On the application of 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] *irst-Fondazione Bruno Kessler Via Sommarive 18, 38123, Trento - Italy*

Abstract

Shifting from a single- to a multi-microphone setting, distant speech recognition (DSR) can be benefited from the multiple instances of the same utterance in numerous ways. A valid approach is channel selection (CS) that is based on the assumption that, for each utterance, there is at least one channel which can improve the recognition results as compared to the decoding of the remaining channels. To perform this selection numerous methods have been proposed in the literature. Nevertheless, no complete, systematic studies that address CS applied for DSR exist. In this work, we perform a detailed experimental analysis of CS under diverse geometric and acoustic conditions. We prove that cepstral distance can be applied for CS and that such a method is strongly related to an oracle selection of the best recognized channels. Moreover, we discuss how the investigated CS method is affected by the geometric characteristics of the acoustic scene. The findings initially identified on synthetic material are verified with the use of real data, for which the cepstral distance based method is shown to outperform the state-of-the-art in CS.

Keywords: Distant speech recognition, channel selection, objective measures, cepstral distance, reverberation

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 the ASR limitations in distant-talking scenarios, the most effective strategies adopt the use of multiple microphones (Wölfel and McDonough, 2009; Brandstein and Ward, 2001).

In a real multi-microphone distant speech recognition (DSR) setting, the sensors can be distributed in an arbitrary fashion. For instance, some microphones can be installed on the walls or the ceiling of the room, others can be embedded in electronic devices, and finally some microphones can be built in the users' personal devices. Therefore, the positions of the sensors are not necessarily known in advance, they may change at any time and finally, the sensors may have different characteristics. As a result, the acquired signals, which are all variants of the uttered speech, are affected by different types and degrees of distortion.

Channel selection (CS) makes the reasonable assumption that among the acquired signals there is one that can lead to better recognition results. As indicated by its name, a meaningful CS should not only consider the attributes of the signal, but also the characteristics of the communication *channel* that shaped the uttered speech, from the source to the sensor. Several measures that quantify the effect of the channel on the speech signals have been proposed in the literature, for example the envelope variance (EV)(Wolf and Nadeu, 2014) and the modulation spectrum ratio (Himawan et al., 2015). In previous work, we showed that objective signal quality measures, and in particular the cepstral distance (CD), can be successfully used to evaluate the available channels, and improve recognition results in a real, multi-microphone setting (Guerrero et al., 2016).

Although there seems to be a variety of CS methods in the literature, no comprehensive study of their behaviour in diverse acoustic conditions has been made so far. For instance, although different attributes of the available channels have been considered (Wolf and Nadeu, 2009) there is no information on how CS is affected by different reverberant situations and microphone distributions. Moreover, even though some well performing ASR systems incorporate CS, and in principle this leads to improved results, there is not a clear understanding of how ASR relates to CS, or what is the best result one can expect.

In order to address this missing link between CS and DSR, we provide a novel experimental framework, which supports the design of CS solutions. We present a methodology that assesses and clarifies the limitations that appear when CS is applied under different acoustic conditions. Specifically, we investigate CS under numerous, diverse reverberant scenarios in a synthetic environment. This facilitates the study of acoustic phenomena with a full control over the experimental conditions. Furthermore, we extend our findings and confirm the benefit of applying CS for DSR, with the use of real data. To the best of our knowledge, this represents the first empirical study that characterizes, from a quantitative standpoint, the overall

[*] Corresponding author. Tel.: +39 046 131 4563
Email addresses: guerrero@fbk.eu (Cristina Guerrero Flores), tryfou@fbk.eu (Georgina Tryfou), omologo@fbk.eu (Maurizio Omologo)

behaviour of a CS system in different environments.

The remaining of this paper is organized as follows. In Section 2 multi-microphone DSR is discussed. An overview of the most relevant existing CS methods is presented in Section 3, while common measures exploited by signal-based CS methods are summarized in Section 4. The CD-based CS methods are elaborated in Section 5. In Section 6 details about the experimental framework are provided. The experimental activities and related 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.

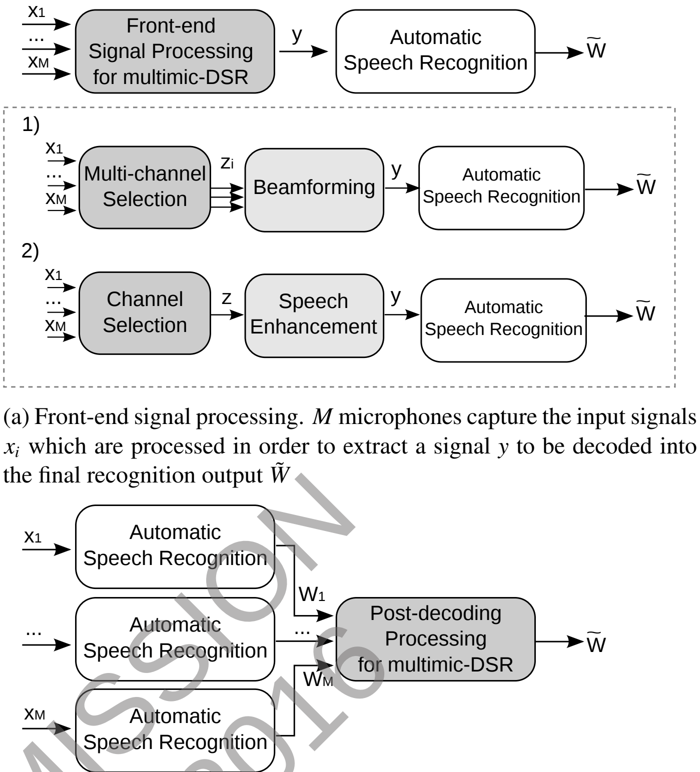
2. Multi-microphone Distant Speech Recognition

In a reverberant environment, when a speaker is uttering a sentence, many delayed and attenuated versions of the original signal arrive at each microphone. Given the particularities of the room, and the position and orientation of the speaker, the channel between the source and the microphone is characterized by an impulse response (IR). 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).

