UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

Information Fusion Approaches for Distant Speech Recognition in a Multi-microphone Setting

Cristina Maritza Guerrero Flores

Advisor

Maurizio Omologo
Fondazione Bruno Kessler

August 2016

This Ph.D. Thesis has been successfully defended on August, 30th 2016 in front of the following evaluation committee:

Prof. Alberto Abad
L2F- Laboratório de Sistemas de Língua Falada
INESC-ID Lisboa, Portugal

Prof. Begum Demir
Remote Sensing Laboratory
Department of Information Engineering and Computer Science
University of Trento, Italy

Prof. Stefano Squartini *(Committee President)*
3MediaLabs - A3LAB
Dipartimento di Ingegneria dell'Informazione
Università Politecnica delle Marche, Italy

Prof. Hugo Van Hamme
Processing Speech and Images Unit
Department of Electrical Engineering
Katholique University of Leuven, Belgium

A mi familia
To my family

Acknowledgements

This achievement would not have been possible without the help and support of many people around me. First, I want to thank my advisor, Maurizio Omologo, for sharing his time and wisdom for this endeavor. Your guidance and encouragement throughout this period helped me to get to this point. I would like to thank the members of the SHINE unit for the scientific and technical support along the development of my work, and for giving me a hand whenever it was needed. Thanks also to the doctoral students and researchers at the lab, for making such a nice scientific and human environment.

My friends at the UNITN, with whom I have shared ideas, pizzas and more; thank you for enriching my life with culture, life points-of-view, tiny doses of drama, and lots of delicious food. Ligia and Saameh, thanks for being here for me. Thanks to the volunteers of association Cachisagua for letting me be part of them; the world would be much better if there were more like you. To my Ecuadorian friends who managed to stay in touch even with my PhD craziness and time differences. To Maria Elena, my little bit of Ecuador in Italy, thanks for your everlasting friendship. My enormous gratitude to Georgina Tryfou. I really appreciate the interest and patience you put in all the scientific, and not so scientific, discussions we had. Thanks for the multiple little pushes you gave me in my professional and personal life, and for being there along this emotional journey. Thank you my friends for the company in this voyage of discovery.

I would like to thank to my family. To my parents, my best example of sacrifice, hard-work, and perseverance, thank you for all your faith and love. To my sister, brother, and their families, for the endless laughter and positivism. To my grandparents, who faced multiple sacrifices to provide a happy life and education to their children; this accomplishment is also yours, it started with you. Thank you all for making me believe that this was possible. I love you so much.

Thank you Bernardo, for being an amazing life partner. Thanks for being by me side in all the crazy adventures I jump into, for taking good care of me, for helping me fight my fears, for standing the hard moments and helping us both to get through them, and for sharing and building dreams with me.

And although you made this thesis a little more complicated to finish, I want to thank you Arianna, my sweet happy little girl. Thanks for making me a new person, for broadening my perspective of life, for renewing my energies, and for showing every morning that wonderful smile while calling me ’mami’. Te amo.

Abstract

It is a well known fact that high quality Automatic Speech Recognition is still difficult to guarantee under conditions in which the speaker is distant from the microphone, due to the distortions caused by acoustic phenomena, such as noise and reverberation. Among the different research directions pursued around this problem, the adoption of multi-channel approaches is of great interest to the community, given the potential of taking advantage of information diversity.

In this thesis we elaborate on approaches that exploit different instances of a sound source, captured by various largely spaced microphones, in order to extract a Distant Speech Recognition hypothesis. Two original solutions are presented, based on information fusion approaches at different levels of the recognition system, one at front-end stage and one at post-decoding stage, namely for the problems of channel selection (CS) and hypothesis combination.

First, a new CS framework is proposed. Cepstral distance (CD), which is effectively applied in other acoustic processing fields, is the basis of the CS method developed. Experimental results confirmed the advantages of a CD-based selection schema under different scenarios. The second contribution concerns the combination of information extracted from the individual decoding processes performed over the multiple captured signals. It is shown how temporal cues can be identified in the hypothesis space, and be beneficial for the elaboration of a multi-microphone confusion network, from which the final speech transcription is derived.

The proposed methods are applicable in a setting equipped with synchronized distributed microphones, independently of the proximity between the sensors. Analysis of the novel concepts were performed over synthetic and real-captured data. Both approaches achieved positive results at the different assessment tasks they were exposed to.

Keywords Distant-talking, distributed microphone network, channel selection, cepstral distance, hypothesis combination, lattice, confusion network.

Contents

List of Abbreviations

AM	Acoustic Model
ASR	Automatic Speech Recognition
CM	Confidence Measures
CN	Confusion Network
CNC	Confusion Network Combination
BF	Beamforming
CS	Channel Selection
DIRHA	Distant-Speech Interaction for Robust Home Applications
DNN	Deep Neural Network
DSR	Distant Speech Recognition
EV	Envelope Variance
HMM	Hidden Markov Model
IR	Impulse Response
LM	Language Model
MMCN	Multi-microphone Confusion Network
POSORI	Position-Orientation
REAL	Real dataset
ROVER	Recognizer Output Voting Error Reduction
SDM	Single Distant Microphone
SIM	Simulated dataset
SLF	HTK Standard Lattice Format
SRILM	SRI Laboratory Language Modeling Toolkit
WER	Word Error Rate
WSJ	Wall Street Journal

List of Tables

List of Figures

Chapter 1

Introduction

It's far easier to start something than it is to finish it.

Amelia Earhart

Towards a natural voice-based interaction

Human beings keep eagerly expecting solutions that allow them to use their voices to speak with a computational system in a way that resembles a natural human conversation. The ultimate goal of voice-based technology is to provide a reliable performance for any purpose; whether it is for voice transcription, command and control, information extraction, or conversational interaction. One of the key elements in this quest is Automatic Speech Recognition (ASR), whose objective is to identify spoken utterances and express them in a textual representation. A recognition engine, or recognizer, is the system in charge of performing this complex task. State-of-the-art speech recognizers[1] achieve highly accurate recognition rates when the speech signal is recorded by microphones used in close proximity to the speaker mouth e.g., a head-set. These microphones are commonly called close-talk microphones.

Thanks to the recent advances in sensor technologies and computing infrastructure, the utilization of such close-talk sensors on smart phones or personal gadgets is invisible for the user. However, in some scenarios the use of a close-talking microphone is not a suitable option, for example when the use of this acoustic sensor disturbs the activity performed by the speaker. With the objective of facilitating a more natural voice-based interaction, microphones located at a certain distance from the speakers are preferred.

[1] See Kaldi `http://kaldi-asr.org/`, HTK `http://htk.eng.cam.ac.uk/`, CMU-Sphinx `http://cmusphinx.sourceforge.net/`, Julius `http://julius.osdn.jp/`

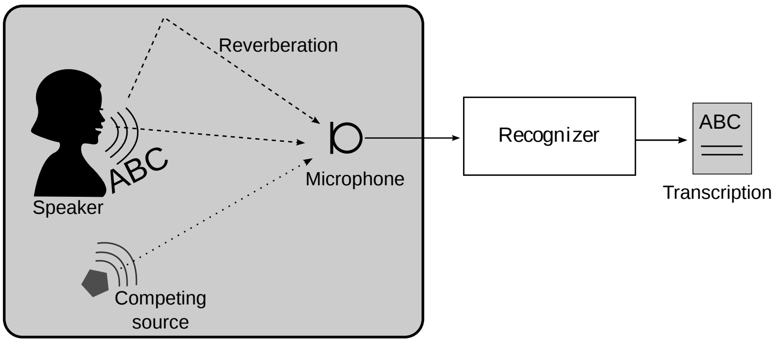


Figure 1.1: Typical distant-talking scenario and conventional processing. Dashed lines indicate some paths between an acoustic source and the microphone.

This condition, also called distant-talking, can be observed for example in meeting rooms where the microphones are located on top of a table. Under such scenarios, multiple issues arise which radically change the speech recognition problem.

1.1 The problem of Distant Speech Recognition

Speech recognition can be considered as a communication process. In broad terms, the speaker sends a message, then the message is received by the recognizer, that processes and transforms it into a transcription. The communication channel, where the message travels through, plays a critical role in the effectiveness of the process. In the case of distant speech communication the channel is severely affected, among other factors, by the nature of acoustic wave transmission and by the distance between the speaker and the microphone. In a distant-talking setting, new challenges are introduced for the recognizer as it must deal with signals coming not only from the desired speech, but also from the background noise, competing speakers, obstacles in the space, or reverberation, i.e., the acoustic reflections on the room surfaces, Figure 1.1. Moreover, not only the position but also the orientation of the speaker may be another unfavorable factor for this problem. All these acoustic variabilities seriously hamper the performance of Distant Speech Recognition (DSR) systems [Gong, 1995; Wölfel and McDonough, 2009], making it notoriously difficult to guarantee high recognition quality.

In the search for a solution that achieves a robust recognition performance amidst the mentioned conditions, numerous efforts, following different strategies, have been undertaken. A first group of methods processes the captured audio signal in order to reduce the effects of the acoustic distortions [Benesty et al., 2005; Droppo and Acero, 2008; Chen et al., 2008; Huang et al., 2008; Pedersen et al., 2008], and then provides, to the recognizer, a cleaner or more manageable single information-stream, e.g., a signal or a feature set. A second category of approaches aims for robust representations of the acoustic information to be later exploited by the recognizer [Hermansky and Morgan, 1994; Kenny, 2012; Hinton et al., 2012; Feng et al., 2014]. The third group focuses on manipulating components of the recognition process, in order to reduce the mismatch between the data on which the recognizer was trained and what is observed in the acoustic scenario [Leggetter and Woodland, 1995; Kolossa et al., 2005; Droppo and Acero, 2008]. Given that the core of standard ASR is a pattern recognition problem, this latter strategy is necessary to guarantee a higher recognition accuracy. Finally, there is a fourth set of methods that performs additional processing over the recognition or decoding output in order to polish the final result [Fiscus, 1997; Mangu et al., 1999; Goel and Byrne, 2000].

A practice frequently observed in the scientific community, which has been reported as effective, is that of integrating various of the previously mentioned approaches in a single processing system. Disadvantageously, the lack of a common evaluation framework complicates the identification and full-comprehension of the impact that each of these methods has on the results.

Another commonly adopted type of approaches was inspired by a biological mechanism, the combination of information captured by two ears as a key element for an enhanced understanding of an acoustic scene. Such approaches rely on the multiplicity of information sources, which can be extracted either from various acoustic channels (e.g., microphones located at different points in the space), or from various recognition systems (e.g., exploiting different data representations), each designed with varied parameters and producing different results. Additionally, an architectured combination of the mentioned sources has demonstrated to positively influence DSR.

Real applications of these strategy lines are evidenced in the results of evaluation campaigns such as REVERB [Kinoshita et al., 2013], CHiME-3 [Barker et al., 2015] or ASpIRE [Harper, 2015]. Many of the most successful systems proposed in these challenges exploit, first, multiple microphones as an audio capture scheme, which is advantageous for further understanding and exploitation of the acoustic scene. And second, they use the combination of the information captured or derived from processing the input signals. Under specific conditions of the sensor network, audio signal processing is performed. Another element that is recurrently used is the combination of recognition outputs [Fiscus, 1997]. Such a standard method, which yields beneficial effects to the combination of recognition transcriptions, presents limitations in terms of the information used and produced. These drawbacks are discussed in Chapter 2. Overcoming these limitations is necessary for further output refinement or other spoken-language based applications.

1.2 Relevant terminology

The focus of this dissertation is on approaches that support speech recognition in a distant-talking condition, dealing with acoustic signals captured by multiple microphones. In order to establish a common understanding of the main directions followed along this work, a relevant terminology is defined in the following.

A *channel* is defined as the signal path created between the acoustic source and the acoustic sensor, the microphone. Likewise, this term is also utilized interchangeably to identify the signal captured by one microphone. In this work, we use the latter definition. In the case of multi-microphone settings, a selection mechanism can be applied, and it is called *channel selection*. Additional information on this specific application is provided in the following chapter.

The output transcription -normally presented as a sentence- produced by the recognizer is called a recognition hypothesis or, simply, a *hypothesis*. The level of information reported in the hypothesis varies according to the decoder requirements, e.g., phone, word. In the present work, the concept of hypothesis concerns word-level hypothesis.

The term lattice is generally associated with a mathematical graphical model. In ASR, a word-lattice, hereafter simply referred to as *lattice*, is a compact and efficient structure for representing a set of word-level hypotheses [Oerder and Ney, 1993], see Figure 1.2. A word lattice is a directed graph $\mathcal{D} = (\mathcal{N}, \mathcal{L})$ where $\mathcal{N}$ is the set of graph nodes, and $\mathcal{L}$ the set of directed links connecting two nodes. A typical lattice format employs the nodes to represent time points, generally discretized to the granularity of a frame. Further, the links are used to represent words, which are associated to various optional scores, e.g., acoustic or language scores.

The whole set of hypotheses included in a representation of hypotheses (e.g., a lattice), is called a *hypothesis space*, Figure 1.2. The final hypothesis that the recognizer delivers as an output is the one that achieves, within the hypothesis space, the highest sentence posterior probability. This probability is computed from a combination of the scores present in the lattices. This measure as well as other different metrics, often computed using acoustic and linguistic scores, are categorized as *confidence measures*. Such measures are used to estimate the reliability of the recognition results.

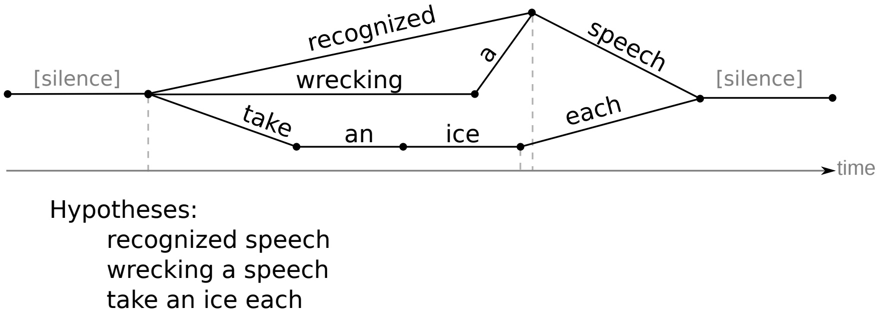


Figure 1.2: Simplified example of a hypothesis space depicted as a word lattice. The hypotheses extracted from it are below the lattice. Dashed lines are used to illustrate the association of nodes to a time instant.

The material used in this study depicts a *multi-microphone setting*, an enclosure in which audio signals are picked up by a number of microphones. The datasets include acoustic signals containing speech uttered in a real or synthetic reverberant environment.

1.3 Motivation

Nowadays, we are experiencing tremendous technological advances which can further advance voice-based solutions. This is evident in the constant reduction in the size and cost of both acoustic sensors and processors empowered with high computing-capabilities. Acoustic sensors, integrated to regular-use devices such as smart-phones, watches or even appliances, could be exploited to facilitate novel computer interfaces in more natural contexts, which at the moment is generally restricted, in terms of distance to the speaker.

On a daily basis, final users demand simple and efficient solutions for tasks, such as search or transcription of voice. Furthermore, elderly and physically impaired people anticipate mechanisms for improving their quality of life. Such mechanisms can be delivered through voice-operated smart domestic spaces, equipped with various sensors and remotely-operated devices.

Given the need for robust solutions to enable voice-based systems, and the opportunities offered by technological advances, we contribute to this direction by investigating distant speech recognition approaches in a setting where microphones are largely distributed in a room. It must be noticed that, when dealing with this scenario, critical issues and open topics are faced, such as the acoustic distortions caused by noise and reverberation or the limitations the existing solutions present when microphones are not geometrically arranged within a short-distance range between them and from the speaker.

1.4 Scope of this thesis

The scope of this thesis is to explore a set of techniques aimed at addressing the problem of distant speech recognition in distributed multi-microphone conditions. These techniques take advantage of the multiple information sources, such as signals or recognition outputs, at different stages of the recognition system. The redundancy and diversity of information, captured by the multiple sensors, are the key points upon which the approaches on this area rely.

The specific objectives of this work are:

- To investigate state-of-the-art approaches for speech recognition in scenarios that feature multiple largely distributed microphones. Two concrete fields of interest are considered: channel selection and hypothesis combination.

- To develop methods for improving the extraction of the recognition hypothesis in a multi-microphone setting.

- Concerning channel selection, the objective is to propose a practical solution to apply in a realistic scenario.

- In the area of hypothesis combination, the goal is to provide a mechanism that exploits the information provided by the multiple microphones.

- To perform evaluations and analyze the behavior of the proposed methods under different variations of the recognition system parameters.

- To use realistic simulations and real material in the different evaluation stages, in order to verify the applicability of the proposed solutions in real environments and conditions.

1.5 Contributions

The specific contributions of this work are:

- A detailed overview of the state-of-the-art methods and technologies for the addressed problem of multi-microphone speech recognition is provided. Strong emphasis is addressed to the problems of channel selection and hypothesis combination.

- For the problem of channel selection, a method based on cepstral distance is proposed. A novel channel selection framework is presented. A signal quality objective measure, effectively employed in other acoustic processing fields, is extended to this problem with positive results.

- A new combination method for hypothesis spaces represented as lattices. In contrast to existing solutions, the combination is performed without using entire hypotheses as the basis, but applying a time framed search. Also, the method presented in this work directly exploits lattices and not particular representations or manipulations of them.

- The performance of the proposed multi-microphone methods, in a standard condition and when facing a set of system feature variations, is investigated.

- The proposed hypothesis combination method is implemented as an extension of a toolkit widely used in the research community, which works on standard lattice format.

- The dissemination of the proposed methods and experimental results:

  Cristina Guerrero and Maurizio Omologo. Exploiting inter-microphone agreement for hypothesis combination in distant speech recognition. In Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European, pages 2385–2389. IEEE, 2014.

  Cristina Guerrero and Maurizio Omologo. Word boundary agreement to combine multi-microphone hypothesis in distant speech recognition. In Hands-free Speech Communication and Microphone Arrays (HSCMA), Joint Workshop. IEEE, 2014.

  Cristina Guerrero, Georgina Tryfou and Maurizio Omologo. Channel Selection for Distant Speech Recognition - Exploiting Cepstral Distance. In INTERSPEECH, 2016.

  Cristina Guerrero, Georgina Tryfou and Maurizio Omologo. "On the Use of Objective Signal Quality Measures for Channel Selection in Distant Speech Recognition", under submission to *"Computer, Speech and Language"* Journal.

1.6 Thesis Outline

The remaining part of this document is organized as follows. Chapter 2 provides background information on the problem of distant speech recognition in a multi-microphone setting. An overview of the possible solutions is presented, which consider additional processing applied to the sources captured or produced by the multiple microphones.

Chapter 3 summarizes the existing strategies applied to the problem of channel selection. Particular emphasis is given to signal-based methods, which are a valid solution for the main problem tackled in this thesis. The main contributions of this research concerning this topic are presented in Chapter 4. First, a novel framework for channel selection is proposed. The scoring mechanism is based on cepstral distance. Second, new evaluation metrics are introduced. Finally, an experimental validation of the method and comparison to state-of-the-art solution is detailed.

Chapter 5, recalls the fundamentals of topics relevant to the problem of hypothesis combination. In Chapter 6 the original work developed on the problem of combination at hypothesis level is presented. Furthermore, an approach for building a compact hypothesis space out of lattices extracted from the multi-microphone signals is proposed. Then, the experimental activities performed in order to evaluate state-of-the-art multi-microphone processing methods and our proposal, are described, together with the different evaluated scenarios. Finally, the conclusions and future directions of this research are drawn in Chapter 7.

Chapter 2

Speech Recognition in a Multi-microphone Setting

> *Mankind has made giant steps forward... However, what we know is really very, very little, compared to what we still have to know.*
>
> Fabiola Gianotti

As previously indicated, an adequate employment of the information captured by multiple sensors has already proved to be positive for acoustic processing tasks. This chapter evolves around the problem of speech recognition in a multi-microphone setting.

First, the architecture of a speech recognition system is described, featuring the scenario in which a single microphone is used for capturing the speech. Then, the role of the multiple microphones in a speech recognition task is presented. We discuss the main directions followed by researchers in the field of DSR within a multi-microphone context, in particular, front-end and post-decoding information fusion approaches. Finally, the multi-microphone experimental setting, in which the main studies conducted in this dissertation were performed, is described.

2.1 Architecture of a single-microphone recognition system

In this section we will explain the operation of the recognition system when a single microphone is used, as a starting point for following discussions on the multi-microphone ASR scenario. First, it is relevant to briefly review some historical milestones that have shaped the evolution of ASR. Initially, the problem of ASR was approached by simple methods able to recognize a small set of sounds, through specifically created rules. By the 50s, a new generation of solutions introduced the use of acoustic-phonetic elements, enabling the ASR systems to recognize phonemes or digit vocabularies. Examples of the most notable systems are [Davis et al., 1952; Forgie and Forgie, 1959; Nagata et al., 1964]. Advances in spectral analysis and representation [Itakura and Saito, 1970; Atal and Hanauer, 1971], pattern recognition and statistical modeling [Vintsyuk, 1968; Linde et al., 1980; Rabiner et al., 1979] were then critical for the rising of the next ASR systems, which had the ability to recognize small to medium-sized vocabularies. After that point, speech dynamics were increasingly better modeled by means of statistical methods, and not only the recognition process but also novel challenges, such as the robustness to multiple acoustic or linguistic phenomena, were targeted [Rabiner and Juang, 1993; Huang et al., 2001; Benesty et al., 2007]. Multiple technological issues were addressed along these last decades in order to provide a stable, practical ASR framework, such as that we know nowadays.

A standard statistical speech recognition system is formed by multiple components, Figure 2.1, that are introduced in the following through the explanation of the entire process.

Pre-processing: The input provided to the system is the digitized acoustic signal. At the *preprocessing* block, transformations are applied to the acoustic signal, in order to enhance the efficiency of subsequent feature extraction and recognition stages, and to achieve, as a consequence, an improvement of the overall system performance. Compensation to acoustic variabilities is one of the tasks addressed by this module [Lim and Oppenheim, 1979; Ephraim and Malah, 1984].

Feature extraction: This component aims at representing the speech in a reduced set of parameters, while preserving information needed to identify spoken units, and other information of the speaker (e.g., accent, emotion), as well as distortion characteristics. These features incorporate concepts from the human auditory processing and perception, and are then used for the recognition process. Diverse types of features have been explored, each with different strengths for recognition [Mporas et al., 2007]. Perhaps the most widely adopted features in ASR are mel frequency cepstral coefficients (MFCC) [Davis and Mermelstein, 1980], and perceptual linear prediction (PLP) [Hermansky, 1990].

Decoding: The *decoding* block generally exploits statistical frameworks for speech processing, at each of the acoustic and linguistic subcomponents [Jelinek, 1997; Huang et al.,

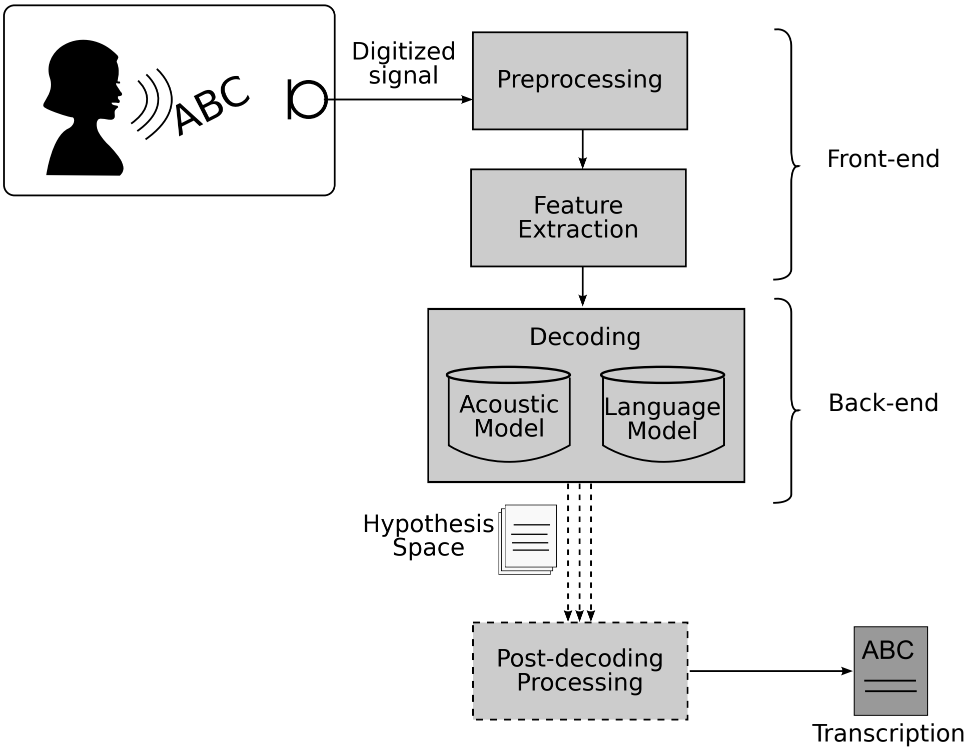


Figure 2.1: Block diagram of a statistical ASR system. Here, a single microphone is used to capture the speech signal. The dashed lines represent optional elements, i.e., hypothesis space representing multiple hypotheses, and a post-decoding processing block. The transcription is the final output of the ASR system.

2001]. Acoustic models are used to incorporate knowledge about how the features are associated to a set of spoken units. These models define the probability that, given that the source emitted the words W, the talker produced the acoustic representation A,

$$P(A|W).$$

A typical estimation of this probability is computed using statistical models such as Hidden Markov Models (HMM). For this type of recognition approaches, a general assumption is made, that the observed sequence of feature vectors was generated by an HMM. Left-to-right models are preferred for these applications, given their coherence with the sequentiality of the speech signal. The parameters of the model are typically estimated from a set of training utterances, feature vectors and their transcriptions [Rabiner, 1989; Bishop, 2006]. Once the models are trained, they are used for the classification of unknown speech segments. In this case, the classifier outputs a limited set of possible hypotheses.

The statistical language model, on the other hand, encodes the characteristics of the language. Different models implicitly convey different characteristics of the language. e.g., grammar, syntax. The objective of this language component is to provide information for predicting the most likely word sequence that was spoken. Language models, $P(W)$, estimate the prior probability of a sequence of words that is most likely to occur given a certain training material [Jurafsky and Martin, 2000]. Two elements shape these models, a vocabulary, and the relations between the words in this vocabulary. The vocabulary lists the words that can be recognized. The relations, commonly learned through a statistical process, indicate how the vocabulary units can be combined to form valid sequences. Statistical language models can be estimated from a training dataset. In such case, the probability provided by the language model can be defined in terms of a word history:

$$P(W) = \prod_{i} P(w_i|w_1 w_2 ... w_{i-1}) \quad i = 1, ..., K \quad , \tag{2.1}$$

where K is the number of words w in the hypothesis W. A common restriction is applied to the word history (i.e., the $N-1$ predecessors of the word) to define N-gram models. The probabilities of these models are estimated counting the occurrences of every sequence of N-words that appears in the training corpus. Additional probabilistic mechanisms are introduced to solve, for example, sequences of words which are not present in the corpus.

Recent advances in the field: The statistical ASR decoding considered the use of HMM to build a unified framework, that joined acoustic and language models in order to identify the most likely sequence of words. Most of the ASR systems relied on HMMs to handle the temporal variability of speech, and Gaussian mixtures to model the states of each HMM, fitting windows of frames to the acoustic input. A significant change was observed in the field of ASR during the last five years with the development of training principles and optimization techniques on deep neural networks (DNN) [Hinton et al., 2012]. Various other technological factors lead to the raise and prevalence of these statistical methods. We can mention, as example, the explosion of data generation and storage capabilities, advances in computing power, and the sustained improvements in machine learning algorithms, all of these critical for the deployment of DNNs. State-of-the-art ASR solutions implement DNNs. The main strengths of these networks is their ability to model non-linear dynamics through intricate connections on which they operate. In multiple occasions, DNNs have shown to outperform statistical HMM-based ASR approaches, making it an attractive and effective option not only for speech processing but for other statistically modeled problems [LeCun et al., 2015; Bellegarda and Monz, 2016].

Post-decoding processing: The outputs produced by the decoding block can adopt different representations, e.g, a sequence of words or a sentence, a list of the best scored sentences, or a lattice. The *post-decoding processing* involves the application of operations on the recognition results, and aims at identifying the best hypothesis from the output. Note that a recognition hypothesis can be extracted from the decoding process directly from an internal decision process. The presence of the post-decoding block is therefore optional, and presented as a refinement over the decoding results. At this stage, confidence measures (CM) are used [Jiang, 2005]. These reliability measures are used in various applications [Kemp et al., 1997; Williams, 1998], as for example keyword spotting whose goal is to detect if a given keyword was spoken in a set of spoken utterances.

2.2 Exploiting multiple microphones for DSR

Biological mechanisms inspire the use of multiple microphones for machine-based acoustic processing [Handel, 1993; Stern and Morgan, 2012]. The ears, our hearing resources, have proved to be crucial elements jointly used for a better general understanding of an acoustic scene and the recognition of voice [Blauert, 1997; Gilkey and Anderson, 2014; Wölfel and McDonough, 2009]. Moreover, for certain tasks, information heard by multiple persons in a space can be exploited for a better understanding, than if done separately. These are natural examples of cooperation of sensors, achievable even with sensors of dissimilar characteristics. In the same manner, an appropriate combination of information, captured by different acoustic sensors or produced by variate acoustic processing systems, may lead to an enhanced performance of different applications, e.g., DSR [Fiscus, 1997; Yu et al., 2004; Lamel and Gauvain, 2005].

Organized sets of sensors are the basis of approaches taking advantage of the availability of multiple microphones, and such sets can adopt different forms. A *microphone array* is a network of sensors, each of them properly set up according to the particular objective of the array. Generally, the target of the array is that of enhancing the sound waves coming from a desired direction through signal combination mechanisms. When designing an array, details such as the characteristics of the sensors, the spacing among them, or the geometry of the array, have been investigated [Huang and Benesty, 2007; Rabinkin et al., 1996, 1997; Brandstein and Ward, 2001] for their optimal performance and the identification of advantages for DSR. Studies in these topics started more than three decades ago [Flanagan et al., 1985; Alvarado, 1990]. It is thanks to their developments that, in the present days, array-based technology has a strong presence in a number of system applications and devices, e.g., video-conference systems.

The presented arrays evolved into largely spaced sensor-networks. The term *distributed microphone network* [Aarabi, 2003] refers to a limited number of microphones localized in space, which are not subject to a reduced geometry distribution, and which are connected to a recording and computing system that ensures a sample-level synchronous processing of the captured signals. As microphone arrays, microphone networks constitute a powerful tool for DSR given their potential of acquiring richer information than that obtained by a single distant microphone. Indoors experimental activities involve at least one microphone array per wall or per room, in order to ensure a broad spatial coverage. Such use of acoustic sensors has been observed in various recent large-scale research projects [CHIL-EU; SWEETHOME-ANR; DIRHA-EU].

Additionally, given the always increasing ubiquity of sensors, in personal or professional devices, the current trend is to exploit input channels which are subject to less restrictions [Bertrand et al., 2015], as for example a specific geometry. There are still open problems to tackle and solve under the conditions of such settings, which can not be addressed in the same way as for arrays.

Handling the different signals captured by the different sensors of a network can be performed in multiple ways. The different ways of processing multiple microphone signals towards a final speech recognition task are presented in the next section.

2.3 Multi-microphone processing

It is straightforward to depict a recognition system when a single microphone is used. When multiple microphones are used, as mentioned in the introductory chapter of this thesis, approaches addressing the fusion of information are necessary, particularly when the target is to produce a single output. Examples of such operations are presented in Figure 2.2.

Numerous system architectures have been proposed for multi-microphone DSR [Kinoshita et al., 2013; Barker et al., 2015; Harper, 2015; Ma et al., 2010]. There are approaches which implement the fusion of the multiple acoustic signals, Figure 2.2-a. A similar solution is provided by feature combination methods, Figure 2.2-b. Such approaches avoid, at the posterior stages, the problem of deciding which of the multiple captured or processed signals to use. These front-end approaches will be covered in Sec-

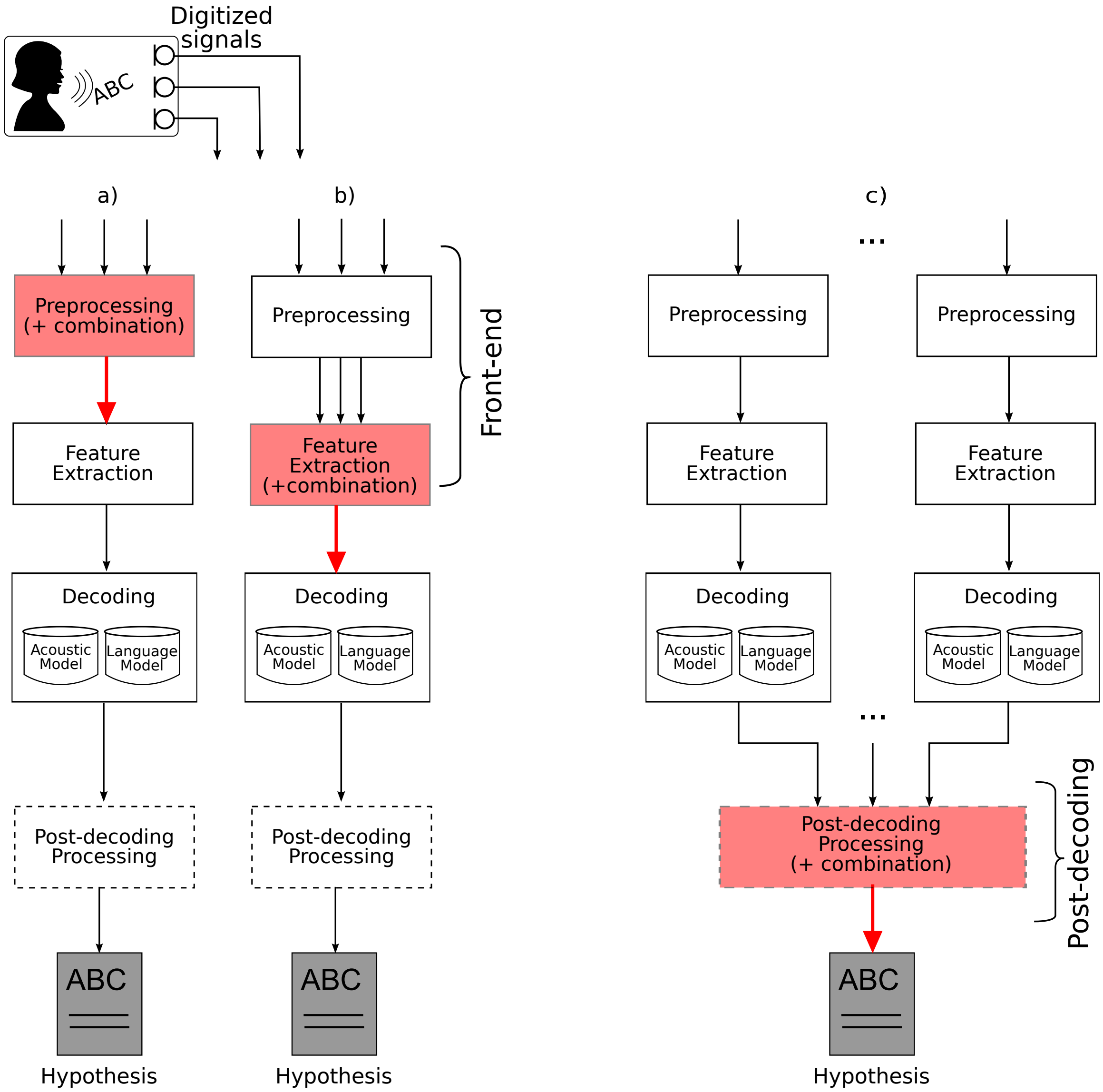


Figure 2.2: Examples of fusion at different stages of the recognition system. a) and b) depict front-end level combinations, and c) shows post-decoding level combination. Red blocks correspond to combination modules. The red arrows indicate the single output generated by an information fusion process.

tion 2.3.1. Another valid fusion approach regards the combination of multiple decoding outcomes, Figure 2.2-c, in order to exploit their complementarity in a later post-decoding stage. These latter topics will be reviewed in the post-decoding processing Section 2.3.2.

Similar to the previously discussed architectures, a selection of a single best performing unit among multiple available ones (e.g., signal, hypothesis) can take place at the different levels of the recognition system. These selection approaches are also considered fusion strategies, since in order to perform a selection a fusion step is necessary. In such a case, a careful consideration of the multiple elements is required, and usually implemented through the exploitation of quality measures. Such scoring techniques, applied for channel selection, will be reviewed in Chapter 3.

