experimental conditions and setup variations, which depend on the sound source position, its orientation, and the microphone network configuration. Based on the use of prior information, we introduce an informed technique to predict CS performance. Experimental results show both the effectiveness of the proposed blind CS method and the value of the aforementioned analysis methodology. The experiments were conducted using different sets of real and simulated data, the latter ones derived from synthetic and from measured impulse responses. It is demonstrated that the proposed blind CS method is well related to the oracle selection of the best recognized channel. Moreover, our method outperforms a state-of-the-art one, especially on real data.

Keywords: Distant speech recognition; Channel selection; Cepstral distance; Reverberation; Direct to reverberant ratio; T60

1. Introduction

Despite the extensive efforts that have been made for reliable automatic speech recognition (ASR), the performance of many voiced based systems is still inadequate under certain conditions. For example, ASR is seriously affected by the presence of reverberation, background noise, and overlapping speakers. In order to overcome these

☆ This paper has been recommended for acceptance by Roger Moore.

* Corresponding author at: SHINE - Fondazione Bruno Kessler. Via Sommarive, 18 I-38123 Trento, Italy.
E-mail address: cristinaguerreroflores@gmail.com (C. Guerrero Flores), tryfou@gmail.com (G. Tryfou), omologo@fbk.eu (M. Omologo).

http://dx.doi.org/10.1016/j.csl.2017.08.003
0885-2308/

limitations in distant-talking scenarios, some of the most effective strategies adopt the use of multiple microphones (Wölfel and McDonough, 2009; Brandstein and Ward, 2001). There are many applications, *e.g.,* in domestic environments, for which a significant improvement in terms of speech recognition rate can be obtained by deploying a large number of microphones, clustered in arrays with specific geometries, and distributed in such a way to cover the whole environment. A sparse distribution of single microphones in space, combined with an automatic channel selection (CS), represents a simple and effective solution to limit the overall complexity of a distant speech recognition (DSR) system.

CS makes the reasonable assumption that among the acquired microphone signals there is one that can lead to a better recognition performance than the others. In order to identify the related microphone, it is worth addressing the attributes of the signal and the characteristics of the communication *channel* that shaped the uttered speech from the source to the sensor, and depends on the speaker location, the head orientation, and the room acoustics. The latter variabilities determine the overall reverberation effects that can be observed in the distant microphone signal. Environmental noise, although it is not the main focus of this work, also represents a relevant issue, in particular when it is more concentrated in some areas, *i.e.,* when it introduces more distortion into a subset of the available microphones.

Various CS methods have been presented in the literature, as reported in the following. Some of them rely on measures that quantify the effect of the channel on the speech signals. Examples of these measures are the envelope variance (EV) (Wolf and Nadeu, 2014) and the modulation spectrum ratio (Himawan et al., 2015). Also energy-based techniques can be applied to CS, in particular under controlled conditions as when a calibrated set of microphones is available (Wolf and Nadeu, 2010).

In a previous work, we presented an initial study of how objective signal quality measures, in particular the cepstral distance (CD), can be successfully applied to CS problem (Guerrero et al., 2016). However, we believe that an important requirement, for a more effective application of these quality measures to our problem, is an in-depth understanding of their relationship with DSR performance. In order to address this missing link between CS and DSR, this work aims to provide a novel methodology for assessing the performance and limitations of CS methods, as far as reverberation effects are concerned. To the best of our knowledge, this represents the first empirical study that characterizes, from a quantitative standpoint, the overall system behavior under parameters such as the distance between the speaker and microphones, the speaker orientation, and the microphone network configuration. Additionally, this work constitutes an extensive and deeper investigation of the CD based technique outlined in Guerrero et al. (2016). We discuss the effectiveness of CD to characterize the reverberation in a room e.g., relating it to the direct-to-reverberant ratio (DRR) feature, supporting its application to CS for DSR. Also, we present evidence that shows that CD based CS is strongly related to an oracle selection of the best recognized channels. Then, the investigated approach is analyzed under variations on the setup that regard the speaker position and orientation, and the microphone network configuration. Finally, we extend our findings and confirm the benefits of applying CS to DSR with the use of real data, on which the proposed method achieves a better performance than an EV based state-of-the-art method.

The remaining of this paper is organized as follows. In Section 2 multi-microphone processing for DSR is discussed. Specific parameters of the room acoustics are presented in Section 3. An overview of the most relevant CS methods is given in Section 4. CD based CS is elaborated in Section 5. In Section 6, details about the experimental framework are provided. The activities and analysis performed on the different experimental settings, and their corresponding results, are presented in Sections 7 and 8. Finally, in Section 9 the conclusions of the study and possible directions for future activities are discussed.

2. Multi-microphone processing for DSR

The problem of DSR in a multi-microphone setting comprises, on one hand, the techniques used for multi-microphone speech processing and, on the other hand, the acoustic properties of the reverberant environments.

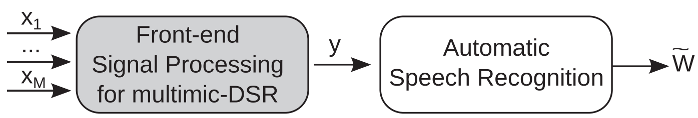
Multi-microphone speech processing approaches have proved their potential to significantly improve DSR performance in comparison to single channel solutions. Various architectures can be adopted to process the multiple inputs in order to derive a single recognition output of a spoken utterance (Wölfel and McDonough, 2009; Kinoshita et al., 2013).

In the most commonly exploited solutions, processing modules operate either at front-end or at post-decoding processing level. These two cases are depicted in Fig. 1. As shown in Fig. 1a, front-end signal processing can be used in order to extract a single signal that is then used as an input for the ASR system. Examples of such techniques are beamforming (Brandstein and Ward, 2001) and CS (Wolf and Nadeu, 2014), both aiming at reducing the number of signals to process at a subsequent recognition level. An effective practice consists in combining front-end processing approaches. As an example, Kumatani et al. (2011) presented a system where a selection of multiple channels was performed for applying beamforming on a reduced set of signals. However, the use of beamforming limits the scope of such methods to scenarios that employ microphone-arrays characterized by a very limited distance between adjacent microphones. Inter-sensor spacing generally affects the resolution of spatial sampling in any array processing application (Van Veen and Buckley, 1988). In particular, this problem becomes critical in distant-speech applications, due to the broadband nature of speech (Flanagan et al., 1985; Ward et al., 1995).

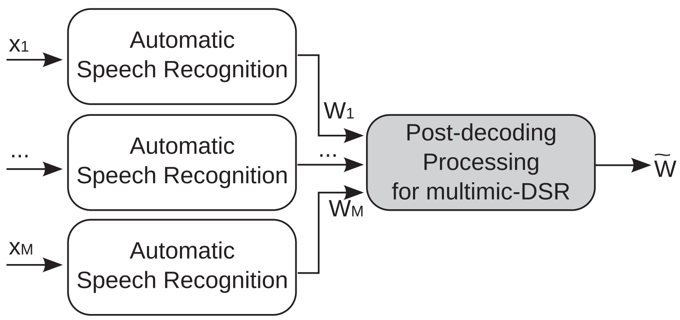
Post-decoding processing approaches perform a combination of information at the last stage of the recognition system, as shown in Fig. 2b. Renown methods, such as ROVER (Fiscus, 1997) and Confusion Network Combination (Evermann and Woodland, 2000), require an individual, parallel recognition of each input signal before applying their combination algorithms. Other methods, such as decoder-based CS (Obuchi, 2004) have also been explored. The complexity and resource demanding nature of post-decoding processing solutions increases with the number of captured channels.

In any of the above cases, the design of effective multi-microphone DSR systems is challenging due to the distortions introduced by reverberation effects. Given the acoustics of the environment, and the position and orientation of the speaker, each channel between the source and one of the distant microphones is described by an acoustic impulse response (IR). IR is the best possible descriptor of the convolutional distortion that is introduced in the source signal, to transform it in a corresponding distant microphone signal. In principle, under non-noisy reverberation conditions the knowledge of the IR would be sufficient to deduce the amount of reverberation characterizing each distant microphone signal, as well as its quality, and as a consequence the possible challenge in its automatic recognition.

In a real experimental environment, the IRs can be measured with the use of the exponential sine sweep (Farina, 2000), as detailed in Ravanelli et al. (2012). Alternatively, IRs can be synthesized through the image method (IM) assuming a shoe-box geometry for a simulated room (Allen and Berkley, 1979). Unfortunately, in a real DSR application blindly estimating the IR from the distant-speech signal is not possible. However, extracting some IR-related



(a) Front-end signal processing. M microphones capture the input signals x_i which are processed in order to extract a signal y to be decoded into the final recognition output $\tilde{W}$



(b) Post-decoding processing. Each signal x_i is decoded, and then individual recognition outputs are processed to extract the final output $\tilde{W}$.