When IRs are available, either as synthetic signals or as real measurements, they can be used as a means of characterizing the room acoustics. Two important parameters that together can describe the sound captured by a microphone in a non-anechoic room are the reverberation time (T_{60}) and Direct-to-Reverberant Ratio (DRR). 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 sound energy that arrives directly at the microphone, to the sound energy that arrives at the microphones after one, or multiple, reflections at the various surfaces (Naylor and Gaubitch, 2010). Both parameters can be directly estimated from the IRs using existing methods, for instance (Naylor and Gaubitch, 2010; Zahorik, 2002). Of course, in practical situations the IRs are not available. Although there are systems that estimate these parameters blindly, the results, particularly for the estimation of DRR are still not satisfactory (Eaton et al., 2016).

Multi-microphone speech processing approaches have proved their potential to significantly improve DSR performance in comparison to single channel solutions. However, the design of effective multi-microphone speech recognition systems is not straightforward due to the large number of variabilities to address under real world conditions. Various architectures can be adopted to process multiple inputs in order to derive a single transcription of a spoken utterance (Wölfel and McDonough, 2009; Kinoshita et al., 2013). In the most commonly exploited solutions, these modules operate either at front-end processing or at post-decoding processing level. These two cases are depicted in Figure 1.

As shown in Figure 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. Such solutions, as for example beamforming (Brandstein and Ward, 2001) and CS (Wolf and Nadeu, 2014), aim at reducing the number of signals to process at a subsequent recognition level. A prevailing effective practice consists in combining front-end processing approaches, *e.g.*,CS for applying beamforming on a reduced set of signals (Kumatani et al., 2011). Examples of such combinations are presented in Figure 1a and Figure 1b.


x1
...
xM
Front-end Signal Processing for multimic-DSR
y
Automatic Speech Recognition
W̃
1)
x1
...
xM
Multi-channel Selection
zi
Beamforming
y
Automatic Speech Recognition
W̃
2)
x1
...
xM
Channel Selection
z
Speech Enhancement
y
Automatic Speech Recognition
W̃
x1
Automatic Speech Recognition
...
Automatic Speech Recognition
xM
Automatic Speech Recognition
W1
...
WM
Post-decoding Processing for multimic-DSR
W̃

(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 x_i is decoded and the individual outputs are processed to extract $\tilde{W}$.

Figure 1: Typical architectures for multi-microphone DSR

On the other hand, post-decoding processing approaches perform a combination of information at the last stage of the recognition system, as shown in Figure 1b. Renown solutions, such as ROVER (Fiscus, 1997) or Confusion Network Combination (Evermann and Woodland, 2000), require an individual, parallel recognition of each input signal before applying their combination algorithms. The complexity and resource demanding nature of post-decoding processing solutions increases with the number of input channels.

3. Channel selection

In the literature on CS, methods are commonly classified as *decoder* and *signal-based* approaches, according to the source of information used to extract the CS measure.

Decoder-based CS methods use information from the decoding process, *e.g.*,likelihoods or posterior probabilities to assess the quality of each channel. It is therefore not possible to apply such techniques independently from the ASR process, for instance as a preprocessing step for speech enhancement. Some representative examples of such methods can be found in (Atal, 1974; De La Torre et al., 2002; Molau et al., 2001; Obuchi, 2004, 2006; Openshaw and Masan, 1994; Shimizu et al., 2000) Although based on the conception that post-decoding measures should present a higher correlation to word error rate (WER) and ASR, this has not been proven in the literature so far (Wolf and Nadeu, 2014). A more detailed review of decoder-based CS can be found in (Wolf, 2013).

On the other hand, *signal-based* approaches operate directly on the signals acquired by the different microphones. The main advantage of such methods is the low computational complexity, since once CS is applied, only one channel will be decoded. We propose the further categorization of signal based CS methods in two groups (i) *informed* and (ii) *blind* methods.

Informed methods, which assume the availability of prior knowledge or reference information, have been explored as the upper-bound mark of a CS measure (Wolf and Nadeu, 2010). In addition, when a CS measure is used in an informed fashion its relation to reverberation properties can be studied in details and certain characteristics can be verified. Although most of the CS measures described in the literature can be easily modified to be used in a blind way, very few authors have performed such an intermediate analysis in the CS literature. In (Wolf and Nadeu, 2009) measured impulse responses (IRs) were used to verify the assumption that ASR 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 are used as further informed CS measures.

Blind methods for CS use scores computed directly from the acquired signals. Such methods share the objective to detect the least reverberated channel(s) among the available ones, assuming that a better match will result between it and the acoustic models of the ASR system. Blind CS measures include the use of energy and SNR, cross-correlation between signals (Kumatani et al., 2011), the variance of the energy envelope (Wolf and Nadeu, 2014), and the modulation spectra of the original and the beamformed signals (Himawan et al., 2015). Some authors focused their efforts in CS for beamforming, where more than one channels are selected for further processing (Kumatani et al., 2011; Himawan et al., 2015). In such methods, the beamformed signal can be used as a reference which can be compared to the acquired signals and rank the latter ones in terms of distortion. Although this idea leads to good CS results, the use of beamforming limits the scope of such methods to scenarios that employ microphone-arrays, *i.e.*,a very limited distance between adjacent microphones. In the most general case, the CS measure is used to select a single least distorted channel (Wölfel et al., 2006; Wolf and Nadeu, 2014).

4. Measures for signal-based CS

In the following sections we summarize state-of-the-art signal-based CS scores.

