EFFECTS OF REVERBERATION ON MONAURAL SPEECH SEPARATION AND RECOGNITION

by
Hakan Kurçenli
B.S., Electronics and Communications Engineering, Yıldız Technical University, 2008

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Electrical and Electronics Engineering
Boğaziçi University
2013

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Assoc. Prof. Murat Saraçlar, for his support, encouragement, guidance and contributions throughout the preparation of this thesis, and for the feedback he provided during the research. He has been a steady influence in the course of my Master's study and is the major contributor in my interest in the field of speech processing.

I would also like to thank Prof. Emin Anarım for his guidance and support throughout the time I spent at Boğaziçi University.

In addition, it has been a privilege for me to get to know and work with many other great people over the last several years and I would like to extend my appreciation to all those who helped me during this journey.

Finally, I am very grateful to my mother, my father who is no longer with us, and my older brother for always believing in me, for their sincere love and infinite support, without which this work would not have been possible.

ABSTRACT

EFFECTS OF REVERBERATION ON MONAURAL SPEECH SEPARATION AND RECOGNITION

Speech recognition is an active area of research with implementations spanning from commercial to medical applications such as hearing aids. Recognition of speech signals is studied for decades and a lot of progress has been shown in this area of research but there is still a lot of room for further research due to the adverse effects of the environmental conditions that contaminate clean speech signals. Such adverse conditions include noise and reverberation. Under such conditions, the recognition of automatic speech recognizers is subject to substantial degradation. Speech separation where the goal is to separate speech signals belonging to more than one talkers speaking at the same time is also an unsolved problem. It gets even harder as adverse environmental conditions are added to the scenario. This research is aimed at studying the effects of reverberation on monaural speech separation and recognition, and increasing the recognition performance of mixed speech signals. In the context of this research, recognition of mixed monaural speech signals with different reverberation levels is implemented and the effects of reverberation are evaluated. Performance increase in recognition is accomplished by training the system with moderately reverberated speech signals and then with speech signals that have predefined constant reverberation levels. For comparison purposes, effects of reverberation on speech recognition in the single-talker scenario are also examined along with the performance increase obtained by training the system with moderately reverberated signals.

ÖZET

TEK KANALLI KONUŞMA AYIRMA VE TANIMADA YANKILAŞIMIN ETKİLERİ

Konuşma tanıma, ticari uygulamalardan işitme cihazları gibi tıbbi uygulamalara varan geniş bir yelpazede kullanım alanı olan, halen aktif bir araştırma konusudur. Konuşma işaretlerinin tanınması yıllardır üzerinde çalışılan ve gelişme kaydedilen bir araştırma alanı olmasına karşın olumsuz çevre koşullarının konuşma işaretlerini bozması nedeniyle bu alanda hala araştırılması gereken birçok açık konu bulunmaktadır. Bu olumsuz çevre koşullarının başlıcaları gürültü ve yankılaşım olarak tanımlanabilir. Bu koşullar, otomatik konuşma tanıma sistemlerinin başarımını ciddi ölçüde düşürmektedir. Birden fazla konuşmacının aynı anda yaptığı konuşmaları ayırmayı hedefleyen konuşma ayırma problemi de henüz tam olarak çözülememiş bir problemdir. Olumsuz çevre koşullarının eklenmesiyle beraber problemin çözümü daha da zorlaşmaktadır. Bu çalışmanın amacı, tek kanallı konuşma ayırma ve tanımada yankılaşımın etkilerinin araştırılması ve karma konuşma sinyallerinin tanınma başarımının artırılmasıdır. Bu tez kapsamında farklı yankılaşım seviyelerine sahip tek kanallı karma konuşma sinyallerinin tanınması gerçekleştirilmiş ve yankılaşımın etkileri incelenmiştir. Sistemin orta derecede yankılaşıma sahip konuşma sinyalleriyle eğitilmesi sayesinde tanıma başarımında artış sağlanmıştır. Ardından sistem, daha önceden belirlenmiş sabit yankılaşım seviyelerindeki konuşma sinyalleri ile eğitilmiş ve sonuçlar incelenmiştir. Karşılaştırma amacıyla yankılaşımın tek konuşmacı durumunda konuşma tanımaya etkileri incelenmiş ve yine sistem orta derecede yankılaşıma sahip konuşma sinyalleriyle eğitilerek başarım artışı sağlanmıştır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AM	Acoustic Model
ASR	Automatic Speech Recognition
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CPP	Cocktail Party Problem
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ICA	Independent Component Analysis
ISA	Independent Subspace Analysis
ITD	Inter-aural Time Difference
LM	Language Model
NMF	Non-negative Matrix Factorization
RIR	Room Impulse Response
RT_{60}	Reverberation Time
SCSS	Single Channel Source Separation
SNR	Signal-to-Noise Ratio
SSR	Signal-to-Signal Ratio
STFT	Short Time Fourier Transform
WER	Word Error Rate