Fig. 1. Typical architectures for multi-microphone DSR.

information from the reverberated speech signal can be of crucial importance as far as the problem of channel selection is concerned, as further discussed in the next sections.

3. Reverberation time and direct-to-reverberant ratio

When available, IRs can be exploited to estimate parameters that characterize the reverberation in a non-anechoic room. Two important parameters are the reverberation time (T_{60}) and DRR (Kuttruff, 2009; Jo and Koyasu, 1975). The T_{60} is defined as the time required for a sound to decay $60dB$ from its initial level, after an abrupt cessation of the source (Kuttruff, 2007). The DRR is defined as the ratio of the sound energy that arrives to the microphone through a direct path, over the sound energy that arrives to the microphones after one, or multiple, reflections on the various surfaces (Naylor and Gaubitch, 2010).

Both parameters can be directly estimated from the IRs (Schroeder, 1965; Naylor and Gaubitch, 2010; Zahorik, 2002; ISO, 2008), and they can be derived in the full-band, expressed as single values representing all the frequencies up to the Nyquist rate, or in ISO-preferred frequency bands[1]. Nowadays, different systems are available to derive T_{60} estimates, either in full-band or in sub-bands, starting from a synthetic or from a measured IR. As discussed in Cabrera et al. (2016), a good agreement between these estimates is generally found, if the IR is characterized by a regular decay curve and low enough noise floor. On the other hand, an accurate estimation of DRR from an IR is a more difficult task, especially in the case of measured IR, due to the uncertainty in deriving the energy of the direct-path (Naylor and Gaubitch, 2010).

Of course, these tasks become much more challenging in real situations in which IRs are not available. Although different approaches have been proposed to estimate the aforementioned parameters blindly (Ratnam et al., 2003), the results are still not satisfactory, particularly for the estimation of DRR. More details on state-of-the-art techniques in this field can be found in Eaton et al. (2016) that is related to the recent Acoustic Characterization of the Environments (ACE) challenge. We believe that improvements in multi-microphone based DSR can also be achieved taking into consideration reverberation parameters as blind DRR.

4. Channel selection

CS methods share the objective to detect the least distorted channel among the available ones, assuming that a better match will result between the selected channel and the acoustic models of the DSR system. CS can be applied either at front-end or at post-decoding level, commonly referred to as *signal based* and *decoder based* approaches, respectively. In both cases, one relies on a specific measure which is optimized for the final selection. According to the type of information exploited for the computation of their measure, these methods can be further categorized into *informed* and *blind methods*.

Informed methods exploit measures computed with the use of prior information. Although not directly applicable in a real scenario, they are interesting because they can be used to study the effectiveness of the related measures under diverse reverberation conditions. In addition, such methods can be explored to derive an upper-bound performance (Wolf and Nadeu, 2010). In particular, the oracle CS which exploits the word error rate (WER) of each recognized signal in order to identify the best channel, can also be seen as an informed decoder-based method to use for reference purposes.

Although most of the measures described in the literature can be easily modified to be used in an informed way, very few authors have performed such a study. In Wolf and Nadeu (2009) measured IRs were used to verify the assumption that DSR can be benefited from IR based CS. In Wolf and Nadeu (2010) the signal-to-noise ratio (SNR) and the position/ orientation of the speaker were used for computing informed measures.

Blind decoder based CS methods use information such as the likelihoods or posterior probabilities, to assess the quality of each channel. Therefore, it is not possible to apply decoder based techniques independently from the ASR process. Some representative examples of such decoder based methods can be found in Shimizu et al. (2000), Obuchi (2004), Obuchi (2006), Wölfel (2007). A detailed review of this topic can be found in Wolf (2013). Although there is the assumption that decoder based measures should present a higher correlation to WER in DSR, this has not been so far proven in the literature (Wolf and Nadeu, 2014).

[1] The experiments conducted in this work always refer to the use of full-band T_{60} and DRR, estimated from synthetic and from measured IRs.

Blind signal based CS methods include, among others, the use of energy and SNR, cross-correlation between signals (Kumatani et al., 2011), and the modulation spectra of the original and the beamformed signals (Himawan et al., 2015). As for energy based methods, for which one can refer to Wolf and Nadeu (2010), a critical issue is related to the need of a preliminary gain calibration. Other measures, e.g. SNR, are generally more robust in non-calibrated settings. One of the most successful measures described in the literature is EV (Wolf and Nadeu, 2014). EV based CS exploits the fact that the reverberation smooths the energy of speech signals. This is observed as a reduction in the dynamic range of the envelope in the speech portions of the input signal. For the calculation of the EV measure, the filter-bank energies (FBE) $X_m(k, l)$ in channel m, sub-band k and time frame l, are first normalized as follows

$$\widehat{X}_m(k,l) = e^{\log X_m(k,l) - \mu_m(k)}, \tag{1}$$

where the mean value $\mu_m(k)$ is calculated over the logarithm of the FBE of the entire speech utterance. The mean normalized sequence of FBE is then compressed with the application of a cube root function, and the variance $V_m(k)$ of each sub-band k, for each channel m, is extracted.

EV based CS selects the channel that maximizes the average variance over all channels:

$$\widehat{M}_V = \arg\max_m \sum_k \frac{V_m(k)}{\max_m \left(V_m(k)\right)}. \tag{2}$$

The application of a different weighting for each channel and sub-band in (2) was proposed in Wolf and Nadeu (2014). However, to the best of our knowledge, no further elaboration of this concept has been described, and no experimental evidence has been derived to support the use of such a weighting scheme.

5. Cepstral distance based CS

Objective signal quality measures have been exploited for many years in various speech processing applications (Quackenbush et al., 1988; Loizou, 2013). Measures such as the CD, the log-likelihood ratio (LLR) (Hansen and Pellom, 1998) and the frequency weighted segmental SNR (fwSSNR) (Tribolet et al., 1978) were initially introduced in the speech coding community (Gray and Markel, 1976; Kitawaki et al., 1988; Furui and Sondhi, 1991) as a means of measuring the amount of distortion introduced by a speech codec. In the 1980s, some of these measures were also diffusely employed in the development of early speech recognition systems (Rabiner and Juang, 1993; Rabiner and Schafer, 2011). Other examples of applications include speech enhancement, noise reduction (Rohdenburg et al., 2005) dereverberation, and robust speech recognition, as more recently addressed under the REVERB challenge (Kinoshita et al., 2013). Perceptually based measures, as PESQ (Rix et al., 2001), and composite techniques combining a subset of the aforesaid measures, also represent very effective approaches to measure speech quality.

In general, as shown in Hu and Loizou (2008) as well as in other works, many of these objective quality measures correlate well with subjective evaluation of signal quality. It is also worth noting that some of these measures were explored in “non-intrusive” speech quality evaluation tasks, which do not require the access to the original clean signal. The ITU-T P.563 recommendation (Malfait et al., 2006) describes a standardized method based on LPC as well as on cepstrum derived parameters.

This background supports the fact that objective signal quality measures and, in particular, CD based ones, can also be used for the selection of the least distorted channel for DSR (Guerrero et al., 2016). In our work, among the various objective measures, CD was selected due to its long known effectiveness and flexibility in different application fields as well as its low computational complexity. Moreover, cepstrum based comparisons are equivalent to comparisons of the smoothed log spectra of the signals (Rabiner and Schafer, 2011), a domain in which the reverberation effect can be viewed as additive (Huang et al., 2001).

In this work, we use the truncated CD definition as in Hu and Loizou (2008):

$$d(\Rightarrow c_x, \Rightarrow c_m) = \frac{10}{\log 10}\sqrt{2\sum_{i=1}^{p}\left[\Rightarrow c_x(i) - \Rightarrow c_m(i)\right]^2}, \tag{3}$$

where p is the number of cepstral coefficients. These coefficients are computed, as usually done in speech recognition, on windowed finite-duration segments of the input signal. As indicated in (3) the 0th order coefficient, associated with the energy of the signal, is disregarded.

The cepstral coefficient vectors $\Rightarrow c_x$ and $\Rightarrow c_m$ correspond to the reference and a distorted signal, respectively. Based on the nature of the reference, the proposed method is characterized either as *informed* or as *blind*.