4.1. Envelope Variance

One of the most successful methods for signal based CS described in the literature, is based on the EV measure (Wolf and Nadeu, 2014). The main idea behind this method is that the reverberation smooths the energy of speech signals, a fact that leads to a reduction in the dynamic range of the envelope of the signal. For the calculation of the EV measure, the filterbank energies (FBE) $x_m(k,l)$ in channel m, sub-band k and time frame l, are first normalized as follows

$$\hat{x}_m(k,l) = e^{\log x_m(k,l) - \mu_{\log x_m(k,l)}} \quad . \tag{1}$$

The mean value $\mu_{\log x_m(k,l)}$ is calculated over the entire speech utterance. The mean normalized sequence of FBE is then compressed with a cube root compression and the variance $V_m(k)$ of each sub-band k, for each channel m, is calculated.

The blind CS method based on the EV measure selects the channel that maximizes the average variance over all channels:

$$\hat{C} = \arg\max_m \sum_k \frac{V_m(k)}{\max_m(V_m(k))} \quad . \tag{2}$$

In (2) the use of a different weighting, for each channel and subband, has been proposed in (Wolf and Nadeu, 2014). However, to our best 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.

4.2. Modulation Spectrum Ratio

More recently, short-time modulation spectrum and bemforming have been adopted in an alternative method to detect a set of best channels (Himawan et al., 2015). The proposed CS measure is based on an assumption very similar to the one done by EV, which is that a clean speech signal will have more modulation than a reverberated one. The difference in terms modulation is formulated as a ratio as follows:

$$\zeta_m(k,f) = 10\log_{10}\frac{|\chi_m(k,f)|^2}{|B(k,f)|^2}, \quad 0 \leq f \leq F \quad , \tag{3}$$

where $\chi_m(k,f)$ and $B(k,f)$ are the modulation spectra of channel m and the beamformed signal, respectively, and f is the highest modulation frequency. As pointed out in previous sections, beamforming cannot be properly applied in an unconstrained distant speech recognition scenario where distributed microphones are used. For this reason, this method is not investigated in our work.

4.3. Objective Quality Measures

When the goal of CS is redefined as the detection of the least distorted channel among the set of signals acquired in a multimicrophone setting, the use of state-of-the-art objective measures of signal quality can be advantageous. Objective quality measures have been consistently exploited for many years in various speech processing applications. Measures such as the CD, the log-likelihood ratio (LLR) (Hansen and Pellom, 1998) and the frequency weighted segmental SNR (fwSNRseg) (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. Similar measures, as for example the PESQ, have been introduced for the quantification of the distortion introduced by speech communication channels (Rix et al., 2001).

The same measures have been reused in numerous applications as for example noise reduction (Rohdenburg et al., 2005) and as metrics for the evaluation in the REVERB challenge (Kinoshita et al., 2013). In (Hu and Loizou, 2008) it was shown that objective quality measures for speech correlate well with subjective evaluation of signal quality. It is therefore reasonable to assume that the use of objective signal quality scores can lead to a meaningful selection of the least distorted channel, among the signals of the distributed microphone network. Particularly, the CD is long known for its effectiveness and flexibility in different application fields (Rabiner and Schafer, 2011)

Cepstral Distance. Perhaps the most intuitive objective measure for signal quality, that applies well in cases of reverberation, is the CD. Cepstrum-based comparisons are equivalent to comparisons of the smoothed log spectra of the signals (Rabiner and Schafer, 2011). In this domain, the reverberation effect can be viewed as additive (Huang et al., 2001). Furthermore, as discussed in (**?**), the CD has a particular frequency domain interpretation in terms of relationship between a set of signals and their geometric mean spectrum. The CD between a clean and a distorted signal is defined as (Hu and Loizou, 2008):

$$d(\vec{c}_x, \vec{c}_m) = \frac{10}{\log_{10}} \sqrt{2 \sum_{k=1}^{p} [\vec{c}_x(k) - \vec{c}_m(k)]^2} \quad , \tag{4}$$

where $\vec{c}_x$ and $\vec{c}_m$ are the cepstral coefficient vectors of the clean and distorted signals respectively, and p is the number of cepstral coefficients used.

5. Cepstral Distance based CS

The application of CD, or any other objective measure for CS requires the use of a reference signal. Based on the nature of the reference the proposed CS is characterized either as *informed* or as *blind*.

5.1. Informed channel selection

Assuming the availability of the close-talk speech signal, $x(t)$, and a multi-microphone setting, let

$$x_m(t) = x(t) * h_m(t) \tag{5}$$

be the signal captured by microphone m, where $h_m(t)$ is the related impulse response (IR), for a given position and orientation of the speaker. Here, $x_m(t)$ is not distorted by environmental noise. Using a signal-based CS measure that expresses the distance $d(x(t), x_m(t))$ between the clean and the m-th reverberated signals, the least distorted channel $\hat{M}_d^x$ can be detected as

$$\hat{M}_d^x = \arg\min_m d(x(t), x_m(t)) \quad . \tag{6}$$

In this work, the distance used for the above minimization is the CD.

5.2. Blind channel selection

In a real scenario, where CS is applied as a means to improve recognition results in a multi-microphone setting, the close-talk signal is not available. In order to overcome this limitation, in (Guerrero et al., 2016) we proposed a non-intrusive way to estimate a set of meaningful CD, upon which a successful CS is performed.

The core of the proposed cepstral based CS lies on the following. When a speaker is oriented towards one of many distributed microphones, and/or is located at a distance lower than the critical distance, the signal acquired by the corresponding microphone demonstrates a relatively higher direct to reverberant ratio. The remaining channels are affected by multiple degrading factors, for example attenuation effects due to both, the multiple reflections and the head of the speaker.