1. INTRODUCTION

1.1. Problem Statement

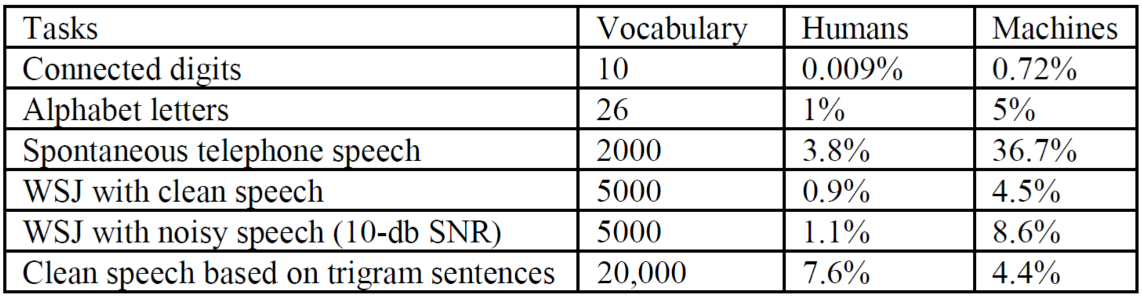
Beginning from the early history to this age, speech has been and will be the dominant mode of interaction and sharing between people. Invention of technologies such as telephony, internet, radio and television has also helped information conveyed by speech to massively grow in size.

Training an automatic speech recognition (ASR) system with clean speech signals that contain no noise causes the system to perform with a degraded performance when tested on real-life speech that does not match the clean training signals. For the system to be called robust, recognition rate of the system should not degrade remarkably in case of the mismatch between training and testing.

Speech signals are subject to transformation as the speech is produced by the speaker until it reaches the ear or it reaches microphone and is digitized. These transformations may be called the acoustical environment as a whole. The two dominant sources that cause distortion of the speech signal can be defined as additive noise and channel distortion. The sound produced by a working fan, by the slam of a door or by other speakers constitutes examples for the additive noise. Thus, it is a common form of noise in everyday life. On the other hand, effects like reverberation, microphone's frequency response or a speech codec are examples of channel distortion. Reverberation, which is the main subject investigated in this thesis work, can change the speech waveforms dramatically.

Information in Table 1.1 indicates that in terms of robustness, humans outperform machines for simple tasks. Rate of error for spontaneous telephone speech recognition is above 35% for machines, a value that is almost 10 times higher than of the humans on the similar task. It should also be noted that recognition error does not increase as abruptly for humans as machines as the noise level increases [1].

Table 1.1. Word error rate comparisons between human and machines on similar tasks [1].



Tasks	Vocabulary	Humans	Machines
Connected digits	10	0.009%	0.72%
Alphabet letters	26	1%	5%
Spontaneous telephone speech	2000	3.8%	36.7%
WSJ with clean speech	5000	0.9%	4.5%
WSJ with noisy speech (10-db SNR)	5000	1.1%	8.6%
Clean speech based on trigram sentences	20,000	7.6%	4.4%

Several challenges must be overcome to enable robust speech recognition in real-life conditions, one of the hardest being the case of multiple sound sources. Speech recognition in the presence of a competing speaker is a topic that has been studied for decades, and different approaches have been proposed for the solution of the problem. Some algorithms try to model the target and the masker speech while others try to separate speech by grouping auditory cues [2].

Without the existence of reverberation, automatic speech recognition after the application of speech separation algorithms perform quite well, almost at the level of human performance. In the case of reverberation in the environment, although speech separation algorithms increase the recognition rate, the overall recognition rate of the separated signals remains well below that of the human listeners for the same task [3].

1.2. Contribution of the Thesis

Contribution of this thesis is the examination of the effects of different levels of reverberation on the separation and automatic recognition of concurrent, overlapping speeches in the two-talker monaural scenario and showing that training the system with reverberated samples can be used as a method to increase the overall recognition performance of the ASR system. The system is also trained with speech samples having predefined constant levels of reverberation and the results are investigated. Performance degradation due to the effects of reverberation in single-talker scenario is also examined for comparison purposes. The experiments are carried out using clean and reverberated utterances in training in order to figure out the performance increase that training on reverberated samples would yield.