5.1. Informed channel selection

In the informed CS method we assume the availability of the close-talk speech signal, $x(t)$. Each distant microphone signal can be expressed as follows:

$$x_m(t) = x(t) * h_m(t) + n(t) \tag{4}$$

where m is the microphone index, $h_m(t)$ is the related IR, and $n(t)$ denotes a possible additive noise. As previously pointed out, in this work we are assuming that $x_m(t)$ is not distorted by environmental noise. Indeed, environmental noise can consist of contributions of very different nature, it can be generated by point vs diffuse sources, it can be stationary or non-stationary, etc. Omologo et al. (1998). As the primary focus of this paper is on reverberated speech, in the following we assume that the last term of (4) drops. Equivalently, in the short-time Fourier transform (STFT) domain each distant microphone signal is expressed as

$$X_m(t,\omega) = X(t,\omega)H_m(t,\omega). \tag{5}$$

The *complex cepstrum* of $X_m(t, \omega)$ is defined as the inverse Fourier transform of its complex logarithm. In practice, as in many speech processing applications, the *complex cepstrum* is replaced here by the *real cepstrum*, which uses the logarithm of the magnitude of $X_m(t, \omega)$. This can be written as

$$\log|X_m(t,\omega)| = \log|X(t,\omega)| + \log|H_m(t,\omega)|. \tag{6}$$

From this representation it can be inferred that the CDs $d(\Rightarrow c_x, \Rightarrow c_m)$ between the close-talk and the reverberated signals are more affected by the set of IRs than by the content of the spoken utterance, that is deemphasized by the subtraction in (3).

Given the set of CDs $d(\Rightarrow c_x, \Rightarrow c_m)$, and assuming that the least distorted channel corresponds to the one *nearest* to the close-talk signal, the selection is performed as follows:

$$\widehat{M_x} = \arg\min_m d(\Rightarrow c_x, \Rightarrow c_m). \tag{7}$$

5.2. Blind channel selection

Since in a real scenario the close-talk signal is not available, in Guerrero et al. (2016) we proposed a non-intrusive way to estimate CDs, from which CS is performed. The method relies on the assumption that one of the distant microphone signals is characterized by a higher DRR. This typically occurs when the speaker is oriented towards that microphone and/or the speaker is located closer than the critical distance. The remaining channels are more affected by several degrading factors, for example attenuation effects due to the multiple reflections and to the head of the speaker.

Based on the above assumption, we proposed to compute a reference as the logarithm of the geometric mean of the signals $x_m(t)$, in the magnitude spectrum domain:

$$\widehat{R}(t,\omega) = \log \prod_m |X_m(t,\omega)|^{1/M} \tag{8}$$

$$= \frac{1}{M}\sum_m \log|X_m(t,\omega)|. \tag{9}$$

where $X_m(t, \omega)$ is the STFT of the signal captured by microphone m, and M is the total number of microphones.

The cepstrum computed from the reference is then used to calculate the distance between the reference and each microphone signal $d(\Rightarrow c_{\widehat{R}}, \Rightarrow c_m)$. The least distorted channel can be selected as the one *furthest* from the reference:

$$\widehat{M}_{\widehat{R}} = \arg\max_m d(\Rightarrow c_{\widehat{R}}, \Rightarrow c_m). \tag{10}$$

In order to better explain the proposed method, we elaborate on (9), which with the use of (6) can be rewritten as:

$$\widehat{R}(t,\omega) = \frac{1}{M}\sum_{m}\left[\log|X(t,\omega)| + \log|H_m(t,\omega)|\right] \quad (11)$$

$$= \log|X(t,\omega)| + \frac{1}{M}\sum_{m}\log|H_m(t,\omega)|. \quad (12)$$

The second term of (12) represents an estimation of the average reverberation that affects the multiple instances of the close-talk signal. Assuming to have a set of microphones uniformly distributed in space, with one characterized by a substantially higher DRR than the others, the resulting reference will be strongly influenced by the latter ones, *i.e.,* it will be far from the former.

Of course, a favourable situation as the one previously outlined can not always be expected. For example, if all channels are equally impinged by reverberation, the selection of a specific channel is not relevant for improving the recognition performance. It is expected that in such cases the decoding of all the microphone signals will result in a similar recognition error rate. For this reason, we focus on scenarios in which CS is meaningful, *i.e.,* scenarios that feature the speaker at favourable positions and/or orientations in relation to at least one of the microphones.

6. Experimental setup

6.1. Multi-microphone environments

In this study, we use two experimental multi-microphone environments, namely the SQUARE and the DIRHA rooms. These two rooms are schematically presented in Fig. 2 and Fig. 3, respectively. Their detailed characteristics are given in Table 1. In both settings, the average distance between the speaker and the microphones fluctuates in the range of 1−4 meters. In contrast to other studies performed in much reduced spaces, the distance explored in this work implies that reverberation significantly affects the degree of signal distortion.

The SQUARE room is simulated using IRs generated with our IM tool, which offers the possibility to set the orientation of the source with a given acoustic directivity pattern. Moreover, it gives a fine control of several other parameters as, for example, T_{60}. We simulated the speaker located at various positions, and at each position we simulated 36 orientations. The obtained rich set of positions/orientations, and microphone configurations constitutes a strong experimental framework for the study of CS in a wide range of scenarios, from the most favourable to very challenging conditions.

The DIRHA room corresponds to the living-room of a real environment, a scenario taken from the DIRHA Project setup (Cristoforetti et al., 2014), see http://dirha.fbk.eu. This room is studied under two modalities: i) as a simulated realistic scenario created with measured IRs (Cristoforetti et al., 2014), and ii) as a real space, with the use of

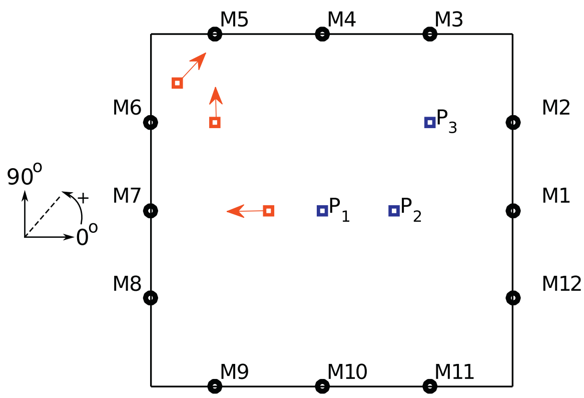


Fig. 2. SQUARE room setting. Black dots indicate the microphone locations, and blue squares show the various positions of speaker. The red squares and arrows indicate the channels used for training the acoustic models in the DSR experiments performed in this room. Orientations are given as depicted on the polar coordinate system on the left. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

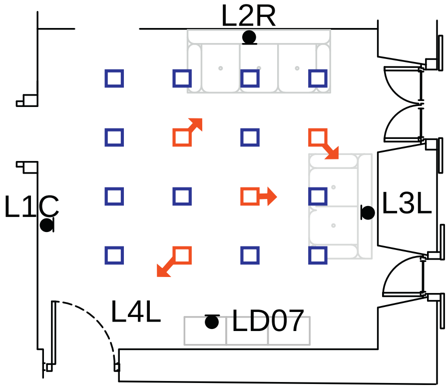


Fig. 3. DIRHA room setting. Black dots indicate the microphone locations, and blue squares show the various positions of the speaker. The red squares and arrows indicate the channels used for training the acoustic models in the DSR experiments performed in this room. L1C, L2R, L3L, and L4L denote high-quality microphones that are fixed on the wall, while LD07 is an electret microphone of a linear array standing on top of a bookshelf. All the microphones are at a height of approximately 2 m. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
The main characteristics of the experimental environments.

	SQUARE	DIRHA
Size (m)	4.80 × 4.80 × 2.74	4.83 × 4.51 × 2.74
T_{60} (s)	0.7	0.75
# microphones	12	5
# positions	5	16
# orientations	36	4
IM IRs	yes	no
Measured IRs	no	yes
Real data	no	yes

real audio recordings (Ravanelli et al., 2015). In both cases, the set-up consists of a set of omnidirectional microphones connected with a professional multi-channel pre-amp and digital recording equipment, operating at 48 kHz/24 bit. The recorded signals are then downsampled to 16 kHz/16 bit for the experiments discussed in the next sections. It is also worth noting that all the input channels are sample synchronized and calibrated at gain level.

6.2. Data sets

For the DSR task, we use the Wall Street Journal (WSJ) corpus. A subset of the clean WSJ (WSJ0-5k) (Garofalo et al., 1993) training set, comprising 7138 utterances, is used as source material for training. As shown in Figs. 2 and 3, the training sets were reverberated using IRs corresponding to a small set of positions and orientations for each experimental room. These position/orientation/microphone combinations describe conditions in which the speaker is directly oriented towards a microphone.