Just for completeness purposes, it is worth mentioning a different kind of approaches called *cross-adaptation*, in which explicit information fusion is not performed. The goal of these approaches is to exploit the output of one system to adapt another one. Such adaptation strategies require modifications on the architecture of the recognition systems. Then, multiple passes are necessary in order to extract the adaptation data at one pass, and then use another pass for the adaptation. In this dissertation, no additional review will be performed on these methods. For further details see [Gales et al., 2007; Liu et al., 2012].

The most relevant research directions on the topic of information fusion are presented in the following.

2.3.1 Front-end processing

The objective of a multi-channel front-end processing approach is to process the acoustic information captured by multiple sensors, in order to produce a suitable input to the recognition and other posterior processing modules, e.g., language understanding. The efficiency of the subsequent processing modules relies on the quality of the product delivered at this level. When referring to multi-channel front-end processing, we consider techniques addressing problems such as beamforming, source enhancement, source separation, source localization, and event detection. Considerable research has been done to tackle each of these specific problems, for more details see [Wölfel and McDonough, 2009; Benesty et al., 2007]. However, for a multi-microphone capture system with the goal of improving the recognition accuracy, it is worth mentioning techniques such as those aiming at the selection or combination of information in order to deliver a single signal or feature-set as output.

Signal and feature combination: Concerning multi-microphone front-end processing, the main types of combination include signal and feature combination.

Techniques such as spatial-temporal filtering, commonly called *beamforming* (BF) [Elko and Meyer, 2008], are implemented on top of multi-microphone audio acquisition mechanisms. The goal of BF is to combine signals from several sensors in order to suppress the interference from undesired sources and emphasize a desired one. Delay-and-sum beamforming [Flanagan et al., 1985], perhaps the most popular approach, uses only geometrical knowledge to combine the signals from several sensors. Although simple as idea, it has proved to give benefits in multiple cases [Barker et al., 2015]. There are several variations of the delay-and-sum approach. Conventional BF algorithms [Kumatani et al., 2012] include Minimum Variance Distortionless Response [Capon, 1969; Basha et al., 2011], Minimum Mean-squared Error, Maximum SNR Ratio. A popular implementation of BF, used mainly in speaker diarization tasks, is BeamformIt [Anguera et al., 2007; Anguera, 2006]. BF techniques are mathematically supported under specific geometrical configuration of a microphone array. Under loosely specified distant microphone configurations their application is possible, however, it is not correct from the theoretical point of view which results in spatial aliasing and other artifacts strongly affecting the output fusion. The use of methods which select the least distorted channel has also been studied for BF, and has been found beneficial [Kumatani et al., 2011]. In [Anguera et al., 2005] the identification of a proper reference channel was also found to have a positive effect in channel weighted delay-and-sum beamforming for speaker segmentation.

Popular examples of other topics exploiting multi-microphone captured signals are *speech enhancement* [Benesty et al., 2005], and *source separation* [Pedersen et al., 2008]. Methods concerning the first topic aim at improving the quality of the captured acoustic signal, compensating the signal degradation. Noise and reverberation are the principal conditions addressed [Droppo and Acero, 2008]. On the other hand, the main objective of sound source separation methods is to extract a target signal, e.g., speech, from a mixture of multiple acoustic signals. With the ongoing development of sensor technologies, many of these solutions are already implemented in applications such as hearing aids or robotics. This set of techniques rely on the assumption that the positions of the microphones are known, and, in many cases, that the position of the desired source is reliably identifiable.

Feature combination approaches are in fact generators of new compact features, and are generally implemented through the concatenation of different feature sets, which are expected to be complementary [Ma et al., 2010]. A common way of approaching this combination problem is the use of principal component analysis or linear discriminant analysis for the reduction of the number of features. The lack of a perfect synchronization among the information streams, i.e., the features representing temporally evolving signals, has shown to severely affect the performance of these algorithms. In the literature, extensive experimental work can be found on the use of this combination approach for single channel settings, with various acoustic (e.g., auditory or articulatory) features [AMI-EU; Hermansky et al., 1996; Zolnay et al., 2005; Schlüter et al., 2006].

The combination of features extracted not only from audio but also from other sources such as video captures, are called multimodal features, which are worth mentioning in this section. This strategy has been evaluated not only for recognition but for other human interaction tasks as well [Potamianos et al., 2003; Neti et al., 2001; GaticaPerez et al., 2005]. Advanced models are exploited in such cases for the implementation of these multimodal features, which pose specific demands in terms of training-data requirements, synchronization of data captured at different rates, or for the combination of different formats [Atrey et al., 2010; Galatas et al., 2012]. Although with less frequency, other combination approaches introduced at the decoding level have also been explored. In [Li and Sim, 2013] a straightforward averaging of posteriors, at state level, is used to combine multiple systems. This approach, particularly tailored for DNN systems, was also used in one of the top performing systems [Du et al., 2015] of the CHiME-3 evaluation campaign.

Best channel identification: *Channel selection* (CS) focuses on picking the signal that produces the best recognition performance, out of the multiple signals captured. The ideal CS criterion, for recognition, should be highly correlated with word error rate and it should be extracted in an unsupervised way. The robust selection of a channel depends on an accurate assessment of the reliability of the channel. The problem here is establishing the terms in which reliability is defined. Several measures have been explored. In [Wolf and Nadeu, 2014] CS approaches were categorized into signal and decoder-based. Signal-based CS methods present the advantage of a lower computational complexity, since recognition is made only once for the selected channel. However, these measures which adjust well to a specific kind of distortion, may fail in different conditions. Signal-to-Noise ratio (SNR), for example, can not reveal any hints about reverberation. Additionally, this ratio is commonly measured in decibels (dB), however the dB fluctuations, due to the relative values compared, lack of significance. Cross-correlation among signals applied to a microphone array [Kumatani et al., 2011], and the envelope variance are other examples of signal-based scores. While signal-based quality measures are directly extracted from the signals characteristics, decoder-based measures do not exploit the degree of signal quality but its effect over the recognition result. Other scores have also been studied for CS, such as room Impulse Response (IR) based measures, which require a priori information about the acoustic environment, or the use of the speaker position and orientation. For the latter, an experimental study was performed for CS, having position and orientation as prior knowledge [Wolf and Nadeu, 2010]. It must be noted that, out of the context of CS, the relation between these speaker features and recognition has been studied [Omologo et al., 1997].

In [Wolf and Nadeu, 2014], various objective measures were explored on CS for DSR. In their work, it was assumed the use of clean speech for training the acoustic models, and that the least distorted signal leads to the highest ASR accuracy. The study used two datasets for evaluations, one simulated with the use of IRs, and another using real recordings. Acoustic conditions in the sets are different; in the first set, no indication is given about noise. Nevertheless, measures such as SNR, which will certainly not be advantageous in one of these studies, are used to confront the performance of others. Further expansion must be done in the portability of such CS findings, first, because the distortion can not always be described by a single measure. And second, because the decoding parameters such as acoustic models are not always trained on clean material. Under the explored experimental settings, the authors reported that signal-based scores outperform decoder-based ones, in terms of word error rate reduction. Additionally, a straightforward combination of score rankings was also implemented and found valuable for the task. However, no study is presented about the heuristics to use in order to perform an optimal selection of measures to combine.

In a recent work [Cohen et al., 2014], a proposed CS method addresses scenarios with loosely located multiple microphone-arrays. The approach takes advantage of power ratios, at local and global level, in order to estimate the level of reverberation at each microphone. The system requires clusters of microphones able to capture audio from different directions. In the literature, this CS approach has been explored for purposes different than that of speech recognition, i.e., for teleconferencing.

2.3.2 Post-decoding processing

The methods of combination and selection of hypotheses take place at a point closer to the decoding process than the previously presented approaches. These methods operate

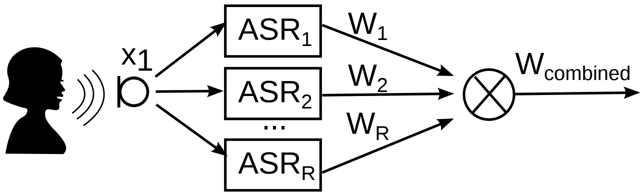


Figure 2.3: Original hypothesis combination architecture with a single-channel. X_1 represents the signal captured by one microphone. ASR_i indicates a particular ASR system, which produces the hypothesis W_i. The resulting hypothesis is $W_{combined}$.

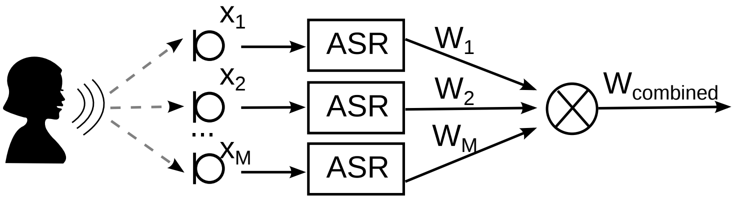


Figure 2.4: Multi-microphone hypothesis combination architecture. X_i represents the signal captured by microphone i. The same ASR system is used to decode each signal. One different hypothesis W_i is extracted from each decoding process. The resulting hypothesis is $W_{combined}$.

on information extracted at the final stage of the recognition system.

Hypothesis combination: These approaches are demanding, in terms of resource consumption, because an individual decoding run is required for each signal or setting, in order to combine the output of each decoding in a post-decoding stage. Though higher in complexity than signal-based combination approaches, hypothesis combination presents a crucial advantage, its capability of exploiting information captured by different sensors without being limited to the specific physical characteristics of the microphone network. In a constantly changing scenario, in which a not so far-in-the-future scene contemplates the availability of sensors in different devices around the space, this advantage constitutes a strong potential of applicability.

Originally, hypothesis combination was proposed to combine the recognition outputs extracted from a single unique acoustic signal passed to multiple recognizers [Fiscus, 1997], Figure 2.3. Such strategy is still applied in work related to robust speech recognition. Later, these single signal-processing based approaches were extended for the combination based on multiple signals captured by a various microphones, Figure 2.4.

Improvements in single-channel hypothesis combination are accomplished primarily due to complementary errors made by different ASR systems, since combination methods exploit the occurrence of these errors in order to make proper decisions. The key point in this case is how to know in advance which variations of the ASR systems would lead to better combinations. There are efforts focused on the extraction [Breslin, 2008] or design of diverse systems for fusion. Normally, a heuristic exploration is applied, modifying one or various of the system modules, in order to identify an optimal configuration. The needed diversity is mostly introduced by using various sets of training models and different acoustic features, or by combining structurally diverse acoustic models such as Gaussian mixture models, HMMs and deep neural networks [Cui et al., 2013]. Recently, a theoretical approach was proposed [Audhkhasi et al., 2014] to explain the link between ASR system diversity and the performance of a hypothesis combination approach.

Large-scale ASR projects [Rabiner and Juang, 1993; Chen et al., 2006; Stallard et al., 2007; Tür et al., 2008; Cui et al., 2013] and systems participating in DSR challenges, have acknowledged that the fusion of hypotheses is a key component for achieving state-of-the-art performance. The application of such methods has shown remarkable improvements on the reduction of recognition errors. However, the manipulation or alignment procedures over the hypotheses leads to an increase of processing time, which is an important feature for some recognition systems. The methods Recognizer Output Voting Error Reduction (ROVER) [Fiscus, 1997] and Confusion Network Combination (CNC) [Mangu et al., 2000; Evermann and Woodland, 2000] constitute the two most relevant approaches on this topic, which have been extensively tested on the combination of systems varying different features; e.g., acoustic channel, front-end, hypothesis from independent systems. In both approaches, hypothesis confidence measures have been used to augment the combination decisions.

The ROVER algorithm produces a composite hypothesis out of multiple recognition hypotheses. This method is based on a voting and weighting scheme. A limitation of the original ROVER method is that only the 1-best hypothesis from each individual system is used. On the contrary, N-best ROVER [Goel et al., 2000] exploits the hypotheses from N-best lists as an attempt to model a wider hypothesis space. Modified versions of ROVER include other machine learning techniques instead of the voting step [Zhang and Rudnicky, 2006], or constraints into the voting, such as language model information [Schwenk and Gauvain, 2000]. [Hillard et al., 2007] exploited a system classifier at each word location in order to identify the most likely word in that position within the hypothesis. The procedure behind [Abida et al., 2011] used a contextual analysis to eliminate the erroneous words from the alignment network used by ROVER before the voting stage. ROVER voting method has also been studied as a starting point for bootstrapping rescoring approaches. In [Fujita et al., 2015] a discriminative word selection method was built on top of a composite word transition network created by a ROVER previous step. A recent work resorts to a mixture of features, some of them aiming at capturing words' pronunciation difficulty, in order to improve ROVER [Jalalvand et al., 2015].

The idea behind the combination of confusion networks is to take as input a set of compact hypothesis spaces, called confusion networks (CN) and then process them through a voting method, in a similar fashion of ROVER but at the level of CNs instead of hypotheses. The standard implementation of this method is available in the SRI Laboratory Language Modeling Toolkit (SRILM) toolkit [Stolcke, 2002]. A step in which lattices are converted into CNs is required. This operation introduces a potential loss of information, although generally not so critical for the final result, due to the lattice manipulation. This effect is observed, for example, when unreachable nodes are discarded in the conversion process.

Both, ROVER and CNC, resort to dynamic programming alignments, which make the methods to be affected by the order in which the elements, i.e., hypotheses or CNs, are processed. In terms of resources, the impact of the additional processing required to compute all the combinations for identifying an optimal one, is not minimal. In a multi-microphone setting, this means that not only the number of microphone but also the sequence in which their derived elements are combined become relevant. Considering a case in which only a subset of microphone-derived hypothesis spaces are combined, the number of possible configurations, K, is:

$$K = \frac{n\,!}{(n-r)\,!} \quad , \tag{2.2}$$

where n is the total number of microphones available to choose from, and r is the number of actually chosen microphones to combine.

CNC has been explored in multi-microphone contexts [Stolcke, 2011; Wölfel et al., 2006; Cossalter et al., 2011]. In [Stolcke, 2011], the author claimed that no significant improvement was found in comparison to signal-based approaches (e.g., Beamforming). The study was conducted on data recorded using tabletop microphones. Even though channel selection criteria have been explored for CNC in a multi-microphone scenario, an optimal solution for selecting or properly assigning weights to the multiple CNs has not been identified yet.

The literature mentions various other efforts in the area of hypothesis level combination. Driven Decoding Algorithm [Lecouteux et al., 2008] presents another combination approach in which a first pass is used to orientate the decoding in a second pass. In [Hoffmeister et al., 2006], the authors proposed a combination that operates at frame level, without altering the structure of the word graph. These experiments were done on a single signal, combining multiple systems features, namely language models. Lattice pruning was applied, before the execution of combination approaches, but its effect was not explored afterwards. No significant differences with ROVER or CNC were found in terms of performance. Previous lattice-based combination approaches [Li et al., 2002], again for single-signal multiple-ASR-systems settings, assumed complementarity between features of type MFCC and PLP. In that work, operations over the lattices were performed in order to merge information elaborated at phone-level. The manipulation of the lattices results in a loss of confidence scores, which is the reason for using a comparison of the approach to ROVER without confidence scores. The authors claimed a better performance than the compared method under the study conditions.

Hypothesis selection

Methods aiming at the selection of one hypothesis, which is more likely to achieve a high recognition accuracy, are included as part of the decoder-based CS approaches, Section 2.3.1. Examples of such approaches are [Wolf, 2013; Stolcke et al., 1997], where an N-best list was used as the hypothesis space from which hypothesis candidates were extracted. Conventional methods resort to estimating, also from the decoding process, the reliability of the candidate hypotheses. Main reliability scores include likelihood-based measures, word posterior probabilities, the normalization of features [Obuchi, 2006], or the separability of classes achieved during the decoding process [Wölfel, 2007]. No specific multi-microphone study has been found on this topic.

2.4 Experimental multi-microphone settings

The problem and target scenario of this thesis are summarized as DSR performed in a room characterized by largely distributed microphones. In this section, first, we discuss the relevance of the experimental scenarios for the research in speech processing related tasks. Then, we describe the smart-room setting that is explored in the main experimental activities for this dissertation.

2.4.1 Experimental scenarios for the ASR progress

Benchmark datasets (e.g., Aurora, AMI-AMIDA, DICIT) and evaluation campaigns (e.g., REVERB, CHiME, ASpIRE) provide important development resources for the scientific community. These resources are crucial for assessing, not only different speech recognition systems and the methods they apply, but also the current progress status of the field of ASR. This global evaluation is only possible due to the varied and rich conditions offered by the mentioned experimental resources. Important factors that support the validity of the experimental findings, on these resources, concern the quality and realism of both the datasets and the processing toolkits. A proper corpora is essential for training and testing the different speech processing, enhancement, and recognition solutions explored by the researchers. For example, the performance of acoustic and language models is greatly influenced by the amount of available training data, and by the similarity between the train and test sets.

The experimental resources, facilitated by contemporary corpora or evaluation campaigns, actively try to incorporate more realism to the exploratory scenarios [Le Roux and Vincent, 2014]. We observe this trend in recent activities including real reverberant speech, and real recorded noise sequences [Barker et al., 2015]. On the other hand, simulations that closely emulate real conditions are exploited, although they have repeatedly been criticized in the past for failing to capture the complexities of real scenarios. Simulations provide inexpensive and efficient means for generating vast amounts of data that, in some cases, are not attainable through real recordings or can not be afforded. In addition, simulations facilitate the elaboration of carefully designed and controlled datasets, upon which focused scientific studies can be performed.

Concerning multi-microphone scenarios, typical settings found in the scientific community include: conference or meeting rooms equipped with microphone arrays on a table or the walls [AMI-EU; Janin et al., 2004], domestic rooms equipped with a distributed microphone network [Vincent et al., 2013; SWEETHOME-ANR], and personal devices with multiple acoustic sensors [Barker et al., 2015]. The numerous systems participating in the mentioned challenges include, in their solutions, recognition components based on different platforms. The REVERB-14 challenge [Kinoshita et al., 2013] provided a baseline ASR system based on the Hidden Markov Model toolkit (see Appendix A.1), while the recent CHiME series challenges exploit the Kaldi toolkit (see Appendix A.2).

2.4.2 DIRHA setting

Concerning the problem of doing research on the DSR problem, different datasets available in the community at the beginning of this work provided favorable research conditions, but at the same time restrictions. In some cases, the space or speaker conditions were biased to a specific application (e.g., a meeting), in other cases the microphone network presented limitations (e.g., the number and distribution of microphones in the room). For this work, we adopted an experimental scenario and datasets, on which we had unrestricted access, and the possibility of extending the resources provided. This section describes the main experimental setting exploited for this dissertation. Specifically, we present the environment and corpora on which the main experiments are conducted. Then, a general description of the speech recognition setup is reported.

Smart-home: Over the last decade, it has been observed an increasing introduction of ASR in several services and application fields. Moreover, with the potential of improving the quality of life of physically impaired people, smart voice-operated domestic spaces equipped with sensors and remotely operable devices have been envisioned. As example, a related recent work is the "Distant-Speech Interaction for Robust Home Applications" (DIRHA) Project [DIRHA-EU] (see http://dirha.fbk.eu), in which a non-intrusive far-field speech-based interaction between a motor-impaired user and an automated house is explored. It must be noticed that such scenarios undergo various critical issues (e.g., spontaneous speech and uncontrolled acoustic conditions). As a result, the development of reliable voice-based interfaces is challenging. In order to investigate solutions for improving the overall robustness on spoken dialogue home automation systems, the project adopted a network of distributed microphones, aiming to mitigate the impact of reverberation and background noises.

The DIRHA smart-home is equipped with 40 microphones that record data synchronously, and that are distributed inside five rooms of a real apartment (approximately 70 m^2). In the living room there are 15 microphones, 13 in the Kitchen, 7 in the Bedroom, 3 in the Bathroom, and the remaining 2 in the Corridor. All microphones are placed on the room walls, with the exception of the Kitchen and Living room that are also equipped with a ceiling array of 6 microphones, arranged in a star-shaped configuration. For our experimental activities, in all datasets exploited, a single active source is considered, located within a single room, namely the living room. The living room is characterized by an average reverberation time about 0.75 seconds. In the case of simulated data, no noise

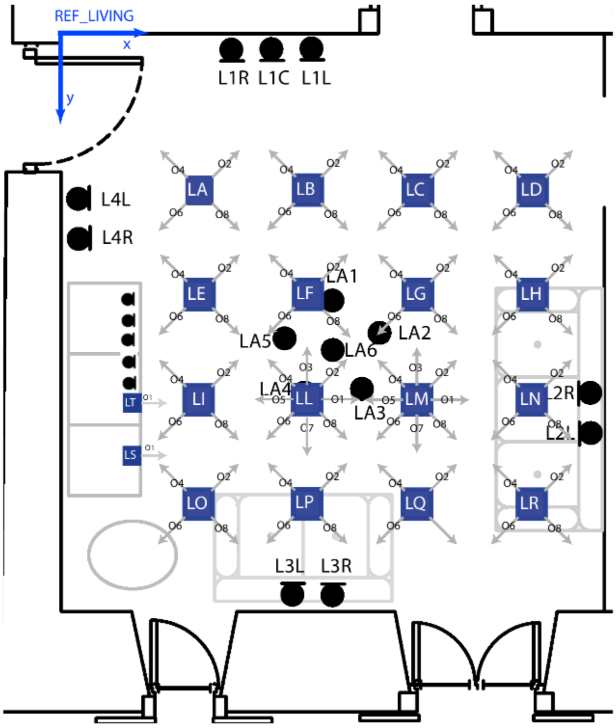


Figure 2.5: General layout of the living room in the DIRHA smart-home. Speaker positions and orientations are marked as blue squares. Microphones are indicated as black circles. Furniture is depicted with gray lines.

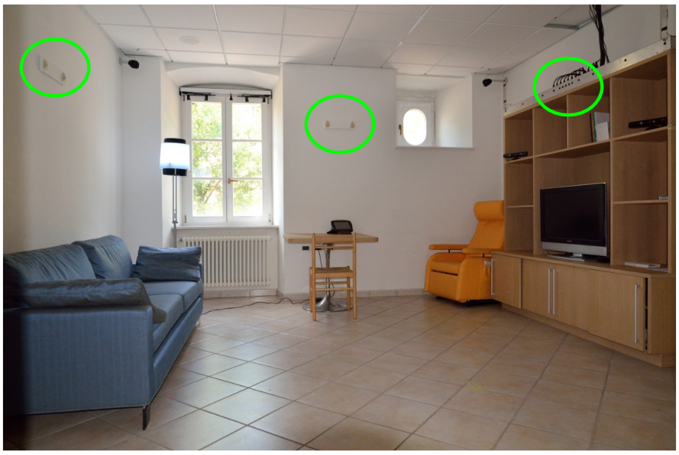


Figure 2.6: A photograph of the living room in the DIRHA smart-home. Some of the microphone arrays are marked by green circles.

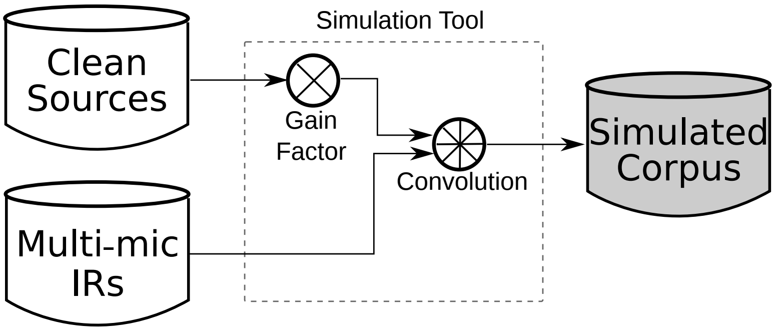


Figure 2.7: Basic schema of the simulation process followed in this work. Clean speech signals are the source, selected from the available clean corpus. Previously measured IRs capture the sound propagation effect of the speaker position and orientation in the space. Then a convolution between the clean signals and the proper set of multi-microphone IRs is performed to account for the room acoustics.

was added, since this phenomena was not within the scope of the studies. For the real recordings, however, a minimal level of noise is present. Audio signals were recorded at a sampling frequency of 48kHz 16 bit PCM. For the experiments reported in this work the signals were converted to a sampling frequency of 16kHz. Figure 2.5 shows the area of the living room; squares and arrows (in color) indicate some of the positions and orientations adopted by acoustic sources in the DIRHA simulated databases. A photograph of the living room is shown in Figure 2.6.

Data: Different databases collected for DIRHA are exploited in this work. DIRHA datasets include clean speech sets recorded both in a recording room and in the appartment described above. More details about the DIRHA datasets can be found in [Cristoforetti et al., 2014; Matassoni et al., 2014; Zwyssig et al., 2015]. Multiple IRs were measured at each room, for a large set of microphone-speaker position configurations [Cristoforetti et al., 2014]. Various datasets were simulated through contamination of clean speech with the measured IRs [Matassoni et al., 2002]. The simulation process adopted in this study is described in the Figure 2.7. Realistic simulations are particularly useful in experiments in which it is required to control certain acoustic phenomena in order to appreciate their effect over the methods studied. Word-level annotations were provided for all datasets. Both, simulations and real recorded reverberant data are used to conduct the studies reported in this dissertation.

Speech recognition: ASR systems are implemented, in real applications, through the use of recognition toolkits or frameworks. In this work, ASR is implemented using HTK (see Appendix A.1) and Kaldi (see Appendix A.2), which are two of the most renowned speech recognition platforms. Details about the configurations adopted are provided at each section where an experimental activity is reported.

Chapter 3

Channel Selection

One never notices what has been done;
one can only see what remains to be done.

Marie Curie

This chapter presents a review of the existing work on the challenge of selecting a channel achieving a good or, hopefully, the best recognition performance. Particular emphasis is provided on mechanisms based on signal processing, which are valid solutions to apply in a scenario as the one targeted in this thesis, where microphones are distributed in space. Finally, we introduce objective quality measures as potential means for CS.

3.1 Overview of channel selection methods

In a setting in which the microphones of the network are largely distributed in a room, a valid alternative to the combination of different signals is given by Channel Selection. As stated before, the goal of CS is to identify, among all the available input signals, the one that leads to the best performance on a given task. CS is frequently associated to the task of speech recognition. In a real application context, CS should work dynamically, selecting the optimal channel at each speech input. Moreover, it must be noticed that, along each spoken utterance, the optimal channel may change due to various possible reasons, such as a possible change in the speaker position or his/her head orientation. A score is generally required to be computed for each input channel in order to apply a decision mechanism. CS methods are commonly categorized according to the stage at which their scores are computed. A limited amount of work is found in the field of CS

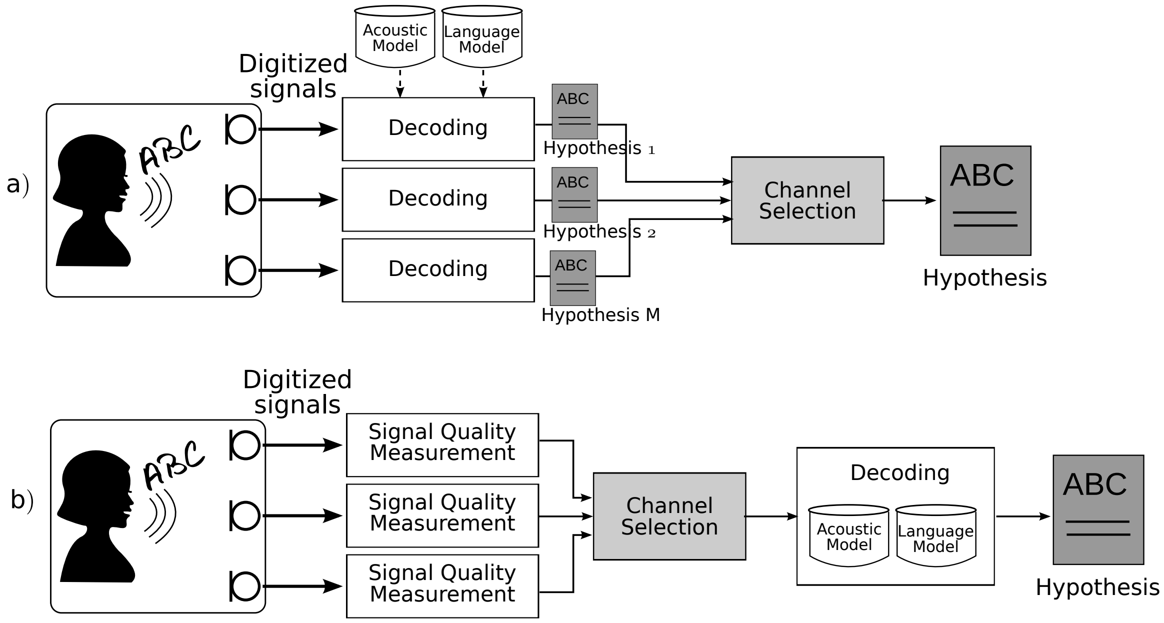


Figure 3.1: Diagram of the operation of CS methods. On top, diagram a) shows decoder based CS, and at the bottom, diagram b) shows signal-based CS.

[Wolf, 2013]. In the literature, CS approaches are commonly categorized as *decoder-based* and *signal-based*, Figure 3.1.

3.1.1 Decoder-based CS

These methods use information from the decoding process, e.g likelihoods or posterior probabilities. Such approaches do not directly estimate the quality of the signals but the plausibility of each decoding result. It is therefore not possible to apply such techniques without requiring first the operation of an ASR over each captured signal. Representative examples of scores employed by these methods are mentioned in the following.

Feature processing: These methods, applied through mean and variance normalization, or histogram equalization, are mechanisms also used in CS [Atal, 1974; Openshaw and Masan, 1994; Molau et al., 2001; De La Torre et al., 2002; Obuchi, 2004, 2006]. The original idea was to take each channel, pass it through a feature processing step, then decode the original and the transformed features, and select the channel that obtained the smallest differences between recognition results. The main deficiency behind this method is the computational burden required by the multiple recognition executions.

Class separability: This approach is a mechanism extended from the field of pattern recognition. In order to apply this concept into CS, class units have to be defined. These units can correspond to phonemes or even words. The concept in this case is to identify the channel in which the decoding process reveals the maximum separability between classes [Wölfel, 2007].

Likelihood: Despite not being an absolute score between individual decoding processes, provides insight about their results. In [Shimizu et al., 2000], this instrument was straightforwardly used for CS. This approach presents serious limitations given that, in the decoding process, it is computed as a non-normalized score. In a multi-microphone context, the posterior probability of the recognition of an utterance, based on acoustic and linguistic components previously introduced at Section 2.1, is defined as:

$$P(W|A_m) = \frac{P(A_m|W)P(W)}{P(A_m)} \quad m = 1, ..., M \quad , \tag{3.1}$$

where m is the index of the microphone and M is the total number of microphones. However, in the practice, it is only approximated, since $P(A_m)$ is neglected. This fact impedes the direct comparison of these measures. In [Wolf, 2013], the authors implemented a pairwise normalization step on top of the likelihoods. The decision mechanism is computed as:

$$C = \arg\max_m \sum_i \frac{P(A_m|W_m)}{P(A_m|W_i)} \quad , \tag{3.2}$$

where m and i are microphone indexes, and C is the selected channel.

Although, based on their nature, post-decoding measures should present a higher correlation to recognition performance. This fact has not been proved in the literature so far [Wolf and Nadeu, 2014].

3.1.2 Signal-based CS

The scoring instruments used in these approaches operate directly on the signals acquired by the different microphones in order to estimate the quality of each channel. The main advantage of such methods is the low computational complexity required, since once CS is applied, only one channel is then processed by the recognizer. It must be noticed that, other acoustic processing applications, as for example BF, may also benefit from multiple CS methods [Kumatani et al., 2011].

Some authors have focused their efforts in CS with the target application of beamforming, where more than one channel is selected for further processing [Kumatani et al., 2011; Himawan et al., 2015]. In such methods the beamformed signal extracted using all channels can be used as a reference which is compared to the acquired signals to rank them in terms of relative 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.

There also other CS methods which are derived from signal level information. In [Wolf and Nadeu, 2009], the authors proposed a CS measure computed as the ratio of the energy from the late reflections of an impulse response, to the energy of the whole IR. Although such a measure operates well when the room IR is known a prior, an uninformed implementation is hard to obtain, since real-time IR estimation in quick changing environments is still an open problem.

The position and orientation of a sound source has also been explored for CS. In [Wolf and Nadeu, 2010], a brief experiment was conducted, using prior knowledge about the source location, for CS. The main limitation of this approach is given by the dependence of CS on a reliable estimation of the source location and orientation. Moreover, existing solutions extract this information exploiting specifically arranged microphone arrays, which will be needed for this CS implementation.

Concerning strictly signal-derived measures that are used for CS, a detailed revision of such instruments is presented in the remaining part of this chapter.

3.2 Signal quality estimation for CS

In a multi-microphone setting, it is assumed that the identification of the least distorted channel leads to optimal ASR results. Signal-based CS can be based on measures that quantify a particular characteristic of the signal, as for example distortion. In the following section we present the measures used in state-of-the-art signal-based CS.

Energy and Signal to Noise Ratio (SNR): Energy is used as a straightforward attempt to identify the least distorted channel. The assumption is that a signal with higher energy is an indication that the speaker was relative closer and better oriented towards the microphone. It is expected that the direct wave shows stronger energy relatively to a reverberated signal, as well as to the background noise. In order for this approach to work, strong assumptions must be made about the conditions in which signals are captured, which are not easily achieved in real life.

SNR, on the other hand, is a useful tool for assessing the level of additive noise in a signal. Limitations of scores based solely on energy could be avoided with a normalization process, for example, computing SNR with energy normalized over energy of the noise in the silent segments. Then, another limitation is observed, since the boundaries between the speech and the silent segments are not clear due to the smearing effect of reverberation. Moreover, even in conditions subject only to noise, particularly for unstationary noise, the estimation of SNR is still an open problem. SNR score was explored in [Obuchi, 2004; Wölfel et al., 2006], where the authors evidenced a problem associated to SNR, that the score does not present a powerful discriminatory ability when facing reverberation.

In [Cohen et al., 2014], a CS method is proposed for the improvement of teleconferencing system, which is based on real-time monitoring of audio signal reverberation. This solution is implemented exploiting the ratios between the powers of the signals captured by microphone clusters. No specific speech recognition evaluation of this method was found in the literature.

Cross-correlation: The concept behind this approach is that of identifying the least correlated channel, assuming that a low correlation indicates that it is the most distorted channel. Then the identified channel is eliminated from the set of candidates. As a result, the CS process selects more than one channel. With the target of improving beamforming, by reducing the set of microphones to those more promising for the outcome, in [Kumatani et al., 2011] the authors showed that, for a given configuration, this reduction in the set of microphones was positive for the task. A strong assumption is done however, on the location of the sensors, since a microphone-array is required. As the distance between microphones increases, the basic assumption, of lacking correlation as a sign of distortion, weakens. This fact restrains the portability of this solution to such distributed microphone settings.

Envelope Variance: One of the most successful literature methods for signal-based CS is based on the envelope variance (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 filter-bank energies (FBE) $x_m(k,l)$ in channel m, a sub-band k and time frame l, are first normalized as follows:

$$\hat{x}_m(k,l) = e^{\log x_m(k,l) - \mu_{\log x_m}(k)} \quad . \tag{3.3}$$

The mean normalized sequence of FBE is then compressed with a cube root compression and the variance $V_m(k)$ of each sub-band and each channel is calculated.

The CS based on the EV measure is based on the selection of the channel that maximizes the average variance over all channels:

$$\hat{C} = \arg\max_{m} \sum_{k} \frac{V_m(k)}{\max\limits_{m}(V_m(k))} \quad . \tag{3.4}$$

In Equation (3.4), the use of a set of different weights for each channel and sub-band has been proposed in [Wolf and Nadeu, 2014], but to our knowledge, no further elaboration of this concept has been described and no relative experimental evidence has been presented, to support the use of such a weighting scheme.

Modulation Spectrum Ratio: A recent method for CS uses the short-time modulation spectrum, and a beamformed signal in order 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 of modulation is formulated as a ratio, as follows:

$$\zeta_m(n, f) = 10 \log_{10} \frac{|\chi_m(n, f)|^2}{|B(n, f)|^2}, \quad 0 \leq f \leq F \quad , \tag{3.5}$$

where n is the index of the discrete frequency, f is the index of the discrete modulation frequency, $\chi_m(n, f)$ and $B(n, f)$ are the modulation spectra of channel m and the beamformed signal, respectively, and F is the highest modulation frequency. It is noted here that the above method for CS is limited by the use of a beamformed signal as a reference since this theoretically can not be applied in an unconstrained distant speech recognition scenario where distributed microphones are used.