At the time of writing, to the best of the knowledge of the author of this thesis, literature survey did not reveal any other study investigating the effects of reverberation on speech separation and recognition using different levels of reverberation although previous work existed for the recognition in the single talker scenario. Only one paper by Mandel *et al.* [3] tries to create an evaluation metric for different source separation techniques performing in an environment with a fixed reverberation time value. The separation method used in this thesis work takes advantage of NMF and exemplar based sparse representation and the channel configuration for the mixed speech signals are chosen to be monaural. The author also encountered no previous record showing the performance increase in the case of training the system with moderately reverberated audio samples compared to clean training material. This work compares the aforementioned effects in the two-talker scenario with the single-talker scenario resulting in a comparative study of the two cases.

1.3. Thesis Outline

Chapter 1 is the introductory part of the thesis. The problem is defined, previous work is given along with the outline followed throughout the thesis.

Chapter 2 gives the necessary background information needed for understanding the problem that is examined in the thesis. It starts by explaining the concepts of automatic speech recognition, continues with the discussion of speech separation and finally concludes with the explanation of reverberation and its effects on speech recognition.

Chapter 3 presents the experiments made to obtain the necessary data needed for evaluating the effects of reverberation on the separation and recognition of speech along with the methods applied for increasing the recognition performance. Methodology that is followed during the experiments is also explained in detail in this section. The results of the experiments are made available and discussion regarding these results is given.

Chapter 4 is the conclusion section of the thesis. It includes a final summary of study and the conclusions reached as a result of the experiments conducted. Possible directions for future work in the topic are also given in this chapter.

2. BACKGROUND

2.1. Automatic Speech Recognition

2.1.1. Defining the Problem

Speech recognition can be defined as the process of transcription of the acoustic signals from the speaker that is captured by a microphone into perceived letters, words or sentences. In the case of applications such as dictation or document preparation, the recognition results may be treated as the final output. Another approach would be feeding these recognition results as input for further processing in terms of linguistics.

Table 2.1 lists some important parameters in the process of speech recognition. The speaker is expected to pause for a short duration after each word in the case of isolated-word speech recognition. On the other hand, it is not necessary to give a pause in between the words for continuous speech recognition system. Conversational speech, which can also be described as spontaneous speech, contains disfluencies and recognition of speech read from a formerly prepared script is easier to recognize than spontaneous speech. Speaker enrollment is another process required by some speech recognizers, where the speakers make available to the system samples of his/her previous speech beforehand. These systems may be defined as speaker-dependent. Speaker-independent systems, on the other hand, do not require speaker enrollment. Some of the parameters listed are task-dependent. For instance, the size of the task's vocabulary being large or having words that sound similar phonetically may pose difficulties in terms of recognition. When order of words in the speech fits into an expected model, the use of language models or grammars to restrict the combination the words can take improves the recognition accuracy. This model may be as simple as a finite state network where the possible words after the use of a specific word are predefined. Or it can be general model approximating natural language in a specific context.

Table 2.1. Typical parameters of ASR systems [4].

Parameters	**Range**
Speaking Mode	Isolated words to continuous speech
Speaking Style	Read speach to spontaneous speech
Enrollment	Speaker-dependent to Speaker-independent
Vocabulary	Small (<20 words) to large (>20,000 words)
Language Model	Finite-state to context sensitive
Perplexity	Small (<10) to large (>100)
SNR	High (>30 dB) to low (< 10 dB)
Transducer	Voice-cancelling microphone to telephone