The test material is extracted from the WSJ0-5k sub-set of the DIRHA-English (Ravanelli et al., 2015) corpus[2]. From this corpus we use the clean material recorded in the FBK recording studio to generate the simulations using measured IRs. Furthermore, from the same corpus we use real distant speech recordings that were captured in the DIRHA apartment. The close-talk signals of the latter real data set were also captured by a head-set worn by the speaker during the recording sessions. An ideal voice activity detection is assumed to be applied over the real data, *i.e.,* ground truth boundaries were used. Finally, it is worth noting that the recordings of real material in the DIRHA apartment were performed under real conditions, *i.e.,* with the environmental noise that typically characterizes a domestic environment and, more critical, with reverberation effects that are represented by a T60 of about 0.75 s.

[2] A public distribution of DIRHA-English WSJ data set is available through LDC. A phonetically rich data set is also available, as described in http://dirha.fbk.eu/English-PHdev.

In the SQUARE room we use a data set which includes 120 sentences, referred to as WSJ120 data set. To create this data set we randomly selected 20 utterances for each of the 6 speakers included in the WSJ0-5k DIRHA-English corpus. Given the fact that each recognition experiment performed in this room is repeated for the whole data set at each position and orientation, a preliminary experiment showed that this is a sufficient number of utterances to consider.

6.3. Speech recognition

All the recognition experiments are performed using the Kaldi speech recognition toolkit (Povey et al., 2011). The experiment recipe is based on the one in Ravanelli et al. (2015), adapted to operate on the explored multi-microphone scenarios. The acoustic features are 13 mel-frequency cepstral coefficients (MFCC), augmented with their first and second order derivatives. Additional feature transformations are applied, which include linear discriminant analysis, maximum likelihood linear transformation, and feature space maximum likelihood linear regression with speaker adaptive training. These feature transformation techniques have been shown to be effective for distant talk speech recognition (Tachioka et al., 2013). The language model is the baseline language model used in CHiME-3 (Barker et al., 2015), which is the standard 5k WSJ trigram. For the recognition results we exploited a decoding system based on Deep Neural Network (DNN) acoustic models, in the following referred to as *dnn*. The *dnn* system is trained following Karel Veselỳ's setup (Veselỳ et al., 2013) included in the KALDI toolkit. For the *dnn* configuration a context window of 11 frames is used as input.

Using DNNs, the recognition performance on the close-talk material captured in the FBK recording studio yields a WER of 3.7%. Acoustic models trained on data simulated with IM IRs were used to decode test data also simulated with the use of IM. Acoustic models trained on simulations generated with measured IRs were used to decode simulated data based on measured IRs and real test data as well.

6.4. Channel selection methods

The following methods are included in the evaluation:

- **oracle** refers to an informed CS method that selects, for each utterance, the microphone corresponding to the channel with the lowest recognition error rate.
- **CD informed** corresponds to the *informed CD based* CS method that uses the close-talk reference, as explained in Section 5.1.
- **CD blind** is the *blind CD based* method that uses the reference described in Section 5.2.
- **EV** is the state-of-the-art CS method, based on Envelope Variance (Wolf and Nadeu, 2013).

For the calculation of the CD, the analysis step and window length are set to 10 ms and 25 ms, respectively, while the order p is equal to 24. As mentioned before, the 0th order coefficient is discarded. The CD of an utterance is calculated as an average over the CD of each analysis frame, a value which is truncated in the range of [0, 10] to minimize the number of outliers, as in Hu and Loizou (2008). The EV measure is estimated with the same analysis step and window length as above. The applied filter-bank consists of 20 sub-bands.

7. Experiments in the SQUARE room

In this section we report the experiments performed in the SQUARE room setting, based on the use of IM generated IRs. Concerning speech recognition, all the experiments were conducted using the *dnn* system detailed in the previous section.

7.1. Relation between CD and reverberation parameters

A first aspect to investigate concerns the relation between CD and reverberation parameters. As earlier outlined, we assume that CD is able to characterize the reverberation present in a channel, in a way consistent with the

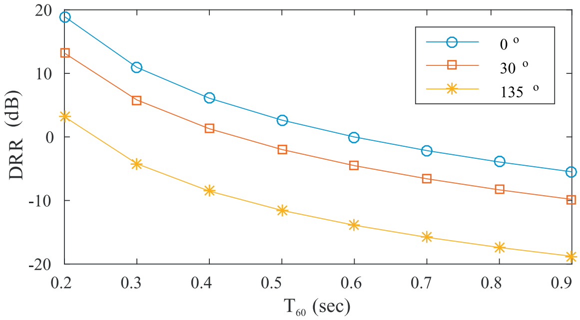


Fig. 4. DRR as a function of T_{60}. T_{60} values range from 0.2 s to 0.9 s. which are reasonable for a domestic environment. The speaker is located at P_1 adopting three different orientations. We consider only the microphone M1.

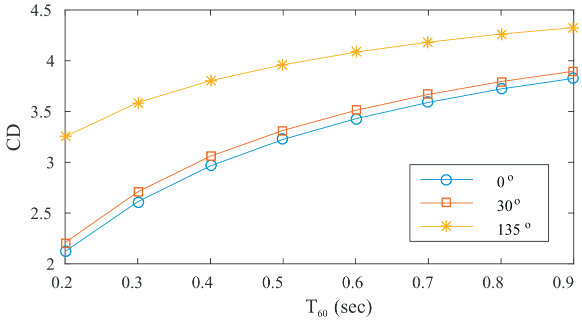


Fig. 5. CD of a reverberated to a close-talk signal, in terms of increasing reverberation time. The WSJ120 data set is used here, with the speaker located at P_1 adopting three different orientations. In this experiment we consider only the microphone M1.

reverberation parameters T_{60} and DRR. As a starting point, in Fig. 4 we present the DRR[3] as a function of T_{60} for three different orientations 0^o, 30^o and 135^o. It is observed that DRR directly relates to the orientation of the speaker towards the microphone, with more directive cases, as for example 0^o and 30^o resulting in higher DRRs. To exemplify a case addressed in the following, when T_{60} is equal to 0.7 s the DRR is higher than −6dB in the range −30: +30 degrees. Moreover, as expected, there is an inverse relation between T_{60} and the corresponding DRR, which is respected in low, average and higher DRR cases, as the different orientations show.

As a next step, for the same cases explored in the previous analysis, the average CD between the clean and the reverberated signals is computed. The results are shown in Fig. 5, where it becomes evident that CD has a behavior very similar to the DRR. The average[4] distance monotonically increases along with the increasing T_{60} , *i.e.,* CD is expected to follow the DRR changes in an inverse fashion.

In order to investigate the previously outlined relation between these parameters, in the following experiments we report the variation of both measures under multiple orientations. The CD between the clean and reverberated signals as a function of different orientations adopted by the speaker is presented in Fig. 6a. The DRR of the IRs used to reverberate the corresponding utterance is shown in Fig. 6b. As expected, it is observed that when the speaker is oriented towards the microphone under consideration, *i.e.,* orientation 0^o, the minimum CD and maximum DRR are measured. In addition, there is a clear inverse behavior between the two measures over all orientations.

Next, we perform the above experiment for a different position, and additional microphones. The set of CDs between the close-talk signal and four reverberated instances, *i.e.,* microphones M1, M4, M7 and M10, as a function

[3] DRR is estimated from the synthetic IRs with the use of the IR_stats toolbox of MATLAB (Zahorik, 2002).

[4] Note that here, and in some of the following experiments, CD is averaged over 120 utterances. Actually, a much smaller number of utterances would be sufficient to converge to the same experimental evidence, in particular for lower reverberation times. Indeed, CD can be influenced by the length and the contents of the utterance, especially with higher reverberation times and/or lower DRRs.