3.3 Objective signal quality measures

Objective quality measures have been consistently exploited for many years in various speech processing applications. Measures such as the cepstral distance (CD), the log-likelihood ratio (LLR) [Hansen and Pellom, 1998] and the frequency weighted segmental SNR [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 error 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 evaluation of speech enhancement [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 objective signal quality scores can lead to a meaningful selection of the least distorted channel, among the signals captured by a distributed microphone network. Particularly, the CD is long known for its effectiveness and flexibility in different application fields [Rabiner and Schafer, 2011].

In this section we present objective signal scores already used in the community for evaluation of speech-enhancement algorithms. No preliminary work was found where these scores were formally used in the field of CS for ASR.

Cepstral Distance (CD): An important concept to introduce here is the cepstrum. The cepstrum was defined as the inverse Fourier transform of the log magnitude spectrum of a signal [Bogert et al., 1963]. Later work showed the connections between cepstrum and the more general concept of homomorphic filtering of signals that are combined by convolution [Oppenheim et al., 1968]. This work introduced the definitions of the cepstrum of a discrete-time signal as well as the extension to the complex cepstrum, and the identification of the complex cepstrum property by which convolution operations can be transformed into additive [Rabiner and Schafer, 2007]. This latter property, exposes the important role of cepstrum in speech analysis, considering that the human model for speech production involves a convolution process.One way for computing cepstral coefficients, based on linear predictive analysis, is as follows [Loizou, 2013]:

$$c(j) = a_j + \sum_{k=1}^{j-1} \frac{k}{j} c(k) a_{j-k} \quad 1 \leq j \leq p \tag{3.6}$$

where a_j are the predictive filter coefficients, and p is the order of the predictive analysis.

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 [Rabiner and Juang, 1993], 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 [Quackenbush et al., 1988; Kitawaki et al., 1988;

Hu and Loizou, 2008]:

$$d(\boldsymbol{c}_x, \boldsymbol{c}_m) = \frac{10}{\log_{10}} \sqrt{2 \sum_{k=1}^{p} [c_x(k) - c_m(k)]^2} \quad , \tag{3.7}$$

where $\boldsymbol{c}_x$ and $\boldsymbol{c}_m$ are the cepstral coefficient vectors of the clean and distorted signals respectively.

Log-Likelihood Ratio (LLR): The LLR is defined as:

$$d_{LLR}(\boldsymbol{a}_x, \boldsymbol{a}_m) = \log \frac{\boldsymbol{a}_x \mathbf{R}_m \boldsymbol{a}_x^T}{\boldsymbol{a}_m \mathbf{R}_m \boldsymbol{a}_m^T} \quad , \tag{3.8}$$

where $\boldsymbol{a}_x$ and $\boldsymbol{a}_m$ are the LPC vectors of the clean and distorted speech signals respectively, and $\mathbf{R}_m$ is the autocorrelation matrix of the reverberated speech signal. LLR has been shown to correlate well with subjective evaluation of signal quality [Hu and Loizou, 2008].

Frequency-weighted segmental SNR (fwSNRseg):

$$fwSNRseg = \frac{10}{M} \cdot \sum_{m=0}^{M-1} \frac{\sum_{j=1}^{K} W(j,m) \log_{10} \frac{|X(j,m)|^2}{(|X(j,m)| - \hat{X}(j,m))^2}}{\sum_{j=1}^{K} W(j,m)}, \tag{3.9}$$

where K is the number of bands, M is the total number of frames, m is the frame index, $W(j,m)$ is the weight given to the jth frequency band, $|X(j,m)|$ is the weighted (by a Gaussian-shaped window) clean signal spectrum in the jth frequency band at the mth frame, and $|\hat{X}(j,m)|$ is the weighted enhanced signal spectrum in the same band.

Speech-to-reverberation modulation energy ratio (SRMR): It can be calculated only from target signals [Falk et al., 2010].

$$SRMR = \frac{\sum\limits_{k=1}^{4} \bar{\varepsilon}_k}{\sum\limits_{k=5}^{K*} \bar{\varepsilon}_k}, \tag{3.10}$$

where k is the modulation band, $\bar{\varepsilon}_k$ is the average per-modulation band energy, and the upper summation bound $K*$ in the denominator is adapted to the speech signal under test.

Perceptual Evaluation of Speech Quality (PESQ): The enhanced speech signals can also be evaluated in terms of the PESQ scores. PESQ requires reference signals. It is computed according to the ITU-T Recommendation P.862 [Recommendation, 2001]. In order to compute the PESQ score, a sequence of processing step is applied. The original clean and degraded signals are first level-equalized, to a standard listening level, and filtered by a filter with response similar to that of a standard telephone handset. Then, the signals are time aligned and processed by an auditory transform to obtain the loudness spectra. Finally, the difference in loudness between the target and processed signals is computed and averaged over time and frequency. The PESQ produces a score that ranges between 1.0 and 4.5; where higher values indicate better quality. High correlations with subjective listening tests were reported by [Rix et al., 2001],and [Hu and Loizou, 2008].

Chapter 4

Exploiting Cepstral Distance for Channel Selection

> *Humans are allergic to change.*
> *They love to say, 'We've always done it this way'.*
>
> Grace Hopper

In this chapter, a novel framework for performing channel selection is proposed and studied. The basis of the proposed method is the use of objective signal quality measures, introduced in the previous chapter. First, we present the elements of the proposed framework: i) a new schema for categorizing CS methods, ii) a CS method exploiting cepstral distance, and iii) new evaluation metrics which are useful for a deeper understanding of CS. Then, an extensive analysis of the performance of cepstral distance based CS is provided. Finally, a set of CS experiments are performed, using synthetic and real datasets. The results obtained point out the validity of the proposed method and its effect on a DSR task.

4.1 Proposed CS framework

4.1.1 CS methods categorization

In [Guerrero et al., 2016], we proposed the categorization of signal-based CS methods into two groups (i) *informed* and (ii) *blind* methods. We call *Informed methods* those approches which assume the availability of prior knowledge or reference information. Such solutions have been explored as the upper-bound mark of a CS measure [Wolf and Nadeu, 2010]. In [Wolf and Nadeu, 2009], measured IRs were used to verify the assumption that ASR can be benefited from IR based CS. In [Wolf and Nadeu, 2010], the SNR and the position/orientation of the speaker are used as further informed CS measures. CS decision mechanisms used in an informed fashion allow a deeper analysis of their relation to the properties of a phenomenon, e.g., reverberation. Although most of the CS measures described in the literature can be easily modified to be used in an informed/blind way, very few authors have performed such an intermediate analysis in the CS literature.

As an example, it is worth reviewing EV measure [Wolf and Nadeu, 2014] (Equation 3.4) which has been exploited only in an uninformed, blind fashion. An extension of this procedure to an informed version, resorting to the variance of the envelope of the clean signal $V_x(k)$, is possible as follows. The informed EV-based CS is then based on the amount of reduction of the dynamics between the clean signal and the reverberated one:

$$\hat{M}_x = \arg\min_m \sum_k \frac{|V_m(k) - V_x(k)|}{\max\limits_m |V_m(k) - V_x(k)|}. \tag{4.1}$$

Blind methods for CS use scores computed strictly from the acquired signals. Such methods share the objective to detect the least distorted channel(s) among the available ones. Blind CS measures include the use of energy and SNR [Obuchi, 2004, 2006], 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].

When reviewing objective measures found in the literature, it can be observed that many of these algorithms rely on the availability of clean signals. The creation of a reference is studied in this work, in order to compute such objective scores without prior knowledge.

4.1.2 Cepstral-Distance based CS

Perhaps the most intuitive objective measure for signal quality estimation, that applies well in cases of reverberation, is the CD. Here, we study the use of CD for channel selection, in an informed and a blind fashion.

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

$$x_m(t) = x(t) * h_m(t) \tag{4.2}$$

be the signal captured by microphone m, where $h_m(t)$ is the related impulse response (IR). Here, $x_m(t)$ is not distorted by environmental noise.

From the set of CDs between the close-talk and all the available channels computed as in Equation 3.7, where $\boldsymbol{c}_x$ and $\boldsymbol{c}_m$ are the cepstral coefficient vectors of the close-talk and a distorted signal respectively, the least distorted signal, based on the distance computed in Equation 3.7, can be selected as follows:

$$\hat{M}_x = \underset{m}{\operatorname{argmin}} \quad d(\boldsymbol{c}_x, \boldsymbol{c}_m). \tag{4.3}$$

Blind CD-based CS In a real scenario, the close-talk signal is not available. Therefore, we propose a non-intrusive method for cepstral-based channel selection, which exploits a multi-microphone distant speech recognition scenario for the estimation of a reference. When the speaker is oriented towards one of the many available distributed microphones, and/or is located at a distance lower than the critical distance [Kuttruff, 2007], it is observed that for the corresponding signal, the direct component is generally stronger than the reverberated part. Other channels, whose energy is attenuated by the head of the speaker and other possible propagation effects, are expected to be more affected by reverberation. Based on this observation, we can average in the log-magnitude spectrum domain as follows:

$$\hat{R}(t,\omega) = \frac{1}{M}\sum_m \log|X_m(t,\omega)| \quad , \tag{4.4}$$

where m is the microphone index, M is the total number of microphones, and $X_m(t,\omega)$ is the short-time Fourier transform (STFT) of the signal captured by microphone m. This represents the corresponding geometric mean spectrum [Rabiner and Juang, 1993] and, using Equation 4.2 this can be rewritten into:

$$\hat{R}(t,\omega) = \log|X(t,\omega)| + \frac{1}{M}\sum_m \log|H_m(t,\omega)| \quad , \tag{4.5}$$

where $X(t,\omega)$ and $H_m(t,\omega)$ are the STFT of the clean signal and m-th IR respectively. In Equation 4.5, the first term is the log-magnitude spectrum of the close-talk signal, and the second term represents an estimation of the average reverberation of the room, based on the available microphone channels. Let us assume that one microphone signal is better than the others in terms of direct to reverberant ratio. The basic assumption is that such a signal will be characterized by a larger distance from the resulting geometric mean spectrum. Therefore, from the set of CDs between the geometric mean spectrum,

$\hat{R}(t, \omega)$, and all the available channels, the least distorted one can be selected as follows:

$$\hat{M}_{\hat{R}} = \arg\max_{m} d(\boldsymbol{c}_{\hat{R}}, \boldsymbol{c}_m) \quad , \tag{4.6}$$

where $\boldsymbol{c}_{\hat{R}}$ and $\boldsymbol{c}_m$ are the cepstral coefficient vectors of the reference and of the microphone m, respectively.

So far, it was assumed that the only source of distortion of the M acquired signals was the reverberation. However, in a real setting this is rarely true, as environmental noise also exists. Given the purpose of this study, we assume that the reverberation effect dominates over the background noise that affects all the microphones.

4.1.3 A new evaluation methodology

CS approaches are traditionally evaluated by means of recognition results using models trained on clean speech [Wolf and Nadeu, 2014; Himawan et al., 2015] and expressed in terms of the recognition error rate. The use of clean acoustic models in CS is justified on the fact that the goal of CS is to find the channel that achieves a recognition performance as close as possible to the ideal clean signal. On the other hand, the assessment of CS is constrained to an indirect result, averaged over the output of another process, i.e., decoding. Such an experimental setup however introduces certain limitations.

First, in complex tasks, as the ones addressed by this work, recognition of distant reverberant speech using clean acoustic models results in a significant performance loss [Barker et al., 2015] compared to training on real and simulated distorted speech. Evaluating a CS method in model mismatching cases, with very inadequate recognition performance levels, makes it hard to individuate its possible advantages. Second, even when clean acoustic models are used, it must be reminded that various mechanisms that aim at the improvement of the single distant microphone recognition accuracy are commonly exploited within the decoding process. The application of these mechanisms often results in more distorted channels having a recognition error rate as low as, or in certain cases lower than, the least distorted available channel. In such a case, the effect of a CS method targeting the identification of the least distorted signal is lost. Also, the use of the recognition error rate does not expose the real capacities of a CS solution. For example, this evaluation approach provides no information about the conditions that are more challenging for a CS method.

In order to obtain a CS evaluation methodology, unrestrained from the above limitations, we propose the use of two evaluation measures, in addition to the recognition error rate: (i) the matching rate to an informed CS, and (ii) the average normalized CD between the selected channel and its close-talk reference.

The **Informed CS Matching (ICSM)** rate is a meaningful evaluation measure in studying how often a certain method succeeds in selecting the least distorted channel. It is defined as

$$ICSM = \frac{\text{\# of matching selections}}{N} \quad , \tag{4.7}$$

where N is the total number of utterances in the test set. In principle, any objective measure can be used to create the matching ground truth; in this study, we use the informed CS based on CD. It is worth noting that such a matching measure can not be computed using recognition performance because, as previously discussed, more than one channel may achieve the same minimum recognition error rate.

The **Average Normalized CD (ANCD)** is computed as follows:

$$ANCD = \frac{1}{N}\sum_{n} \tilde{d}_n(\boldsymbol{c}_x, \boldsymbol{c}_{\hat{M}}) \quad , \tag{4.8}$$

where $\boldsymbol{c}_x$ and $\boldsymbol{c}_{\hat{M}}$ are the cepstral coefficient vectors of the close-talk signal, x, and the channel, $\hat{M}$, selected by a certain CS method respectively. Here, $\tilde{d}_n(\boldsymbol{c}_x, \boldsymbol{c}_{\hat{M}})$ refers to the cepstral distance, normalized over the maximum CD from all signals, for the utterance n. Therefore, this measure will be bounded within the ANCD of the informed CS method and 1:

$$ANCD \in [\min_m(\tilde{d}_n(\boldsymbol{c}_x, \boldsymbol{c}_m)), 1] \quad . \tag{4.9}$$

4.2 Analysis of CD-based CS

4.2.1 Experimental setting

Multi-microphone room In this section we consider multi-microphone simulated scenarios, which feature talkers speaking in reverberant environments. Noise distortion is not addressed here. In the explored settings, the average distance between the speaker and the microphones is approximately 2 meters. In contrast to other CS studies, performed in reduced spaces or constraining the location of the speaker, our conditions imply a scenario where signals are significantly affected by reverberation.

The synthetically created environment is a square room (4.80m by 4.80m), Figure 4.1. The simulations use IRs generated with an implementation of the image method

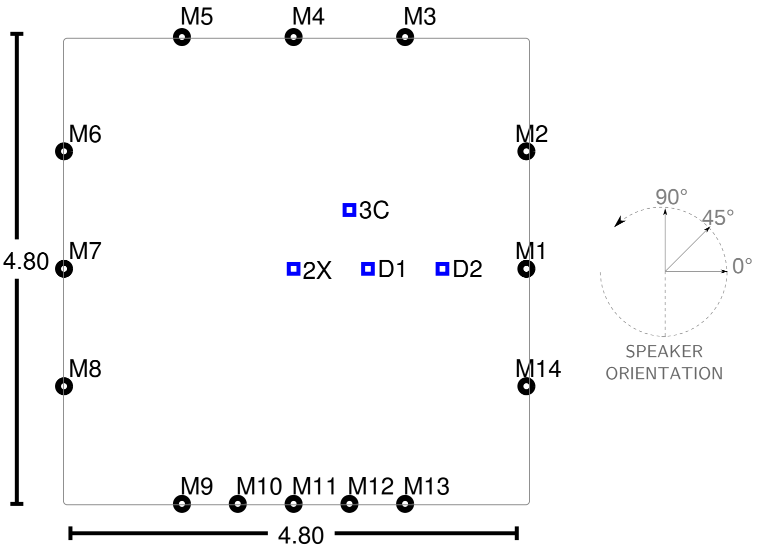


Figure 4.1: Diagram of the ideal square room setting. Room dimensions are presented in meters. Black circles indicate the location of microphones, and blue squares show the various locations of the speaker. On the right, an angle equivalence of the speaker orientation is shown.

(IM) [Allen and Berkley, 1979; Peterson, 1986; Brutti et al., 2013]. We resort to such a synthetic room because it facilitates the study of multiple acoustic phenomena under controlled parameters. The most relevant benefits of using this scenario are the ability to study CS under different reverberation characteristics of the room, exploring multiple speaker orientations, and with varied geometries of the microphone network.

The square-room includes a set of 14 microphones located on the walls, Figure 4.1. These microphones are used under various microphone network configurations, which will be specified at the different experiments performed in the following section. A total of 3 positions were used to study the effect of the speaker location. At each position, the speaker orientation was varied along the 360 degrees.

Speech recognition For ASR experiments, each of the signals captured by the microphones is decoded with a recognizer implemented with the Kaldi speech recognition toolkit [Povey et al., 2011]. The language and lexicon models are built according to the s5 recipe included in the Kaldi WSJ configuration. The recognition is based on Karel's recipe [Veselỳ et al., 2013], on top of MFCC features transformed with Linear Discriminant Analysis (LDA), Maximum Likelihood Linear Transform (MLLT), and feature space Maximum Likelihood Linear Regression (fMLLR) -a technique widely used for speaker adaptation.

The recognition task is designed around the Wall Street Journal corpus. The training data consists of 7138 simulated reverberant utterances, derived from the full clean Wall Street Journal (WSJ0-5k) [Garofalo et al., 1993] training set. All IRs used to create this set consider only channels in which the speaker position/orientation (POSORIs) is direct towards a microphone. For the test set simulations, clean material is extracted from the WSJ0-5k sub-set of the DIRHA-English [Ravanelli et al., 2015] corpus. For this analysis, both training and test sets are simulated using IM IRs.

ASR evaluation The performance of the recognition system is measured in terms of word error rate (WER). In order to compute the WER, the output of the recognizer is aligned with the reference transcription, and the errors are counted. The performance of an ASR system is measured in terms of three types of errors [Bahl and Jelinek, 1975]:

- Deletions: Words missing in the resulting hypothesis.
- Insertions: Additional words which appear in the resulting hypothesis.
- Substitutions: Words that were misrecognized or confused with others.

The overall error of the system, WER, is calculated as:

$$WER(W_n) = \frac{L(W_n, R_n)}{|R_n|}, \tag{4.10}$$

where W_n and R_n are the transcribed and reference sentences, respectively, $L(W_n, R_n)$ is the Levenshtein edit distance [Levenshtein, 1966] between the decoded sentence and the reference, and $|R_n|$ is the number of words (symbols) in the sentence R_n.

4.2.2 Interactions between objective measures and DSR

Reverberation characterization Objective scores present different discrimination power to reverberation characteristics. One reverberation descriptor is the reverberation time. The reverberant sound in an enclosure decays along time, as the energy of the sound is absorbed by the interactions with the surface of objects or walls of the room. A highly reflective room is characterized by sounds taking longer time for the sound to fade out. On the contrary, a very absorbent room will cause sounds to fade out rapidly. An objective measure of this time is defined as the time for the sound to be reduced to a level 60 decibels below its original energy level, giving it the name of *T60*.

Experimentally, it was observed the change of the distance of a reverberated signal to its close-talk version, measured with three objective measures, under varied *T60* values, Figure 4.2. For this purpose, a subset (i.e., 120 utterances) of the test clean material was contaminated using IRs generated with the IM tool. The results presented here depict only the case for which the speaker was located at position D1, illustrated in Figure 4.5, and oriented towards microphone M1. Figure 4.2 shows the discrimination power of CD, LLR and fwSNRseg to T60. A normalization was applied over the different measures for visualization purposes, without altering its original trend. For different reverberation conditions, different scores show different competences. It is observed here that fwSNRseg discriminatory power for low T60s is not as meaningful as CD or LLR, and that, along the curve, CD and LLR expose similar behavior.

Recognition The relation between objective scoring of a signal and its recognition performance is not straightforward to analyze. There are multiple components inside a recognition engine which may affect the recognition results and guide to misleading conclusions. Some of the studies conducted in CS assume the use of clean material for training the acoustic models. In acoustic conditions such as those explored in this study, such approach would result in very high WERs and making unclear the benefits of a selection procedure. Moreover, the use of contaminated or real reverberated material for training acoustic models has shown to be more positive for DSR tasks, with respect to training on clean speech only [Barker et al., 2015].

Assuming the use of reverberant material for training acoustic models, recognition was performed over multiple channels. For each channel, its corresponding distance to the close-talk signal was computed using CD, LLR and fwSNRseg. Figure 4.3 presents the objective scores as a function of the WER obtained by the recognition process. A clear correlation is observed between the degree of reverberation/distortion determined by the objective measure of a signal and its corresponding recognition results. Moreover, these results confirm the capability of objective measures to characterize the acoustic conditions of the room, which justifies the similarities to the curves observed in Figure 4.2.

4.2.3 CD-based CS and the recognition oracle

An oracle-based CS is an ideal unrealistic approach. This method selects, among all the evaluated channels, the one that achieved the lowest WER. It must be noticed, that this selection is not possible in a real scenario, since the recognition performance of a channel

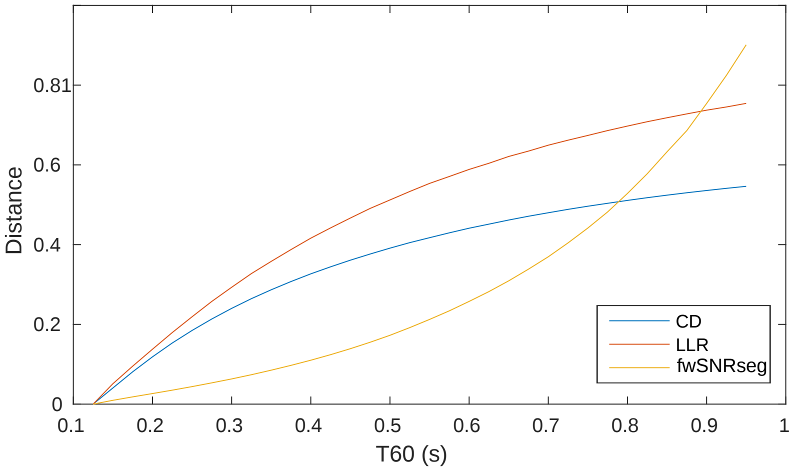


Figure 4.2: Objective distances between reverberant and close-talk signals, computed at different reverberation times (T60s). For visualization purposes, distances have been normalized.

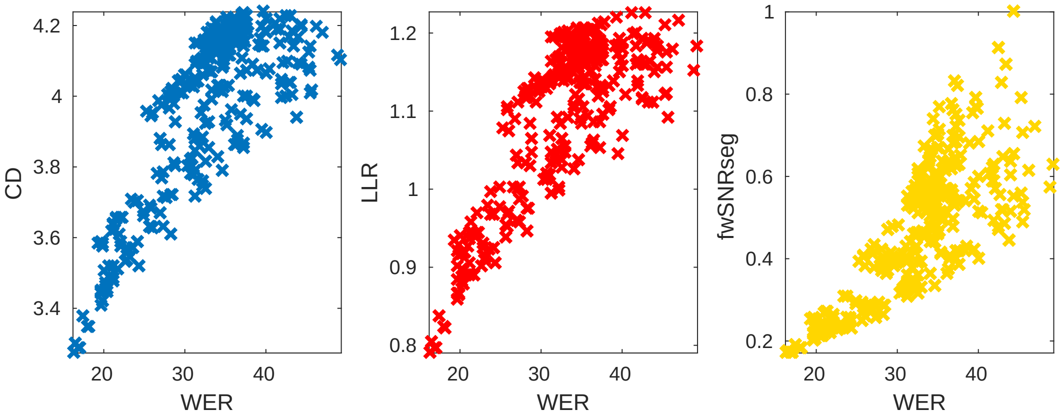


Figure 4.3: Objective distances between reverberant and close-talk signals, as a function of the WER obtained by decoding reverberant signals. For visualization purposes, fwSNRseg curve has been normalized.

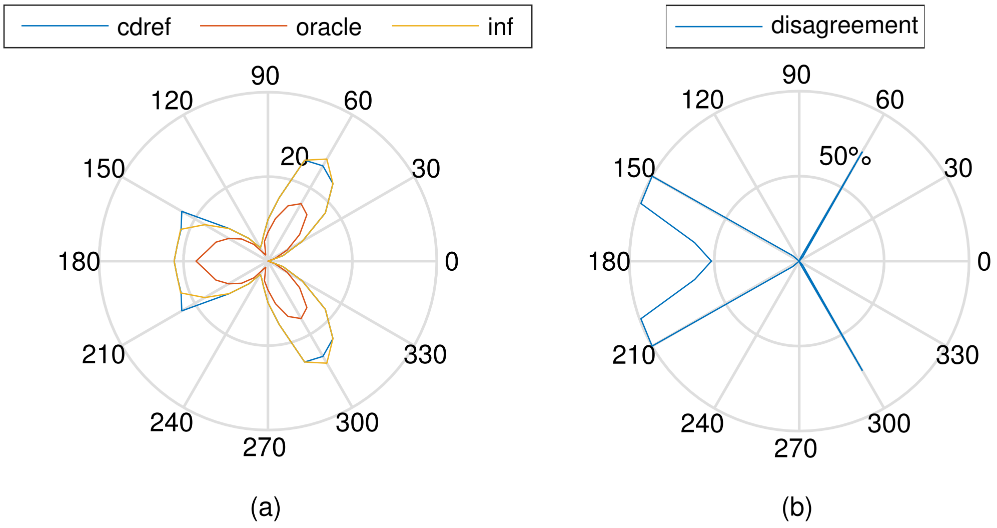


Figure 4.4: Polar representations of (a) the average WER for different CS methods and (b) the total disagreement of the proposed blind CS method and the informed one. Plot (a) reports 3 CS methods: oracle, informed CD-based (inf) and blind CD-based (CDref). The points around the polar plots represent orientation degrees, e.g., speaker oriented at 30 degrees. The scale inside the plot (a) corresponds to WER, while on the plot (b) it shows a percentage of disagreement.

can not be known in advance. It is, anyway, useful to set an upperbound for CS in terms of recognition performance.

For the study presented in this section, the speaker is located at position D1 in the room, adopting a varied set of orientations (0 to 350 degrees). Polar plots are used to depict the results obtained at the different speaker orientations. Figure 4.4 presents two parts. The plot on the left (a) shows the average WER obtained by decoding the channel selected by 3 CS methods: oracle, blind CD based CS (CDref), and informed CD based CS (inf). The lobes in the graph indicate the occurrence of the errors. It can be observed how the different CS methods follow the same error-trend.

The plot on the right *(b)* illustrates a different score, the disagreement between the CS achieved with the blind CD-based method (i.e., distance to the generated reference) and the informed CD-based method (i.e., distance to the close-talk signal). First, it is worth noting the strong agreement between both polar plots which constitutes an important finding. This fact verifies that the conclusions derived from the signal-based analysis can be extended to speech recognition targeted experiments.

Moreover, these results allow us to identify and investigate the recognition performance at the successful or highly error-prone orientations. The region between orientations 300 degrees and 30 degrees clearly show the effect of a good orientation and distance from speaker to microphone. On the other hand, at the specific orientations of 60, 180, and 300 degrees a deeper examination is needed. When the speaker is oriented at 60 and 300 degrees, he is facing two corners of the room, while when he is oriented at 180 degrees, he is facing the most distant microphone (M7). These are already intuitive reasons to understand the results obtained. Looking at the individual microphone recognition results for these problematic orientations, it is observed that the microphones obtain similar recognition results. Therefore, in these cases, a selection among them is not meaningful. If we take a closer look at the CD scores obtained by these microphones, we also identify other irregularities. At orientations 60,180, and 300 degrees, we observe more than one microphone achieving similar CD scores. For these cases, a signal level CS is not meaningful. This finding strengthens the connection of the signal-based analysis to recognition. These results are important for experimental studies, since recognition experiments can be indirectly studied through signal-based CS, which is less resource and time demanding than a recognition-based analysis.

4.2.4 Influence of the physical setting on CD-based CS

Specific characteristics of the acoustic scene affect the performance of CS methods. We resort to a set of specific cases where location of the speaker and geometry arrangement of the microphones are varied. These cases can help us understand the effects of each of these variations in the CS process.

Speaker position: For this analysis we perform CD-based CS, both in an informed and blind fashion, in an ideal symmetric microphone arrangement microphones, locating the microphones at the same relative positions at each wall. Four microphones, each located at the center of each wall (i.e., M1, M4, M7, M11) are considered, see Figure 4.5.

Figures 4.6 and 4.7 report the microphones selected by CD-based CS methods, both in an informed and in a blind fashion, with the speaker located at positions 2X, D1, and D2. On the left, Figures 4.6a, 4.7a and 4.7c show the CS in an informed fashion. On the right, the blind version of CD-based CS is presented. Figures 4.6a and 4.6b feature the speaker located at the center of the room. In such a case, a very intuitive CS result is observed. Minimal differences between the informed and the blind version are reported.

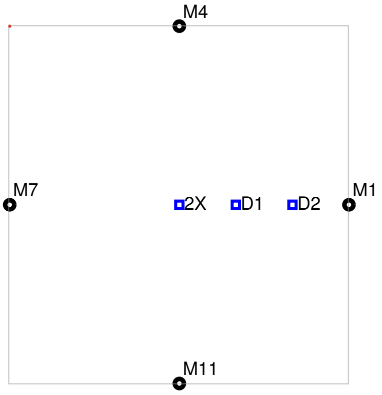
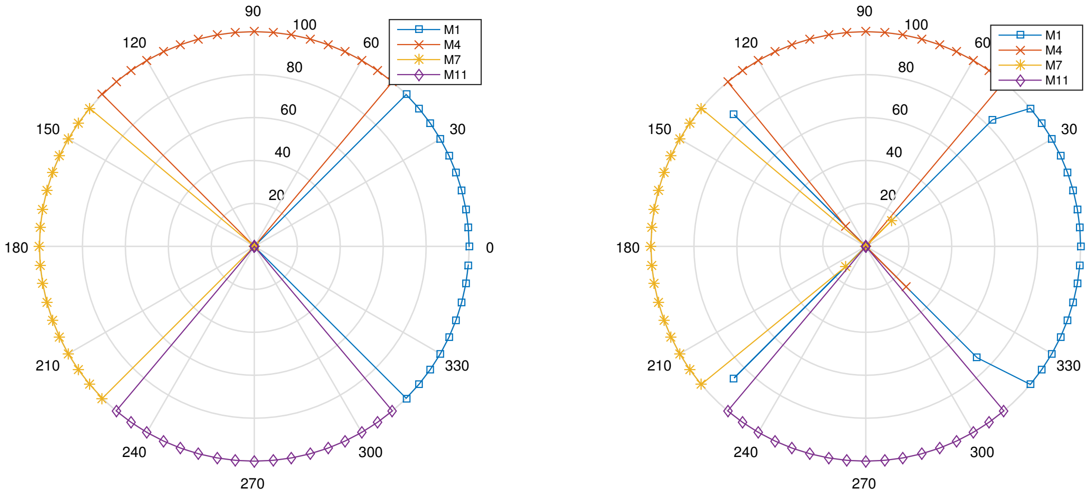


Figure 4.5: Diagram of the square room setting with symmetric 4 mics. Black circles indicate the location of microphones, and blue square show the location of the speaker.

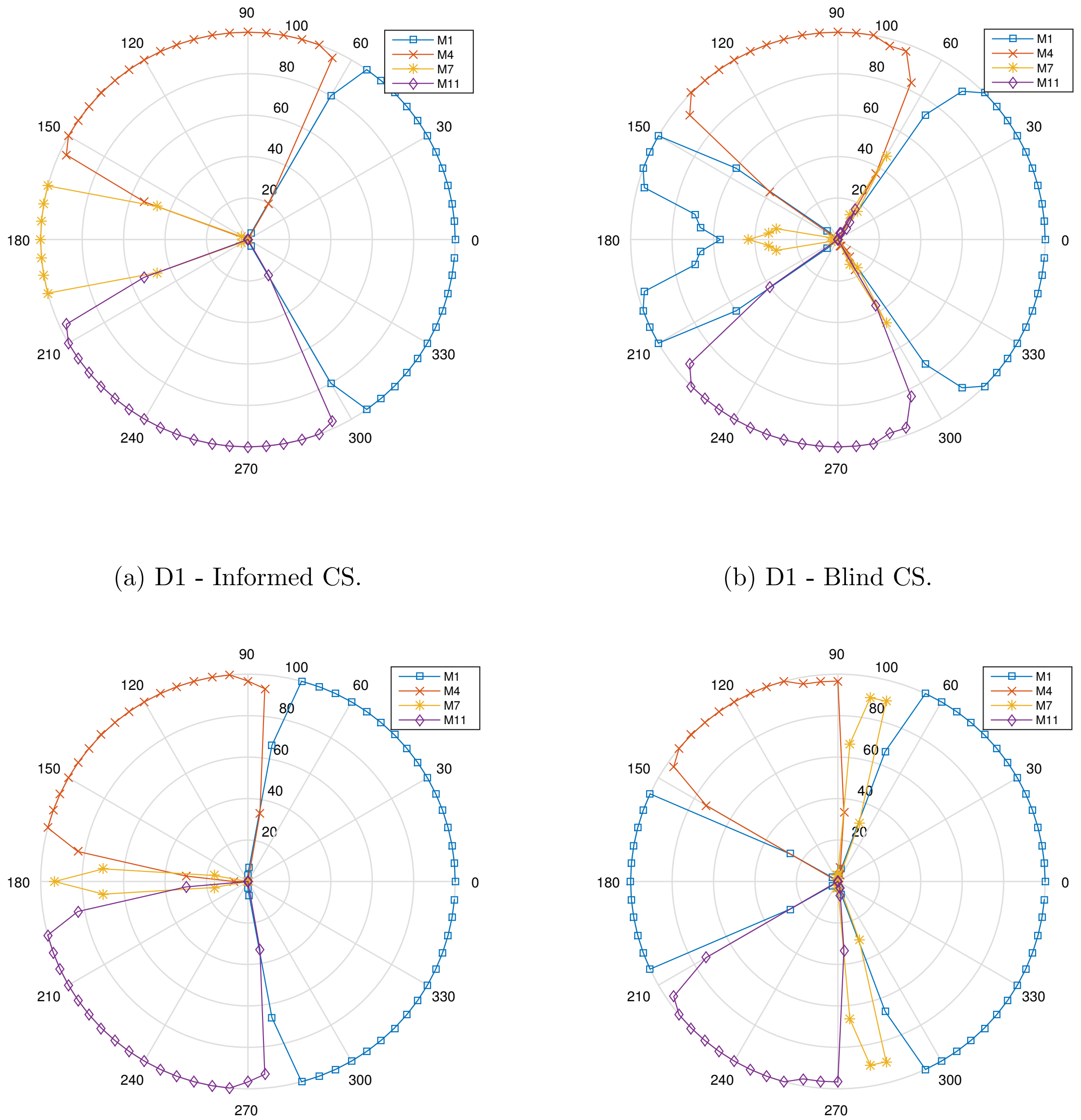


(a) 2X - Informed CS.

(b) 2X - Blind CS.

Figure 4.6: Microphones selected by CD-based CS methods in a square room equipped with 4 symmetrically set microphones. Only the case where speaker is located at the center of the room, i.e., position 2X, is reported. Subfigures are labeled with the speaker location and CD based CS method. The scale within the plots represents the number of times a microphone is selected, as a percentage. The orientation step was 5 degrees.

It must be noticed that these erroneous regions correspond to the cases where the speaker is oriented towards the corners, and the measured CDs are comparable for more than one microphone, leading to a confused selection.



(a) D1 - Informed CS.

(b) D1 - Blind CS.

(c) D2 - Informed CS.

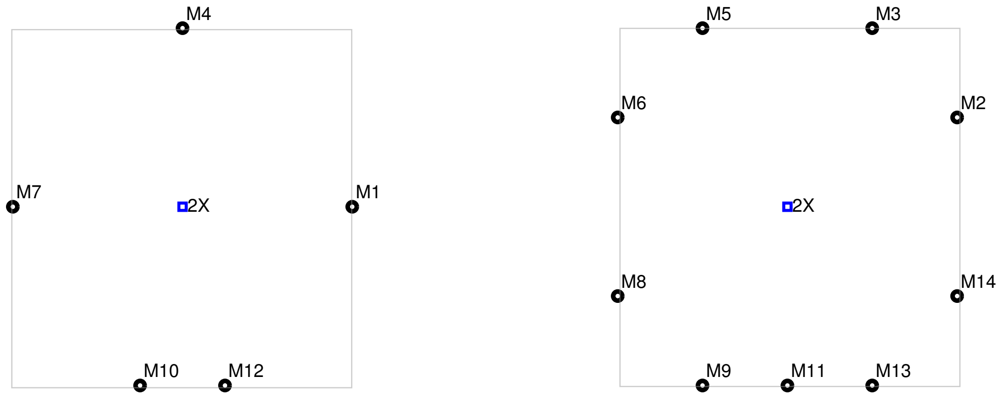
(d) D2 - Blind CS.