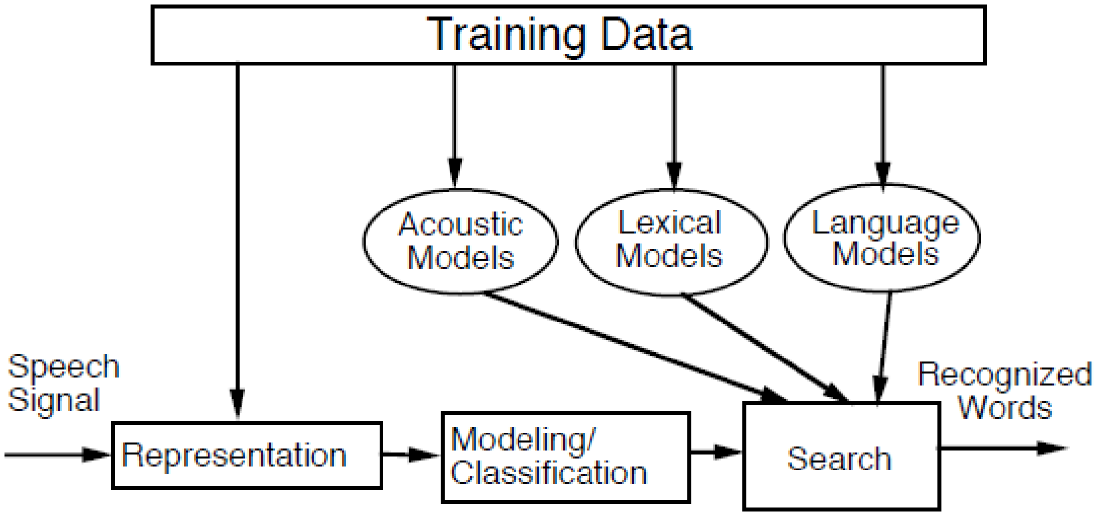


Figure 2.1. Components of a typical speech recognition system [4].

Figure 2.1 summarizes the major building blocks of a typical ASR system. At a fixed rate of 10-20 ms., the speech is transformed into meaningful features. Then using these features as the basic representation of speech, search is initiated to find the most probable word by using the acoustic, lexical and language models. Designation of the model parameters is accomplished by the training process [4].

The aforementioned sources of variability are represented in several ways by ASR systems. State-of-the-art ASR systems usually maintain speaker-independency through the use of extensive multi-speaker training [5].

Statistical methods using large amounts of training data are the preferred way of modeling variabilities at the acoustic level to reach the optimal settings in the search procedure. Speaker adaptation is also used to produce speaker-dependent models out of the speaker-independent models. Context dependent acoustic modeling is a method that encompasses context-dependent training for phonemes by producing a separate model for each context that the phoneme can be used and it can be utilized to maintain acoustic to lexical mapping.

Networks called “pronunciation networks” is used to combat the word level variability where a word can be pronounced in several different ways. These pronunciation networks layout the alternate pronunciations for the words in the lexicon. Variations due to alternate pronunciations and different accents form alternate paths in the network that the search algorithm can take. The most likely sequence of words is found using n-gram statistical language models [4,6].

Hidden Markov Models (HMM) is currently the agreed upon paradigm in speech recognition. An HMM is a stochastic model, where the generation of the underlying phonemes and the acoustic features are represented probabilistically as Markov processes. In hybrid systems, neural networks are integrated into the HMM based system.

There are two competitive approaches in terms of segment identification. Speech segments might first be identified and then scored for the recognition of the words. In the case of frame based HMMs, segments are identified during the search process. Both approaches produce similar recognition performance [4].

2.1.2. Bayesian Model For Speech Recognition

The main objective of the probabilistic noisy channel model for speech recognition can be described by the following question:

"Given the acoustic input O, what is the most probable sentence among the sentences possible in the language L*?"*

The acoustic input O is a sequence of individual observations where each observation is obtained by taking 10 ms. frames off the input and representing them by the energy in its frequency bands. The index i in o_i indicates the number of the time interval Each index then represents the corresponding time interval, and consecutive observations make up the input:

$O = o_1, o_2, o_3, \ldots, o_t$

A sentence can be represented by consecutive words in a similar manner:

$W = w_1, w_2, w_3, \ldots, w_n$

These representations can be thought of as simplifications. If we are modeling a group of words rather than individual words, then representing the sentence as consecutive words would be a detailed division. On the other hand, if we are dealing with the morphology, it would be a broad division.

When described in the probabilistic framework, the aforementioned way can be mathematically expressed as follows:

$$\widehat{W} = \operatorname{argmax}_{W \in L} P(W|O) \tag{2.1}$$

Equation 2.1 is results in the optimal sentence W. The problem now can be defined as the computation of P(W|O) for a given sentence W and the acoustic sequence O. Bayes' rule can be used to break down any probability P(x|y) into its constituents as follows:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \tag{2.2}$$

When Equation 2.2 is substituted into Equation 2.1, we get the resulting equation as follows:

$$\widehat{W} = \mathrm{argmax}_{W \in L} \frac{P(O|W)P(W)}{P(O)} \tag{2.3}$$

The probabilities P(O|W), P(W), P(O) in Equation 2.3 are usually easier to compute than P(W|O). For example, the prior probability P(W) is estimated using the n-gram language models. Estimating the probability P(O|W) is easy as well. But estimating the probability of the acoustic observations, P(O), is harder to estimate but it can be ignored.