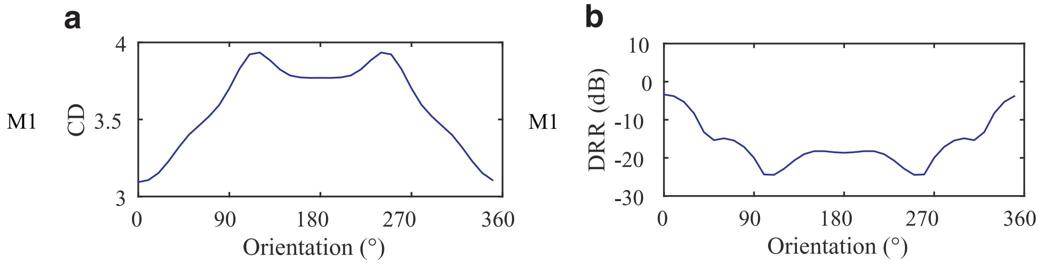


Fig. 6. CD and DRR as a function of different orientations, for an utterance simulated with the speaker located at P_1. Results are presented here for the microphone M1.

of different orientations are shown in Fig. 7a. In Fig. 7b we present the DRRs of all the corresponding IRs. These results confirm the previous insights concerning the relation between CD and DRR. In addition, this case illustrates how these parameters vary under more complex conditions, for instance when the speaker is oriented towards a microphone, but at a considerably larger distance. As an example, in Fig. 7, we can compare the curves computed

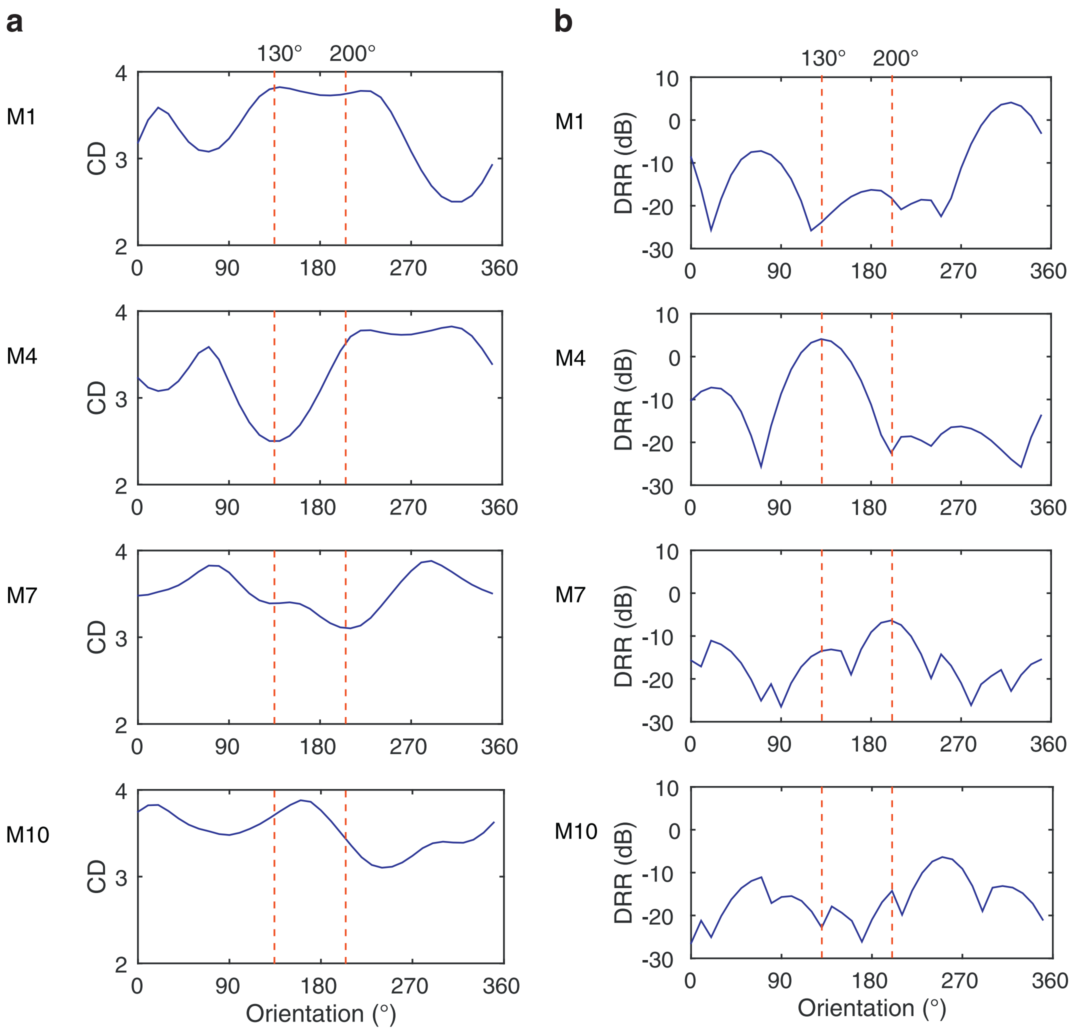


Fig. 7. CD and DRR as a function of different orientations for an utterance simulated with speaker located at P_3. Results are presented, from top to bottom, for the microphones M1, M4, M7 and M10. Dashed lines emphasize two orientations of interest in the related discussion.

from signals captured by microphones M4 and M7, which represent cases of a near and a far microphone, respectively. When the speaker is oriented at around 130°, *i.e.,* direct towards M4, in the related sub-figures there is a clear distinction of the lowest CD and highest DRR both over all orientations and over all microphones. On the other hand, when the speaker is oriented at around 200°, *i.e.,* direct towards M7, although the curves are characterized by a minimum CD and a maximum DRR, the distinction of these values is not so clear. This can also be related to the average DRR decrease, and exposes a complex non discriminative scenario for the identification of the least distorted channel, even with the exploitation of prior information.

7.2. Relation between CD and WER

The above findings provide an important basis for the use of CD as a means for selecting the least reverberant channel. However, a second assumption that has to be verified is that higher WERs correspond to reverberated signals which present larger CDs from the clean signals. In order to study this aspect, in Fig. 8 we present the scatter graph for CD and WER value pairs. Each point relates the average CD between the close-talk and the reverberated signals for the WSJ120 data set, and the average WER that results from decoding the reverberated signals.

It is evident that CD is not only related to the reverberation time, as previously discussed, but also to the recognition rate. In fact, a clear trend can be observed that an increasing degree of signal distortion, as measured by CD, corresponds to an increasing WER. For cases in which a speaker is directly oriented toward a microphone, associated to lower CDs, lower WERs are observed. As for the large fluctuations in WER that can be observed in the plot, in case of CD higher than 4, it is worth noting that the speaking style of some subjects as well as some more critical orientation angles cause a mismatch between input signals and acoustic model, which only sometimes could be compensated for by the language model. Furthermore, from this experiment we can extract some useful observations concerning the application of CD based CS for speech recognition using reverberated speech based acoustic models. In the literature, clean acoustic models were used in order to evaluate the detection of the least distorted signal (Wolf, 2013). Under such conditions, even an oracle CS results in a very low performance. However, the results reported in Fig. 8 prove that the use of reverberated acoustic models, which guarantee a better overall performance, is a valid choice, as already shown in past work (Matassoni et al., 2002).

7.3. Relation between CD based CS and oracle CS

Another interesting study concerns the relation between CD based CS and oracle CS. In Fig. 9, we present results obtained using WSJ120 data set, in terms of average WER on selected channels, with three different CS methods: (i) oracle, (ii) CD informed and (iii) CD blind. We observe that the latter two methods follow the same trend as the oracle.

In order to better understand the curves in Fig. 9, let us associate them with the angles highlighted in Fig. 10. It becomes evident that lower error rates are achieved when the speaker is directly oriented towards one of the closer

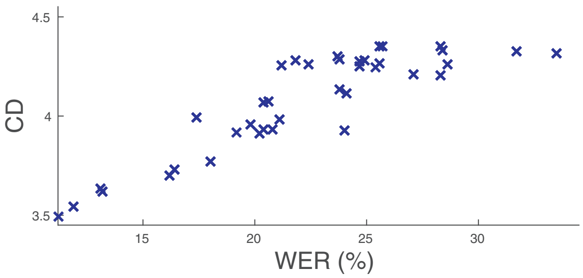


Fig. 8. Distribution of the average CD, between close talk and reverberated signals, with relation to the average WER achieved by the reverberated signals. Each mark corresponds to different channels, *i.e.,* M1, M4, M7, M10 of various orientations at position P_2. Silence segments at the start/end of the signal were removed to compute the CD.

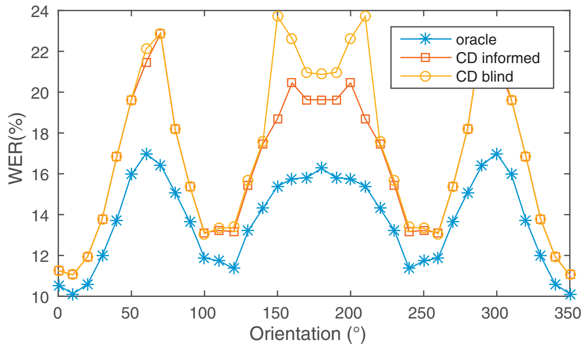


Fig. 9. WER for different CS methods when the speaker is located at the position P_2 of the SQUARE room, with microphones M1, M4, M7 and M10.