Of course, not always a favourable situation as the one previously outlined can be expected. For example, all channels may be equally impinged by reverberation and in this case the decoding of all the microphone signals will result in a similar recognition error rate. Therefore, the selection of a specific channel is not relevant for improving the recognition performance. For this reason, in this work we focus on scenarios featuring the speaker at favourable positions and/or orientations, in which CS is meaningful. We proposed to estimate a reference signal which represents the average distortion that characterizes each specific scenario, and, using objective quality measures, to detect the signal that is the most distant from this reference. When CD is employed, the reference can be estimated as the geometric mean spectrum of the acquired signals, in the log-magnitude spectrum domain

$$\hat{R}(t, \omega) = \frac{1}{M} \sum_m \log |X_m(t, \omega)| \quad , \tag{7}$$

where $X_m(t, \omega)$ is the short-time Fourier transform (STFT) of the signal captured by microphone m, and M is the total number of 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 room. These two rooms are schematically presented in Figure 2 and Figure 3, respectively. Their detailed characteristics are given in Table 1. In both settings the average distance between the speaker and the microphones fluctuates around 1-4 meters. In contrast to other CS studies performed in much reduced spaces, this condition implies that reverberation significantly affects the degree of the signal distortion.

The SQUARE room is simulated using IRs generated with the IM. This tool offers the possibility to set the orientation of the source with a given acoustic directivity pattern. 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.

Table 1: The main characteristics of the experimental environments.

	SQUARE	DIRHA
Size (m)	4.80×4.80	4.83×4.5
T_{60} (s)	0.7	0.75
# microphones	12	5
# positions	5	16
# orientations	36	4
IM IRs	yes	yes
Measured IRs	no	yes
Real data	no	yes

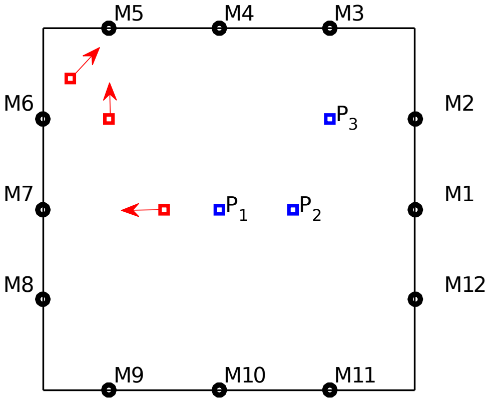


Figure 2: SQUARE room setting. Black dots indicate the microphone positions, and blue squares show the various locations of speaker. The red squares and arrows indicate the channels used for training the acoustic models in the DSR experiments performed in this room.

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 three modalities: i) as a simulated synthetic setting, with the use of IM, ii) as a simulated realistic scenario, with the use of measured IRs (Cristoforetti et al., 2014), and iii) as a real space, with the use of real audio recordings (Ravanelli et al., 2015). An ideal voice activity detection is assumed to be applied over the real data, *i.e.*,ground truth boundaries were used.

6.2. Datasets

The Wall Street Journal (wsj) corpus is used for the distant speech recognition task. A subset of the clean wsj (WSJ0-5k) (Garofalo et al., 1993) training set, comprising 7138 utterances is used as training data. As shown in Figure 2 and Figure 3, the training set was reverberated using a small set of positions and orientations for each experimental room. The used positions/orientations correspond to channels 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[1]. From this corpus we use the clean material recorded in the FBK recording studio to generate all the simulations that use either IM or measured IRs. Furthermore, from the same corpus we use the real distant speech recordings that were captured in the DIRHA room and the corresponding close-talk signals that were captured by a head-set worn by the speaker during the recording sessions. An ideal voice activity detection is assumed to be applied to the real data.

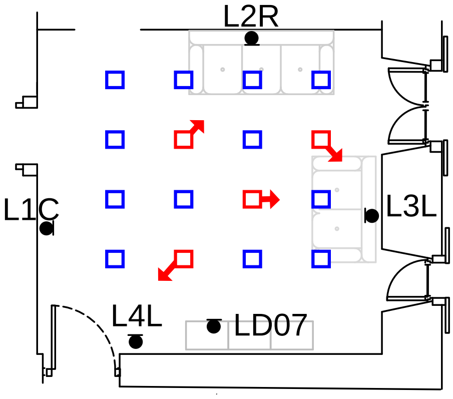


Figure 3: DIRHA room setting. Black dots indicate the microphone positions, and blue squares show the various locations of the speaker. The red squares and arrows indicate the channels used for training the acoustic models in the DSR experiments performed in this room.

In the SQUARE room we use a dataset which includes 120 sentences, referred to as wsj120 dataset. To create this dataset we randomly selected 20 utterances for each of the 6 speakers. Given the fact that each recognition experiment performed in this room is repeated for the whole dataset at each position and orientation, a preliminary experiment showed that this is a sufficient number of utterances to consider. In the DIRHA room dataset, which includes mixed positions and orientations, we use the complete set of available utterances both for simulated and real data.

6.3. Speech recognition

Each of the microphone signal is decoded using the Kaldi speech recognition toolkit (Povey et al., 2011) . The language and lexicon models are built according to the default WSJ s5 recipe. The recognition is based on Karel's recipe (Veselỳ et al., 2013), on top of MFCC-LDA-MLLT-fMLLR transformed features. The performance of the recognition tasks are measured in terms of WER. The recognition performance on the close-talk material captured in the FBK recording studio yields a WER of 6.2%.

6.4. Channel selection methods

The following CS methods are included in the evaluation:

- **oracle** reports an informed CS method that exploits prior knowledge of the WER achieved by each of the single microphone signals. In this CS method the selection of the microphone corresponds to the channel with the lowest recognition error.

[1] The distribution of WSJ data set is under discussion with LDC.