Since maximization is carried out over all possible sentences, the expression being computed will be:

$$\frac{\mathrm{P(O|W)P(W)}}{\mathrm{P(O)}} \tag{2.4}$$

for each sentence in the language. Since for each sentence, the same observation sequence *O* is examined, the value of P(O) does not change from sentence to sentence. Thus:

$$\widehat{W} = \mathrm{argmax}_{W \in L} \frac{P(O|W)P(W)}{P(O)} = \mathrm{argmax}_{W \in L} P(O|W)P(W) \tag{2.5}$$

In short, the most likely sentence *W* given the observation sequence *O* can be determined by taking the product of two probabilities given above for each sentence, and finding the sentence for which this product is greatest. The language model related component of the recognizer computes $P(W)$, the prior probability, the acoustic model related part computes $P(O|W)$, the observation likelihood.

The language model (LM) prior $P(W)$ is a measure of how probable a given sequence of words is a source sentence of English. An *N*-gram grammar lets us assign a probability to a sentence by computing:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k | w_{k-N+1}^{k-1}) \qquad (2.6)$$

Given the acoustic model and language probabilities, the probabilistic model can be utilized in a search algorithm in order to calculate the maximum probability word sequence for a given acoustic input.

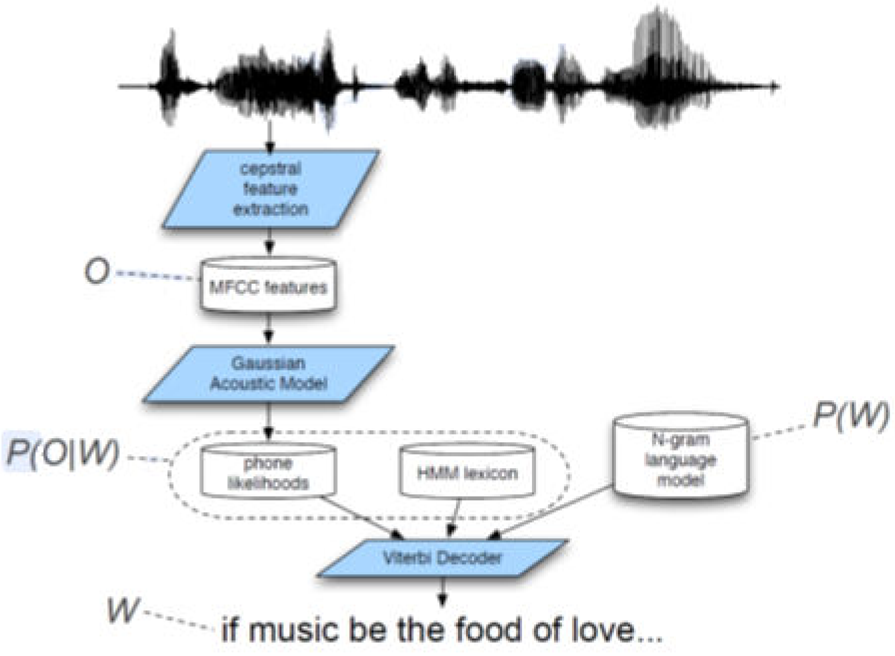


Figure 2.2. Schematic architecture for a simplified recognizer [7].

Figure 2.2 shows the components of an HMM speech recognizer. The recognizer in the figure processes a single utterance, indicating the computation of the prior and likelihood. The recognition process in the figure is carried out in three steps. The feature extraction is mainly a signal processing stage where the acoustic waveform is divided into frames of usually of 10-20 ms. in length. These frames are then transformed into spectral features. Each frame is represented by a vector of around 39 features representing this spectral information as well as information about energy and spectral change. In the phone recognition step, likelihood of the observed spectral feature vectors is calculated given the linguistic units like words or phones. In the decoding stage, most likely sequence of words are produced as output. Viterbi algorithm is the preferred method in the ASR domain for decoding and it speeds up the decoding process by means of pruning, fast match and tree-structured lexicons [7].