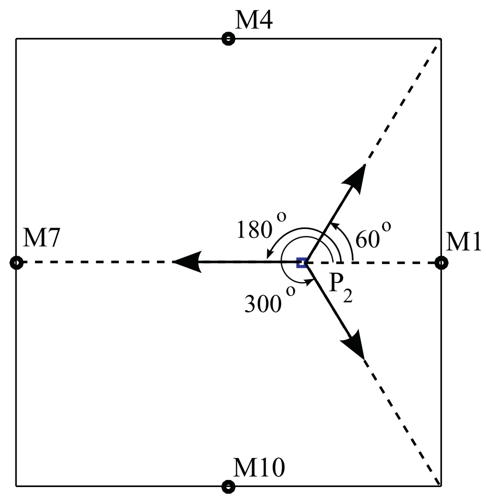


Fig. 10. When the speaker is located at the position P_2 of the SQUARE room the orientations 60^o, 150^o, 210^o and 300^o correspond to the corners of the room. When the microphones M1, M4, M7 and M10 are considered the speaker is directed towards one of them at the orientations 0^o , 120^o, 180^o and 240^o, respectively.

located microphones. Opposite to that, there are certain regions where an increase of WER is observed. These regions correspond to the following geometric conditions:

- The speaker is directed towards a corner of the room, and/or
- The speaker is directed towards a microphone that is clearly more distant than the remaining ones, and/or
- Due to the symmetry of the geometrical problem (e.g. speaker in P2 directed towards M7) the microphone is impinged by a more significant contribution in terms of strong early reflections.

Table 2 presents a subset of the single distant microphone (SDM) recognition results. The first two rows correspond to cases in which the speaker is oriented towards M1. Notice how this condition is reflected into a much lower WER for the indicated microphone. The next set of orientations, around 60^o, corresponds to cases in which the speaker is oriented towards the top-right corner of the room. The last set of orientations, around 180^o, corresponds to cases in which the speaker is directed towards a more distant microphone. For the latter two angular regions, a slightly better performance is provided with M1 and M7, respectively. However, all the available channels produce very similar WER. Therefore, it can be argued that any type of CS, even the oracle one, is not relevant here. For this reason, in the following of this work we will give less emphasis to such situations.

Table 2
SDM WER (%) for speaker position P_2 and microphones M1, M4, M7, M10 of the SQUARE room.

orientation	M1	M4	M7	M10
0^o	11.2	24.1	25.4	24.1
10^o	11.1	20.2	25.5	27.1
...				
50^o	19.5	21.6	27.0	23.8
60^o	20.2	24.0	28.4	23.8
70^o	20.0	22.8	30.6	24.6
...				
170^o	25.3	22.4	19.5	26.2
180^o	25.7	23.8	19.2	23.8
190^o	25.3	26.2	19.5	22.4
...				

7.4. Effects of the setup on CD based CS

Here, we examine the effects that different acoustic scenes have in the performance of the investigated methods. We consider two different aspects of the geometry of the SQUARE room which are first, the position and orientation of the speaker and, second, the configuration of the microphone network. The WSJ120 data set is simulated at each of the positions mentioned and with 36 different orientations.

7.4.1. Position and orientation of the speaker

A set of results that can be understood in an intuitive way is presented in Fig. 11. For this, we introduce a polar representation in which the angle corresponds to the speaker orientation, and the radius to the rate, normalized to 1, at which each channel is selected. Horizontally, each row of polar graphs corresponds to a different position of the speaker, with the leftmost and center polar graphs showing the results of CD informed and CD blind, respectively, and the rightmost graph showing the selection achieved by the oracle. In the latter one, it must be noticed that for some cases the same WER was achieved by more than one microphone. In such cases all the selected channels contribute equally to the rate represented in the polar graph.

Focusing first on the left column concerning the informed method, the results can be explained in a very intuitive way: the best channel corresponds to the microphone towards which the speaker is roughly directed. For example, at position P_1 the selected microphone changes every 90^o, with the region at which a microphone is selected centred around this microphone. A broad agreement between the CD based CS and the oracle is observed as well. When the speaker moves closer to M1 (position P_2) the region at which this microphone is selected is symmetrically expanded around it. An interesting observation results from position P_3, where the behavior of the informed CS is different from the above cases, but can be related to reflections that arrive at the selected channel. For instance, for the orientation of 60^o , the selection of M1 can be explained by the first set of reflections that arrive at this microphone from the top wall. It also must be noticed that the selection at orientations between $180^o{-}270^o$, which is associated to situations with high error rates, shows nevertheless a high agreement between the oracle and the informed CS. Related to this case, a further check of the analysis shown in Fig. 7 confirms that such uncertainty can be explained by the foreseen corresponding CD and DRR.

On the central column of Fig. 11 notice how well the blind CS agrees with the informed one when the speaker is located at positions P_1 and P_2. Disagreements start appearing at orientations that correspond to the corners of the room, or to more distant microphones, for the same reasons discussed in Section 7.3. Such areas of disagreement, that become wider for the positions P_3, still correspond to orientations directed towards very distant microphones.

In order to understand how the above polar plots correspond to recognition results, for each position the WER averaged over all orientations is presented in Table 3. For comparison purposes, we also report the results obtained with EV, which show the improvements provided by the proposed method in two out of three positions. From a detailed analysis of the results in the position, P_3, we found that the improvement obtained with EV is associated to the orientations 180^o -270^o . This evidences an important difference between the two methods. When the microphone signals are all similarly impinged by a high degree of reverberation, EV efficiently selects the least distorted

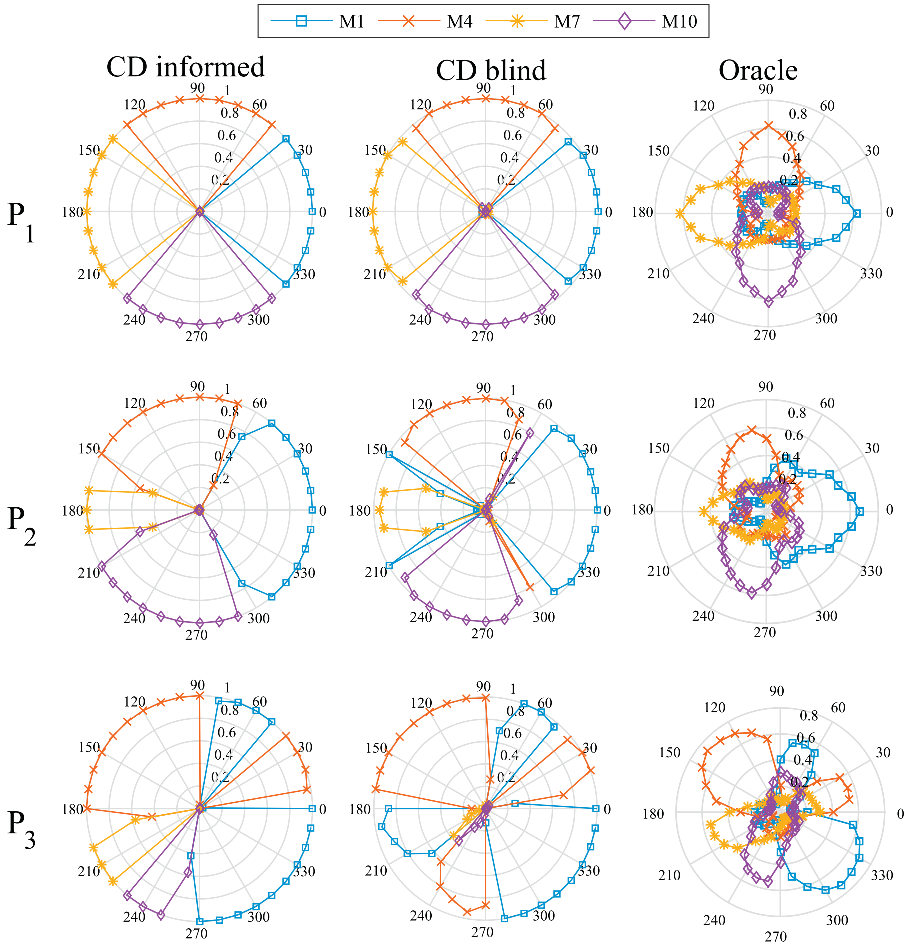


Fig. 11. Channel selection with the informed (leftmost column), blind CD based method (center) and oracle (right), for multiple positions and orientations of the speaker, and the use of the microphones M1, M4, M7 and M10.

one. However, when one less reverberant channel exists, the CD based method is more accurate in selecting it. Next experiments on real data were conceived to further check this specific aspect.

7.4.2. Configuration of the microphone network

Here we discuss the behavior of CS for a set of 7 different microphone configurations, which involve a varying number of microphones. In Fig. 12, we show the WER of the channel selected using either the informed or the blind method, as a function of the WER of the corresponding oracle selection. Each microphone configuration is represented by a data point on the figure, and the average WER is calculated over all the studied positions and

Table 3
WER (%) for different positions, averaged over all orientations. The CS is performed on a configuration over the microphones M1, M4, M7 and M10.