Figure 4.7: Microphones selected by CD-based CS methods in a square room equipped with 4 symmetrically set microphones. The cases where speaker is located at positions D1 and D2 are reported. Subfigures are labeled with the speaker location and CD based CS method. The scale within the plots represents the number of times a microphone is selected, as a percentage. The orientation step was 5 degrees.

The cases depicting the speaker located at D1 or D2 present broader disagreement regions between the informed and blind methods. It must be remarked that the errors occur at cases where the distance towards the microphones are increased and therefore a higher reverberation distortion will be observed, e.g., at angle 180. Still, for the cases where the speaker is distant from the microphone, at around 2m, the selection is successfully achieved.

Microphone network setting: The goal of these experiments is to understand the impact of the microphone network configuration into the proposed CS. This study is of interest for the CD-based CS, particularly for the blind version of the method, where a reference to compute the distance is created from the multiple signals. Two microphone configurations are studied with the speaker located in the central point of the room (2X): a) having 5 microphones (Figure 4.8a), and b) having 9 microphones (Figure 4.8b) distributed in the room. These configurations represent an unbalance over the symmetrically distributed microphones, explored in the previous section.

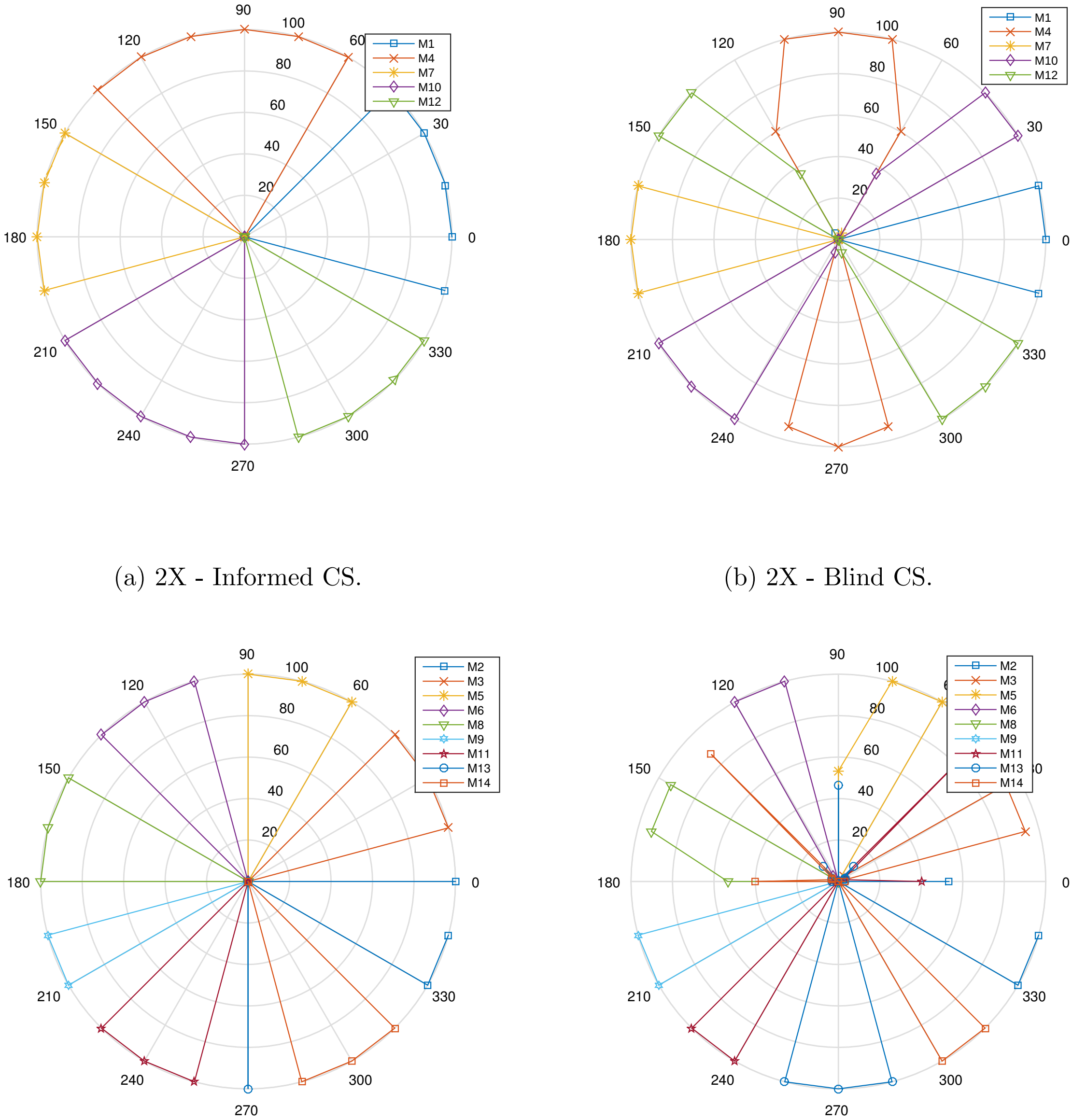
Comparing Figures 4.9a to 4.9b, the informed to the blind CD-based methods, it is evident that, in this experimental setting, the unbalance in the microphone distribution affects the performance of blind CD-based CS. The agreement between informed and blind methods is reduced due to the computation of the reference required in the blind CD-based CS method. However, a similar unbalance in the microphone network is presented, this time with 9 microphones. See the agreement between Figures 4.9c and 4.9d. According to the results of the latter experiment, the sensitivity of the proposed solution to the



(a) 5 mics.

(b) 9 mics.

Figure 4.8: Diagram of the square room with 5 and 9 mics. Black circles indicate the location of microphones, and blue square show the location of the speaker.



(a) 2X - Informed CS.

(b) 2X - Blind CS.

(c) 2X - Informed CS.

(d) 2X - Blind CS.

Figure 4.9: Microphones selected by CD-based CS methods in a square room equipped with a,b) 5 microphones, and c,d) 9 microphones. Only the case where speaker is located at the center of the room, i.e., position 2X, is reported. Sub-figures are labeled with the speaker location and CD based CS method. The scale within the plot represents the number of times a microphone is selected, as a percentage.

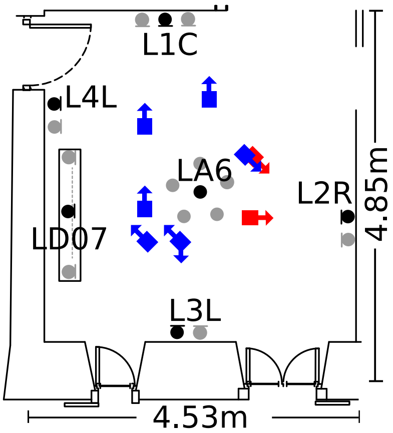


Figure 4.10: DIRHA Setting used in the CS experiments. Black circles indicate the microphones used. The arrows show position/orientation of the speakers for the Direct scenarios, in red for simulated data, and in blue for real data.

unbalance in the multi-microphone setting can be circumvented by defining a network with a wider coverage of the space.

4.3 CS Experiments

4.3.1 Experimental setup

The next experimental scenario is taken from the DIRHA Project setup, previously described in Section 2.4.2. A subset of 6 microphones are selected for this study, as shown in Figure 4.10.

The training data consists of 7138 simulated reverberant utterances, derived from the full clean WSJ0-5k [Garofalo et al., 1993] training set. This training set was simulated using recorded IRs [Cristoforetti et al., 2014], which consider only channels in which the speaker position/orientation is direct towards a microphone.

The test material is extracted from the WSJ0-5k sub-set of the DIRHA-English [Ravanelli et al., 2015] corpus, which includes data recorded in the real living room. Concerning the test sets, in order to focus the analysis on the different CS methods, two scenarios are considered:

- In the first one, the speaker POSORI is always direct in respect to one microphone, i.e., speaker is looking at one microphone. Such a setting narrows the DSR problem, allowing us to perform an intuitive analysis of the correlation between signal distortion and recognition performance. For this scenario, simulated data is generated under two specific POSORI configurations (**DirSim**). Additionally, real data is extracted based on a set of 7 different POSORIs (**DirReal**). Figure 4.10 depicts the POSORIs used for the first scenario.

- The second scenario incorporates a set of 36 mixed POSORIs, for each of the simulated and real cases (**MixSim** and **MixReal**). In this scenario, the adopted POSORIs are distributed in the room and are not only direct.

For all the simulated data, the close-talk signals were recorded in the FBK recording studio, while for the real data, these were captured by a head-set. The DirSim, MixSim and MixReal datasets are composed by 410 utterances each. DirReal dataset is composed by 82 utterances. An ideal voice activity detection is assumed to be applied over real and simulated data.

4.3.2 Channel selection methods

The following CS methods are included in the evaluation:

- **CDi** is the proposed informed CS method that uses the close-talk reference, as explained in Section 4.1.2. We adopted the standard implementation distributed by the REVERB 2014 challenge[1].

- **CDref** is the blind proposed CD method that uses the geometric mean spectrum as a reference, as described in Section 4.1.2. For both CD based methods the CD parameters were assigned the following values: $framesize = 0.025 seconds, shift = 0.01, window - type = hanning, order = 24$.

- **EV** [Wolf, 2013] is the state-of-the-art CS method, based on Envelope Variance, as introduced in Section 3.2. This algorithm uses filter-bank outputs extracted by the speech recognition system, in order to compute EVs.

- **Random** is a random selection of a channel performed at each utterance.

For completeness, the recognition performance of decoding each of the single distant microphones (**SDM**) is also presented.

[1] See http://reverb2014.dereverberation.com

4.3.3 Speech recognition

Each of the signals captured by the microphones is decoded with a recognizer implemented with the Kaldi speech recognition toolkit [Povey et al., 2011], see A.2, with the following configuration. The language and lexicon models are built according to the s5 recipe included in the Kaldi WSJ configuration. The recognition uses deep neural networks, trained according to Karel's recipe [Veselỳ et al., 2013], on top of MFCC features transformed with LDA, MLLT, and fMLLR. The network architecture is shaped by 6 hidden layers of 1024 neurons, with a context window of 11 consecutive frames (5 before and 5 after the analysis frame), and an initial learning rate of 0.008. The recognition performance on the close-talk material yields a word error rate, WER, of 3.7%.

4.3.4 Results and Discussion

In this section, we analyze the performance of the proposed CS methods using the previously described corpora and evaluation criteria. First, we present the CD of the 6 different microphones to the close-talk signals. Second, the proposed ICSM rate and ANCD measures are displayed. Third, recognition results are reported.

Table 4.1 reports the average CD between the close-talk signal and each of the SDMs used in the study. For the direct simulated case (DirSim), first it is worth noting in Figure 4.10 the speaker orientations used, indicated in red. In such case, the channel that is intuitively identified as optimal (L2R) has the lowest CD to the close-talk. In the remaining cases (DirReal, MixSim, MixReal), the same trend is not evident because of the averaging among the multiple POSORIs adopted by the speakers. However, in a per-utterance analysis, it is clear that when a direct path between the speaker and one of the microphones exists, the corresponding signal has the lowest CD among all the microphones.

Proposed CS evaluation In Table 4.2, the informed CS matching rate, ICSM, is presented for EV and CDref. The proposed blind CS significantly outperforms both EV and Random CS, which in this experimental setup, for the 6 microphones used, would achieve an ICSM rate of $1/6 \approx 16\%$. CDref achieves a relatively low ICSM rate for the MixReal case, which can be attributed to the fact that this case considers more complex situations, comprising multiple non-direct POSORIs. This type of setup comes in contrast to the original assumption of the proposed method concerning the availability of a direct

Table 4.1: Average CD of the distributed microphones.

	Direct		Mixed	
SDM	DirSim	DirReal	MixSim	MixReal
L1C	3.92	2.98	3.79	3.09
L2R	3.25	3.14	3.71	3.15
L3L	3.74	3.05	3.75	3.13
L4L	3.93	3.05	3.81	3.12
LA6	3.78	2.97	3.73	3.04
LD07	3.87	2.89	3.73	3.01

Table 4.2: Informed CS Matching Rate (ICSM) (%).

CS	DirSim	DirReal	MixSim	MixReal
EV	47.92	31.70	39.36	39.85
CDref	75.00	81.70	75.30	52.32

channel. Moreover, even for an informed CS method such schemes can not be properly addressed, since a selection among highly distorted channels is not always relevant.

In Table 4.3, the Average Normalized CD, ANCD, is presented for CDi, EV and CDref. It is recalled here that the ANCD of CDi is the upper-bound for a blind CS method. Furthermore, it can be viewed as an indication of the complexity of the conditions of each dataset. As an example, the high ANCD (0.88), for the CDi in MixReal, evidences the inclusion of more unfavorable cases than in DirReal (0.84). This confirms the previously discussed observations concerning the complexity of the MixReal dataset. The proposed blind CS method achieves an average distance closer to the one reached by the informed method, see for example for ANCD of DirReal with EV is 0.89, while with CDref is 0.86. This fact, as indicated in the ICSM rate evaluation, occurs because the two CD-based methods repeatedly select the same channel.

Speech recognition results Concerning the recognition performance, Table 4.4 reports the WER for the recognition of the SDM for each experiment. First, it is worth reminding the different dataset conditions, see Figure 4.10. For the simulated dataset which features speakers oriented towards a microphone, *DirSim*, all the simulations present a condition

Table 4.3: Average Normalized CD (ANCD) between the selected channel and its clean reference.

CS	DirSim	DirReal	MixSim	MixReal
CDi	0.82	0.84	0.85	0.88
EV	0.91	0.89	0.91	0.91
CDref	0.88	0.86	0.89	0.89

that results favorable specifically for the microphone L2R. The real dataset, which features speakers oriented to a microphone, *DirReal*, there is not one specific microphone that is favored by the speaker POSORI. In this latter case, the speakers adopted various POSORIs during the recordings. For this reason, a direct comparison between DirSim and DirReal can not be applied in terms of channel WER trends. If DirReal results are explored at a per-utterance level, it can be observed that in most cases, the microphone favored by the speaker POSORI achieves the minimum WER. Observe the column of DirSim. An interesting result concerns the low WER achieved by the intuitively best channel (L2R) in the DirSim case. This is correlated with the CD scores previously presented. However, there is no direct agreement in the channel ranking given by the objective SDM scoring and the SDM WER.

Finally, Table 4.5 presents the average recognition performance of the CS methods for each dataset. It is recalled here, that Random CS roughly corresponds to the average of the SDM WER. Note that the average CS WER achieved by EV is improved with the proposed blind method in all cases, as shown in Table 4.5, where the corresponding relative improvement (Rel. Imp.) is reported.

When observing the proposed evaluation measures in addition to the recognition accuracy, one can gain a deeper understanding of the strength of the proposed blind method above the EV based one. See for example the case of DirReal, where for both CS methods WER is reduced in comparison to SDM. However, ICSM rate of CDref is significantly closer to a perfect matching rate, a fact not evidenced from the WER. These remarks indicate the previously discussed gap in the way CS is traditionally evaluated, by means of WER, and the need for evaluation measures similar to the ones introduced in this paper.

Table 4.4: WER [%] of the distributed microphones.

SDM	DirSim	DirReal	MixSim	MixReal
L1C	16.6	14.4	16.0	14.8
L2R	10.8	19.2	15.8	16.2
L3L	13.6	15.8	16.5	15.2
L4L	15.0	16.3	17.0	15.1
LA6	16.5	15.1	17.7	14.9
LD07	14.8	14.2	16.4	14.7
Avg	14.5	15.8	16.6	15.2

Table 4.5: WER [%] by various CS methods.

CS	DirSim	DirReal	MixSim	MixReal
CDi	10.8	12.0	12.8	12.6
EV	12.7	14.7	14.6	13.9
CDref	12.1	12.5	14.1	13.7
Random	14.5	15.9	16.8	15.3
Rel. Imp.	4%	14%	3%	1%

4.4 Conclusions

In this chapter, we presented a CS framework which exploits CD, both as a channel scoring function and as a means of detailed evaluation. A detailed description of the elements of the proposed solution is presented. Moreover, the extension of the method to other objective measures is introduced.

An analysis performed on the relations between the objective measures and reverberation time, and also recognition, under multiple scenarios and conditions, expose not only the limitations but also the solid benefits of the proposed CS method. Concerning the reverberation time analysis, a similar examination can be extended to different reverberation characteristics, as for example direct-to-reverberant ratio, which is useful for various research lines other than CS.

Finally, through a series of experimental cases, we have proved that the proposed blind

CS method (i) improves in all cases the average SDM WER, (ii) consistently outperforms the state-of-the-art EV-based CS method and (iii) successfully selects the least distorted channel when sufficient room coverage is provided by the microphone network. Furthermore, it is illustrated how the standard evaluation of CS, based solely on WER, hides the strengths and weaknesses of different methods. So far, we have considered reverberation to be the main source of degradation of distant speech, however, in a real scenario, environmental noise significantly affects the captured signals.

A future research line considers an extension of this work to scenarios which include different types of noise, and with different SNR. In addition, it is interesting to study the use of other objective speech processing measures for CS, as the ones also reported here, both in an informed and blind fashion. Another open topic derived from this study, concerns finding more effective solutions when facing complex conditions, that involve unfavorable speaker positions and/or orientations. A possible direction towards this goal is to detect these cases and replace the CS, given by existing blind methods, with novel techniques.

Chapter 5

Hypothesis Combination

Science doesn't always go forwards. It's a bit like doing a Rubik's cube. You sometimes have to make more of a mess with a Rubik's cube before you can get it to go right.

Jocelyn Bell Burnell

The scenario targeted in this dissertation is characterized by multiple distant microphones. In order to extract a recognition hypothesis in a multi-channel context, two aspects are considered critical: the recognition of a speech signal, and the processing of multiple inputs. This chapter evolves around the paradigm of combining information produced by each recognition process in order to obtain an overall improved ASR performance. Starting from a general view of ASR, details about the representation of the hypothesis space are presented. The mechanisms by which the recognition hypothesis is extracted by the decoder are then described. Later, we give an overview of ASR related confidence measures. Finally, hypothesis combination approaches are detailed at the end of this chapter.

5.1 Statistical speech recognition

ASR concerns the problem of finding the most likely string of words $\hat{W}$ given a sequence of acoustic observations A, which characterize the received speech waveform. A source-channel mathematical model is often used to formulate the ASR problem [Huang et al., 2001; Jelinek, 1997], and is described as follows. In a two person interaction, a speaker and a listener, the speaker starts the interaction deciding the source word sequence W to be delivered. The source is passed through a noisy communication channel shaped by

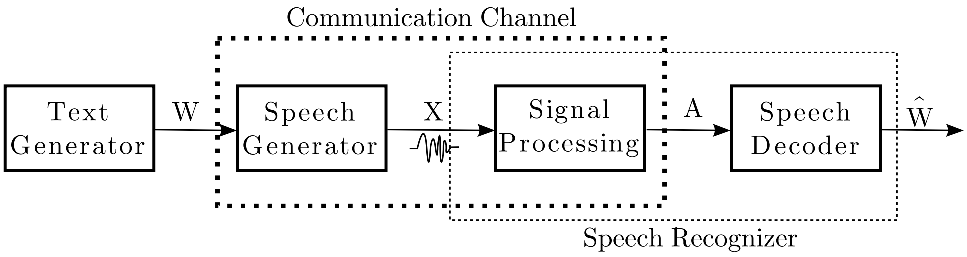


Figure 5.1: Source-channel model for a speech recognition system.

the vocal apparatus of the speaker that produces the speech waveform, and the auditory signal processing component of the listener (speech recognizer). At the end of the chain, the goal of the speech decoder is to convert the acoustic signal A into a word sequence $\hat{W}$. The objective of the transmission is to recognize a sequence of words as close as possible to the original one, this is to achieve $P(\hat{W} = W)$ asymptotically close to 1.

As previously introduced, initial ASR efforts focused on pattern recognition. Those approaches were based on templates and the temporal alignment of the patterns as a function of spectral distance [Myers et al., 1980; Rabiner and Juang, 1993]. Later, a statistical framework, which until recently constituted the standard approach, was developed. The formulation behind that statistical approach describes a method to extract the most likely sequence of words W given the acoustic observation A, which will be described in the following section. The basis of such framework are the HMMs. The HMMs present a clear advantage over previous methods because of the structure of the network, which allows modeling the speech and the language within a single statistical framework.

From the previously depicted model, two components of the statistical ASR system are identified: the signal processing component and the decoder. At signal processing level, the speech signal is transformed into a discrete sequence of feature vectors in a process called *feature extraction*. The decoder exploits a) acoustic models (AM), which concentrate knowledge about the acoustic representation of the message, the speaker and the environment, and b) language models (LM), which tell the decoder which words/units are likely to occur and in what sequence. These components were introduced in Section 2.1. A coordinated operation of these components produces as output a recognition hypothesis. This latter step, generally based on search algorithms (e.g., Viterbi beam search), often resorts to pruning techniques over the hypothesis space, keeping only the more likely relevant information for posterior processes, as detailed in Section 5.1.2.

5.1.1 Hypothesis space

It is important for many post-processing systems, such as machine translation or dialogue applications, to receive as input a set of most likely outputs from the recognition system, instead of a single output. Some spoken language understanding applications, for example, use a cascaded approach where multiple hypotheses are part of the input, which are jointly processed with additional information sources. Techniques such as hypothesis re-scoring, use a set of hypotheses in an intermediate processing unit before providing the final best hypothesis. Even ASR systems can resort to intermediate data representations, or hypothesis spaces, before providing a single final hypothesis.

When the final hypothesis is extracted through intermediate processing stages, it is said that the recognition system includes multi-pass decoding. Under such a schema, commonly, in the first pass, efficient information knowledge and methods are applied. While in the subsequent passes, more sophisticated resources are exploited. This section reports some of the representations that the hypothesis spaces can adopt.

N-best List

We call an N-best List to a list which includes the N most likely hypotheses, commonly ordered by likelihood. The acoustic likelihood and language model probability associated with each hypothesis can also be included. The list is produced from the hypothesis search algorithm, which must apply a mechanism for pruning out the not so likely options [Young, 1984]. This representation provides the most simple and straightforward way for keeping a subset of hypotheses, however some limitations are associated to it [Jurafsky and Martin, 2000]. One problem is the lack of variability or complementarity of the hypotheses, which is normally expected by a post-decoding process in order to extract the final hypothesis. Additionally, when the number of hypotheses to list is large, the complexity of keeping relevant information in the search, make the algorithm inefficient [Schwartz and Chow, 1990].

Lattice

Word graphs or lattices are one of the most popular representations to approximate the hypothesis spaces [Oerder and Ney, 1993; Aubert and Ney, 1995]. They provide a representation of a subset of the search space, richer than the N-best list, which can be used for a posterior rescoring or processing for the extraction of a recognition hypothesis. For-

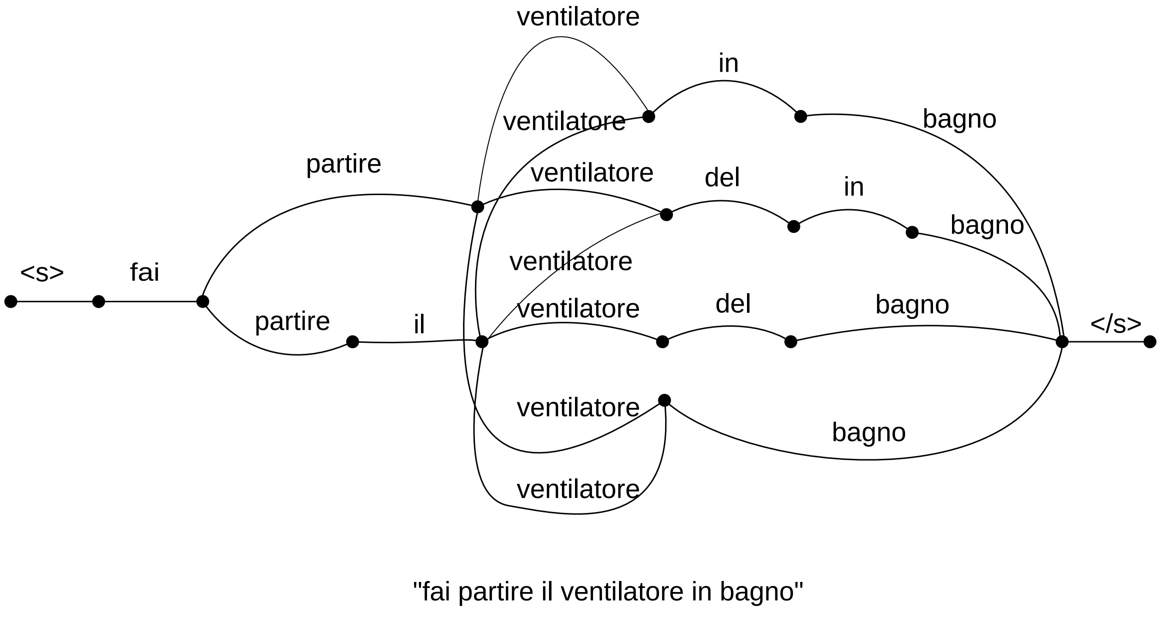


Figure 5.2: Example of a lattice or word graph. In this example, no temporal information or acoustic/linguistic scores are included. The nodes labeled $< s >$ and $< /s >$ refer to the start/end of the utterance, respectively. The spoken utterance is indicated below the graph.

mally, a lattice G is a directed, acyclic, weighted graph whose *nodes* correspond to discrete points in time. Typically, each of the *links* connecting the nodes carries information about the hypothesized word w, and the starting s and ending e nodes. Moreover, these links are characterized by scores, such as acoustic, pronunciation or linguistic scores. During the recognition stage, the same word can be associated to different starting/ending times. For this reason, multiple links can be characterized with the same word, but with slightly different time points. As previously mentioned, some pruning methods are required for producing manageable lattices, implementing, for example, the merge of similar occurrences of a word.

These word structures are also useful to estimate confidence measures [Wessel et al., 2001], as will be specified in Section 5.2.1. The internal architecture of the lattices is strongly influenced by the language/grammar component. A sentence hypothesis is extracted from a path f_1^J that traverses the lattice, from the initial node 1 to the final node J. An example of this graph can be seen in Figure 5.2.

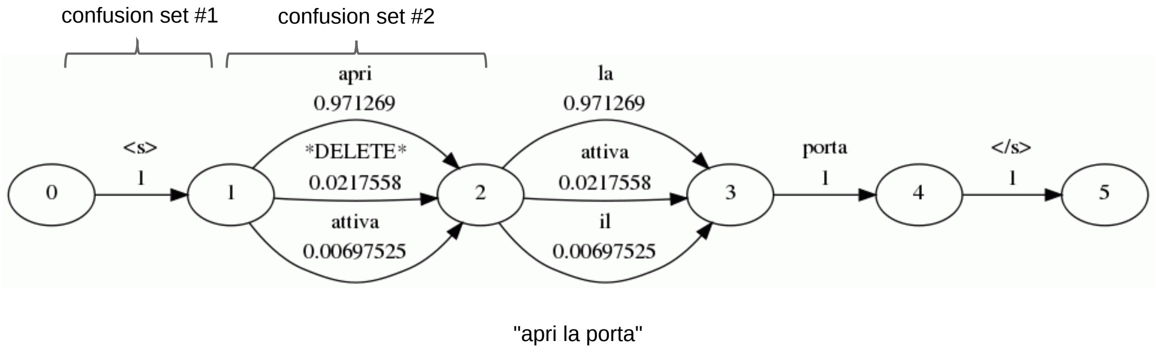


Figure 5.3: Example of a confusion network. A total of 5 confusion sets are observed. The links labeled $< s >$ and $< /s >$ refer to the start/end of the utterance, respectively. The spoken utterance is indicated below the graph.

Confusion Network/ Consensus Decoding

These networks, proposed in [Mangu et al., 2000], produce a relatively small hypothesis space in comparison to lattices, Figure 5.3. In the same manner as lattices, a CN is a directed, linear graph, defined by a set of nodes. The CN follows a set of particular properties: a) the general network is formed by a sequence of confusion sets or bins, bounded by an starting and an ending node, b) each confusion set is composed by one or more word candidates (or a NULL symbol), c) each candidate in a confusion set has a posterior probability, d) the sum of the posterior probabilities of the candidates in a confusion set is equal to 1, e) the best hypothesis of the CN is extracted by selecting the word with the highest posterior probability at each confusion set.

In the standard CN extraction procedure [Stolcke, 2002], more closely related to [Hakkani-Tür et al., 2006], the best path in the lattice is selected as the basis frame for the final CN. Then, an iterative alignment method optimizes the decision of assigning a word to a confusion set, or inserting it in a new one inside the final network. These new representations may, in some cases, modify the original hypothesis space. In contrast to lattices, CN nodes are not associated to specific time instants, since the CNs are created from the alignment of words that do not occur in the same starting/ending points.

In the CN example, Figure 5.3, we can identify 5 confusion sets, defined by a starting and an ending node, e.g., confusion set I starts at node 0 and ends at node 1 and is characterized by the symbol $< s >$ that represents the beginning of the sentence. In this sample case, the hypothesis corresponds to the path in the CN with the highest posterior probability, but this is not always true in the real application.

5.1.2 Searching the recognition hypothesis

ASR systems count on a search module in order to provide the best possible recognition hypothesis. The role of this module is to make an exhaustive search of the sequence of words whose corresponding model sequence is closest to the observed sequence of acoustic features. This search exploits different acoustic and linguistic knowledge sources, e.g., grammar, pronunciation, context dependency decision trees. Given that no information is known in advance about the number of words emitted, nor their segmentation, this is a complex task. It can be efficiently implemented through a series of constraints or assumptions.

Maximum A Posteriori Decoding The goal of Maximum A Posteriori (MAP) approach for speech recognition [Bahl et al., 1983] is to identify the sequence of words that maximizes the posterior probability $P(W|A)$ of the sequence W given a set of acoustic observations A,

$$\hat{W} = \arg\max_{W} P(W|A). \tag{5.1}$$

Through the Bayes formula of probability theory, the Equation 5.1 can be written as:

$$\hat{W} = \arg\max_{W} \frac{P(A|W)P(W)}{P(A)}, \tag{5.2}$$

where $P(A|W)$ corresponds to the acoustic model, $P(W)$ to the language model, and $P(A)$ is the probability of an observation sequence, which can be ignored since it is fixed for all the sequences. Therefore, the recognition problem is reformulated as:

$$\hat{W} = \arg\max_{W} P(A|W)P(W). \tag{5.3}$$

The recognized word sequence depends on the contributions both of the acoustic and language models. In practice, the acoustic model likelihood is not normalized, which would result in a disproportionate influence in comparison to that of the language model. For this reason, the language model is often scaled by means of an empirically determined constant, also called the language model weight, and with a word insertion penalty.

The generative model provided by HMMs can be seen as a probabilistic finite-state machine that makes a transition from a state to another with a certain probability, and emits a feature vector based on a certain distribution. The models, representing for example words formed by sequences of linguistic units, can be concatenated to represent a word sequence W. These representations are used to estimate the likelihood of $P(A|W)$.

Decoding proceeds through the use of a search algorithm that incorporates a pruning technique. Such techniques are necessary, because the size of the search space would otherwise result unmanageable, even in the case of a medium-large vocabulary. Intermediate outputs produced during the decoding phase can be used to generate a resulting recognition hypothesis space. Given a specific set of parameters that enable recording the historical trace of the decoding process, unit graphs can be generated, e.g., a lattice.

Consensus Decoding The bayesian decision theory behind MAP maximizes a decision for the whole sentence hypothesis, a sequence of words, whereas the common evaluation metric, the word error rate focuses on the edit per-word distance between the hypothesis and the reference. In a Minimum Bayesian Risk (MBR) decoding, the hypothesis selection is based on the minimization of the expected word error $\mathcal{L}$, as:

$$\hat{W} = \arg\min_{\hat{W}} \sum_{W} P(W|A)\mathcal{L}(W, \hat{W}). \tag{5.4}$$

One of the MBR initial works [Stolcke et al., 1997], presented word error minimization by bounding the search space to an N-Best list [Goel and Byrne, 2000].

Since lattices offer a combinatorial number of hypotheses, much richer than an N-Best list, these representations were later used to address the word error minimization problem. Such approaches resort to the use of a specific word sequence, also called an alignment. By aligning each word in the lattice to a particular alignment, and modifying the error distance computation, an approach known as Consensus or Confusion Network Decoding was proposed [Mangu et al., 1999]. A CN presents a compact representation of a hypothesis space. From this network, a hypothesis with a lower WER, also called the consensus hypothesis, can be extracted by selecting, at each point in the alignment, the word with the highest score. Generally, this approach is applied with a low probability word-level pruning before the extraction of the word posterior probabilities.

The elaboration of the CN is achieved through iterative alignments of all lattice arcs. The topology of the lattice, time information and scores are all factors exploited by the algorithm. Lattice arcs are merged first by word and similarity. The similarity for merging lattice arcs A_1 and A_2, occurring in a temporal cluster, is:

$$sim(A_1, A_2) = \max_{w_1 \in A_1, w_2 \in A_2} e(w_1, w_2)P(w_1|A_1)P(w_2|A_2), \tag{5.5}$$

where w_n is a word from the cluster A_n, $P(w|A)$ is the posterior probability of w in the cluster A, and $e(w_1, w_2)$ is the word-length normalized overlap between the arcs. Posterior probabilities of the same word are added. Then, arcs with different words are clustered. Clusters become the confusion sets. The original work proposed, for this step, a similarity function based on the expected phonetic similarity E:

$$sim(A_1, A_2) = E[sim_p(A_1, A_2)]. \tag{5.6}$$

Because of the way CNs are constructed, many paths which were originally not present in the original lattice may appear in the derived CN. This particular characteristic is considered potentially beneficial for the identification of the final hypothesis. Other authors showed an experimental work in which a small absolute oracle word error rate was reduced introducing some rescoring over the CN [Deoras and Jelinek, 2009]. Variations on CN, for a faster CN generation or a modification on the final network have been proposed [Xue and Zhao, 2005].

Other forms of MBR approaches have been studied [Doumpiotis and Byrne, 2004], offering other representations of the hypothesis space according to the target task they were created for.

5.2 Confidence measures in ASR

ASR systems are still unable to guarantee a perfect transcription of speech. For some tasks, it is critical to understand how reliable are the recognition results. It is in such cases that a confidence measure (CM) is required. Diverse mechanisms have been investigated on this specific topic, theoretically and experimentally [Jiang, 2005]. In the 90s, motivated by the interest arose out of the quick expansion of dialogue systems, a great amount of research was devoted to the identification of a reliable CM [Schaaf and Kemp, 1997; Kemp et al., 1997; Stolcke et al., 1997; Goel and Byrne, 2000; Wessel et al., 2001]. CM studies have focused on the search of the best measure for scoring a hypothesis in a single channel scenario.

Researchers have given different organization schemes for CMs, but the most accepted one is the one based on the nature of its computation, as high or low-level [Guo et al., 2004]. Low-level CMs exploit sources of information also used in the recognition process, e.g., acoustic and language models. Common low-level CMs include word posterior probabilities, N-best counting, or Likelihood Ratio testing. High-level CMs are estimated with the use of additional information sources, e.g., incoherence of transcription given task related semantics. Additionally, various combinations of these CMs have been studied. High-level CMs are not always permissible, since they are subject to the domain or task.

It is relevant to indicate that word posterior probabilities are the best low-level utterance verification confidence scores. Also, the potential correctness of these scores is relative to each individual system, whereas an absolute multi-microphone score has not yet been proposed.

5.2.1 Posterior probability

Originally, a normalized word scoring was proposed to detect misrecognition and out-of-vocabulary words for continuous speech recognition. This measure was computed using an all-phone recognition system [Young, 1994; Young and Ward, 1993]. The posterior probability in the standard maximum a posteriori schema offers a good estimation of hypothesis reliability, however it is hard to be precisely estimated due to the normalization term in the denominator. In the MAP decision rule, as seen in Equation 5.2, the term $P(A)$ is ignored, because it is constant across different W. This works for the identification of the most likely hypothesis, the one with the maximum posterior probability $P(W|A)$, but not as a CM since it is not normalized. A normalization factor, $P(A)$ is required in order to compute the posterior probability. In theory, this normalization factor should be computed as:

$$P(A) = \sum_W P(A, W) = \sum_W P(A|W)P(W), \tag{5.7}$$

where W denotes any hypothesis for A, and the summation is done over all possible hypotheses. Without any constraint, it is unfeasible to list the whole set of possible hypotheses. Many approaches have been proposed for the approximation of this score.