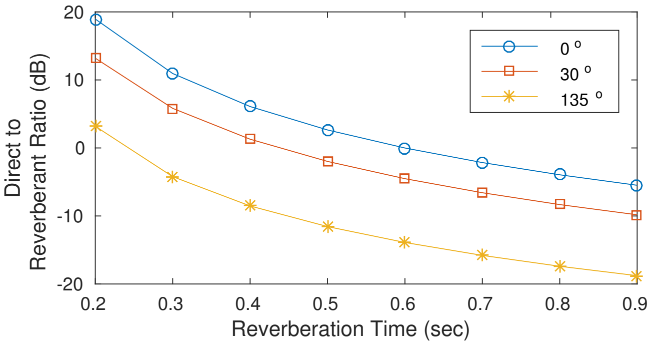


Figure 4: DRR as a function of T60 at the position P_1, with three different orientations.

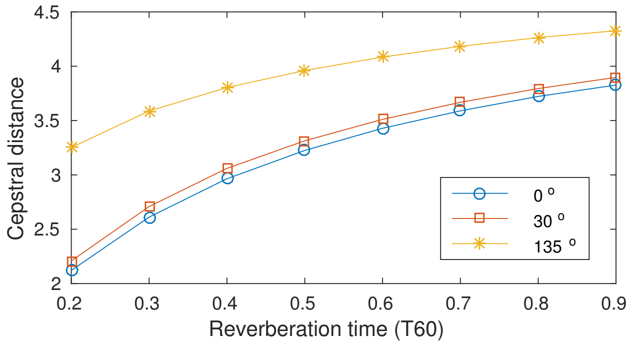


Figure 5: CD of a reverberant to a close-talk signal, in terms of increasing reverberation time at the position P_1, with three different orientations.

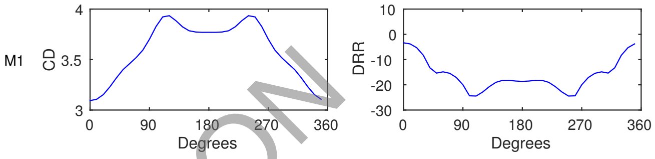


Figure 6: CD and DRR variation for an utterance simulated with speaker located at P_1 adopting different orientations. Results are presented for microphone M1.

- **CD informed** corresponds to the *informed CD based* CS method that uses the close-talk reference, as explained in Section 5.

- **CD blind** is the *blind CD based* method that uses the geometric mean spectrum as a reference, as described in Section 5.

- **EV** is the state-of-the-art CS method, based on Envelope Variance (Wolf and Nadeu, 2013).

7. Experiments in the SQUARE room

In this section we report experiments performed in the SQUARE room setting, based on the use of IM generated IRs.

7.1. Relation between CD and reverberation

A first aspect to investigate concerns the ability of CD to characterize reverberation. T_{60} and DRR are considered as two critical parameters for characterising the sound captured by distant microphones. The IM tool used in our experiments offers a full control of T_{60} which allows us to generate IRs that correspond to different T_{60} values in the range from 0.2sec to 0.9sec. Since no control over the DRR is given, we acquire different DRR cases, by manipulating the orientation of the speaker towards the microphone. Therefore, the wsj120 set is simulated with the speakers at the position P_1 under different orientations. For each IR, the DRR was calculated with the use of the IR_stats toolbox of MATLAB (Zahorik, 2002).

In Figure 4, we present the DRR 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. Moreover as expected, there is an inverse relation between T_{60} and the corresponding DRR, which is respected in all DRR cases.

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 Figure 5, where it becomes evident that CD provides a meaningful characterization of the reverberated signals, *i.e.*,the average distance monotonically increases along with the increasing reverberation time. Moreover, a clear separation is observed among the CDs that are obtained for the same T60, from simulations corresponding to different orientations. This is supporting the original assumption that CD can be applied for CS, since it is confirmed that in a single environment this objective measure can characterize the degree of reverberation, in a similar manner to the DRR.

Next we report, for a single utterance, the variation of CD and DRR when the speaker adopts various orientations at a specific location. In the first case, Figure 6, the speaker is located at position P_1. As intuitively expected, it is observed that when the speaker is oriented towards microphone M1, orientation 0^o, the minimum CD and maximum DRR are measured with relation to such a microphone. Due to the symmetrical conditions of the speaker and microphone locations, the same results were observed for the other microphones with a shifting caused by the relation between the orientation and the direct microphone. We also report a second case which features the speaker located at position P_3. Figure 7 shows the results presented for the microphones M1, M4, M7 and M10. First, these results confirm the previous observations concerning CD and DRR relation with the speaker location and orientation. Moreover, for conditions in which the speaker is oriented towards a microphone at a considerably large distance, e.g., at 200^o oriented towards M7, the dynamics of CD and DRR are reduced. This is translated into a lower discrimination power for the identification of the least distorted channel, even with the use of prior information. These experimental results confirm the power of CD for identifying DRR characteristics of a signal captured by multiple distant microphones in a room.

The above findings are important basis for the use of CD as a means of selecting the least reverberated channel. However, a second assumption has to be verified concerning the fact that a higher WER corresponds to a larger average distance between the clean and the reverberated signals. In order to study this

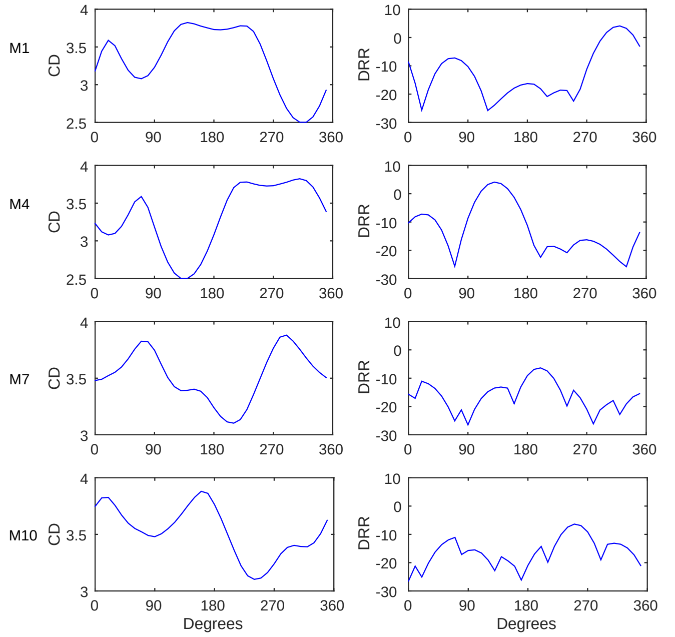


Figure 7: CD and DRR variation for an utterance simulated with speaker located at P_1 with different orientations. Results are presented for microphones M1, M4, M7 and M10.