2.1.3 Applying the Hidden Markov Model to Speech

An HMM includes two stochastic processes, a hidden Markov chain that is responsible for the temporal variability, and an observable process that is responsible for the spectral variability. This combination is able to cope with the most important sources of speech variability, and allows the implementation of recognition systems with very large vocabularies [4].

In the speech domain, the hidden states of HMM correspond to phones, subphones or words. Each observation of HMM map into the energy present at the specific spectral bands of the feature vector belonging to the waveform at a certain time interval and the decoding stage translates this acoustic information into phones and words.

Acoustic feature vectors in speech recognition make up the observation sequence. Each vector represents information regarding the amount of energy present in separate frequency bands at a particular time frame. It should be noted that each observation in the sequence is made up of a vector consisting of 39 real-valued features regarding the spectral information. Each observation typically corresponds to a time interval of 10 milliseconds, hence 1 second of speech is represented by nearly a hundred feature vectors of length 39.

Speech may be modeled in several different ways by using the hidden states of HMMs. For trivial recognition tasks, like the recognition of digits starting from zero up to ten or for the type of recognition where the input from the speaker is a simple yes or no, an HMM can be formed with words corresponding to states. For most larger tasks, however, the hidden states of the HMM usually map to phone-like units in the case of large recognition tasks and words are represented by sequences of these hidden states.

Since speech can be classified as sequential, HMM models concerning speech place strong constraints on transitions. Transitions can happen from one state to the next or to itself, not to the previous state.

Since a single phone can take up a variable amount of duration depending on the person speaking, self-loops are used to cope with this source of variation. Duration of the phone may change depending on the phone, the speaker's rate of speech or the context.

It is generally preferred to use a three-state HMM model including a state each for the beginning the middle and the end. As shown in Figure 2.3, phones modeled this way include three emitting states along with two non-emitting states placed right at the beginning and at the end of the three states. So when the term "phone model" is mentioned, it usually refers to this 5 state representation. The term "HMM states" is usually used to denote the three emitting states corresponding to the subphones found in the middle. [7]

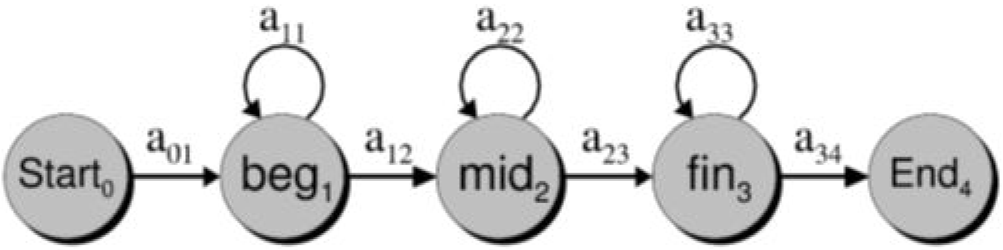


Figure 2.3. A standard 5-state HMM model for a phone [7].

2.1.4. State of the Art in ASR

When explaining the state-of-the-art in speech recognition, it is important to make distinctions between the different types of recognition tasks with different difficulty levels and with different constraints on the task. Different recognition tasks may require different methods. For instance, in the case of small vocabulary tasks, it would be suitable to use a word level HMM model. Whereas this method would not be suitable for a large vocabulary continuous speech recognition task where it would be more appropriate to use phone level HMMs. Performance of ASR systems is usually expressed in terms of word error rate (WER), which is defined as:

$$E = \frac{S+I+D}{N} 100 \tag{2.7}$$

where N represents the total number of words in the test set, S the number of substitutions, I the number of insertions and finally D the number of total deletions.

There has been remarkable progress in the field of speech recognition when the past decade is considered. Due to the important advances in the basic technology, performance increase has been achieved in the cases speaker independence, continuous speech, and large vocabularies. One of the reasons behind this progress is the integration of HMM into speech recognition. By using the large amounts of training data collected, it has been possible to train the parameters of the speech models automatically.

As just mentioned, training the HMM models requires large training data to be collected and the past decade has witnessed the preparation of large speech corpora for training, development and testing of the ASR systems. These large corpora usually include tens of thousands of sentences. By the use of such corpora, statistical analysis for the determination of the parameters of the ASR systems is made possible.

Another area of progress in the field was in terms of defining consistent standards for measuring the performance of the systems built for speech recognition. The use of locally collected data for training was also a problem. These factors caused difficulty in comparison of performance of different recognition systems. It was also a source of degradation in performance when these ASR systems were subject to speech much different than the ones they were trained on. After large corpora are made available publicly and certain standards are set in comparing the ASR systems, consistency has been reached in the evaluation of the systems' performance.