Word Posterior Probabilities estimated on Lattices

Various approaches have been proposed to estimate a word posterior probability in a lattice. In general terms, such methods propose, first the computation of the posterior probability $P(l|A)$ of every link l, and then the combination of the probabilities of the links that correspond to the same word w. Figure 5.4 shows a lattice, and its link posterior probabilities. Observe the occurrence of the same words at different time instants.

It must be noticed that the estimation of a word posterior probability, as previously introduced, faces a problem, since w can take place at different starting/ending times. Certain approaches have been proposed to address this temporal misalignment. Nevertheless, if the word posterior probabilities were to be used in the original lattice, it would result in an unbalanced probability mass, Figure 5.5.

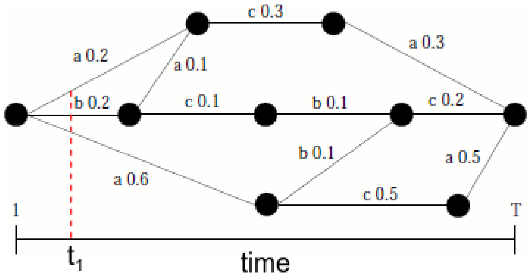


Figure 5.4: Example of a lattice. Each link features its word label (e.g, "a") and the link posterior probability. A time instance, t1, is also marked. The sum of the probabilities sum up to one at t1, as well as at any other time point.

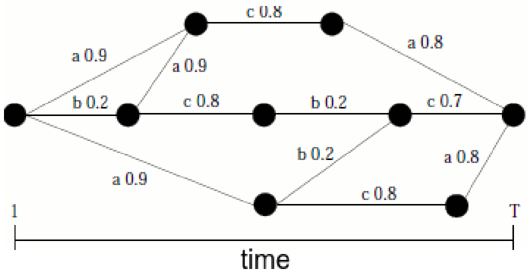


Figure 5.5: Example of a lattice(from the previous lattice). This graph reports word posterior probabilities. The max-probability method was used to compute word posterior probabilities. Note the distribution of probabilities at each time instant not adding up to 1.

The joint probability of a whole path Q, a sequence of words, and the acoustic observations A is computed from:

$$P(Q, A) = P(A|Q)P(Q). \tag{5.8}$$

The language model is scaled by the constant α, and a word insertion penalty β is added, which in the logarithmic domain results into:

$$logP(Q, A) = logP(A|Q) + \alpha logP(Q) + \beta. \tag{5.9}$$

The posterior probability of a link l is computed as the sum of the joint probability of all paths that pass through the link, i.e., the set of links Q_l:

$$P(l|A) = \frac{\sum_{Q_l} P(Q, A)}{P(A)}. \tag{5.10}$$

The normalization factor $P(A)$ is computed as:

$$P(A) = \sum_{k} P(W^k)P(A|W^k), \tag{5.11}$$

where k comprises all the hypotheses in the hypothesis space generated by the recognizer, and W^k denotes the k^{th} hypothesis.

The second step concerns the merging of the links characterized by the word w. In [Wessel et al., 2001], the maximum of the sum of time-frame posterior probabilities of these links was proposed to yield the word posterior probability.

$$P(w) = \sum P(l = w, A). \tag{5.12}$$

Word Posterior Probability in Consensus Decoding

The acoustic models underestimate the emission probabilities due to invalid independence assumptions, if a scaling is not used the best path dominates the estimation of posterior probabilities. This is not a problem in MAP decoding, since a maximization is applied. In CN however, this issue must be avoided. For this reason, LM scores are left unscaled, converting Equation 5.9 into:

$$logP(Q, A) = \frac{1}{\lambda} logP(A|Q) + logP(Q), \tag{5.13}$$

where λ is equal to the language model weight. In the original work, the use of word insertion penalty β was not found to be useful.

In the original consensus decoding work [Mangu et al., 2000], the word posterior probabilities estimated on the lattices are computed by summing up the posterior probabilities of time overlapping links with the same word, Equation 5.12.

5.3 Hypothesis combination

Concerning a multi-microphone scenario, combination of information can occur at different levels. Previously mentioned approaches included signal, feature and hypothesis based methods. Hypothesis combination encompasses approaches that aim at combining the different word-level outputs to produce a hypothesis that achieves a word error rate lower than that of the individual combined elements.

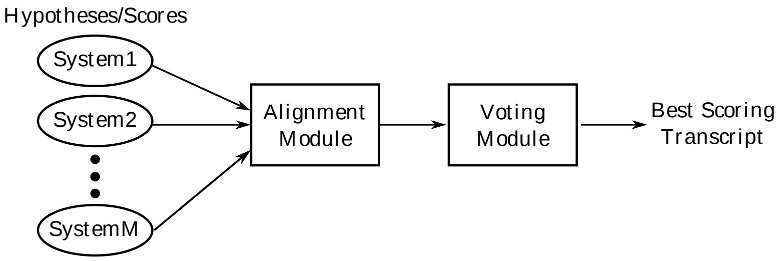


Figure 5.6: ROVER Architecture. The ovals marked with $System_i$ correspond to hypotheses generated by an ASR system and their scores.

5.3.1 ROVER

The National Institute of Standards and Technology (NIST) developed an algorithm [Fiscus, 1997] to produce a composite hypothesis out of multiple recognition hypotheses. Originally, ROVER was used to exploit hypotheses derived from a single acoustic signal, through various recognition systems. In such studies, the variety or complementarity of the hypothesis was then produced by the variation of features. As seen in Figure 5.6, the system presents two stages: the alignment of the hypotheses, and the voting. No context, forward or backward, is considered in the voting decision. If no word scores are used, a simple frequency of occurrence takes place.

The alignment phase produces a word transition network, similar to a CN, which has multiple correspondence sets. Each set, is formed by the aligned words or the NULL symbol. A general scoring is computed as:

$$ROVER(w) = \alpha \left(\frac{N(w,i)}{M} \right) + (1-\alpha)C(w,i), \tag{5.14}$$

where $N(w,i)$ is the accumulate occurrence of word w in the correspondence set i, M is the number of systems to combine, $C(w,i)$ is the confidence score for w in the set i, and α is a parameter that balances the importance of either the occurrences or the confidence values in the final score. The authors suggested to train α for an optimal performance.

5.3.2 Confusion Network Combination

The idea behind Confusion Network Combination (CNC) [Evermann and Woodland, 2000] is to take as an input a set of CNs and then process them through a voting method, in a similar fashion of ROVER but at the level of CNs instead of hypotheses. Experimental evidence has found that in most cases CNC provides an improvement over ROVER, mainly because the combination is performed on a hypothesis space, rather than individual words.

In order to apply CNC, the lattices generated by the individual recognizers are transformed into CNs. Once the CNs have been extracted, an alignment and voting method is applied at a final stage to combine the different CNs into a single CN. A weight can be assigned to the individual CNs before their combination, which can have a considerable impact on the resulting network. Another factor affecting this process, is the order in which the networks are combined.

Hypothesis combination was initially adopted as an approach for exploiting the complementarity of different ASRs over a single signal. With time, this approach was extended to consider multiple signals instead of a single one, in the search of exploiting the complementarity of the different perspectives of a perceived source. In this context, CNC has been explored [Stolcke, 2011; Wölfel et al., 2006; Cossalter et al., 2011] showing no significant improvement in comparison to signal-based approaches.

Chapter 6

Building a Multi-microphone Confusion Network

> *Invention, it must be humbly admitted,*
> *does not consist in creating out of void, but out of chaos.*
>
> Mary Shelley

As previously stated, in a multi-microphone setting, one approach for extracting the final recognition hypothesis is given by the fusion of the information derived from decoding the different captured signals. First, we study the hypothesis space, characterized as a word lattice, in the multi-microphone context. Then, we present a new method for performing multi-microphone hypothesis combination. Later, the proposed method as well as other state-of-the-art methods, are subject to experimental activities, under various system configurations.

6.1 Information in lattices extracted from multiple microphones

Hypothesis combination approaches use the words within the hypotheses, and in some cases their temporal information and scores, in order to execute the final combination. In the case of ROVER, the single hypothesis is the basic unit to combine. In the case of CNC, the unit is the Confusion Network. Both of these units are derived from the lattices, through additional processing steps. It is worth examining the lattices and the information they convey, before any manipulation is applied to them.

We explored word lattices in the multi-microphone context. This section gives an overview of the main elements observed, that lead us to the elaboration of a multi-microphone CN. Experimental activities allowed us to understand the information shared among microphones under different constraints. The findings derived from these studies helped us determine key properties for the target, bounded confusion network. The observations presented in this section were shared in all the lattices derived from the different microphones of a same room, with various language/grammar structures. The experimental multi-microphone scenario is defined by the living room, previously described in Figure 2.5. The word annotations (what was actually said) and time boundaries of each word, which are provided by the ground-truth transcriptions, were exploited.

6.1.1 Pruning and temporal information

On one hand, word lattices offer a considerable amount of information, which is provided not only by the recognized elements, but also by the general characteristics of the network. For example, time points indicating the start/end of a word are reported in a lattice. This fact constitutes an advantage over other hypothesis space representations. On the other hand, in order to generate manageable lattices, that facilitate their automatic manipulation, certain constraints, such as pruning, must be imposed.

It was observed in the literature, that defining a measure of size for word-graphs is not a trivial task. A good indicator of the complexity of processing the lattice may be given by the number of arcs. The number of nodes has no greater impact on the computation of the lattice, with reference to the processing required for search operations. Another measure studied is the lattice density, which is given by the number of arcs divided by the total number of words in the target transcription [Ortmanns et al., 1997]. According to the target task where the lattice is exploited, there are different processes to be applied. For example, for re-scoring tasks, the time-alignments of words is not critical and it is desirable to reduce the size of the lattice without altering the set of hypotheses. This means to apply a minimization process that reduces a lattice keeping most of the hypothesis space.

When pruning is applied, partial hypotheses that are relatively unlikely, are discarded. This affects the final results by guiding the search to more potentially accurate hypotheses. Pruning introduces a set of decisions taken by the decoder. If too rigid, the pruning can be detrimental for a posterior lattice elaboration. In the end, pruning affects the resulting combination with other lattices too. Exploiting pruned lattices could be preferred mainly to spare resources, which are required when dealing with large un-pruned lattices. On the other hand, slightly or not pruned lattices, could be useful to explore pieces of information

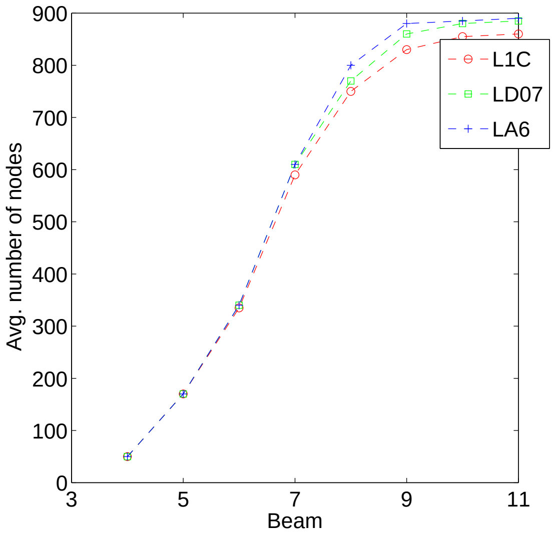


Figure 6.1: Variation of the number of nodes as a function of the lattice pruning beam factor. Although similar results were observed for other microphones, only 3 of them, on different locations, are reported here: L1C -on a wall-, LD07 -on a furniture-, and LA6 -on the ceiling-.

which could be missing in a pruned scenario. In our case, a lattice reduction process could eliminate relevant information for our combination algorithm. For this reason, one of the conditions explored was the degree of pruning to apply, in order to explore if it was necessary to tune the pruning to an optimal level.

We measured the number of nodes and links present in a lattice under different pruning scales, also called beams [Liu et al., 2003]. In the Figure 6.1 it can be seen that above a certain level of pruning, the breakpoint, the number of nodes minimally varies. This behavior was also found in the different microphones of the network, independently of the language/grammar conditions used for the recognizer. Likewise, a similar trend was observed for the links.

6.1.2 Content agreement: boundaries and words

Perhaps the most relevant aspect for the proposed method is the agreement, both of time information and words, between microphones, which is hypothesized to highlight relevant information for the combination. As part of this multi-microphone coherence analysis we performed a bi-dimensional analysis, focused on the occurrence of the ground-truth words in the lattices, considering also the ground-truth boundaries. For this purpose, the ground-truth transcription of the utterance was used, which includes the time boundaries

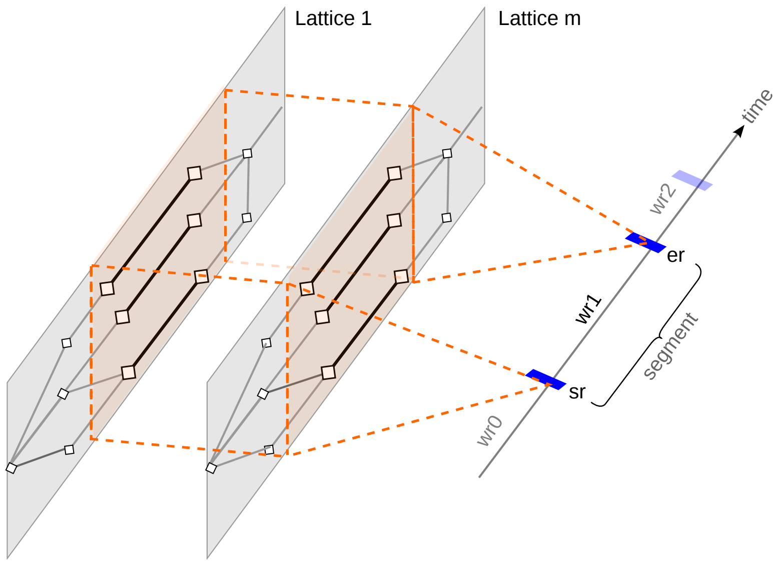


Figure 6.2: Graphical representation of the search for agreement among multi-mic derived lattices. Ground-truth boundaries of the word $wr1$ are marked as sr/er. The temporal segment under analysis is defined by the sr/er boundaries.

for each word.

At each transcription, each word marks a temporal segment of analysis, limited by the time boundaries of the word. Figure 6.2 emphasizes, for one temporal segment, the links that are analyzed in the various lattices. We compared locating words from the ground-truth transcription in single lattices, to locating them in *at least one* or more lattices. In order to clarify the notations, we define here the middle point of link as the center temporal point between the starting and ending times. For such analysis, the lattice links were selected when their middle point was anywhere within the ground-truth boundaries as:

$$s_r \leq \frac{(s_l + e_l)}{2} \leq e_r$$

where s_r and e_r are respectively the starting and ending times of a word in the ground-truth transcription, and s_l and e_l are the starting and ending times of a link under exploration.

Figure 6.3 shows the results obtained for different speakers in a real dataset.Various observations are identified. First, the singular characteristics of the speaker (e.g., speech rate, pronunciation) affect the recognized hypothesis space, and therefore the search of the word under the temporal constraints imposed in the analysis. Second, a ground-truth word is more likely to be found when a joint consideration of lattices is performed, than

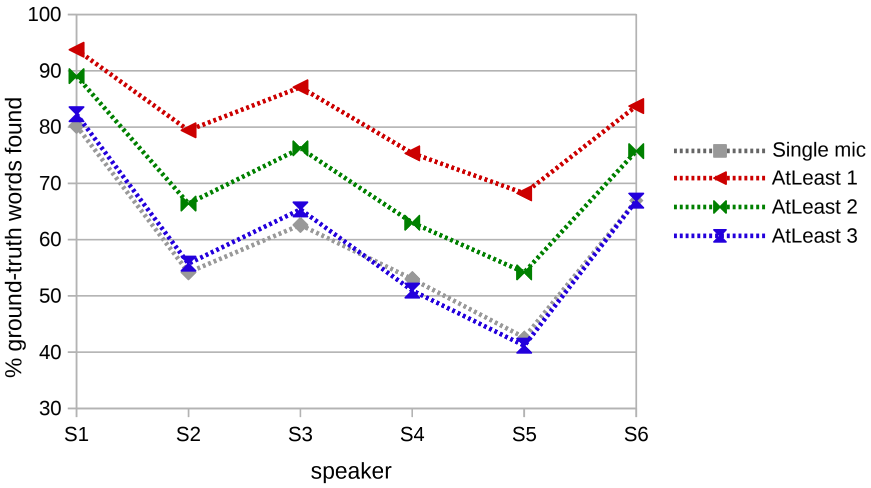


Figure 6.3: Percentage of ground-truth words found, according to 4 search constraints, for a set of speakers S_i. Word search constraints included are: 1) in a single mic, 2) in at least one mic, 3) in at least 2 mics, 4) in at least 3 mics. In the cases 2,3,4) the word search performs a joint lattice consideration.

when the single, individual lattices are considered. Third, the stronger the restriction on the search of the ground-truth word, the lower the times a word is considered found. For example, if we jointly consider all the lattices, but we look for a ground-truth word occurring in *at least 3 lattices*, the likelihood of considering that ground-truth word as found is similar as the search over a single individual lattice.

The selection of links to measure the agreement, was another key point evaluated. Even when the ground-truth time boundaries are available from the transcription, there may be a myriad of approaches to apply for the selection of the links. We explored various selection approaches that included the use of tolerance range around the time boundaries provided by the ground-truth transcription. In order to understand how these selection methods work, we investigated the following: How often does a link labeled with the ground-truth word occur in at least one lattice (microphone)?. We explored such issue under the following conditions:

a) the link's middle point is anywhere between the ground-truth starting/ending boundaries,
b) the link's middle point is anywhere between the ground-truth boundaries when a tolerance range is permitted -allowing the middle point to occur at Δ before the start or after the end boundary-,
c) the link's middle point is anywhere between the ground-truth boundaries, and the

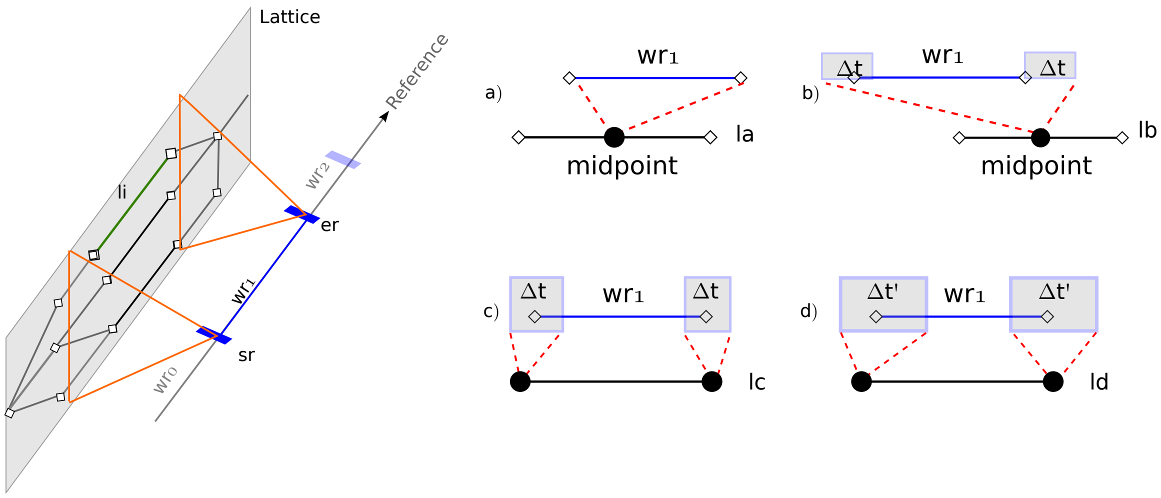


Figure 6.4: Link selection approaches. a) and b) refer to midpoint-based selection, whilst c) and d) refer to selection based on the starting/ending boundaries given a fixed Δt or a dynamic $\Delta t'$. On the left, a correspondence of the lattice link analysis according to the ground-truth word boundaries, is presented.

start and ending points of the link are within a certain dynamic tolerance $\Delta Start$ or ΔEnd,

d) the link's middle point is anywhere between the ground-truth boundaries, and the start and ending points of the link are within a certain fixed tolerance calculated from the length of the segment between ground-truth boundaries.

A general schema of these selection conditions are depicted in Figure 6.4. Various tolerance ranges and values were explored. A brief summary of the results is presented in Figure 6.5. It is observed that, a fixed tolerance range (d), conditioning the start and end points of the links, appears as the least efficient approach for the selection of links. The other selection methods produce comparable results. It must be reminded that, in these experiments, the ground-truth boundaries are used. If this information is not accurate, a middle-point-based approach could be more prone to loss in the final application of the technique than using a starting/ending based approach.

6.1.3 Content agreement: scores

Up to this point, and with the support of the ground-truth transcriptions, we have identified the links characterized by a ground-truth word, within certain time boundaries, and we have shown the importance of a joint consideration of multiple lattices. However,

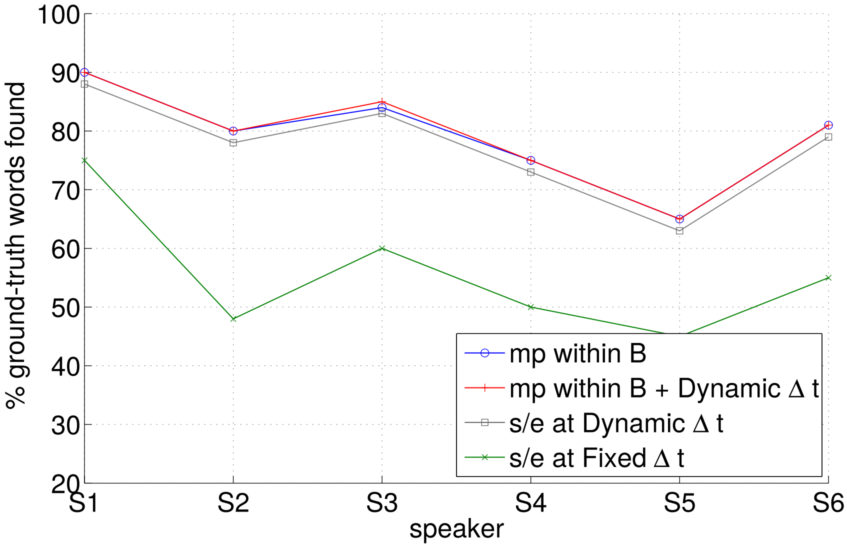


Figure 6.5: Percentage of ground-truth words found under various selection constraints. Again, a set of 6 speakers is considered. The word is found if it appears in the jointly considered lattices and the link respects the selection constraints concerning the link middple point (mp) or the starting/end times (s/e), with relation to the ground-truth boundaries (B).

nothing has been said about the scores associated to these links. For the final elaboration of a joint word graph, the score of the words identified in the analyzed temporal segments was another pivotal aspect considered. In this case we define a scoring hierarchy: intra-microphone and inter-microphone scoring. For the intra-microphone score, we refer to the work of [Wessel et al., 2001], in which different word confidence measures, derived from word graphs, were studied. According to this work, within a single-word lattice, the sum of the posterior probabilities of all parallel word links, hypothesized at a specific time instant t, is equal to one:

$$\sum_{\substack{[w;\tau,t]:\\ \tau\leq t'\leq t}} p\left([w,\tau,t]\,|x_1^T(t')\right) = 1 \quad \forall t' \in \{1,..,T\}\,, \tag{6.1}$$

where $p\left([w,\tau,t]\,|x_1^T(t')\right)$ is the posterior probability of word w occurring within boundaries τ, t, and $x_1^T(t')$ are the observations at instant t', for the utterance of length T. Extending the computation of a word posterior probability found in the referred work, we adopt the sum of all the posterior probabilities of the links identified with a word label, over a temporal segment. This score is computed at each individual microphone lattice, and it is referred in the following as the intra-microphone score.

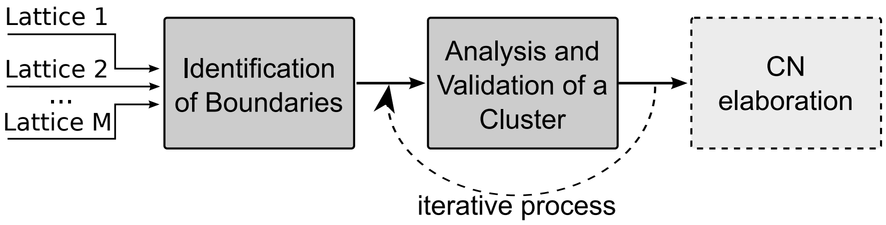

Lattice 1
Lattice 2
...
Lattice M
Identification of Boundaries
Analysis and Validation of a Cluster
CN elaboration
iterative process

Figure 6.6: Block diagram of the proposed MMCN method. Dashed line indicates an iterative process. The block labeled as CN elaboration requires the previously validated cluster.

For the inter-microphone score, no previous work concerning its computation was found. After exploring the performance of maximum identification and averaging of intra-microphone scores, the averaging of intra-microphone scores reflected a more reliable performance. A weighted scheme can also lead to valid outcomes, but this introduces the uncertainty of how to estimate the weight of each microphone and its associated lattice.

6.2 A method for extracting a multi-microphone confusion network

This section provides the general formalization of the proposed method for extracting a Multi-microphone Confusion Network (MMCN). As done in the original CN work [Mangu et al., 2000], we exploit a heuristic approach based on lattice topology definitions, as well as the time information associated with word hypotheses [Guerrero and Omologo, 2014a,b]. As mentioned in Section 5.1.1, a confusion network is defined by a chain of nodes, which denote confusion sets, each characterized by a set of word candidates and their scores. Because CNs align words occurring at different time instants in order to build each confusion set, these sets do not strictly have specific time starting/ending points. In the previous Section, we observed the lattices in the multi-microphone context, and performed a set of analysis over temporal segments, each bounded by starting/ending time boundaries. A key idea behind our approach is to build a CN, in which there is a corresponding pair of time boundaries associated to each confusion set.

A general block diagram of the operation of MMCN is shown in Figure 6.6. The process can be summarized as follows. First, a temporal analysis is performed on the multi-microphone derived lattices in order to identify a set of candidate time boundaries, that are likely to define word boundaries in the final hypothesis. These boundaries are used to define temporal clusters which are validated, according a procedure that is later detailed. The elaboration of the final CN is performed depending on the result of the previous step of validation of the cluster.

Let L_m be the set of links of a word lattice λ_m, with each link l identified by a word w, a starting $Snode(l)$, an ending node $Enode(l)$, and a posterior probability $p(l)$ (or a set of scores to compute this link probability). Let f_m be a partial path in λ_m, characterized by the words:

$$Words(f_m) = \{w \quad | \quad \exists l \in f_m\} . \tag{6.2}$$

Let B be a set of time boundaries. Then each pair of boundaries (b_s, b_e) will be used to analyze temporal segments, also called clusters. A cluster under analysis starts at b_s and ends at b_e. For the whole set of M microphones, and their corresponding lattices, we explore inter-microphone clusters, each defined by the time boundaries (b_s, b_e). A cluster is characterized by the words:

$$Words(F_{m=1}^{M}, b_s, b_e) = \{w \quad | \quad \exists l \in F_m : Snode(l) >= b_s, Enode(l) <= b_e\} , \tag{6.3}$$

where F_m is the set of paths associated to the lattice of the microphone index m. The posterior probabilities are computed for each word in the cluster. Then, a NULL posterior probability heuristic is used to accept or discard the explored clusters. Each accepted cluster represents a confusion set for the final CN. When a cluster is discarded, a new pair of boundaries is explored following the same procedure described above. An example of the procedure detailed above is also depicted in Figure 6.7.

In contrast to the original CN work, where by definition an ordered link equivalence defines the goal to be achieved, our proposal does not imply to strictly follow all the alignments in the original lattice. Consecutive words can be assigned to the same cluster when long temporal segments are explored. However, given the sequentiality of the boundaries, a partial order on the lattice links is respected among clusters. No specific initialization of the final network is required for the proposed approach. The algorithm explores the remaining unexplored pairs of boundaries, until the final boundary is reached. Given the temporal inaccuracies present in the different lattices, we consider a tolerance range for the boundaries that determine the selection of the valid links.

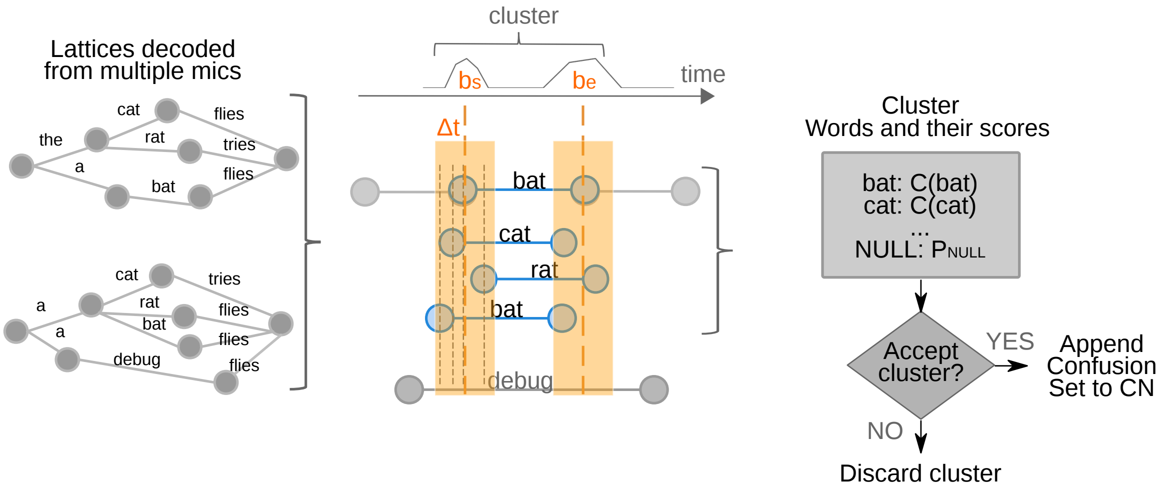


Figure 6.7: Example of the analysis of a cluster, for the extraction of a confusion set in the final CN. From left to right, this diagram describes the MMCN flow, starting from the lattices to be combined. A cluster, defined by the temporal boundaries b_s-b_e, is used for selecting links. The links selected from the various lattices are used to identify the cluster words and to compute their associated probabilities. Also the NULL probability is computed. Finally, the acceptance or discarding of the cluster is performed.

6.2.1 Boundary identification

The MMCN approach introduced in this section relies on the assumption that a reasonably accurate set of word boundaries can be identified. This set extraction is done in two stages, first a general set of boundaries are extracted, then, in a consecutive stage, these boundaries are refined. In order to identify a first set of boundaries to refine, an initial naive approach was to build a node-function out of the occurrence of all temporal nodes of the whole set of microphone lattices, as shown in Figure 6.8. Then the peaks from this extracted function were used to populate the list of candidate boundaries. Additionaly, the use of a threshold, a minimum posterior probability of the links to consider for the elaboration of the node-function, was evaluated showing no positive results. Once established the list of candidate boundaries, these were passed for a posterior refinement stage. This approach of boundary identification has its flaws. A considerably high number of false peaks appear, leading to a higher number of validations, and greater misidentifications of valid segments.

Enhanced approach: In order to improve the temporal curves from where the candidate boundaries were extracted, also the words in the links were considered to derive a new

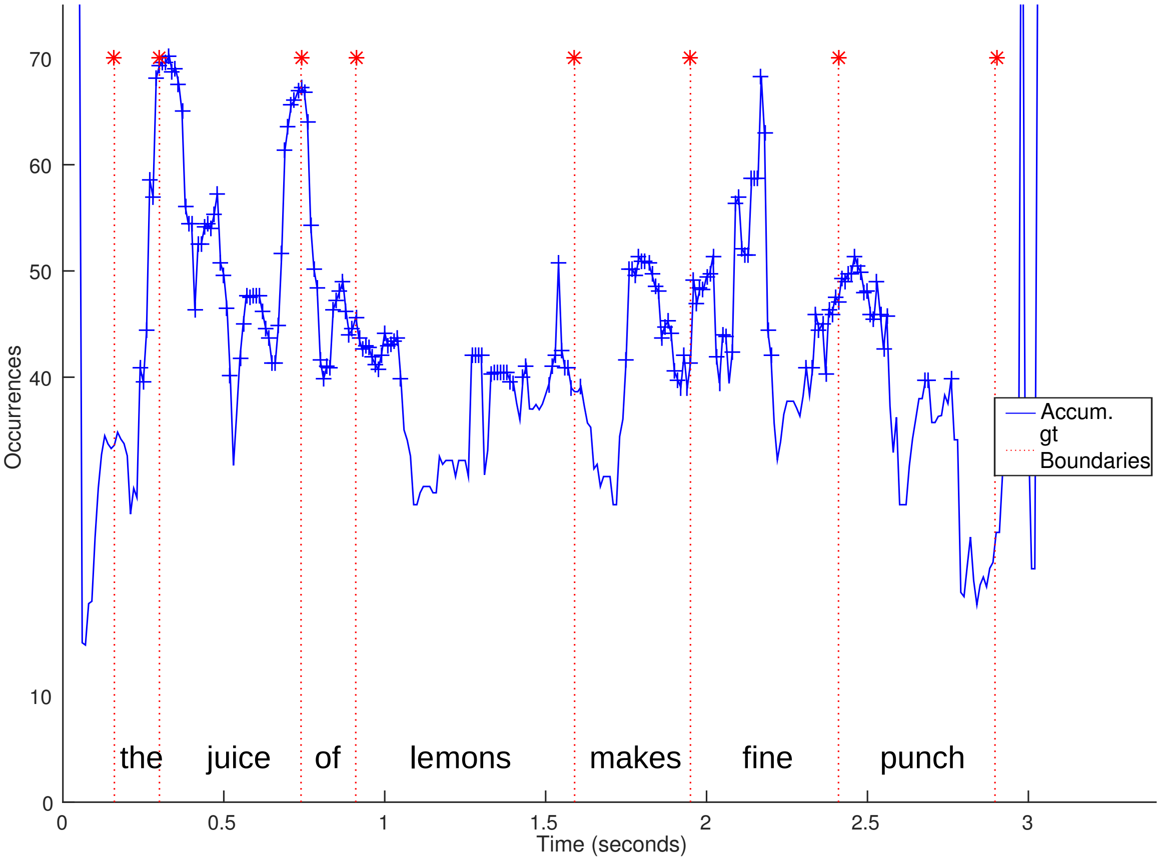


Figure 6.8: Example of a straightforward function for performing boundary identification. The node-function results from the accumulation of time nodes, presented in blue. The target ground-truth (gt) boundaries are marked in red.

node-function. There are two components of this new function, the departure function d and arrival function a links curves. Each curve was built, evaluating one time instant at a time according to the lattice resolution step (e.g., 10ms). For both link curves, a time threshold Δt was used, at each time instant.

Function d): In order to compute d, at each time instant t, the words of all links departing/starting from the window $t \pm \Delta t$ are collected. This procedure is applied collecting information from all the microphone lattices. Then for each of these words, its occurrence in each lattice is used for the final frame-score. If no words take place in the frame window, the score is 0. When all the words, of the temporal window under analysis, occur in all the microphone lattices, the score is 1. No posterior probabilities or likelihoods are used. The resulting function is defined as:

$$d(t) = \begin{cases} 0 & N_W = 0 \\ 1 & f(w_i) = M, i = 1 : N_W \\ \left[\prod_i^{N_W} \frac{f(w_i)}{M}\right]^{1/N_W} & \text{otherwise} \end{cases} \tag{6.4}$$

where W is the list of words starting from the temporal window around instant t, N_W is the size of this list, $f(w_i)$ measures in how many microphone lattices the word w_i occurred in the explored links, and M is the number of microphone lattices used. This latter computation reflects the joint occurrence of a word in the lattice set. This curve is not as noisy as the one previously introduced (counting all occurrences from all microphones). However, still some ghost peaks appear, and some correct peaks are lost.

Function a): This function is computed in a similar way as d, but this time using the links arriving at the temporal window around instant t. The curve is different than d. Particularly differences were found at the end of the sentence, due to the multiple link arrival-points represented in the lattices.

If a loose pruning setting is used (i.e, leaving more nodes in the lattice), more peaks appear in both curves. A combination of the curves a and d produced a smoother version of the node-function, defined as:

$$y = a + d. \tag{6.5}$$

Peak extraction: From the previously defined function, Equation 6.5, a set of candidate CN nodes are identified exploiting a peak selection process. In practice, we resort to an approximation of the derivative of y, as:

$$y'(t) = \frac{\sum_{i=-k}^{k} iy(t+i)}{\sum_{i=-k}^{k} i^2}, \tag{6.6}$$

where k defines the window length. Each zero crossing point of y' corresponds to a candidate CN node. Observe y at Figure 6.9, and the time points extracted from the derivative of y at Figure 6.10, y' is extracted with $k = 3$.