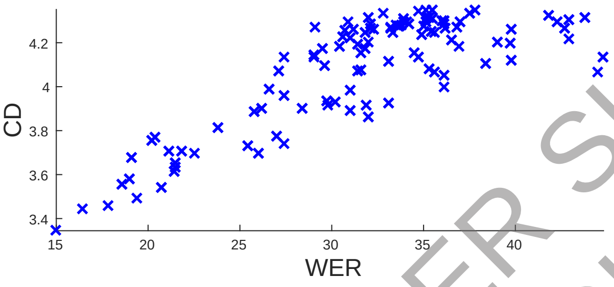


Figure 8: Distribution of CD, between close talk and reverberant signals, with relation to the WER achieved by the reverberant signal. In this experiment, a single T60 is explored, i.e., 0.75s. The acoustic models are trained on contaminated material. Silence segments at the start/end of the signal were removed to compute the CD.

aspect, we used the IM to generate a set of IRs which simulate a speaker at the position P_2 of the square room. In Figure 8 each point relates a CD to a WER, both averaged over the wsj120 dataset simulated under different orientations. The CDs are calculated between the clean and reverberated signals, and the WER is the result of decoding the reverberated signals.

From Figure 8, it is evident that CD is not only related to the reverberation time but also to recognition rate. Furthermore, from this experiment we can extract some useful observations concerning the application of cepstral based CS for speech recognition using contaminated acoustic models. In the literature on CS, experiments are normally performed with clean acoustic models, in order to evaluate the detection of the most clean signal. In this case, even an oracle CS results in a very low performance. However, the results reported in Figure 8 prove that the use of contaminated acoustic models, which guarantee a better overall performance, is a better choice for a CS application. We observe a clear trend that an increasing degree of signal distortion, as measured by CD, corresponds to an increasing WER.

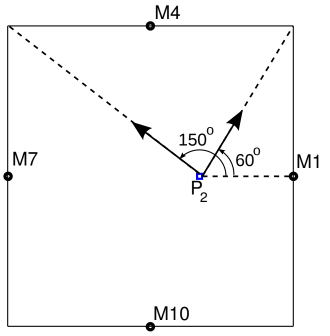


Figure 9: When the speaker is located at the position P_2 of the square room the orientations 60°, 150°, 210°, 300° correspond to the corners of the room. When the microphones M1, M4, M7 and M10 are considered the speaker is directed towards one of them at the orientations 0°, 120°, 180°, and 240°.

7.2. Relation between CD-based CS and oracle CS

Another interesting study concerns the relation between CD based CS and the oracle CS. The oracle constitutes an upper-bound limit, and provides a meaningful CS which leads to an important reduction in WER. In this study, we consider the wsj120 data, simulated at the position P_2 of the square room, as shown in Figure 9. The speaker adopts different orientations, from 0^o to 360^o , relative to the reference system. In order to visualize the results, we use a polar representation in which the angle corresponds to the speaker orientation and the radius to the various depicted measures.

The CS is performed over the microphones M1, M4, M7 and M10 with three different methods (i) oracle, (ii) CD informed and (iii) CD blind. As shown in Figure 10 the different CS methods follow the same trend as the oracle, with lower errors achieved when the speaker is directly oriented towards one of the closer located used 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 the corners of the room, or
- the speaker is directed towards a microphone that is clearly more distant than the remaining ones.

Table 2 presents a subset of the single distant microphone (SDM) recognition results. The first 2 orientations correspond to the cases where the speaker is directed towards M1. Notice how this condition is reflected into a much lower SDM for the indicated microphone. The next set of orientations, around 60^o , correspond to the top-right corner of the room. For this region, all the available channels produce very similar SDM WER. Therefore, it can be argued that any type of CS, even the oracle one, is not meaningful here. This is a very important observation to keep in mind when the errors of any CS method are analysed: at a region where CS itself is not meaningful, no attempt to understand errors and inconsistencies should be made.

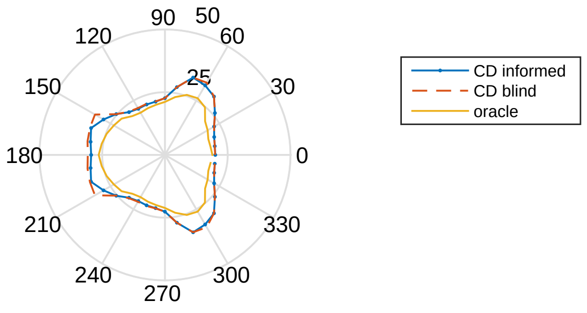


Figure 10: Polar representation of the WER for different CS methods

Table 2: SDM WER for P_2 and four microphones.

orientation	M1	M4	M7	M10
0^o	19.4	33.1	35.6	33.1
10^o	20.7	26.6	35.4	39.8
...				
50^o	30.0	31.0	36.4	34.0
60^o	31.9	33.1	37.4	33.8
70^o	31.1	32.5	39.5	34.6
...				
170^o	35.2	32.4	30.0	36.8
180^o	35.1	34.7	29.8	34.7
190^o	35.2	36.8	30.0	32.4

7.3. The effects of the setup on the proposed CD-based CS method

Here, we examine the effects that different acoustic scenes have in the performance of CS. 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 dataset is simulated at each of the positions mentioned and 36 different orientations, with a step of 10^o . The set of microphones used in the different experiments is recognized by labels in the form "1.4.7.10" with each number referring to the index of a microphone as specified in Figure 2.