	P_1	P_2	P_3
SDM	22.46	22.40	21.59
oracle	14.92	13.67	12.72
CD informed	17.05	16.44	15.38
CD blind	18.18	17.22	16.88
EV	18.88	17.30	15.42

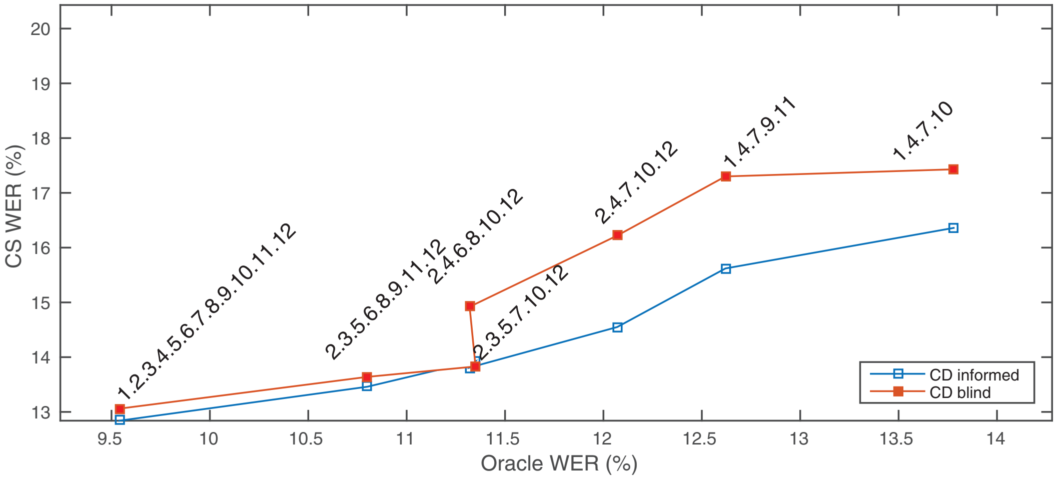


Fig. 12. The WER of the proposed CS methods as a function of the oracle WER. The different data points correspond to the microphone network configurations, for each CS method. The numbers indicated in the text labels refer to the microphone indices as shown in Fig. 2. Each point corresponds to an average over all studied positions and orientations.

orientations. As expected, the oracle WER decreases when more microphones are used, since there is more often a direct path between the speaker and one of the available microphones. The important finding however regards the behavior of the proposed blind method, which is shown to follow the performance of the oracle CS. It can also be observed that the CD informed is always slightly better than the CD blind method, which is reasonable given the fact that it exploits information extracted from the close-talk signal.

Apart from the number of the microphones, their spatial distribution is proved to be another important aspect. For instance, notice the performance with settings 2.3.5.7.10.12 and 2.4.6.8.10.12. Although both settings comprise a total of 6 microphones, we attribute the better performance obtained applying CD blind, with the former setting, to the presence of microphone 3 that is closer to the speaker position.

8. Experiments in the DIRHA room

This section is concerned with recognition experiments in the DIRHA room, which involve the use of two data sets. The first one consists of reverberated speech generated by convolving the IRs measured in the real environment and the clean speech acquired in the FBK recording studio. The second one includes real data recorded in a reverberant room, as reported in Section 6.2. The speech recognition results presented in this section were produced using the *dnn* system detailed in Section 6.3.

The data sets addressed in the previous section result from the application of IM tool to a limited number of source positions, *i.e.,* three. The two new data sets include a large number of position and orientations. However, as outlined before, the main focus of these experiments is to investigate reverberant situations in which one of the captured signals is relatively less affected by reverberation. With this purpose, we selected only a subset of speaker positions and orientations, each determining at least one channel with a DRR higher than −6 dB (*i.e.,* overall exploring 35 position-orientations, characterized by a range of distances between speaker and most directed microphone from 1.5 to 3 meters).

For both data sets, DSR experiments were conducted with the following CS methods: (i) oracle, (ii) CD informed, (iii) CD blind, and (iv) EV. Table 4 presents the WERs obtained with the SDM and with these four methods. The improvement over SDM recognition performance when CS is applied becomes evident, both on synthetic and on real data. First, the results achieved by the oracle highlight the potential recognition improvement that could be achieved with an ideal CS. For this specific experiment, a reduction of around 40% over SDM WER is provided with both data sets. A relevant experimental result is the relative improvement of about 3−4% of CD blind over EV, which is observed with both test data sets. An extended comparison between the different methods in terms of WER

Table 4
WER (%) of the SDM and various CS methods, for data sets characterized by DRRs higher than $-6dB$. The last four rows show the relative WER reduction (%) of blind CS methods over SDM and over the oracle.

	Measured IRs	Real
SDM	18.0	16.60
oracle	9.95	9.68
CD informed	12.60	12.04
CD blind	13.47	12.90
EV	14.03	13.26
	Relative WER reduction (%)	
CD blind to SDM	25.17	22.31
EV to SDM	22.06	20.14
CD blind to oracle	−35.38	−33.26
EV to oracle	−41.01	−36.98
CD blind to EV	3.99	2.71

reduction rate is also available in the table. However, a deeper experimental analysis on the EV method would be required in order to fully confirm this interpretation.

9. Conclusions and future directions

This work has proposed an effective approach to study CS for DSR. The focus was given to CS based on objective quality measures, and particularly on the use of CD in an informed and a blind fashion. With the use of simulated material we studied the relation between the CD and specific characteristics of the acoustic conditions. It was shown that CD is closely related both to T_{60} and to DRR, a finding that endorses the use of CD measure in the context of CS. Furthermore, CD was found to be related to the recognition rate as obtained by decoding reverberated signals. The behavior of the proposed method was investigated through a series of experiments that cover the possible source orientations, in a thorough way, under various speaker positions and microphone network configurations. The informed CD based CS demonstrated an intuitive selection. The performance of the blind method was presented for various conditions, showing a strong agreement with its informed version. Concerning the microphone network configuration, the use of more microphones installed on the walls closer to the possible locations of the speaker was found to improve both the oracle and the CD based approaches. Finally, certain limitations of CS were outlined, for example when a clearly best channel is not available. The applicability and effectiveness of the proposed method under real-world conditions were then addressed. Experimental tasks of gradually increasing degree of realism were tackled, which involve the use of measured IRs and real distant speech material. In both cases, the proposed blind CD based method outperforms a state-of-the-art EV based one.

Some preliminary simulation experiments are being conducted, concerning speech signals deteriorated by both reverberation and additive noise, which suggest that the proposed CD CS technique outperforms EV also in the latter case. Future studies are planned to better investigate on it and on a possible joint use of the proposed CS technique and of different strategies of multi-condition training for acoustic modeling (Huang et al., 2014; Yin et al., 2015). Another possible direction to further improve noise robustness is the use of temporal modulation filters, as explored in Moritz et al. (2016).

Other experimental results also show that the EV based CS performs better under more critical conditions characterized by lower DRRs. This fact suggests a possible next direction of study based on a joint use of CD and EV, to take benefits of the strengths of both processing techniques.

Our future work also envisages the study of variants of the basic proposed method, including weighted CD, and other acoustic features that can well represent the spectral contents of distant speech input. Moreover, other objective signal quality measures, e.g., PESQ, and other features related to room acoustics, e.g., blindly estimated DRR, represent our next steps towards a deeper understanding of the possible achievements that can be obtained by CS methods in DSR.

References

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

Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2015. The third CHiME speech separation and recognition challenge: dataset, task and baselines. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 504–511.

Brandstein, M., Ward, D., 2001. Microphone Arrays: Signal Processing Techniques and Applications. Springer Science & Business Media.

Cabrera, D., Xun, J., Guski, M., 2016. Calculating reverberation time from impulse responses: a comparison of software implementations. Acoust. Aust. 44 (2), 369–378.

Cristoforetti, L., Ravanelli, M., Omologo, M., Sosi, A., Abad, A., Hagmüller, M., Maragos, P., 2014. The DIRHA simulated corpus. In: Proceedings of International Conference on Language Resources and Evaluation, pp. 2629–2634.

Eaton, J., Gaubitch, N.D., Moore, A.H., Naylor, P.A., 2016. Estimation of room acoustic parameters: the ACE challenge. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (10), 1681–1693.

Evermann, G., Woodland, P., 2000. Posterior probability decoding, confidence estimation and system combination. In: Proceedings of Speech Transcription Workshop, Vol. 27.

Farina, A., 2000. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In: Proceedings of 108-th Audio Engineering Society Convention.

Fiscus, J.G., 1997. A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER). In: Proceedings of IEEE Workshop on Automatic Speeck Recognition and Understanding, pp. 347–354.