The occurrence of ghost peaks found in the resulting function is mainly caused by inaccuracies in the lattice content. Further improvements on the boundary identification algorithm are possible, however, it is remarked that a satisfactory performance is achieved with this approach in the proposed method.

6.2.2 Intra/Inter microphone scoring

Inspired by the confidence measure used in [Falavigna et al., 2002], for each microphone j and starting boundary B index i, the intra-microphone posterior score C assigned to the l^{th} word W_{lij} is computed as:

$$C\left([W_{lij}, B_i, B_{i+1}]\right) = \sum_{\substack{[w;\tau,t]: \\ [B_i-\Delta\leq\tau\leq B_i+\Delta], \\ [B_{i+1}-\Delta\leq t\leq B_{i+1}+\Delta]}} P\left([W_{lij}, \tau, t]\,|x_1^T(j)\right). \tag{6.7}$$

where $P\left([W_{lij}, \tau, t]\,|x_1^T(j)\right)$ corresponds to the posterior probability of the link characterized by the word W_{lij} given the observation sequence $x_1^T(j)$ related to the lattice derived from the j^{th} microphone.

The resulting intra-microphone scores are then averaged over all the channels as follows:

$$C\left([W_{li}, B_i, B_{i+1}]\right) = \frac{1}{M}\sum_{j} C\left([W_{lij}, B_i, B_{i+1}]\right), \tag{6.8}$$

where W_{li} denotes the l^{th} word of the i^{th} confusion set.

The NULL word is assigned a posterior that is complementary to the sum of the posteriors of the other words hypothesized for the segment under analysis. Once a segment is declared as valid, the inter-microphone score is used as the final posterior probability assigned of a word in a confusion set.

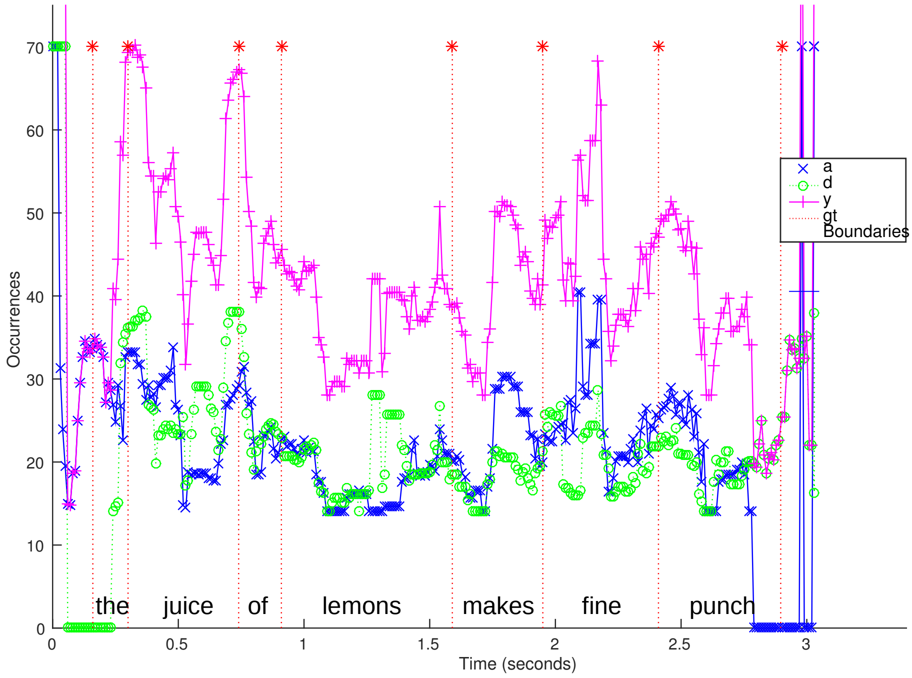


Figure 6.9: Functions used for boundary identification: a, d and their combination (y). The ground-truth boundaries are also indicated.

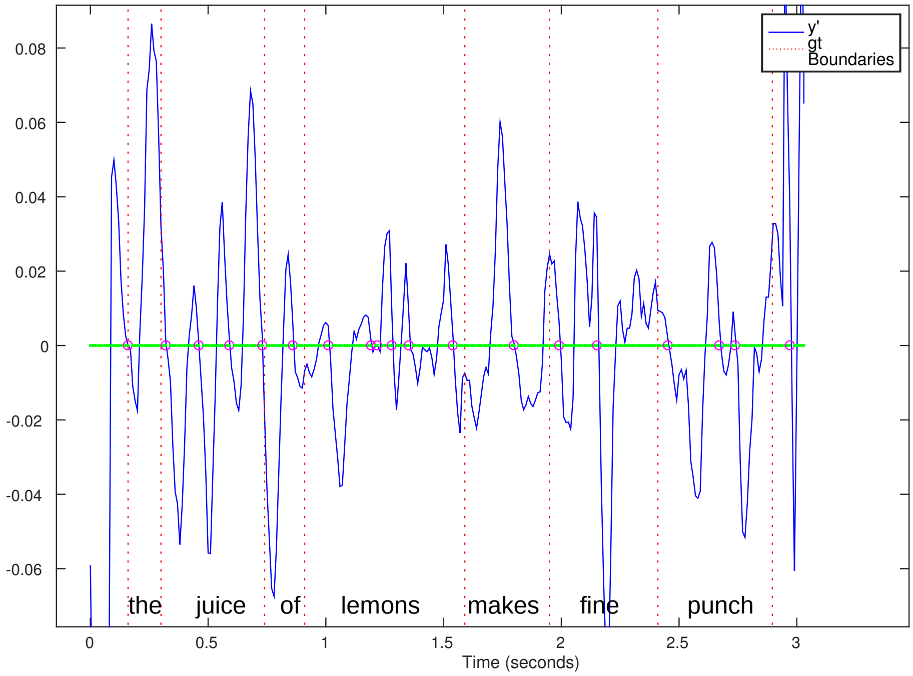


Figure 6.10: Example of the use of the derivative of a node-function for the identification of boundaries. Boundaries identified from y' are marked with circles. The ground-truth boundaries are also indicated.

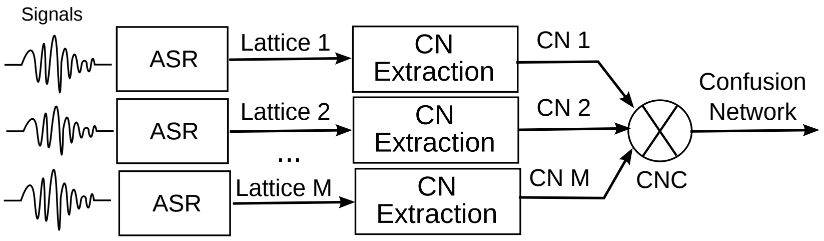


Figure 6.11: Schema of a multi-microphone recognition system with CNC.

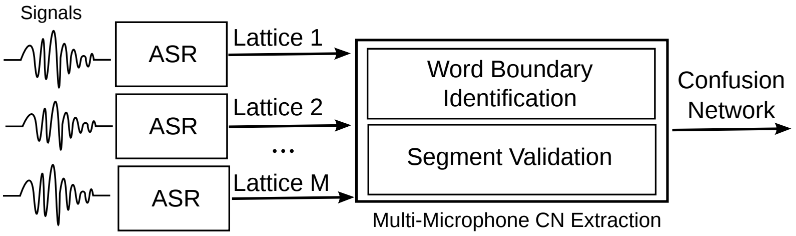


Figure 6.12: Schema of a multi-microphone recognition system with MMCN.

6.2.3 Segment validation

At every iteration a pair of boundaries defines a segment. This segment is validated in order to confirm that words, with a meaningful posterior probability, are occurring in the explored segments as to form a confusion set. If the single NULL posterior probability surpasses half of the total set probability, the segment or cluster is declared invalid, and a new segment is explored. Once a segment is declared valid it is inserted into the CN as a confusion set, and the ending boundary of the accepted segment is used as the start boundary of the next segment to explore. This is a left-to-right procedure.

6.2.4 MMCN: differences with CNC

MMCN extraction algorithm was implemented on top of the SRILM toolkit [Stolcke, 2002] (available at http://www.speech.sri.com/projects/srilm/). Therefore it can be directly compared to the SRILM implementation of CNC, which is currently the standard. This version of CNC is more closely related to the pivot version [Hakkani-Tür et al., 2006] of CN extraction than to Mangu's original version. The pivot approach aims at normalizing the topology of the word graph, according to a canonical form.

CNC computation In the CNC approach, the process to extract CNs starts from the identification/selection of the pivot hypothesis (e.g., Viterbi generated). This pivot defines the states, or confusion sets, in an initial CN template. This CN is updated iteratively every time a path (subpath or link) is aligned to the CN. Contextual information (e.g., precedence of the link) is taken into consideration. Posterior probabilities are used to update the final scores, using the "allies/competitors" approach [Falavigna et al., 2002]. Once the CNs have been extracted, the combination is applied over them. One by one, the CNs are merged into the final CN. First, the confusion sets of the first CN are inserted in a blank CN template. Then, this template is updated aligning the second CN to the template. This process is repeated with all the CNs. The alignment of the CNs resorts to the computation of the costs of aligning the words in the confusion set to those in the template. Moreover, additional rules take place in this stage.

Diagrams of the operation of CNC and MMCN are presented in Figures 6.11 and 6.12. More specifically differences among CNC and MMCN methods include but are not limited to:

- CNC explores topologically ordered lattices, while MMCN decomposes the lattices to achieve its combination.

- CNC starts the alignments with full word-paths, therefore, the context is explicitly considered for the combination of hypotheses. Subsequently, sub-paths (with the remaining link-nodes) are added using the alignment of context words. In MMCN the context is not explicitly considered. This information is embedded from the use of specific temporal boundaries.

- MMCN does not join CNs, but directly lattices. Hence, no confusion sets (columns) are considered to be aligned, as in the pivot approach.

- Time information is not used nor required in CNC. This fact may introduce discontinuities in the hypotheses extracted from the resulting CN. Only approximate locations or the first temporal occurrence of a word are used to generate timed CNs.

- In CNC the alignment is performed through the estimation of alignment costs. It computes all the alignment errors (Substitutions, Deletions, Insertions) between columns, and selects the one with minimum cost.

- CNC aligns links only once; MMCN re-uses them according to the segment boundaries explored. If an erroneous assignation of a link to a confusion set is done, MMCN's strategy can recover in a consecutive cluster analysis.

- Mangu's version aims at WER reduction (biased to ASR), while SRILM does not.

- CNC extraction exploits dynamic programming alignments, which leads to different results according to the order in which the elements are combined. This introduces a new variance into the problem, since no a priori mechanism is known to estimate in advance the order in which the CNs must be combined.

It must be reminded that time from lattices is not a completely reliable feature; these come from estimations of the most likely identified units (phones). Noise and reverberation confuse the recognizer about the word boundaries.

The algorithm was implemented over SRILM toolkit, supporting the combination of Standard Lattice Format lattices, described in Appendix A.

6.3 Multi-microphone processing Experiments - Decoding with N-gram LMs

In this section, three experimental tasks are explored. The first task, which in the following we identify as Case I, deals with the recognition of home-control commands in Italian. The second task, Case II, concerns Wall Street Journal (WSJ) material in English. This latter corresponds to a large vocabulary continuous speech recognition task. With these two experimental cases we wanted to explore the issues concerning the use of a large number of microphones, the benefits of using such numerous input sources, and the performance of the proposed method in comparison to other signal or hypothesis based methods. Another aspect evaluated was the performance of two algorithms for boundary identification in MMCN. The use of a bigram and a tri-gram language models was also investigated. Finally, a third task, Case III, includes the recognition of read commands. The main concern of this latter case is the exploration of the oracle gap, identified in the preliminary experiments. Part of the results described in this section appear in: [Guerrero and Omologo, 2014b,a].
The following activities are reported:
i) ASR performance results for various multi-microphone processing methods, for Cases I and II.
ii) Experimental analysis of the gap found between the oracle and the average single distant microphone, SDM, recognition. This analysis is reported as Case III.

6.3.1 Experimental setup specifications - Case I

The experimental scenario exploited is the living room in the DIRHA smart-home setup, introduced in Section 2.4.2. A total of 15 synchronized microphones were used; 6 on the ceiling (LA*) and 9 on the walls (L1*, L2*, L3*, L4*).

APASCI database [Angelini et al., 1994] was contaminated [Matassoni et al., 2002], and this version was used to extract the training dataset for the acoustic models. This database includes 20 phonetically rich sentences spoken by 164 speakers.

The test material was extracted from the DIRHA corpus, see Section 2.4.2. Example of sentences in the test material are:

- *"imposta il telefono in modalità silenziosa"*
- *"apri la velux in camera"*

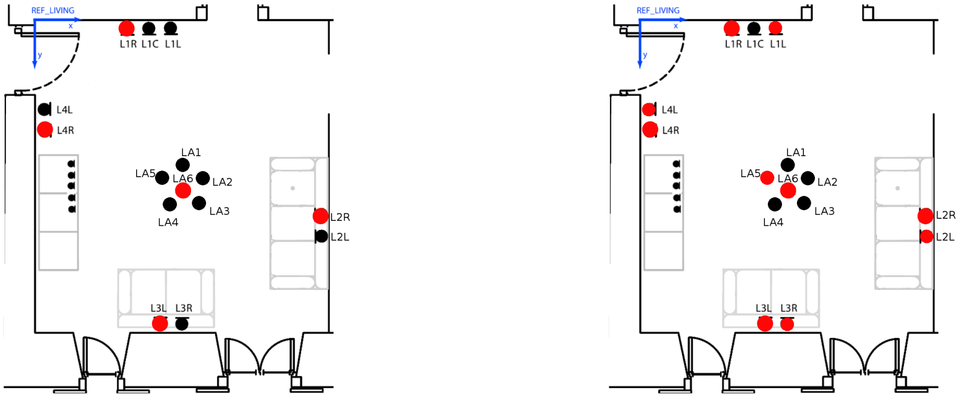
The simulations required a set of measured IRs. The selection of source POSORIs to generate the development and test simulated material, was performed randomly. There is no overlap between development and test sets, as far as utterances spoken by a speaker at a certain location. The development set (devset) was composed of 61 phrases, spoken by 27 speakers at 43 position/orientations. In the test set there were 2245 phrases in total, spoken by 30 speakers, at 74 position/orientations.

Multi-microphone configurations Three microphone configurations are presented in the tables, which are denoted as C_5, C_{10} and C_{15}. The microphones composing the configurations C_5 and C_{10} are displayed in Figure 6.13, while C_{15} includes all 15 microphones in the room.

Multi-microphone processing methods

The following methods are evaluated:

- For comparison purposes, the results of applying Beamforming (**BF**) [Anguera et al., 2007] to the distant microphones is reported.
- Hypothesis combination methods: **CNC, ROVER and MMCN**. The results of word level hypothesis combination ROVER were derived with the SCTK Toolkit[NIST, 2009]. In order to apply CNC for hypothesis combination, the SRILM toolkit[Stolcke, 2002] was used. All the CNs were assigned a uniform weight.



(a) C_5 (b) C_{10}

Figure 6.13: Microphones corresponding to each configuration appear in red.

Concerning ROVER and CNC, the order of the elements used in the combination affects the final hypothesis. According to the number of microphones to combine, there would be a large number of possible permutations to explore (e.g., given 5 microphones, there are 120 possible permutations). For this reason, in this work, given N microphones, only N permutations were addressed with ROVER and CNC. Each permutation is created in a cyclic fashion, starting at element k, with $k = 1, .., N$. The results for ROVER and CNC present the average performance of these permutations.

The MMCN boundary identification algorithm considered the naive approach, counting all nodes, for the Case I.

A general overview of the experimental framework is shown in Figure 6.14.

Speech recognition and evaluation

The speech recognition system used in this work was built on the HTK toolkit. A standard front-end processing was employed, with a pre-emphasis step and a feature vector composed of 12 MFCC plus the energy, and their first and second derivatives. Mean and energy normalization were applied.

A set of 27 phonemes was employed. Acoustic context-independent phone units, modeled with three states and 32 Gaussian mixtures per state, were trained.

The language model was a bigram, trained on a mixture of read and spontaneous commands, collected under the DIRHA project. The size of the dictionary was of 380 words. Language model scale and word insertion penalty were optimized in the devset,

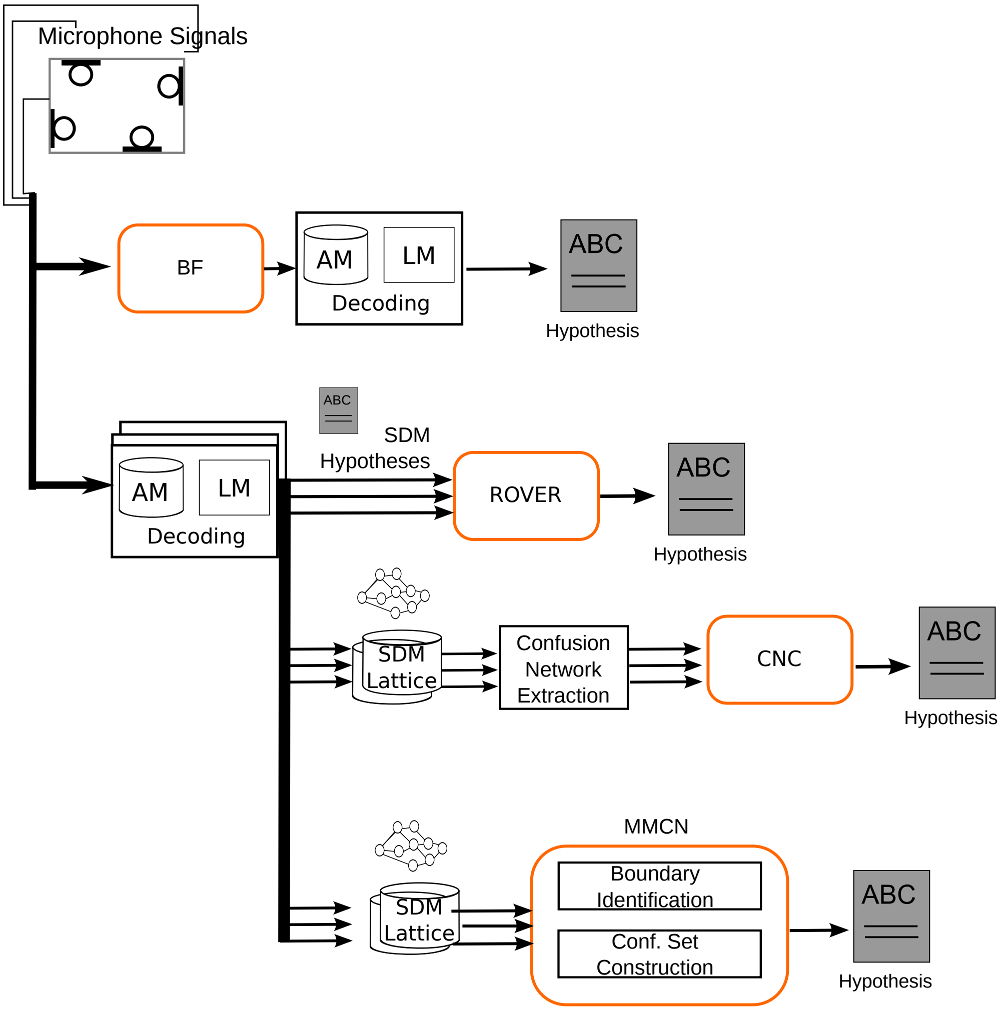


Figure 6.14: General diagram of the experimental framework.

using only one microphone per each microphone group (e.g., mic LA6 in the group LAx). A single combination of parameters (Language model scale of 11, word insertion penalty of 16) was then used, for all the microphones in the test set.

Various lattice pruning beams were explored in order to understand the effect of this parameter in the resulting performance of the combination methods. Here we contrast the results of applying beam-search with a beam of 80, 100 and no beam, this latter being a more computational demanding setting.

The recognition of the individual signals, or the processed ones, are measured in terms of WER. SDM reports the performance of decoding each single microphone, for the whole set of utterances evaluated. The ORACLE is computed a posteriori. Per each utterance, after computing the WER of each microphone, the microphone with the lowest WER is selected. This oracle is a “cheating” measure which is useful to provide an upperbound performance.

6.3.2 Experimental setup specifications - Case II

The experimental scenario exploited is also the living room in the DIRHA smart-home setup, introduced in Section 2.4.2. A total of 5 synchronized microphones were used; 1 on the ceiling (LA*) and 4 on the walls (L1*, L2*, L3*, L4*). Moreover, two microphone arrays, one linear-array on a wall and one circular-array on the ceiling, were available in this dataset. These additional arrays were used for a specific signal processing task, indicated in the following. The microphones in the linear harmonic array have different characteristics to the other sensors [Ravanelli et al., 2015; Cristoforetti et al., 2014].

The training data consists of 7138 simulated reverberant utterances, derived from the full clean WSJ0-5k [Garofalo et al., 1993] training set. This training set was simulated using 12 measured IRs [Cristoforetti et al., 2014]. The test material is extracted from the WSJ0-5k sub-set of the DIRHA-English [Ravanelli et al., 2015] corpus, which includes data recorded in the real living room. The sentences include large variations in terms of number of words, as shown in these examples:

- *“it didn’t elaborate”*
- *“link resources corporation estimates the electronic mail market at about two hundred ninety six million dollars a year and voice mail at about seventy six million dollars”*

Four different datasets were evaluated, 2 development and 2 test sets, with simulations (SIM) and recorded (REAL) data.

Multi-microphone processing methods

For this case, we evaluate the performance of **BF** and hypothesis combination methods, i.e., **ROVER, CNC and MMCN**.

A subset of all the possible combination of sensors were evaluated for the cases of ROVER and CNC. These combinations, identified by c_n, correspond to a circular ordering, as presented in the previous experiments.

Speech recognition and evaluation

The speech recognition system used in this work was built on the Kaldi platform.

Standard MFFC based feature extraction was applied. The vectors were composed of 13 MFCCs per frame, augmented with first and second order derivatives. Acoustic models (AM) as extracted by DNN in Karel's recipe [Veselỳ et al., 2013] were composed of 6 hidden layers of 1024 neurons, and a window of 11 context frames. AMs were trained on a contaminated version of the TIMIT [Garofolo et al., 1993] corpus.

6.3.3 DSR results for Case I and Case II

Results of DSR for both Case I and II are presented here. The complete results are detailed in Tables B.1, B.3 and B.4 in the Appendix B. As previously noted, in the Case I various pruning beams are explored. It was experimentally evidenced that the application of different pruning beams affects the search space from which the final hypothesis is extracted. Another factor explored was the microphone set. We found that the use of a different subset of microphones changes the results achieved with all the different techniques explored. In Figures 6.15 and 6.16, the performance of the different techniques studied, for the Case I, is presented, both for development and test sets. Figure 6.17 provides similar information for the Case II.

The curves corresponding to SDM, report the average of the subset of microphones used in each experiment. They correspond to a sort of baseline, where no combination or processing has been applied. Experimental results show that hypothesis combination approaches achieve better performance than SDM and even BF. Furthermore, with the optimal set of parameters, MMCN achieves a performance comparable to CNC. Small variations on the trends given by the different beam degrees are observed for the different beam settings in the Case I. These were caused by the limited size of the development set, on which recognition parameters were tuned. Nevertheless, the observations among

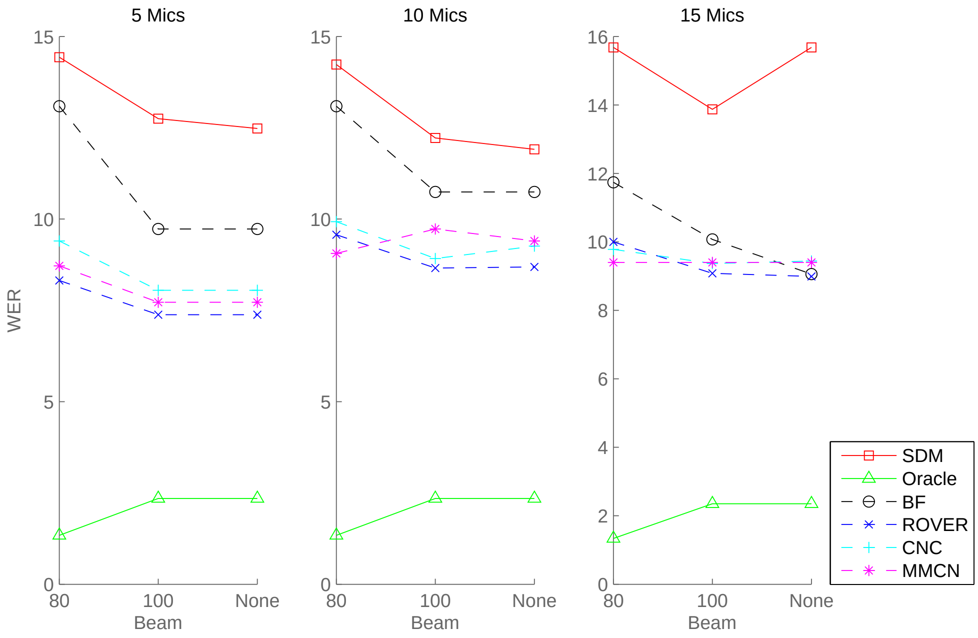


Figure 6.15: WER obtained for the Case I with different microphone (mic) sets. These results correspond to the Devset.

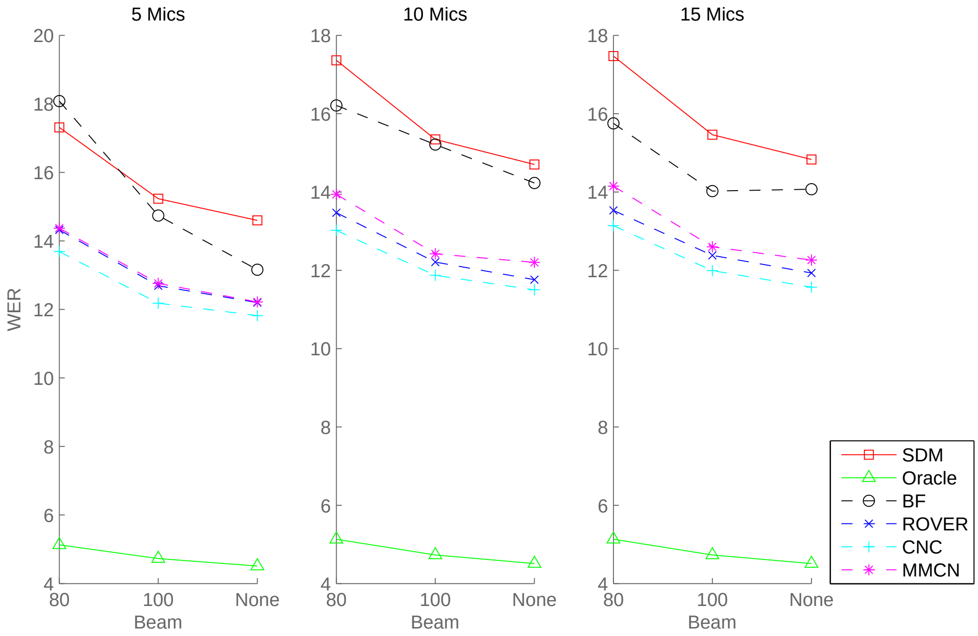


Figure 6.16: WER obtained for the Case I with different microphone sets. These results correspond to the Testset.

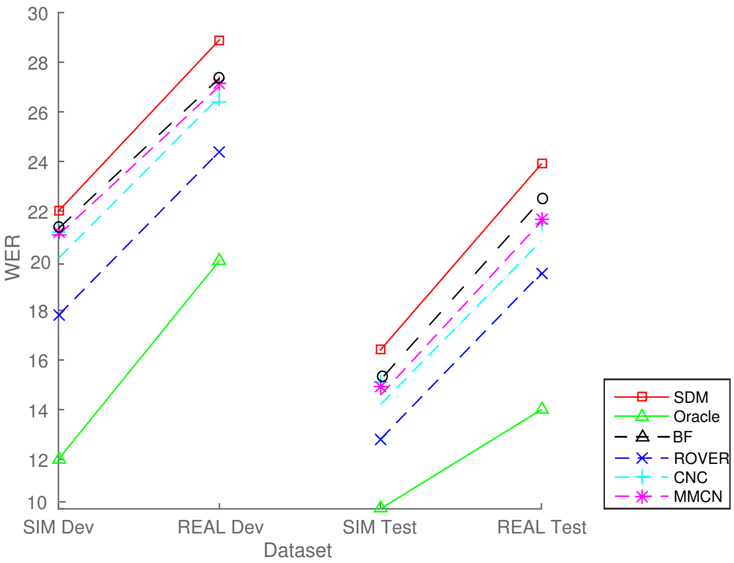


Figure 6.17: WER results for the Case II, on simulated and real dev and test sets.

the techniques were confirmed by both datasets. These observations were also confirmed by the results in the Case II. For this latter Case, the BF results reported in the Figure 6.17 are not computed over the complete set of distant microphones. These are presented in the same figure for comparison purposes. In the experiments, in Case II, two BF operations were performed, one on a linear array and another on a circular array. The Figure 6.17 reports the average of these BF cases. Detailed results can be found in the Appendix B.

Impact of Microphone Permutations

The order in which the microphones are incorporated affect standard combination approaches. In order to evidence the effect of this issue, we explored, for the Case I, other microphone configurations, such as one based on ranking microphone-group WERs. Three groups were created: X_6) the microphones on the ceiling (6 microphones), then X_{10}) those in X_6 plus one mic of each wall group (10 microphones), and finally X_{15}) adding the remaining wall sensors (15 microphones). Note that in this case only one permutation is reported for each microphone configuration.

With CNC, in the case of X_{15}, its performance is different from the average reported in C_{15} Table B.1. The specific combination given by X_{15} is one of the numerous configurations that can be explored by CNC, with the same set of microphones. This illustrates the impact of the arrangement of microphones on CNC, which is not an issue for MMCN.

We could also expose the effect of adding more microphones to the combination. In this analysis, the performance achieved by the first configuration, which includes lattices that achieve the worst group performance (i.e., the highest SDM WER), can be improved with the balanced inclusion of microphones with better performance. In Figure 6.18, it can be observed that, for most of the explored mic-configurations, MMCN shows a lower WER than CNC.

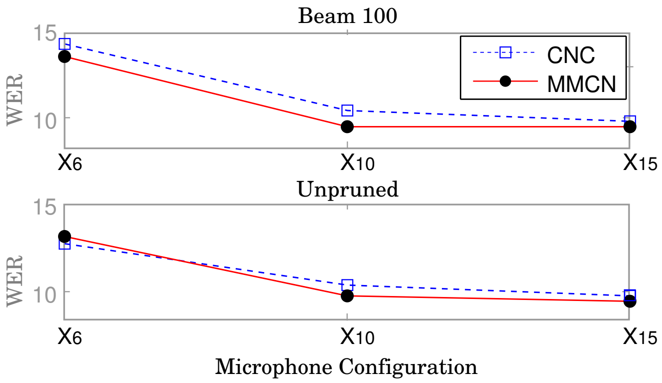


Figure 6.18: WER variation with different mic-configurations.

6.3.4 Oracle observations in Case III

It can be observed, from the previous experimental results, the existence of a considerable gap between the results achieved with the oracle and those achieved through hypothesis combination approaches. With the intention of increasing our understanding about the origin of this gap, additional analysis was performed on a dataset with similar characteristics to those of the Case I.

Experimental setup specifications - Case III

A total of 5 synchronized microphones present in the experimental scenario, i.e., living room in the DIRHA smart-home, were used.

The train dataset is the same as the one described in Case I. The test material was extracted from the DIRHA corpus, but this set was directly recorded at the real living room. In this case, the position and orientation of the speakers where fixed, and indicated to the speaker during the experiment, Figure 6.19. These specific locations allowed us to isolate the analysis only as a function of the source location.

The speech recognition was configured as in the Case I, and built over the HTK toolkit.

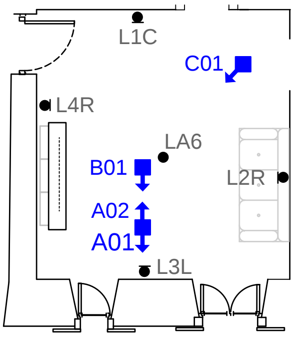


Figure 6.19: Setting used for the oracle analysis. Speaker positions and orientations are shown in blue. Microphones are indicated as black circles.

Oracle ranking

It must be noticed that the Oracle is derived from the recognized hypothesis of a single microphone; the one with the lowest WER from an a posteriori evaluation. However, the WER can be the same for more than one microphone. If the hypotheses from the different microphones share the same errors or the same number of errors, it makes them all top-scored microphones. Nevertheless, at every utterance, only one microphone is selected. Numerically, this selection results in the same amount of errors in the dataset evaluation as if another hypothesis, with the same amount of errors, had been selected. The selected first-best oracle hypothesis is the oracle-best (OB).

If the standard CN-based hypothesis combination technique is applied over the top-scored oracle microphones we observed that:

- if a CN is extracted from the OB lattice, the WER of the hypothesis derived from that CN is higher than the one obtained by the OB hypothesis.
- if all the oracle top-microphones are selected, the WER from their CNC is lower than in the previous case, but anyway higher than the OB.
- if a maximum number of microphones is selected, for example at most the top 3 out of the 5 available microphones, its performance is similar to selecting all the microphones.
- if a fixed number of microphones is selected, for example the highest 3 from the oracle ranking, its WER is higher than that obtained by the single top microphone.

In addition to the bigram LM, a simple grammar in which all words have the same

Table 6.1: Percentage of utterances with min-WER that are shared among a number of microphones, for each position/orientation (POSORI).

	Num. Microphones in Oracle				
POSORI	1	2	3	4	5
A01	45.37	20.37	13.89	10.65	9.72
A02	37.96	22.22	14.35	10.19	15.28
B01	39.35	19.91	14.82	13.43	12.50
C01	38.89	23.61	18.06	6.49	12.96

likelihood of occurring, was used. The findings were evidenced for both LMs. Moreover, for the bigram case, different beam-pruning settings were studied.

In all the combinations of Oracle-CNC cases, the WER was higher than the OB. This indicates that in some cases the combination introduces relevant errors, probably caused by the alignment forced to the lattice(s). However, the combination is still valuable; a WER reduction was observed when multiple oracle microphones were combined.

Moreover, in comparison to the combination of all the microphones, the pre-selection provided by the oracle showed an WER reduction. The questions to solve for such a selection in an automatic fashion are: is there a number of microphones required to achieve such an optimized combination?, how can this selection be done?. It was shown that a fixed number was not optimal.

Influence of Position/orientation

Given the availability of Position/Orientations of speakers in the test set, it was also possible to analyze if the oracle-combination issue was dependent of the specific speaker location. In table 6.1, we can see that the number of microphones indicated as best, top-scored microphones, by the oracle ranking was less dependent on the POSORIs evaluated, and more on the utterance content.

CNs and posterior pruning

It is also interesting to study the relationship between the oracle and the hypothesis combination, and observe where the loss of performance takes place. In order to explore

Table 6.2: WER per method combination for each position/orientation (POSORI).

	POSORI			
Methods	A01	A02	B01	C01
CNC-Oracle 1	53.85	60.7	57.27	58.18
CNC- Oracle 1 PPruned	52.73	59.02	56.15	57.34
CNC-Oracle 3max	52.73	58.88	56.08	57.41
CNC- Oracle 3max PPruned	52.03	57.69	55.45	56.5
CNC- Oracle 3fix	56.99	62.38	59.3	59.93
CNC-Oracle (all)	52.73	58.81	56.15	57.55
CNC- Oracle (all) PPruned	51.89	57.62	55.31	56.36
Oracle	**52.10**	**56.50**	**55.31**	**56.29**

this issue in depth, we applied Posterior-Pruning to the lattices, as suggested by Mangu's work. Table 6.2 presents the performance of combining the results provided by the Oracle and CNC. At the end of the table, for reference, the Oracle performance is also reported. It can be observed that for the different combinations, the addition of Posterior-pruning to the lattices reduces the WER, approaching the resulting WER to that of the Oracle. This is slightly different for the position-orientation A02 though. The justification can be found in the analysis of the oracle-ranks. Position A02 has the highest variance of number of microphones with minimum WER, and also the highest average WER from the first top-scored mic. This means that for this POSORI, there will be more elements (lattices) involved in the combinations, and that the likelihood of having errors introduced by these elements is higher than in the other POSORIs.