7.3.1. Position and orientation of the speaker

A set of CS results that can easily be understood in an intuitive way (experiment 1.4.7.10) are presented in polar form in Figure 11. Horizontally, each pair of polars corresponds to a different position of the speaker, with the left polar showing the results of the CD informed CS and the right one the results of the CD blind CS. Each point of the polars represents the frequency with which the corresponding channel was selected when the speaker was oriented at the angle at which the point is shown.

Focusing first on the left column, the informed CS results can be explained in a very intuitive way: the best channel corresponds to the microphones 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. When the speaker moves closer to M1 (positions 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 multiple reflections that take place at the closely located corner of the room cause an informed CS that is not intuitively understood at the corresponding region (see at the bottom left polar, the region 0^o -90^o).

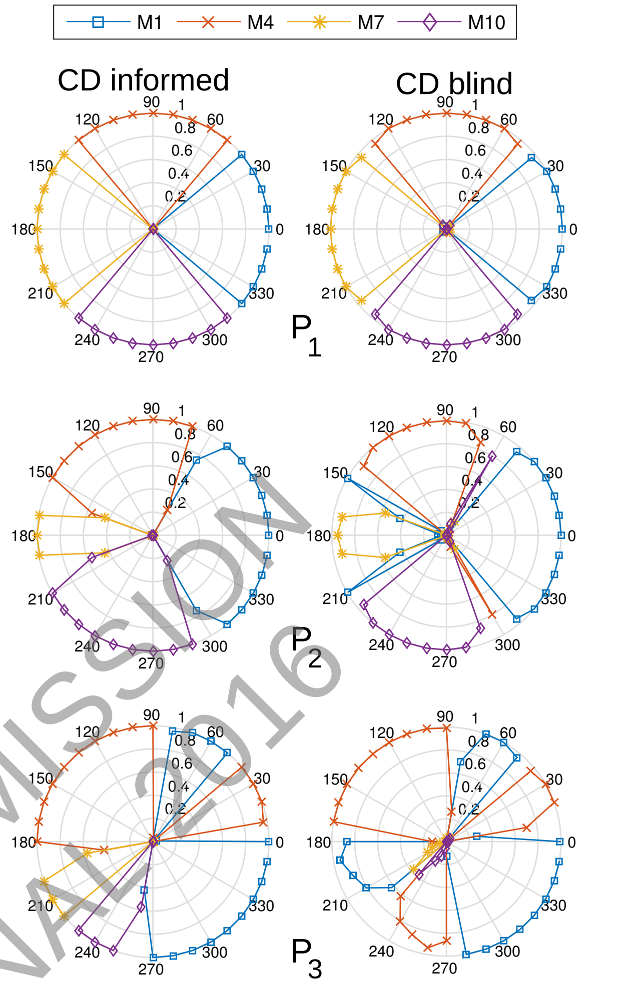


Figure 11: Channel selection with the informed and blind CD based method, for multiple positions and orientations of the speaker, and the use of four microphones (1.4.7.10)

On the right column of Figure 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 very distant microphones, both cases can be attributed to the reasons discussed in the previous section. 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, in Table 3, for each position the average WER over all orientations is presented.

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	32.59	32.22	31.49
oracle	24.11	22.28	20.72
CD informed	27.19	25.84	23.84
CD blind	27.20	26.27	25.35

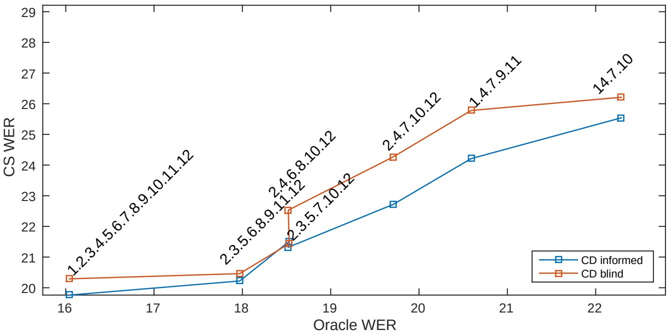


Figure 12: The WER of the proposed CS methods as a function of the oracle WER. The different data points correspond to the microphone network configurations, for each CS method. The numbers indicated in the text labels refer to the microphone indices.

7.3.2. *Configuration of the microphone network*

Here we discuss the behaviour of Cs for a set of 7 different microphone configurations, which consider a varying number of microphones per wall, from 1 to 3. In Figure 12 the WER of CS experiments as a function of the WER of the corresponding oracle CS is shown. Each microphone configuration is represented by a data point on the figure, and the average WER is calculated over all the studied positions and orientations (*i.e.*,a total of 144 recognition experiments). 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 behaviour of the proposed CS methods, which is shown to follow the performance of the oracle CS. The proposed CS is not limited by the characteristics of the room, for instance the T_{60} or the positions and orientations assumed by the speakers.

Apart from the number of the microphones, their spatial distribution is proved to be another important aspect. For instance, notice the WER results for CS performed in the settings 2.3.5.7.10.12 and 2.4.6.8.10.12. Although both settings comprise a total of 6 microphones, the fact that in the first one more microphones are located on the walls which are closer to the speaker, seems to result in an important decrease on the WER of the blind CS.

8. Experiments in the DIRHA room

This section is concerned with recognition experiments in the DIRHA room, which comprise three sets of data, more realistic than what was considered so far. First, we use the IM to generate a set of IRs that simulate the DIRHA room. Second, the measured IRs of the same environment are used. Finally, DSR is performed on the real, reverberated data recorded in the DIRHA room. In this section, apart from the oracle and the objective based CS we also include the state-of-the-art EV CS method.