One of the very important catalysts in the advances regarding speech technology is advances made in computer technology. Inexpensive mass storage made it possible to store and share the large bodies of corpora among the researches and as the computers got faster, processing of this large body of data became possible in a realistic amount of time. It was also important in the sense that implementation and testing of theoretic ideas became possible. Now, ASR systems can run in real time with acceptable performance levels.

Among the tasks with low perplexity, digit recognition stands out as a popular one. As of now, digits spoken continuously in English on the telephone can be recognized speaker independently with a WER of 0.3%. Another popular task with moderate perplexity was the Resource Management task, where inquiries are made about naval vessels in the Pacific Ocean. It is a task constrained with 1000 words. When a word-pair language model is used in recognition, the word error rate for this task is less than 4%. Recognition of conversational speech is another active area of research where a WER of less than 3% is achieved.

Dictation applications usually require a vocabulary with thousands of words. Beginning from the 1990s, very large vocabulary continuous speech recognition has been at the center of attention. These systems had a high perplexity and they were usually designed to be speaker independent [4]. WER of 9.9% was achieved in the transcription of broadcast news in English in the latest NIST benchmark [8].

Due to the improvement in performance over the years, speech recognition is being employed also by telecommunications sector. Many automated services are starting to prefer voice as the driving input instead of touch tone. Speaker-dependent voice dialing of phone numbers can be given as an example. Keyword spotting is another technology used to understand and route the customers' inquiries as they speak spontaneously explaining what they want.

Despite the steady progress in the field, there is still a lot of room for the recognition of conversations [4]. WER of 16.7% is achieved for the recognition of phone conversations in English using the Switchboard cellular conversational telephone-based speech in the latest NIST benchmark [8]. Implementation of an unlimited vocabulary, speaker-independent continuous speech recognition with the performance comparable to that of humans is not a very realistic expectation in the near future [4].

2.1.5. HTK (HMM Toolkit)

HTK is a toolkit mainly designed for providing HMM based speech processing and recognition capabilities although it can be used to build Hidden Markov Models (HMMs) in various other areas of research. Hence, attention in HTK is given to the task of speech recognition.

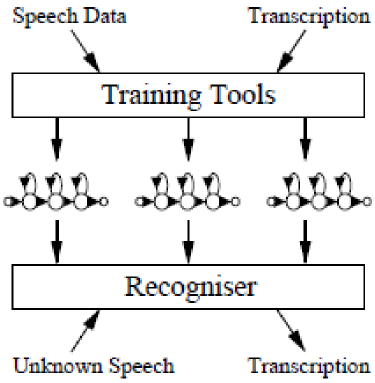


Figure 2.4. Processing stages of HTK [9].

As depicted in Figure 2.4, there are two major processing steps involve training and recognition. In the first stage, utterances in the training set with corresponding labels are utilized for the estimation of the parameters of HMMs. In the second stage, utterances in the test set are transcribed using the recognition tools.

Library modules contain much of the functionality. Consistency in how each tool interfaces with the other is maintained using these modules by also using common functions provided. Figure 2.5 shows the architectural structure of the HTK along with the interfaces.