We wanted to explore if posterior pruning could be beneficial for MMCN and CNC when all the microphones were used. In this case, no Oracle ranking was used. The performance of both MMCN and CNC was similar to the unpruned cases. For MMCN, in fact, the performance was worse than the unpruned version. This happens due to the lack of information, relevant for the MMCN approach, which is removed by the pruning step.

6.3.5 Conclusions

From the different setups evaluated, it can be observed that the number of microphones is not the unique factor affecting hypothesis combination approaches; the quality of the source lattices, an information not available a priori, is also relevant. MMCN leads to WER reductions in some of the explored cases, and moreover, its performance is invariable independently of the arrangement of the microphoness. This is an advantage over CNC in a multi-microphone scenario where an analysis of a proper ordering is hard to achieve. Note that the algorithms in MMCN are just at starting level, concerning for example the algorithms used for boundary identification. This leaves a window for improvement concerning the identification of optimal parameters.

Additionally, some insights were extracted from the analysis performed on the investigation of the gap between the Oracle results and those of the combination approaches. Experimental results confirmed that combining all the microphones is not always a good strategy. In an enclosure, with microphones redundantly located in space, combining only a set of them -the least noisy ones- seems to lead to reductions in the WER. A deeper study on different ranking and selection techniques could be exploited for further investigation. For example, CS techniques addressed in this work, in Chapter 4, can be explored together with hypothesis combination methods.

6.4 Multi-microphone processing Experiments - Decoding with a word-loop grammar

In this experiment, the task was the recognition of phonetically rich sentences in English language. One of the factors we wanted to explore in this experiment was the influence that a 0-gram grammar would have on the hypothesis combination approaches. Additionally, we wanted to have a better understanding of the robustness of the proposed hypothesis combination approach, when facing specifically introduced variations.

The following activities are reported:
i) Performance comparison of SDM recognition to selection and combination methods.
ii) Evaluation of MMCN's sensitivity to specific issues. The MMCN boundary identification algorithm used in this case incorporated the boundary identification approach based on weighted arrivals and departures at time nodes.

6.4.1 Experimental setup specifications

The experimental scenario is again the DIRHA living room, introduced in Section 2.4.2. A subset of 5 microphones was mainly used, four of them located at the walls (L1C, L2R, L3L, L4L) and one on the ceiling (LA6).

The traning dataset corresponds to a contaminated version of the TIMIT [Garofolo et al., 1993] training data. IRs describing 16 randomly selected POSORIs were used to contaminate the training material. The test material is extracted as a subset of the DIRHA-US dataset, see Section 2.4.2. The sentences do not correspond to spontaneous speech; they were read by the speakers. Examples of sentences found in the test material:

- *"the birch canoe slid on the smooth planks"*
- *"glue the sheet to the dark blue background"*

Two types of test data were used: synthetic simulations (SIM) and real-recorded (REAL) data. For the SIM data, a random selection of multiple speaker POSORIs was performed. The REAL dataset was registered in the real experimental scenario, see Section 2.4.2. Each test set considers the participation of six different speakers, 3 male and 3 female.

Multi-microphone processing methods

Multi-microphone processing evaluated methods include:

- Channel Selection based on an un-weighted implementation of the Envelope-Variance [Wolf, 2013]. This **CS** method was selected because it is reported in the literature as the state-of-the-art.
- Hypothesis combination methods: **CNC, ROVER and MMCN**.

The methods of ROVER and CNC employ 5 microphones to compute the combination. As previously indicated, these methods are affected by the dynamic alignment of elements, for which different results are achieved for each arrangement of the combined microphones. In this experiments, only 5 out of the 120 different microphone arrangements were evaluated, see Table 6.3.

Table 6.3: Microphone arrangements explored for ROVER and CNC

Config.	MICs
c1)	L1C-L2R-L3L-L4L-LA6
c2)	L2R-L3L-L4L-LA6-L1C
c3)	L3L-L4L-LA6-L1C-L2R
c4)	L4L-LA6-L1C-L2R-L3L
c5)	LA6-L1C-L2R-L3L-L4L

Speech recognition and evaluation

The ASR system used in this experiment was built with the Kaldi toolkit. A phoneset of 48 phones was used. Mono-phones were adopted for the acoustic models. Word-loop or 0-gram grammars were trained on each speaker material. Each grammar included a vocabulary size of approximately 250 words.

The recognition of the individual signals, or the processed ones, are measured in terms of WER. As in the previous experiments, SDM and ORACLE results are reported.

6.4.2 DSR results

In Appendix B, Tables B.6 and B.7 detail the results obtained with the ORACLE, CS, and the hypothesis combination methods, for the simulated and real datasets respectively. In Figures 6.20 and 6.21, as highlighted in Section 6.3.4, the existing gap between the Oracle and the different evaluated methods is evidenced. In this experimental case the recognition task is much more complex than the one presented in the previous settings, although the vocabulary size is smaller. This justifies the low performance results achieved by all methods. The influence of speaker features, e.g., speech rate, is observed in these curves. The CS approach did not lead to further improvement over SDM. A probable cause of this results can be the CS method, i.e., EV, being affected by the speech content characteristics. Further exploration on this issue can be performed in a future study. Still, hypothesis combination approaches show a reduction in the WER in comparison to SDM decoding, for all the different conditions.

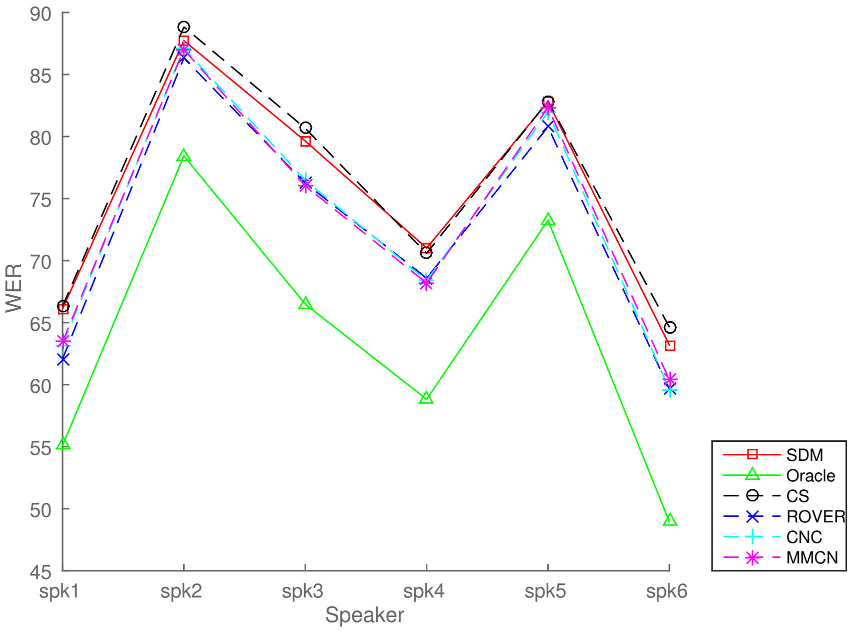


Figure 6.20: WER results of the different evaluated methods over the simulated dataset.

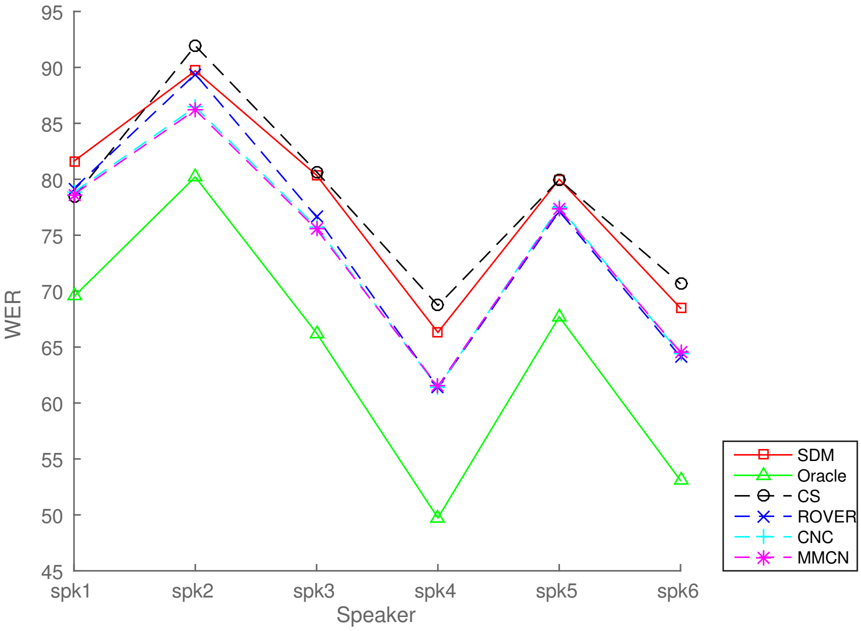


Figure 6.21: WER results of the different evaluated methods over the real dataset.

6.4.3 MMCN sensitivity tests

In the following sub-sections, the performance of MMCN under different boundary misplacement errors is measured. These errors correspond to potential inaccuracies that can occur during the operation of the method. The results to the analysis are reported for each of the 6 speakers available in the datasets, in order to evidence how their specific acoustic characteristics could affect the performance of the algorithms evaluated.

Boundary shift

The first test measures the sensitivity of the approach to boundary misplacement. Positive or negative shifts are applied to the ground-truth boundaries. Shifts are applied to one boundary at a time. The procedure, for every utterance, is as follows. For every shift-step, the resulting score is the average over all the possible shifts applied. Shifts are applied if: a) they respect the starting/ending boundaries of the utterance, and b) they respect neighbouring segments (greater than the previous boundary, lower than the next boundary). Table B.8 reports the results on SIM and REAL datasets. In this case, the validation of the segment is also performed. Figure 6.22 reports the results of this evaluation only on the SIM dataset.

Not using the segment validation mechanism was also considered. This operation allowed us to isolate the "link selection" sensitivity, independent from the rest of the other modules of the method. The relative variations of WER are observed in Figure 6.23.

Boundary loss

This test measures the sensitivity of the approach when boundaries are removed from the list of ground-truth boundaries used for posterior CN building. In this case, only segment validation module of MMCN is being evaluated. Table B.9 details the results of SIM and REAL datasets, and the relative losses can be observed in Figure 6.24.

Boundary addition

This test evaluates the sensitivity of the approach when fake boundaries are added to the ground-truth boundaries. Additions are applied between two ground-truth boundaries. The challenge for MMCN in this case is to identify the false segments, and discard them. As in the previous test, only the segment validation module of MMCN is being evaluated.

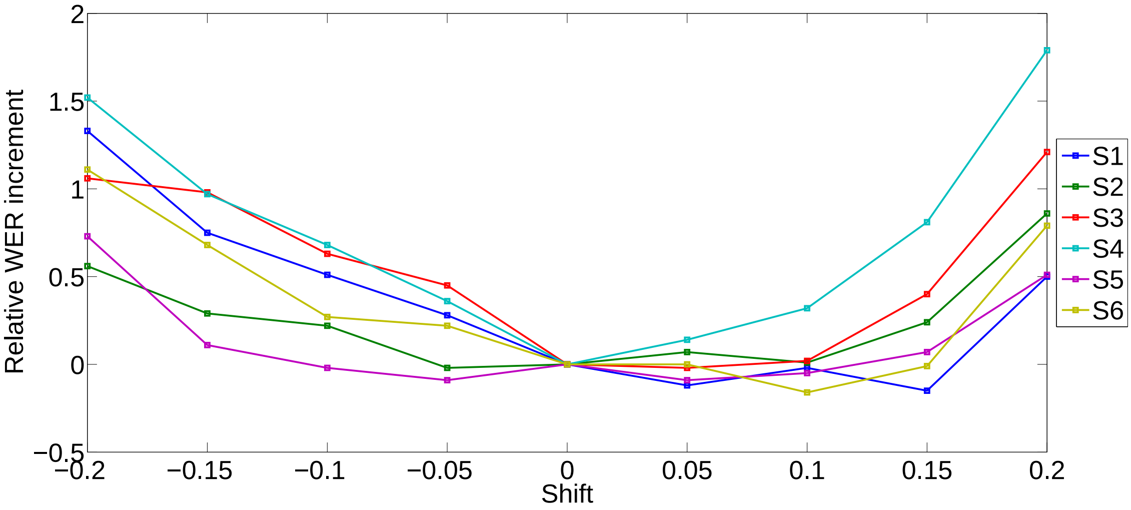


Figure 6.22: WER variation as a result of introducing Boundary Shifting + Segment Validation. Results are presented for each speaker S_i. These results correspond to experiments on simulated data.

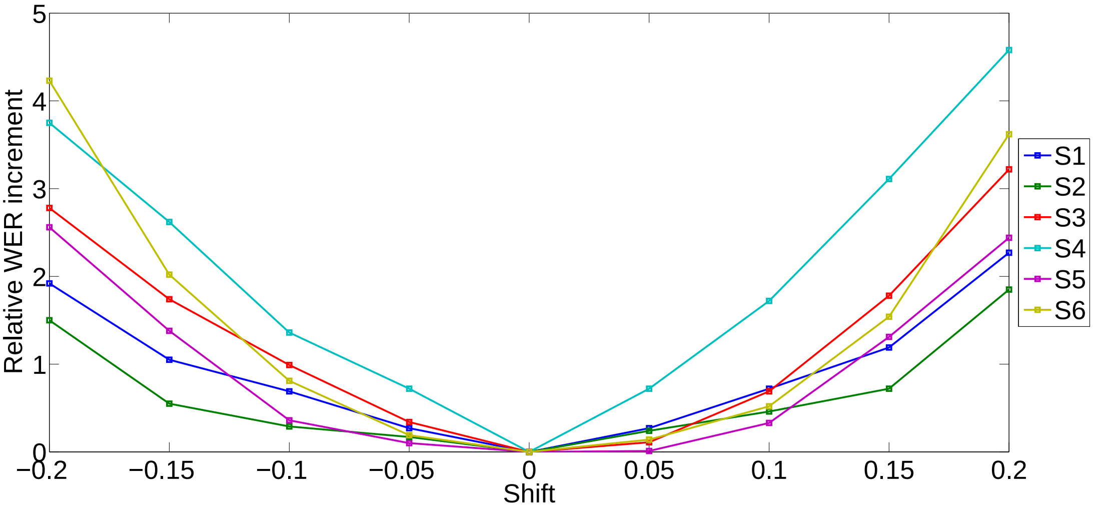


Figure 6.23: WER variation as a result of introducing Boundary Shifting. Results are presented for each speaker S_i. These results correspond to experiments on simulated data.

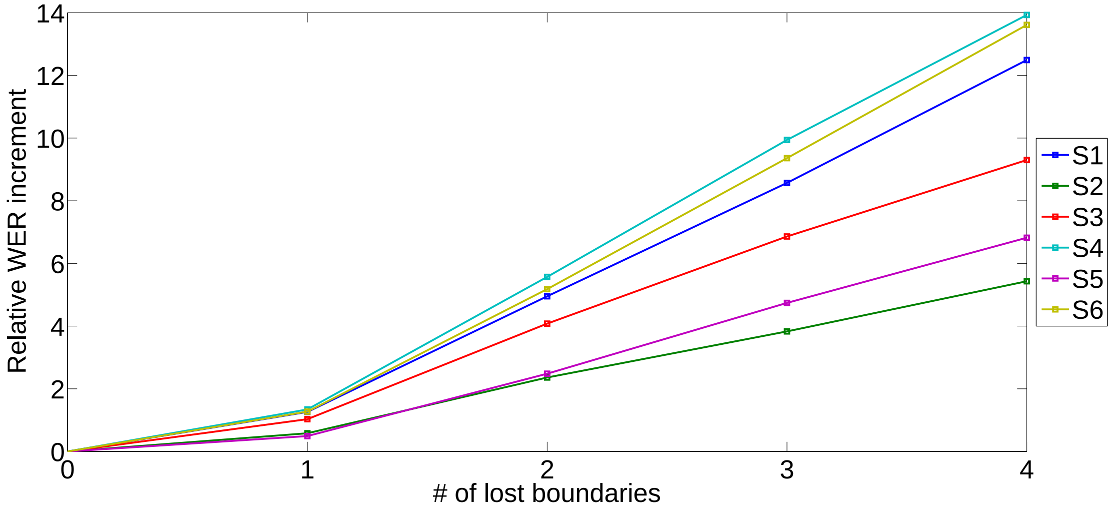


Figure 6.24: WER variation as a result of Boundary Loss + Segment Validation. Results are presented for each speaker S_i. These results correspond to experiments on simulated data.

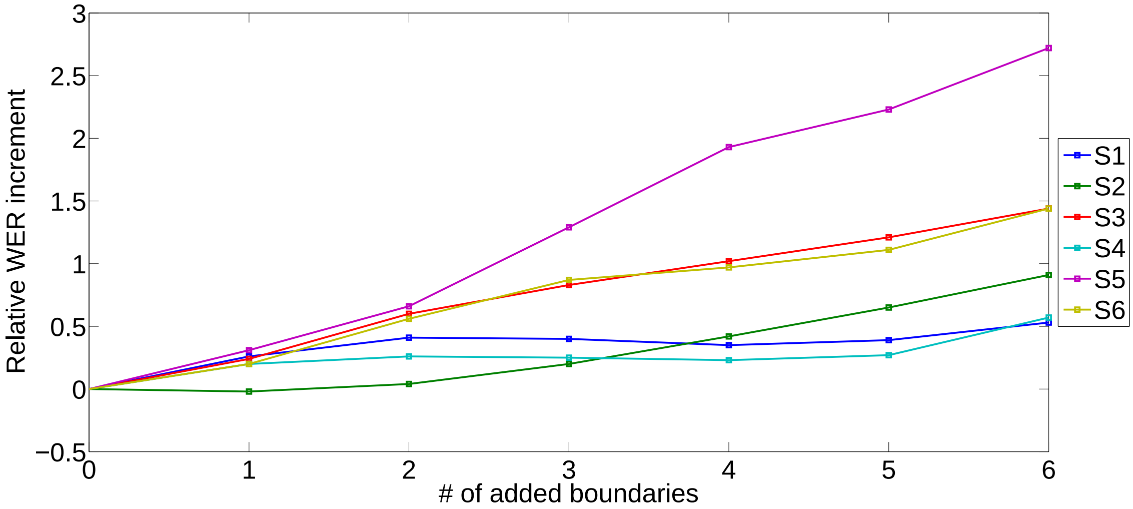


Figure 6.25: WER variation as a result of Adding Boundaries + Segment Validation. Results are presented for each speaker S_i. These results correspond to experiments on real data.

Table B.10 reports the results of SIM and REAL datasets, which are summarized in Figure 6.25.

Results

As described in Chapter 5 in the description of MMCN, after the identification of the boundaries, the validation of the segments defined by these boundaries is performed. Although no specific recovery is implemented for the case of a lost boundary, the analysis on which MMCN is based evidenced the presence of additional boundaries in the identification step. Therefore, even if a precise boundary is lost, the likelihood of having another boundary in the vicinity of the correct one is high. This assumption was confirmed in the sensitivity test of boundary addition, where the increment of errors is not so pronounced. The other evaluated condition, boundary shifting, showed that it also does not critically affect the performance of the algorithm. In fact, in boundary addition or shift conditions, the MMCN algorithm recovers partial information from the lattice content.

In the detailed results, found in Appendix B, it can be observed that the performance of blind MMCN, when no ground-truth boundaries were used, lead to a performance reduction of only 1% under the one achieved when the whole set of ground-truth boundaries was used.

6.4.4 Conclusions

The results confirmed those found in the previous related, experimental activities. MMCN performance is comparable to that of other hypothesis combination methods, and in all cases exceeded the average performance of the recognition of individual channels. Although CS was not the main core of this study, some results are presented with a signal-based CS method. An open issue arose from the results found in this set of experiments concerning a future study of the influence that speech content may have over CS methods.

Concerning the sensitivity of MMCN to specific issues, we found that loosing boundaries was more detrimental than identifying additional ones. For the latter, the Segment Validation mechanism provides a partial recovery of the correct boundaries. The performance of no-cheating MMCN, when no ground-truth boundaries were used, decreased only 1% than the one achieved when the whole set of ground-truth boundaries was used. MMCN performed similar to loosing only one of the ground-truth boundaries, and having the rest of the boundaries correctly identified. Shifting boundaries is even less detrimental than adding boundaries. Boundary variations, derived from a decoding process, are inherent to speaker characteristics (e.g., speech rate).

6.5 Summary

In this chapter, we presented MMCN, a method for extracting a confusion network through the combination of multiple lattices derived from various microphone signals. MMCN enables the estimation of temporal information to associate to each of the confusion sets. Unlike standard approaches for hypothesis combination, MMCN does not rely on dynamic programming based alignments, and therefore is not affected by the order in which the elements are combined.

Experimental evaluations performed with the proposed method, and other state-of-the-art selection and combination techniques, are summarized. The comparison analysis included signal and hypothesis based methods. The experimental setup was a domestic environment equipped with multiple largely distributed microphones. The datasets used in the evaluation were extracted from the DIRHA Project.

The experimental results suggest that MMCN performance is comparable to state-of-the-art techniques such as ROVER or CNC, in a multi-microphone setting. The variation of the language model, which has a direct effect over the structure of the word graph, did not affect the general trend of performance of the different techniques. The same observation applies to other recognition parameters explored, such as vocabulary size or spoken language. No critical differences were found concerning the lattice content when using different ASR toolkits. This evidence suggests that the proposed method can be exploited also with other speech recognition system parameters and configurations.

Finally, various sensitivity tests were performed on the proposed method, MMCN. These results evidenced the robustness of the algorithm against issues such as the misidentification of word boundaries , but also reflected its limitations for recovering from a loss of boundary.

Chapter 7

Conclusions and future work

A selective memory for remembering the good things,
... and defiant optimism for facing the future.

from *"The sum of our days"* by Isabel Allende

Extensive literature supports the paradigm of fusion of information for the improvement of the automatic transcription of a spoken phrase pronounced at a certain distance from the microphone. Conventionally, these sources of information to fuse were extracted from a single signal, through operations performed by multiple processing systems. More recently, information extracted from multiple microphones has been exploited. This latter approach has proven to benefit not only speech recognition, but also other acoustic processing areas. This dissertation elaborates on information fusion approaches for the enhancement of distant-talking recognition in a distributed multi-microphone setting. It is argued that such an improvement can be achieved through a proper manipulation of the information extracted from each microphone at different stages of the recognition system. Two research problems are investigated: channel selection and hypothesis combination.

The different components of a recognition system, and the different processing stages that occur during the automatic transcription of speech, are examined. Understanding these system characteristics is fundamental for studying the problem of multi-microphone speech recognition. The most relevant work and the conventional solutions that have been explored in the scientific community are presented and discussed. Among the different existing research lines, particular emphasis is given to channel selection and hypothesis combination, which constitute valid approaches for the microphone setting that is addressed in this work.

Channel Selection

A framework which exploits cepstral distance is proposed to address the problem of CS. The proposed framework introduces various novelties, i) the classification of CS methods into informed and blind methods, ii) evaluation metrics that exploit informed CS methods as upperbounds for assessing CS performance, which results in a deeper understanding of the strengths and limitations of CS methods, and iii) the consideration of an objective measure, concretely cepstral distance, for the implementation of signal-based CS. Moreover, the extension of the method to other objective measures is introduced. A set of experimental analyses are presented to investigate the interactions between objective measures and acoustic characteristics, such as reverberation time, and also to study the relation between objective measures and speech recognition. Such explorations were performed under multiple scenarios and acoustic conditions. The results of these experiments show the solid benefits of the proposed CS method.

Through a series of experimental cases, we have proved that the proposed CS method can be considered a suitable solution for CS in largely distributed multi-microphone settings at reverberant scenarios. The main achievements of the proposed blind CS method are: it improves in all cases the average SDM WER, consistently outperforms the state-of-the-art EV-based CS method and successfully selects the least distorted channel, when sufficient room coverage is provided by the microphone network. The new CS metrics provide important information to the analysis of the performance of a CS method. Moreover, it is discussed how the standard evaluation of CS, based solely on WER, hides the strengths and weaknesses of different signal-based CS methods.

Various research activities can be organized from this point. First, a similar experimental analysis as the one presented here for reverberation time can be extended to other reverberation characteristics, such as for example direct-to-reverberant ratio, which is useful for various research lines other than CS. As a second point, it must be noticed that in the experiments only reverberation was considered to be the main source of degradation of distant speech, however, in a real scenario, environmental noise should also be addressed. In order to approach more realistic conditions, the proposed solution as well as other existing CS methods, can be studied at scenarios which include different types of noisy conditions. In addition, given the varied strengths presented by the different objective quality measures found in the literature, it is interesting to study the use of other measures for CS, such as those also reported here, both in an informed and blind fashion. Another topic derived from this study, concerns the exploration of complex conditions, e.g., speakers directed to corners or areas where no microphone is present. Understanding at signal and recognition level the various implications of such scenarios, in order to propose robust CS solutions is still unresolved. A possible direction towards this goal involves a multi-step approach where, first, these complex conditions are identified, and then, novel techniques for CS are applied.

Hypothesis Combination

Although higher in complexity than signal-based approaches, hypothesis combination methods offer an undeniable flexibility for allowing the combination of information sources, independent of the physical arrangement of the microphones. This means also a high applicability potential, of hypothesis combination approaches, into future smart-scenarios, where sensors are largely distributed in a space. State-of-the-art solutions operate on the individual hypotheses or transformation of the hypothesis-spaces. This fact introduces limitations or processing errors to the resulting outcome. In this context, we proposed MMCN, a mechanism for combining hypothesis spaces directly at lattice level.

The proposal is based on four main elements: the time points, the contents of the lattices, the confidence scores, and the agreement between the different microphone-derived lattices. The rationale of how these elements are integrated into a novel combination approach, is provided. For practical purposes, the proposed method is implemented as an extension of the SRILM toolkit platform, which can be applied over SLF lattices. The lattice pruning degree was found to be a relevant factor affecting the proposed method. Strong pruning reduces the information available in the hypothesis space on which the method operates, a fact that results in a performance reduction.

The sensitivity tests applied on MMCN allow us to understand how well the method reacts to inaccuracies introduced by its different processing modules. A stronger negative effect was observed by the loss of boundaries, since the method does not offer any recovery mechanism for them at the present moment. One of these blind sensitivity tests evidenced that MMCN performs similarly as having all but one of the reference boundaries. This evidence reveals the efficiency of the identification of boundaries, but also shows an open window for improvement particularly on the validation of segments. Given that the method operates on top of the outcomes resulting from a decoding process, features such as the speaker speech-rate affect the performance of the solution.

Experimental evaluations of MMCN and of other hypothesis and signal combination or selection techniques are discussed. The goal of these evaluations is to measure the performance of the different techniques under variations of recognition system parameters. The use of a LM or a simple word-loop, which directly influences the architecture of a hypothesis space, is one of the main parameters adopted in the tests. This feature has a significant impact over the performance of the recognition process, but interestingly the performance of the proposed method shows no variation. Languages, phoneme sets, synthetic or real datasets, are all conditions in which the different methods are evaluated. In all of these conditions, the results achieved by the proposed hypothesis combination method were comparable to those of state-of-the-art information fusion methods. While benefits are limited in terms of WER reduction, there are advantages over other approaches, particularly with relation to the complexity of running multiple combinations.

The use of the agreement of temporal information among microphones shows positive effects for the proposed method. It is possible to extract the resulting CN featuring a pair of time boundaries associated to each confusion set. Due to inaccuracies inherent to the nature of the elaboration of lattices, the temporal boundaries extracted from the inter-microphone agreement are not precise though. A solution to this problem is not clearly devised, since it relies on the approximations made by the decoding process. Physical phenomena such as reverberation make the identification of intra-word boundaries prone to errors in DSR.

The proposed information fusion method requires no particular training, and is flexible in the sense that it can be extended to any multi-microphone condition, without any specific a priori information about the arrangement of the sensors. Although a single experimental scenario was employed for this study, various microphone settings are adopted for the evaluations, in order to explore and understand the impact of having multiple sensors in the combinations. The quantity of sensors was not the main relevant factor, but the quality of the information incorporated in the lattices. This left an open area to explore for further improvement, the selection of the best sources of information which is of high relevance for hypothesis combinatorial solutions. Studying decoder based hypothesis-ranking or selection approaches could take combinatorial approaches to a further improvement. Such measures could also be used for weighting the participation of the different sources into the combination process. This however is not an easy task, since no common reference exists among multiple microphone lattices.

Results obtained with the proposed method prove that a coherence-based approach, operating directly over lattices, can lead to significant improvement of recognition, when compared to the performance of the individual microphone decoding results. Still some work is left to resolve the limitations not only of the proposed method but of the DSR recognition problem. The proposed approach advances the state-of-the-art as it provides a simplified method for extracting a recognition hypothesis, to the problem of DSR in a multi-microphone setting.

Perspectives on future work

The different results, presented in this dissertation, report the positive accomplishments of the multi-microphone information fusion methods proposed, to address the problems of channel selection and hypothesis combination. As previously discussed, there is still room for improvement within the approaches proposed, that will benefit not only the specifically addressed areas but also acoustic processing ones.

Further research is needed to clarify and formalize the relations between the acoustic setting and the reference we create in the proposed CS method. This particular topic, CS, can be explored not only targeting speech recognition but other signal related tasks. It is also relevant to expand CS work introduced in this thesis, to other objective signal quality measures. Concerning the proposed MMCN extraction, a wide spectrum of advance machine learning or statistical algorithms can support the development of smarter algorithms than those already included.

One possible immediate direction to follow, based on the proposed approaches developed in this thesis, considers the investigation of a hybrid/integrated solution for hypothesis combination that relies on the ranking of the CS methods. This rank information can be used, for example, to weight or arrange the microphone lattices or confusion networks. Other future directions can exploit the novel approaches for other acoustic related tasks.

All the different acoustic and speech processing solutions are intertwined. Acoustic scene analysis approaches, such as speaker localization and tracking or acoustic event detection, and channel selection methods share partial views of the acoustic scenario. Although much of the existing work on these an other related areas has progressed independently, more integrative approaches are foreseen. Solutions, as those proposed in this work at front-end level, can operate together with acoustic scene analysis methods in order to achieve a better realization of the acoustic complex reality, and therefore propose novel solutions for the multiple problems still open.

Bibliography

Aarabi, P. The fusion of distributed microphone arrays for sound localization. *EURASIP Journal on Advances in Signal Processing*, 2003(4):1–10, 2003.

Abida, K., Karray, F., and Abida, W. cROVER: Improving ROVER using automatic error detection. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 1753–1756. IEEE, 2011.

Allen, J. B. and Berkley, D. A. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

Alvarado, V. M. *Talker Localization and Optimal Placement of Microphones for a Linear Microphone Array Using Stochastic Region Contraction.* PhD thesis, Brown University, Providence, 1990.

AMI-EU. “Augmented Multi-party Interaction (AMI) EU Project”. [online] Available: http://www.amiproject.org/.

Angelini, B., Brugnara, F., Falavigna, D., Giuliani, D., Gretter, R., and Omologo, M. Speaker independent continuous speech recognition using an acoustic-phonetic italian corpus. *Proc. of International Conference on Spoken Language*, pages 1391–1394, 1994.

Anguera, X. Beamformit. http://www.xavieranguera.com/beamformit/, 2006.

Anguera, X., Woofers, C., and Hernando, J. Speaker diarization for multi-party meetings using acoustic fusion. In *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, pages 426–431. IEEE, 2005.

Anguera, X., Wooters, C., and Hernando, J. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7): 2011–2022, 2007.

Atal, B. S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America*, 55(6):1304–1312, 1974.

Atal, B. S. and Hanauer, S. L. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50(2B):637–655, 1971.

Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. Multimodal fusion for multimedia analysis: A survey. *Multimedia systems*, 16(6):345–379, 2010.

Aubert, X. and Ney, H. Large vocabulary continuous speech recognition using word graphs. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 49–52. IEEE, 1995.

Audhkhasi, K., Zavou, A. M., Georgiou, P. G., and Narayanan, S. S. Theoretical analysis of diversity in an ensemble of automatic speech recognition systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(3):711–726, 2014.

Bahl, L. R. and Jelinek, F. Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition. *IEEE Transactions on Information Theory*, 21(4):404–411, 1975.

Bahl, L. R., Jelinek, F., and Mercer, R. L. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):179–190, 1983.

Barker, J., Marxer, R., Vincent, E., and Watanabe, S. The Third CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines. In *2015 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU*, 2015.

Basha, T. G., Sridevi, P., and Prasad, M. G. Enhancement in gain and interference of smart antennas using two stage genetic algorithm by implementing it on beam forming. *International Journal of Electronics Engineering*, 3(2):265–269, 2011.

Bellegarda, J. R. and Monz, C. State of the art in statistical methods for language and speech processing. *Computer Speech & Language*, 35:163–184, 2016.

Benesty, J., Makino, S., and Chen, J. *Speech Enhancement*. Springer, Berlin, Germany, 2005.

Benesty, J., Sondhi, M. M., and Huang, Y. *Springer Handbook of Speech Processing.* Springer Science & Business Media, 2007.

Bertrand, A., Doclo, S., Gannot, S., Ono, N., and Van Waterschoot, T. Special issue on wireless acoustic sensor networks and ad hoc microphone arrays. *Signal Process.*, 107 (C):1–3, February 2015. ISSN 0165-1684. doi: 10.1016/j.sigpro.2014.10.001.

Bishop, C. M. Pattern recognition. *Machine Learning*, 128, 2006.

Blauert, J. *Spatial hearing: the psychophysics of human sound localization.* MIT press, 1997.

Bogert, B. P., Healy, M. J., and Tukey, J. W. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In *Proc. of the symposium on time series analysis*, volume 15, pages 209–243. chapter, 1963.

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

Breslin, C. *Generation and combination of complementary systems for automatic speech recognition.* PhD thesis, University of Cambridge, Cambridge, 2008.

Brutti, A., Omologo, M., and Svaizer, P. An environment aware ML estimation of acoustic radiation pattern with distributed microphone pairs. *Signal Processing*, 93(4):784–796, 2013.

Capon, J. High-resolution frequency-wavenumber spectrum analysis. *Proc. of the IEEE*, 57(8):1408–1418, 1969.

Chen, J., Benesty, J., Huang, Y. A., and Diethorn, E. J. Fundamentals of noise reduction. In *Springer Handbook of Speech Processing*, pages 843–872. Springer, 2008.

Chen, S. F., Kingsbury, B., Mangu, L., Povey, D., Saon, G., Soltau, H., and Zweig, G. Advances in Speech Transcription at IBM under the DARPA EARS Program. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1596–1608, 2006.

CHIL-EU. "The CHIL (Computers in the Human Interaction Loop) EU Project". [online] Available: http://chil.server.de/ http://shine.fbk.eu/it/node/14.

Cohen, I., Berdugo, B., and Marash, J. Real-time microphone selection in noisy reverberant environments for teleconferencing systems. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume Show and Tell. IEEE, 2014.

Cossalter, M., Sundararajan, P., and Lane, I. R. Ad-hoc meeting transcription on clusters of mobile devices. In *Proc. of INTERSPEECH*, pages 2881–2884, 2011.

Cristoforetti, L., Ravanelli, M., Omologo, M., Sosi, A., Abad, A., Hagmüller, M., and Maragos, P. The DIRHA simulated corpus. In *9th edition of the Language Resources and Evaluation Conference*, pages 2629–2634, Reykjavik, Iceland, 2014.

Cui, J., Cui, X., Ramabhadran, B., Kim, J.-H., Kingsbury, B., Mamou, J., Mangu, L., Picheny, M., Sainath, T. N., and Sethy, A. Developing speech recognition systems for corpus indexing under the iarpa babel program. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 6753–6757. IEEE, 2013.

Davis, K., Biddulph, R., and Balashek, S. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952.

Davis, S. B. and Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

De La Torre, A., Segura, J. C., Benitez, C., Peinado, A. M., and Rubio, A. J. Non-linear transformations of the feature space for robust speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages I–401. IEEE, 2002.

Deoras, A. and Jelinek, F. Iterative decoding: A novel re-scoring framework for confusion networks. In *IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU*, pages 282–286. IEEE, 2009.

DIRHA-EU. "The DIRHA (Distance-Speech Interaction for Robust Home Applications) EU Project". [online] Available: http://dirha.fbk.eu/.

Doumpiotis, V. and Byrne, W. Pinched lattice minimum bayes risk discriminative training for large vocabulary continuous speech recognition. In *Proc. of INTERSPEECH*, 2004.

Droppo, J. and Acero, A. Environmental robustness. In *Springer Handbook of Speech Processing*, pages 653–680. Springer, 2008.