Flanagan, J., Johnston, J., Zahn, R., Elko, G., 1985. Computer-steered microphone arrays for sound transduction in large rooms. J. Acoust. Soc. Am. 78 (5), 1508–1518.

Furui, S., Sondhi, M.M., 1991. Advances in Speech Signal Processing. Electrical and Computer Engineering. Marcel Dekker Inc.

Garofalo, J., Graff, D., Paul, D., Pallett, D., 1993. Continuous speech recognition (CSR-I) Wall Street Journal (WSJ0) News Complete. Linguistic, LDC93S6A. DVD. Data Consortium Philadelphia.

Gray, A., Markel, J., 1976. Distance measures for speech processing. IEEE Trans. Acoust. Speech Signal Process. 24 (5), 380–391.

Guerrero, C., Tryfou, G., Omologo, M., 2016. Channel selection for distant speech recognition - exploiting cepstral distance. In: Proceedings of Interspeech - Annual Conference of the International Speech Communication Association, pp. 1986–1990.

Hansen, J.H., Pellom, B.L., 1998. An effective quality evaluation protocol for speech enhancement algorithms. In: Proceedings of International Conference on Spoken Language Processing, ICSLP, Vol. 7, pp. 2819–2822.

Himawan, I., Motlicek, P., Sridharan, S., Dean, D., Tjondronegoro, D., 2015. Channel selection in the short-time modulation domain for distant speech recognition. In: Proceedings of Interspeech - Annual Conference of the International Speech Communication Association.

Hu, Y., Loizou, P.C., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 16 (1), 229–238.

Huang, X., Acero, A., Hon, H.-W., Foreword By-Reddy, R., 2001. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR.

Huang, Y., Slaney, M., Seltzer, M.L., Gong, Y., 2014. Towards better performance with heterogeneous training data in acoustic modeling using deep neural networks. In: Proceedings of Interspeech - Annual Conference of the International Speech Communication Association, pp. 845–849.

ISO, E.N., 2008. Acoustics Measurement of Room Acoustic Parameters- Part 2: reverberation Time in Ordinary Rooms. International Organization for Standardization, Brussels, Belgium 3382−2.

Jo, T., Koyasu, M., 1975. Measurement of reverberation time based on the direct-reverberant sound energy ratio in steady state. In: Proceedings of INTER-NOISE and NOISE-CON Congress and Conference Proceedings. Institute of Noise Control Engineering, pp. 579–582.

Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., Maas, R., 2013. The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech. In: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 1–4.

Kitawaki, N., Nagabuchi, H., Itoh, K., 1988. Objective quality evaluation for low-bit-rate speech coding systems. IEEE J. Sel. Areas Commun. 6 (2), 242–248.

Kumatani, K., McDonough, J., Lehman, J.F., Raj, B., 2011. Channel selection based on multichannel cross-correlation coefficients for distant speech recognition. In: Proceedings of Joint Workshop on Hands-free Speech Communication and Microphone Arrays, pp. 1–6.

Kuttruff, H., 2007. Acoustics: An Introduction. CRC Press.

Kuttruff, H., 2009. Room Acoustics. CRC Press. 1973 fifth.

Loizou, P.C., 2013. Speech Enhancement: Theory and Practice. CRC Press.

Malfait, L., Berger, J., Kastner, M., 2006. P. 563 & 8212; the ITU-T standard for single-ended speech quality assessment. IEEE Trans. Audio Speech Lang. Process. 14 (6), 1924–1934.

Matassoni, M., Omologo, M., Giuliani, D., Svaizer, P., 2002. Hidden Markov model training with contaminated speech material for distant-talking speech recognition. Comput. Speech Lang. 16 (2), 205–223.

Moritz, N., Kollmeier, B., Anemüller, J., 2016. Integration of optimized modulation filter sets into deep neural networks for automatic speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (12), 2439–2452.

Naylor, P.A., Gaubitch, N.D., 2010. Speech Dereverberation. Springer Science & Business Media.

Obuchi, Y., 2004. Multiple-microphone robust speech recognition using decoder-based channel selection. In: Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing.

Obuchi, Y., 2006. Noise robust speech recognition using delta-cepstrum normalization and channel selection. Electron. Commun. Jpn (Part II: Electron.) 89 (7), 9–20.

Omologo, M., Svaizer, P., Matassoni, M., 1998. Environmental conditions and acoustic transduction in hands-free speech recognition. Speech Commun. 25 (1), 75–95.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The Kaldi speech recognition toolkit. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding.

Quackenbush, S., Barnwell, T., Clements, M., 1988. Objective measures of speech quality. Prentice-Hall. Signal Processing Series, Englewood Cliffs, NJ.

Rabiner, L., Juang, B.-H., 1993. Fundamentals of Speech Recognition. Prentice Hall.

Rabiner, L.R., Schafer, R.W., 2011. Theory and Application of Digital Speech Processing. PEARSON.

Ratnam, R., Jones, D.L., Wheeler, B.C., OBrien Jr, W.D., Lansing, C.R., Feng, A.S., 2003. Blind estimation of reverberation time. J. Acoust. Soc. Am. 114 (5), 2877–2892.

Ravanelli, M., Cristoforetti, L., Gretter, R., Pellin, M., Sosi, A., Omologo, M., 2015. The DIRHA-English corpus and related tasks for distant-speech recognition in domestic environments. In: Proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 275–282.

Ravanelli, M., Sosi, A., Svaizer, P., Omologo, M., 2012. Impulse response estimation for robust speech recognition in a reverberant environment. In: Proceedings of 20th European Signal Processing Conference, pp. 1668–1672.

Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P., 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 749–752.

Rohdenburg, T., Hohmann, V., Kollmeier, B., 2005. Objective perceptual quality measures for the evaluation of noise reduction schemes. In: Proceedings of 9th International Workshop on Acoustic Echo and Noise Control, pp. 169–172.

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

Shimizu, Y., Kajita, S., Takeda, K., Itakura, F., 2000. Speech recognition based on space diversity using distributed multi-microphone. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1747–1750.

Tachioka, Y., Watanabe, S., Le Roux, J., Hershey, J.R., 2013. Discriminative methods for noise robust speech recognition: a CHiME challenge benchmark. In: Proceedings of the CHiME 2013 International Workshop on Machine Listening in Multisource Environments.

Tribolet, J.M., Noll, P., McDermott, B.J., Crochiere, R.E., 1978. A study of complexity and quality of speech waveform coders. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 586–590.

Van Veen, B.D., Buckley, K.M., 1988. Beamforming: a versatile approach to spatial filtering. IEEE ASSP Mag. 5 (2), 4–24.

Veselỳ, K., Ghoshal, A., Burget, L., Povey, D., 2013. Sequence-discriminative training of deep neural networks. In: Proceedings of Interspeech - Annual Conference of the International Speech Communication Association, pp. 2345–2349.

Ward, D.B., Kennedy, R.A., Williamson, R.C., 1995. Theory and design of broadband sensor arrays with frequency invariant far-field beam patterns. J. Acoust. Soc. Am. 97 (2), 1023–1034.

Wolf, M., 2013. Channel selection and reverberation-robust automatic speech recognition. Ph.D. thesis.Universitat Politècnica de Catalunya.

Wolf, M., Nadeu, C., 2009. Towards microphone selection based on room impulse response energy-related measures. In: Proceedings of Workshop on Speech and Language Technologies for Iberian Languages, pp. 61–64.

Wolf, M., Nadeu, C., 2010. On the potential of channel selection for recognition of reverberated speech with multiple microphones. In: Proceedings of Interspeech - Annual Conference of the International Speech Communication Association, pp. 80–83.

Wolf, M., Nadeu, C., 2013. Channel selection using N-Best hypothesis for multi-microphone ASR. In: Proceedings of Interspeech - Annual Conference of the International Speech Communication Association, pp. 3507–3511.

Wolf, M., Nadeu, C., 2014. Channel selection measures for multi-microphone speech recognition. Speech Commun. 57, 170–180.

Wölfel, M., 2007. Channel selection by class separability measures for automatic transcriptions on distant microphones. In: Proceedings of Interspeech - Annual Conference of the International Speech Communication Association, pp. 582–585.

Wölfel, M., McDonough, J., 2009. Distant Speech Recognition. Wiley.

Yin, S., Liu, C., Zhang, Z., Lin, Y., Wang, D., Tejedor, J., Zheng, T.F., Li, Y., 2015. Noisy training for deep neural networks in speech recognition. EURASIP J. Audio Speech Music Process. 2015 (1), 1–14.

Zahorik, P., 2002. Direct-to-reverberant energy ratio sensitivity. J. Acoust. Soc. Am. 112 (5), 2110–2117.