Table 4: WER for various CS methods, for the DRR-6 dataset. The last two rows show the relative WER increase of CS methods over the oracle.

	IM IRs	Measured IRs	Real
SDM	27.84	25.6	26.81
oracle	15.45	15.53	16.92
CD informed	19.92	19.37	20.22
CD blind	19.86	19.99	21.58
EV	18.70	20.35	23.15
rel.CD blind	28.5	28.7	27.5
rel.EV	21.0	31.0	36.8

The IM generated IRs were used to calculate the DRR for each position/orientation. Based on the results presented in Figure 4, for the T_{60} of the DIRHA room, channels with a maximum of 30^o between the speaker and the microphone correspond to DRR values higher than $-6dB$. Since, as discussed earlier, the CS is only relevant when a direct channel is available, experiments were performed on the "DRR-6" subset of the available data, that consists of all the positions/orientations that correspond to a DRR higher than $-6dB$, for at least one of the considered microphones. In Table 4 the WER for different CS methods are presented. The improvement of the SDM recognition results when CS is applied becomes evident, both on synthetic and on real data. It is relevant to notice that shifting from IM generated IRs, to measured IRs and finally to real data, the CD based CS reduces the gap to the oracle CS as shown by the relative WER increase. On the other hand, the EV method seems to loose performance as more realistic data is used.

9. Conclusions

This work has proposed an effective method to study CS for DSR. The focus was on CS based on objective quality measures, and particularly the CD in an informed and a blind fashion. With the use of artificial material the relation between the CD and certain characteristics of the acoustic conditions was studied. It was shown that CD is closely related both to T_{60} and DRR, a finding that endorses the use of this measure in the context of CS. Furthermore, CD was found to be related to the recognition rate as obtained with the decoding of data with contaminated acoustic models. This supports our strong belief that the use of contaminated acoustic models is a suitable choice for CS based DSR.

The behaviour of CD based CS was investigated through a series of experiments that cover in an exhaustive way the possible orientations of the speaker, under diverse positions and microphone network configurations. 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 CS. Finally, certain limitations of CS were outlined, as for example its inability to improve WER when a clearly best channel is not available.

In order to verify the validity of the above findings in real material, scenarios of gradually increasing degrees of realism were employed. A similar trend in recognition performance was observed when IM extracted IRs were replaced with measured IRs and then, when real data was used. Furthermore, in the latter experiments it was shown that the regarded CD based CS method outperforms the state-of-the-art EV method.

References

Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. The Journal of the Acoustical Society of America 65, 943–950.

Atal, B.S., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. The Journal of the Acoustical Society of America 55, 1304–1312.

Brandstein, M., Ward, D., 2001. Microphone arrays: signal processing techniques and applications. Springer Science & Business Media.

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

De La Torre, A., Segura, J.C., Benitez, C., Peinado, A.M., Rubio, A.J., 2002. Non-linear transformations of the feature space for robust speech recognition, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. I–401.

Eaton, J., Gaubitch, N.D., Moore, A.H., Naylor, P.A., 2016. Estimation of room acoustic parameters: The ace challenge. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24, 1681–1693.

Evermann, G., Woodland, P., 2000. Posterior probability decoding, confidence estimation and system combination, in: Speech Transcription Workshop.

Farina, A., 2000. Simultaneous measurement of impulse response and distortion with a swept-sine technique, in: 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: IEEE Workshop on Automatic Speeck Recognition and Understanding, pp. 347–354.

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. Continous speech recognition (CSR-I) Wall Street Journal (WSJ0) News Complete. LDC93S6A. DVD. Linguistic Data Consortium, Philadelphia .

Gray, A., Markel, J., 1976. Distance measures for speech processing. IEEE Transactions on Acoustics, Speech, and Signal Processing 24, 380–391.

Guerrero, C., Tryfou, G., Omologo, M., 2016. Channel selection for distant speech recognition - exploiting cepstral distance, in: Annual Conference of the International Speech Communication Association, pp. 1–1.

Hansen, J.H., Pellom, B.L., 1998. An effective quality evaluation protocol for speech enhancement algorithms., in: International Conference on Spoken Language Processing, 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: Annual Conference of the International Speech Communication Association.

Hu, Y., Loizou, P.C., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Transactions on Audio, Speech, and Language Processing 16, 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.

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: 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 Journal on Selected Areas in Communications 6, 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: Joint Workshop on Hands-free Speech Communication and Microphone Arrays, pp. 1–6.

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

Molau, S., Pitz, M., Ney, H., 2001. Histogram based normalization in the acoustic feature space, in: IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 21–24.

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: 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. Electronics and Communications in Japan (Part II: Electronics) 89, 9–20.

Openshaw, J.P., Masan, J., 1994. On the limitations of cepstral features in noise, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. II–49.

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: IEEE Workshop on Automatic Speech Recognition and Understanding.

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

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: 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: 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: 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: 9th International Workshop on Acoustic Acho and Noise Control, pp. 169–172.

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

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

Veselỳ, K., Ghoshal, A., Burget, L., Povey, D., 2013. Sequence-discriminative training of deep neural networks., in: INTERSPEECH, pp. 2345–2349.

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: 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: INSTERSPEECH, pp. 80–83.

Wolf, M., Nadeu, C., 2013. Channel selection using n-best hypothesis for multi-microphone asr, in: 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 Communication 57, 170–180.

Wölfel, M., Fügen, C., Ikbal, S., McDonough, J.W., 2006. Multi-source far-distance microphone selection and combination for automatic transcription of lectures, in: International Conference on Spoken Language Processing, pp. 361–364.

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

Zahorik, P., 2002. Direct-to-reverberant energy ratio sensitivity. The Journal of the Acoustical Society of America 112, 2110–2117.