Du, J., Wang, Q., Yan-Hui, T., Bao, X., Dai, L.-R., and Lee, C.-H. An information fusion approach to recognizing microphone array speech in the CHiME-3 challenge based on a deep learning framework. In *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, pages 430–435. IEEE, 2015.

Elko, G. W. and Meyer, J. Microphone arrays. In *Springer Handbook of Speech Processing*, pages 1021–1041. Springer, 2008.

Ephraim, Y. and Malah, D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121, 1984.

Evermann, G. and Woodland, P. Posterior probability decoding, confidence estimation and system combination. In *Proc. NIST Speech Transcription Workshop*, volume 27. Baltimore, 2000.

Falavigna, D., Gretter, R., and Riccardi, G. Acoustic and word lattice based algorithms for confidence scores. In *Proc. of INTERSPEECH*, 2002.

Falk, T. H., Zheng, C., and Chan, W.-Y. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1766–1774, 2010.

Feng, X., Kumatani, K., and McDonough, J. The CMU-MIT REVERB challenge 2014 system: description and results. In *Proc. of REVERB Challenge Workshop, p1*, volume 9. Citeseer, 2014.

Fiscus, J. G. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Workshop on ASR and Understanding*, pages 347–354. IEEE, dec 1997. doi: 10.1109/ASRU.1997.659110.

Flanagan, J., Johnston, J., Zahn, R., and Elko, G. Computer-steered microphone arrays for sound transduction in large rooms. *The Journal of the Acoustical Society of America*, 78(5):1508–1518, 1985.

Forgie, J. W. and Forgie, C. D. Results obtained from a vowel recognition computer program. *The Journal of the Acoustical Society of America*, 31(11):1480–1489, 1959.

Fujita, Y., Takashima, R., Homma, T., Ikeshita, R., Kawaguchi, Y., Sumiyoshi, T., Endo, T., and Togami, M. Unified ASR system using LGM-based source separation, noise-robust feature extraction and word hypothesis selection. In *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, pages 416–423. IEEE, 2015.

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

Galatas, G., Potamianos, G., and Makedon, F. Audio-visual speech recognition incorporating facial depth information captured by the kinect. In *Proceedings of the 20th European Signal Processing Conference, EUSIPCO*, pages 2714–2717. IEEE, 2012.

Gales, M. J., Liu, X., Sinha, R., Woodland, P. C., Yu, K., Matsoukas, S., Ng, T., Nguyen, K., Nguyen, L., Gauvain, J.-L., et al. Speech recognition system combination for machine translation. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 4, pages IV–1277. IEEE, 2007.

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

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. DARPA TIMIT acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon Technical Report N*, 93:27403, 1993.

GaticaPerez, D., Lathoud, G., Odobez, J.-M., and McCowan, I. Multimodal multispeaker probabilistic tracking in meetings. In *International Conference on Multimodal Interfaces*, pages 183–190, 2005. doi: 10.1145/1088463.1088496.

Gilkey, R. and Anderson, T. R. *Binaural and spatial hearing in real and virtual environments.* Psychology Press, 2014.

Goel, V. and Byrne, W. J. Minimum bayes-risk automatic speech recognition. *Computer Speech & Language*, 14(2):115–135, 2000.

Goel, V., Kumar, S., and Byrne, W. Segmental minimum bayes-risk ASR voting strategies. In *Proc. of INTERSPEECH*, pages 139–142, 2000.

Gong, Y. Speech recognition in noisy environments: A survey. *Speech Communication*, 16(3):261–291, 1995.

Gray, A. and Markel, J. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5):380–391, Oct 1976.

Guerrero, C. and Omologo, M. Exploiting inter-microphone agreement for hypothesis combination in distant speech recognition. In *Proceedings of the 22nd European Signal Processing Conference, EUSIPCO*, pages 2385–2389. IEEE, 2014a.

Guerrero, C. and Omologo, M. Word boundary agreement to combine multi-microphone hypothesis in distant speech recognition. In *Joint Workshop Hands-free Speech Communication and Microphone Arrays, HSCMA*. IEEE, 2014b.

Guerrero, C., Tryfou, G., and Omologo, M. Channel selection for distant speech recognition - exploiting cepstral distance. In *Proc. of INTERSPEECH*, 2016.

Guo, G., Huang, C., Jiang, H., and Wang, R.-H. A comparative study on various confidence measures in large vocabulary speech recognition. In *Chinese Spoken Language Processing, 2004 International Symposium on*, pages 9–12. IEEE, 2004.

Hakkani-Tür, D., Béchet, F., Riccardi, G., and Tur, G. Beyond ASR 1-best: Using word confusion networks in spoken language understanding. *Computer Speech & Language*, 20(4):495–514, 2006.

Handel, S. *Listening: An introduction to the perception of auditory events.* The MIT Press, 1993.

Hansen, J. H. and Pellom, B. L. An effective quality evaluation protocol for speech enhancement algorithms. In *International Conference on Spoken Language Processing, ICSLP*, volume 7, pages 2819–2822. Citeseer, 1998.

Harper, M. The automatic speech recognition in reverberant environments (ASpIRE) Challenge. In *IEEE Workshop on Automatic Speech Recognition & Understanding.* IEEE, 2015.

Hermansky, H. Perceptual Linear Predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

Hermansky, H. and Morgan, N. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994.

Hermansky, H., Tibrewala, S., and Pavel, M. Towards ASR on partially corrupted speech. In *Proc. Fourth International Conference on Spoken Language, 1996.*, volume 1, pages 462–465. IEEE, ISCA, 1996. URL `http://www.isca-speech.org/archive/icslp_1996/i96_0462.html`.

Hillard, D., Hoffmeister, B., Ostendorf, M., Schlüter, R., and Ney, H. i ROVER: improving system combination with classification. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 65–68. Association for Computational Linguistics, 2007.

Himawan, I., Motlicek, P., Sridharan, S., Dean, D., and Tjondronegoro, D. Channel selection in the short-time modulation domain for distant speech recognition. In *Proc. of INTERSPEECH*, number EPFL-CONF-209075, 2015.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Hoffmeister, B., Klein, T., Schlüter, R., and Ney, H. Frame based system combination and a comparison with weighted rover and cnc. In *Proc. of INTERSPEECH*. Citeseer, 2006.

Hu, Y. and Loizou, P. C. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238, 2008.

Huang, X., Acero, A., and Hon, H.-W. *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice Hall PTR, 2001.

Huang, Y. A. and Benesty, J. *Audio signal processing for next-generation multimedia communication systems*. Springer Science & Business Media, 2007.

Huang, Y. A., Benesty, J., and Chen, J. Dereverberation. In *Springer Handbook of Speech Processing*, pages 929–944. Springer, 2008.

Itakura, F. and Saito, S. A statistical method for estimation of speech spectral density and formant frequencies. *Electron. Commun. Japan*, 53(1):36–43, 1970.

Jalalvand, S., Negri, M., Falavigna, D., and Turchi, M. Driving ROVER with segment-based ASR quality estimation. In *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 1095–1105, 2015.

Janin, A., Ang, J., Bhagat, S., Dhillon, R., Edwards, J., Macias-Guarasa, J., Morgan, N., Peskin, B., Shriberg, E., Stolcke, A., Wooters, C., and Wrede, B. The ICSI Meeting Project: Resources and Research. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICAASP -Meeting Recognition Workshop (NIST RT-04 Spring Recognition Evaluation)*, 2004.

Jelinek, F. *Statistical methods for speech recognition*. MIT press, 1997.

Jiang, H. Confidence measures for speech recognition: A survey. *Speech communication*, 45(4):455–470, 2005.

Jurafsky, D. and Martin, J. H. *Speech & language processing*. Pearson Education India, 2000.

Kemp, T., Schaaf, T., et al. Estimating confidence using word lattices. In *EuroSpeech*, 1997.

Kenny, P. A small footprint i-vector extractor. In *Odyssey*, pages 1–6, 2012.

Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., and Maas, R. 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*, pages 1–4. IEEE, 2013.

Kitawaki, N., Nagabuchi, H., and Itoh, K. Objective quality evaluation for low-bit-rate speech coding systems. *IEEE Journal on Selected Areas in Communications*, 6(2): 242–248, Feb 1988. ISSN 0733-8716. doi: 10.1109/49.601.

Kolossa, D., Klimas, A., and Orglmeister, R. Separation and robust recognition of noisy, convolutive speech mixtures using time-frequency masking and missing data techniques. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 82–85. IEEE, 2005.

Kumatani, K., McDonough, J., and Raj, B. Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. *Signal Processing Magazine, IEEE*, 29(6):127–140, 2012. ISSN 1053-5888. doi: 10.1109/MSP.2012.2205285.

Kumatani, K., McDonough, J., Lehman, J. F., and Raj, B. Channel selection based on multichannel cross-correlation coefficients for distant speech recognition. In *Joint*

Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA), pages 1–6. IEEE, 2011.

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

Lamel, L. and Gauvain, J.-L. Alternate phone models for conversational speech. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 1005–1008, 2005.

Le Roux, J. and Vincent, E. A categorization of robust speech processing datasets. Technical Report TR2014-116, Mitsubishi Electric Research Labs, Cambridge, MA, USA, August 2014. v2014-09.

Lecouteux, B., Linares, G., Esteve, Y., and Gravier, G. Generalized driven decoding for speech recognition system combination. In *International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 1549–1552. IEEE, 2008.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015.

Leggetter, C. J. and Woodland, P. C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9(2):171–185, 1995.

Levenshtein, V. I. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707, 1966.

Li, B. and Sim, K. C. Improving robustness of deep neural networks via spectral masking for automatic speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, pages 279–284. IEEE, 2013.

Li, X., Singh, R., and Stern, R. M. Lattice combination for improved speech recognition. In *Proc. of International Conference on Spoken Language Processing*, 2002.

Lim, J. S. and Oppenheim, A. V. Enhancement and bandwidth compression of noisy speech. *Proc. of the IEEE*, 67(12):1586–1604, 1979.

Linde, Y., Buzo, A., and Gray, R. An algorithm for vector quantizer design. *IEEE Transactions on communications*, 28(1):84–95, 1980.

Liu, X., Gales, M. J., and Woodland, P. C. Language model cross adaptation for lvcsr system combination. *Computer Speech & Language*, 2012.

Liu, Y., Harper, M. P., Johnson, M. T., and Jamieson, L. H. The effect of pruning and compression on graphical representations of the output of a speech recognizer. *Computer Speech & Language*, 17(4):329–356, 2003.

Loizou, P. C. *Speech enhancement: theory and practice.* CRC press, 2013.

Ma, C., Kuo, H., Soltau, H., Cui, X., Chaudhari, U., Mangu, L., and Lee, C.-H. A comparative study on system combination schemes for LVCSR. In *International Conference on Acoustics Speech and Signal Processing, ICASSP*, pages 4394–4397. IEEE, 2010.

Mangu, L., Brill, E., and Stolcke, A. Finding consensus among words: lattice-based word error minimization. In *EUROSPEECH.* ISCA, 1999.

Mangu, L., Brill, E., and Stolcke, A. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4):373–400, 2000.

Matassoni, M., Omologo, M., Giuliani, D., and Svaizer, P. Hidden Markov Model training with contaminated speech material for distant-talking speech recognition. *Computer Speech & Language*, 16(2):205–223, 2002.

Matassoni, M., Astudillo, R. F., Katsamanis, A., and Ravanelli, M. The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones. In *Proc. of INTERSPEECH*, pages 1613–1617, 2014.

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

Mporas, I., Ganchev, T., Siafarikas, M., and Fakotakis, N. Comparison of speech features on the speech recognition task. *Journal of Computer Science*, 3(8):608–616, 2007 2007.

Myers, C., Rabiner, L., and Rosenberg, A. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(6):623–635, 1980.

Nagata, K., Kato, Y., and Chiba, S. Spoken digit recognizer for japanese language. In *Audio Engineering Society Convention 16.* Audio Engineering Society, 1964.

Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., and Vergyri, D. Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins Summer 2000 Workshop. In *Fourth Workshop on Multimedia Signal Processing*, pages 619–624. IEEE, 2001.

NIST. NIST-SCTK Speech Recognition Scoring Toolkit. http://www.nist.gov/speech/tools/, 2009.

Obuchi, Y. Multiple-microphone robust speech recognition using decoder-based channel selection. In *ISCA Tutorial and Research Workshop (ITRW) on Statistical and Perceptual Audio Processing*, 2004.

Obuchi, Y. Noise robust speech recognition using delta-cepstrum normalization and channel selection. *Electronics and Communications in Japan (Part II: Electronics)*, 89(7): 9–20, 2006.

Oerder, M. and Ney, H. Word graphs: an efficient interface between continuous-speech recognition and language understanding. In *International Conference on Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE*, volume 2, pages 119–122 vol.2, April 1993. doi: 10.1109/ICASSP.1993.319246.

Omologo, M., Matassoni, M., Svaizer, P., and Giuliani, D. Microphone array based speech recognition with different talker-array positions. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 227–230 vol.1, Apr 1997. doi: 10.1109/ICASSP.1997.599610.

Openshaw, J. P. and Masan, J. On the limitations of cepstral features in noise. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 2, pages II–49. IEEE, 1994.

Oppenheim, A. v., Schafer, R., and Stockham, T. Nonlinear filtering of multiplied and convolved signals. *IEEE Transactions on Audio and Electroacoustics*, 16(3):437–466, 1968.

Ortmanns, S., Ney, H., and Aubert, X. A word graph algorithm for large vocabulary continuous speech recognition. *Computer Speech & Language*, 11(1):43–72, 1997.

Pedersen, M. S., Larsen, J., Kjems, U., and Parra, L. C. A survey of convolutive blind source separation methods. In *Springer Handbook of Speech Processing*. Springer Press, nov 2008.

Peterson, P. M. Simulating the response of multiple microphones to a single acoustic source in a reverberant room. *The Journal of the Acoustical Society of America*, 80(5): 1527–1529, 1986.

Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A. Recent advances in the automatic recognition of audiovisual speech. *Proc. of the IEEE*, 91(9):1306–1326, 2003.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, number EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

Quackenbush, S. R., Barnwell, T. P., and Clements, M. A. *Objective Measures of Speech Quality*. Prentice-Hall, Englewood Cliffs, NJ, 1988.

Rabiner, L., Levinson, S., Rosenberg, A., and Wilpon, J. Speaker-independent recognition of isolated words using clustering techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(4):336–349, 1979.

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

Rabiner, L. and Juang, B.-H. Fundamentals of speech recognition. *Signal Processing Series. Prentice Hall, Englewood Cliffs, NJ*, 1993.

Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, 1989.

Rabiner, L. R. and Schafer, R. W. Introduction to digital speech processing. *Foundations and trends in signal processing*, 1(1):1–194, 2007.

Rabinkin, D. V., Renomeron, R. J., French, J. C., and Flanagan, J. L. Estimation of wavefront arrival delay using the cross-power spectrum phase technique. In *132nd Meeting of the Acoustical Society of America*, volume 100, page 2697. Citeseer, 1996.

Rabinkin, D. V., Renomeron, R. J., French, J. C., and Flanagan, J. L. Optimum microphone placement for array sound capture. In *Optical Science, Engineering and Instrumentation'97*, pages 227–239. International Society for Optics and Photonics, 1997.

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

Recommendation, I. Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs ". *ITU-T Recommendation*, page 862, 2001.

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 2, pages 749–752. IEEE, 2001.

Rohdenburg, T., Hohmann, V., and Kollmeier, B. Objective perceptual quality measures for the evaluation of noise reduction schemes. In *9th international workshop on acoustic echo and noise control*, pages 169–172, 2005.

Schaaf, T. and Kemp, T. Confidence measures for spontaneous speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 2, pages 875–878. IEEE, 1997.

Schlüter, R., Zolnay, A., and Ney, H. Feature combination using linear discriminant analysis and its pitfalls. In *Proc. of INTERSPEECH*. Citeseer, 2006.

Schwartz, R. and Chow, Y.-L. The n-best algorithms: an efficient and exact procedure for finding the n most likely sentence hypotheses. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 81–84. IEEE, 1990.

Schwenk, H. and Gauvain, J.-L. Combining multiple speech recognizers using voting and language model information. In *Proc. of INTERSPEECH*, pages 915–918, 2000.

Shimizu, Y., Kajita, S., Takeda, K., and Itakura, F. Speech recognition based on space diversity using distributed multi-microphone. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 3, pages 1747–1750. IEEE, 2000.

Stallard, D., Choi, F., Kao, C.-L., Krstovski, K., Natarajan, P., Prasad, R., Saleem, S., and Subramanian, K. The bbn 2007 displayless english/iraqi speech-to-speech translation system. In *Proc. of INTERSPEECH*, pages 2817–2820, 2007.

Stern, R. and Morgan, N. Hearing is believing: Biologically-inspired feature extraction for robust automatic speech recognition. *IEEE Signal Processing Magazine*, 29(34-43): 170, 2012.

Stolcke, A. SRILM – An extensible language modeling toolkit. In *Proc. of International Conference on Spoken Language Processing*, pages 901–904, 2002.

Stolcke, A. Making the most from multiple microphones in meeting recognition. In *Acoustics, Speech and Signal Processing, 2011 IEEE International Conference on*, pages 4992–4995. IEEE, 2011.

Stolcke, A., Konig, Y., and Weintraub, M. Explicit word error minimization in N-best list rescoring. In *Eurospeech*, volume 97, pages 163–166, 1997.

SWEETHOME-ANR. "Sweet Home (Système Domotique d'Assistance au Domicile) ANR Project". [online] Available: http://sweet-home.imag.fr/.

Tribolet, J. M., Noll, P., McDermott, B. J., and Crochiere, R. E. A study of complexity and quality of speech waveform coders. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 3, pages 586–590. IEEE, 1978.

Tür, G., Stolcke, A., Voss, L. L., Dowding, J., Favre, B., Fernández, R., Frampton, M., Frandsen, M. W., Frederickson, C., Graciarena, M., et al. The CALO meeting speech recognition and understanding system. In *SLT*, pages 69–72, 2008.

Vesely̌, K., Ghoshal, A., Burget, L., and Povey, D. Sequence-discriminative training of deep neural networks. In *Proc. of INTERSPEECH*, pages 2345–2349, 2013.

Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., and Matassoni, M. The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing, ICAASP*, pages 126–130. IEEE, 2013.

Vintsyuk, T. K. Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis*, 4(1):52–57, 1968.

Wessel, F., Schlüter, R., Macherey, K., and Ney, H. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9:288–298, 2001.

Williams, G. A study of the use and evaluation of confidence measures in automatic speech recognition. Technical report, 1998.

Wolf, M. *Channel selection and reverberation-robust automatic speech recognition.* PhD thesis, Universitat Politècnica de Catalunya, Barcelona, 2013.

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

Wolf, M. and Nadeu, C. Towards microphone selection based on room impulse response energy-related measures. In *Proc. of I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, pages 61–64, Porto Salvo, Portugal, 2009.

Wolf, M. and Nadeu, C. On the potential of channel selection for recognition of reverberated speech with multiple microphones. In *Proc. of INTERSPEECH*, pages 80–83, Tokyo, Japan, 2010.

Wölfel, M. Channel selection by class separability measures for automatic transcriptions on distant microphones. In *Proc. of INTERSPEECH*, pages 582–585. Citeseer, 2007.

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

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

Xue, J. and Zhao, Y. Improved confusion network algorithm and shortest path search from word lattice. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 853–856. IEEE, 2005.

Young, S. R. Detecting misrecognitions and out-of-vocabulary words. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, volume 2, pages II–21. IEEE, 1994.

Young, S. R. and Ward, W. Learning new words from spontaneous speech. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 590–591. IEEE, 1993.

Young, S. Generating multiple solutions from connected word DP recognition algorithms. *Proc. of the Institute of Acoustics*, 6(Part 4):351–354, 1984.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. *The HTK book*, volume 2. Entropic Cambridge Research Laboratory Cambridge, 1997.

Yu, H., Tam, Y.-C., Schaaf, T., Stüker, S., Jin, Q., Noamany, M., and Schultz, T. The ISL RT04 mandarin broadcast news evaluation system. In *EARS Rich Transcription Workshop*, 2004.

Zhang, R. and Rudnicky, A. I. Investigations of issues for using multiple acoustic models to improve continuous speech recognition. *International Conference on Spoken Language Processing*, 2006.

Zolnay, A., Schlüter, R., and Ney, H. Acoustic feature combination for robust speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 457–460. Citeseer, 2005.

Zwyssig, E., Ravanelli, M., Svaizer, P., and Omologo, M. A multi-channel corpus for distant-speech interaction in presence of known interferences. In *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pages 4480–4484. IEEE, 2015.

Appendix A

Speech recognition toolkits

This section describes the speech recognition platforms used in the experimental activities conducted for this dissertation.

A.1 HTK

The Hidden Markov Model Toolkit (HTK) is a toolkit that supports building HMMs, see http://htk.eng.cam.ac.uk/ [Young et al., 1997]. HTK's first version was developed at the Speech Vision and Robotics Group of the Cambridge University Engineering Department (CUED) in 1989. The initial target research area was speech recognition, however since its release it has been used for various applications, such as speech synthesis, and DNA sequencing.
The toolkit is composed by a set of C library modules and tools, which are designed to run with a traditional command line style interface. The tools facilitate speech analysis, HMM training, testing and analysis of decoding results. The platform supports the use of HMMs with continuous density mixture Gaussians or discrete distributions.
In the latest released version, HTK v3.5, the platform includes, among other resources, support for artificial neural network models, and the decoding of recurrent neural network language models.
Setting up the system requires the development of scripts calling HTK modules and tools. Basic examples and tutorials provided by the tool in order to implement a whole recognition system setup.

Lattices

Lattices provided a compressed mechanism for storing multiple hypotheses. In HTK the lattices are used as output of a speech recognition process, and to specify finite state syntax networks for recognition. HTK presented a Standard Lattice Format (SLF) which incorporates a core set of common features. Among other features, SLF includes: header definitions reporting information about the recognition process configuration (e.g., location of the language model, language model scale factor, word insertion penalty), node definitions -with optional fields such as time-, link definitions comprising start and end node -plus optional fields such as word identity, score information.

A.2 Kaldi

Kaldi is an open-source framework for ASR based on finite-state transducers. It is licensed under the open Apache License v2.0. The first version of Kaldi was released on 2011 [Povey et al., 2011], and it can be downloaded from http://kaldi.sf.net/.

One of the main characteristics of Kaldi is the modularity it offers, which allows the developers to configure specific settings, according to the needs of each project. Moreover, the modularity of the platform allows the extension of the toolkit. The main modules are implemented on C++. A general diagram of Kaldi's architecture is presented in Figure A.1.

The decoding core is based on finite-state transducers, for which linear algebra support is provided.

In its release, Kaldi includes several sets of scripts called *recipes*. These recipes facilitate the implementation of a complete speech recognition system, from training to testing routines, and are already adjusted to operate on widely available corpora and commonly addressed recognition tasks, e.g., Wall Street Journal, TIMIT.

Kaldi operations are multiple, including: speaker adaptive training (SAT), maximum likelihood linear regression, feature-space MLLR and maximum mutual information. Gaussian Mixture Models (GMM) and Subspace GMM capabilities are also provided. In addition, the training of Deep Neural Networks on top of GMM models with layer-wise pre-training based on Restricted Boltzmann Machines, per-frame cross-entropy training, and sequence-discriminative training are included in the toolkit. Considering the high demands required for the neural network training, Kaldi facilitates the implementation of parallel computing which significantly reduces the processing time.

Another clear advantage offered by Kaldi to the researchers and developers, is the active supporting community.

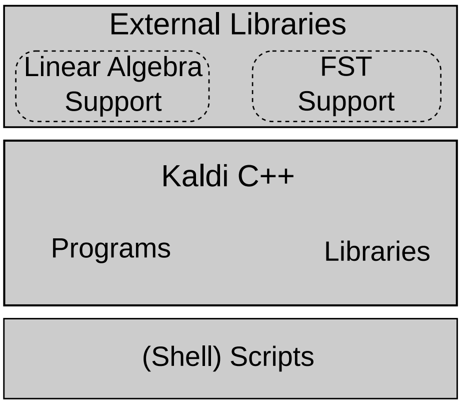


Figure A.1: Simplified view of the architecture of the Kaldi toolkit.

Appendix B

Multi-microphone processing methods: Experimental results

Here we present the results obtained with the different multi-microphone processing methods on the ASR experiments described in Chapter 6.

B.1 Decoding with N-grams: Case I

These tables report WER results for: SDM, Oracle, BF, CNC, and MMCN. Processes applied to sets of microphones are labeled 5', 10', 15' indicating the number of sensors. CNC reports 2 combinations per each microphone set.

Table B.1: WER results on the development set (part 1)

	Mic	beam80	beam100	unpruned
SDM	L1L	13.76	11.41	11.41
	L1C	17.11	17.11	16.11
	L1R	17.79	16.11	16.11
	L2L	10.07	9.40	9.40
	L2R	16.11	11.07	11.07
	L3L	9.06	6.38	6.38

Table B.2: WER results on the development set (part 2)

	Mic	beam80	beam100	unpruned
	L3R	13.76	10.74	10.07
	L4L	15.10	16.78	16.44
	L4R	14.09	12.08	12.08
	LA1	18.79	17.45	17.45
	LA2	15.44	15.44	15.44
	LA3	21.81	19.46	18.46
	LA4	19.13	17.45	17.45
	LA5	18.46	14.77	13.76
	LA6	14.77	12.42	12.42
Oracle	-	1.34	2.35	2.35
BF	5’	12.08	10.74	10.74
	10’	13.09	10.07	10.07
	15’	11.74	10.4	9.06
	5	8.72	7.72	7.72
CNC	5’	9.06	8.05	8.05
	10	9.73	8.72	9.06
	10’	10.07	9.73	10.07
	15	9.06	8.72	8.72
	15’	9.73	9.73	9.73
	5	8.72	7.72	7.72
MMCN	10	9.06	9.73	9.40
	15	9.40	9.40	9.40

Table B.3: WER results on the test set

	Mic	beam80	beam100
SDM	L1L	17.00	14.86
	L1C	17.41	15.07
	L1R	16.67	14.62
	L2L	16.90	14.66
	L2R	16.22	14.32
	L3L	17.18	14.68
	L3R	16.68	14.58
	L4L	18.14	16.52
	L4R	19.43	17.42
	LA1	18.59	16.22
	LA2	18.44	16.34
	LA3	17.09	15.14
	LA4	17.67	16.24
	LA5	17.06	15.7
	LA6	17.60	15.55
	Avg.	17.47	15.46
Oracle	-	5.13	4.73
BF	5’	18.78	14.74
	10’	16.21	15.21
	15’	15.75	14.02
CNC	5	13.72	12.22
	5’	13.71	12.18
	10	13.12	11.89
	10’	13.03	11.83
	15	13.16	12.02
	15’	13.22	12.03
MMCN	5	14.39	13.07
	10	14.18	12.72
	15	14.15	12.72

B.2 Decoding with N-grams: Case II

This table reports WER results for: SDM, Oracle, BF, BF on a circular array (BC), BF on a linear array (BL), ROVER, CNC, and MMCN. For ROVER and CNC the combinations are labeled $c1, .., c6$.

Table B.4: WER results on the simulation/real development/test sets. (part 1)

		SIM		REAL	
		Dev	Test	Dev	Test
SDM	L1C	21.8	17.5	29.6	25.8
	L2R	21.8	16.6	33.3	26.5
	L3L	22.1	17.6	30.3	23.9
	L4L	22.7	17.1	29.3	27.2
	LA6	22.8	16.7	29.9	25
	LD07	22.6	16.8	29.1	24
Oracle	-	12.4	9.4	19.5	13.7
BF	BC	20.7	14.5	27.3	23.4
	BL	20.8	14.7	26.4	19.3
ROVER	c1	17.8	12.8	24.9	19.5
	c2	17.8	12.8	25	19.8
	c3	17.7	12.9	24.6	19.7
	c4	17.7	12.8	24.5	20
	c5	18	12.9	24.3	19.6
	c6	17.7	12.8	24.2	19.5
	Avg	17.8	12.8	24.4	19.5

Table B.5: WER results on the simulation/real development/test sets (part 2).

		SIM		REAL	
		Dev	Test	Dev	Test
CNC	c1	20.1	14.2	26.6	20.9
	c2	20.2	14.3	26.7	20.9
	c3	20.1	14.2	26.7	20.7
	c4	20.1	14.0	26.7	20.7
	c5	20.2	14.1	26.6	20.8
	c6	20.1	14.0	26.4	20.5
	Avg	20.1	14.2	26.6	20.8
MMCN	-	20.2	14.3	26.5	20.8

.

B.3 Decoding with a word-loop

These tables report WER results for: SDM, Oracle, CS, ROVER, CNC, and MMCN. ROVER and CNC report 5 combination labeled c1..c5. Results are presented per speaker.

Table B.6: WER results on the development set - SIM

	Mic	Speaker					
		1	2	3	4	5	6
	CT	37.08	53.77	37.63	27.27	50.13	33.51
SDM	L1C	64.23	87.27	80.15	70.86	81.56	63.56
	L2R	68.15	88.05	79.12	70.32	83.9	61.97
	L3L	66.6	87.01	79.38	71.12	82.34	62.5
	L4L	66.06	88.31	82.22	72.99	82.86	66.49
	LA6	65.54	88.05	77.32	69.79	83.38	61.44
	AVG	66.116	87.738	79.638	71.016	82.808	63.192
ORACLE	-	55.09	78.44	66.49	58.82	73.25	48.94
CS	-	66.32	88.83	80.67	70.59	82.86	64.63
ROVER	c1	61.36	86.23	77.1	67.65	81.3	60.37
	c2	62.14	85.97	75.8	68.45	80.52	59.31
	c3	62.92	85.8	75.52	68.72	80.26	59.84
	c4	61.88	87.3	76.55	69.52	81.04	59.04
	c5	62.14	87.01	76.5	68.45	81.3	59.31
CNC	c1	63.45	86.75	76.8	67.91	82.34	59.84
	c2	63.45	87.27	76.8	68.18	81.82	59.31
	c3	63.19	87.27	76.8	68.45	82.08	59.57
	c4	63.71	87.27	75.52	68.98	82.34	60.11
	c5	63.45	87.01	76.03	68.18	82.34	60.37
MMCN	All	**62.92**	**87.01**	**74.74**	**64.97**	**82.34**	**59.84**

Table B.7: WER results on the development set - REAL

	Mic	Speaker					
		1	2	3	4	5	6
	CT	38.8	71.7	44.33	33.69	60.16	31.47
SDM	L1C	79.17	91.41	78.09	62.3	78.39	65.07
	L2R	84.9	91.67	84.79	66.04	81.25	78.4
	L3L	82.29	90.1	81.96	68.18	79.17	68
	L4L	81.51	86.98	78.35	68.45	79.95	68.8
	LA6	80.21	88.28	78.61	66.58	81.25	62.13
	AVG	81.62	89.69	80.36	66.31	80.00	68.48
ORACLE	-	69.53	80.21	66.24	49.73	67.71	53.07
CS	-	78.39	91.93	80.67	68.72	79.95	70.67
ROVER	c1	79.7	90.1	75.3	60.7	77.08	64.53
	c2	78.9	89.84	77.84	61.5	77.6	64.8
	c3	78.91	89.58	76.8	61.23	76.82	63.73
	c4	79.2	88.1	76.55	62.03	77.08	63.47
	c5	79.4	88.28	76.55	60.16	77.08	63.47
CNC	c1	78.91	86.2	75.77	61.76	77.6	65.33
	c2	78.65	85.94	75.77	62.03	78.12	64
	c3	78.39	85.94	75.52	61.76	77.86	64.27
	c4	78.91	85.94	75	61.23	77.08	64.8
	c5	78.65	86.2	75.52	61.5	77.34	64.53
MMCN	All	**77.6**	**86.98**	**74.23**	**64.71**	**79.17**	**66.93**

B.4 MMCN Sensitivity to Boundary Misplacement

The following tables report a set of sensitivity tests applied to MMCN.

Table B.8: WER increase by boundary shifting

SIM								
Shift(s)	spk1	spk2	spk3	spk4	spk5	spk6	avg	**Increm**
-0.2	62.4	86.8	75.6	65.2	81.2	60.4	71.9	**1.06**
-0.15	61.8	86.5	75.5	64.7	80.6	59.9	71.5	**0.63**
-0.1	61.6	86.4	75.2	64.4	80.4	59.5	71.2	**0.38**
-0.05	61.4	86.2	74.9	64.1	80.4	59.4	71.1	**0.20**
0	61.0637	86.2	74.5	63.7	80.5	59.2	70.9	**0**
0.05	60.9	86.3	74.5	63.9	80.4	59.2	70.9	**-0.01**
0.1	61.0	86.2	74.5	64.0	80.4	59.1	70.9	**0.02**
0.15	60.9	86.4	74.9	64.5	80.5	59.2	71.1	**0.23**
0.2	61.56	87.05	75.72	65.50	80.95	60.01	71.80	**0.94**

REAL								
Shift(s)	spk1	spk2	spk3	spk4	spk5	spk6	avg	Increm
-0.2	76.2	87.9	75.2	64.9	77.5	64.9	74.5	**1.12**
-0.15	75.9	87.7	74.9	64.5	77.2	64.2	74.0	**0.72**
-0.1	75.6	87.6	74.3	63.7	76.9	64.1	73.7	**0.39**
-0.05	75.3	87.3	74.0	63.6	76.8	63.9	73.5	**0.18**
0	75.2	87.3	73.8	63.4	76.5	63.8	73.3	**0**
0.05	75.2	87.3	73.7	63.0	76.5	63.8	73.3	**-0.08**
0.1	75.1	87.1	74.1	63.2	76.5	63.9	73.3	**-0.03**
0.15	75.6	87.1	74.4	64.1	76.7	64.3	73.7	**0.35**
0.2	76.5	87.7	75.6	65.2	77.6	65.1	74.6	**1.28**

.

B.5 MMCN Sensitivity to Boundary Loss

Table B.9: WER when one or more reference boundaries are ommitted.

SIM								
Lost	spk1	spk2	spk3	spk4	spk5	spk6	avg	**Increm**
0	61.1	86.2	74.5	63.7	80.5	59.2	70.9	**0**
1	62.3	86.8	75.5	65.0	80.9	60.5	71.9	**0.99**
2	66.0	88.6	78.6	69.3	82.9	64.4	74.9	**4.10**
3	69.6	90.0	81.4	73.6	85.2	68.6	78.1	**7.21**
4	73.6	91.6	83.8	77.6	87.3	72.8	81.1	**10.26**
REAL								
Lost	spk1	spk2	spk3	spk4	spk5	spk6	avg	**Increm**
0	75.2	87.3	73.7756	63.4	76.5	63.8	73.3	**0**
1	75.9	87.8	74.9	64.7	77.1	65.1	74.3	**0.93**
2	78.1	88.9	77.6	68.7	79.5	68.1	76.8	**3.52**
3	80.8	90.1	80.7	73.3	82.3	71.8	79.8	**6.51**
4	83.8	91.5	83.3	77.6	84.9	75.7	82.8	**9.48**

B.6 MMCN Sensitivity to Boundary Addition

Table B.10: WER when one or more boundaries are added to reference boundaries.

SIM								
Added	spk1	spk2	spk3	spk4	spk5	spk6	avg	**Increm**
0	61.1	86.2	74.5	63.7	80.5	59.2	70.9	**0**
1	61.2	86.3	74.7	63.8	80.5	59.4	70.9	**0.11**
2	61.5	86.5	74.9	64.0	80.4	59.9	71.2	**0.32**
3	61.7	86.8	75.0	64.3	80.2	60.4	71.4	**0.54**
4	61.9	87.1	75.3	64.5	80.1	60.7	71.6	**0.75**
5	62.2	87.2	75.6	64.9	80.2	60.8	71.8	**0.96**
6	62.6	87.5	75.6	65.1	80.5	61.0	72.0	**1.18**
REAL								
Added	spk1	spk2	spk3	spk4	spk5	spk6	avg	**Increm**
0	75.2	87.3	73.8	63.4	76.5	63.8	73.3	**0**
1	75.5	87.3	74.0	63.6	76.8	63.9	73.5	**0.19**
2	75.6	87.4	74.4	63.6	77.2	64.4	73.8	**0.42**
3	75.6	87.5	74.6	63.6	77.8	64.6	73.9	**0.64**
4	75.6	87.7	74.8	63.6	78.5	64.7	74.2	**0.82**
5	75.6	87.9	74.9	63.6	78.8	64.9	74.30	**0.98**
6	75.7	88.2	75.2	63.9	79.3	65.2	74.6	**1.27**