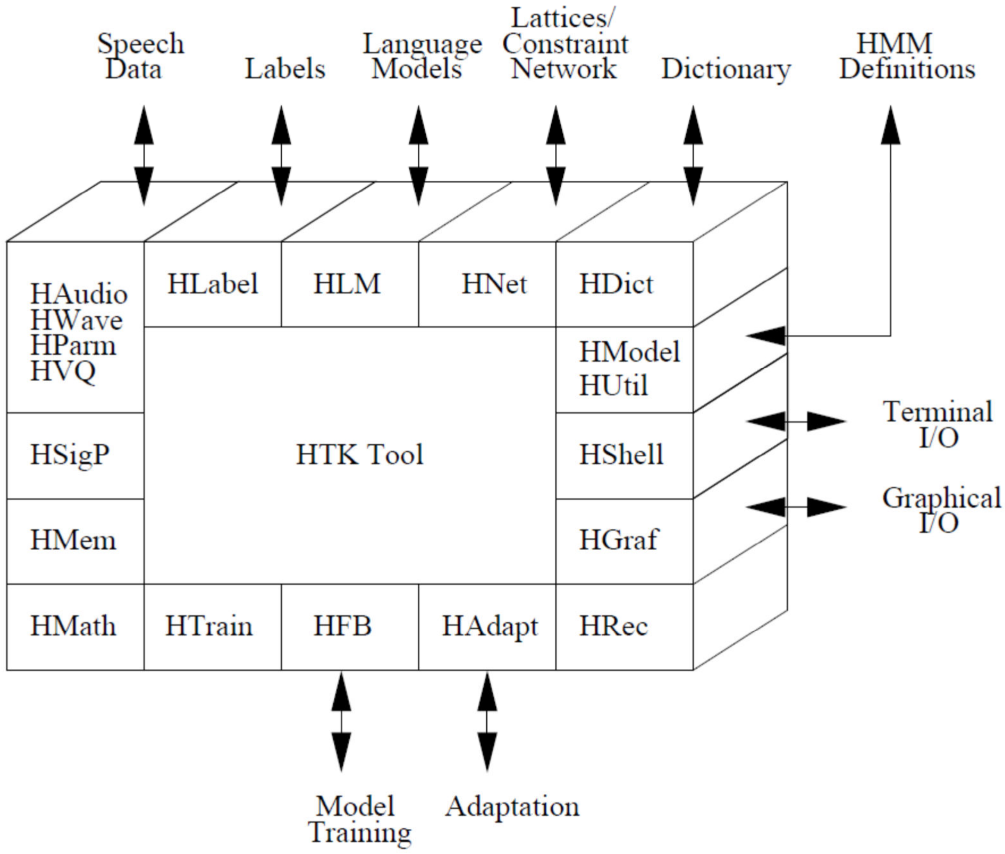


Figure 2.5. Software architecture of HTK [9].

HSell module controls the user input/output and interacts with the operating system. HMem takes care of the memory management related functions. HMath provides the mathematical backbone required, whereas HSigP provides the support for the core signal processing operations needed during the acoustic analysis phase. Each file type has a dedicated interface module:

HLabel → label files

HLM → language model files

HNet → networks and lattices

HDict → dictionaries

HVQ → VQ codebooks
HModel → HMM definitions

Waveform level support is provided by HWave parameterised level support is given by HParm. For the purpose of importing data from various other sources, HWave and HLabel modules allow for multiple file formats to be used. Direct audio from an audio capture device is possible by the use of the HAudio module HGraf provides simple interactive graphics functionality. Configuration files are used for the necessary adjustments of these library modules [9].

2.2. Speech Separation

2.2.1. Problem Statement

Colin Cherry is the first to set the definition of the cocktail party problem (CPP) in a paper published in 1953. This problem refers to the interesting psychoacoustic phenomenon people have the remarkable ability to single out and recognize only one source of auditory input in a noisy environment, where noise may be due to the presence of competing speakers or other noise sources thought to be independent of each other [10].

The task of speech separation in complex environments, such as separating and recognizing a single speaker in a cocktail party is a very difficult task. Usually speech enhancement or separation algorithms cannot handle the task well when the properties of both the target and the masker are very similar [11].

2.2.2. Separation Systems

There exists numerous source separation algorithms in literature but they are unable to perfectly separate the target speech from a reverberant mixture so that the recognition accuracy of ASR will not drop when fed with the separated signal instead of the original, clean signal [12].

Since the objective of this thesis work is constrained with the monaural case, single channel source separation methods will be examined in the next sections. In general, single channel source separation (SCSS) algorithms include independent component analysis (ICA), non-negative matrix factorization (NMF), source-driven and model-driven methods.

2.2.3. ICA and ISA

ICA can be considered as a special case of blind source separation [13]. ICA is a method proposed by Hyvarinen *et al.* that is used to find a linear representation of non-Gaussian multivariate signal with the assumption that these components are statistically independent, or as independent as possible [14]. As long as the following conditions are met, ICA can separate the sources from the mixture completely:

- The mixing matrix must be full-rank.
- The number of observations should be larger than or at least equal to the number of unknown sources in the mixture.
- The independence assumption should hold true regarding the components of the mixture.
- The number of sources in the mixture should be known in advance.

The listed conditions act as limiting factors in the applicability of ICA algorithms for source separation in the single channel case [15].

Maximum likelihood approach is proposed in a paper by Jang and Lee as an extension to blind source separation for single channel source separation. The proposed method performs successfully for mixtures containing both speech and music but the performance degrades for mixtures that consist of two speech signals. The reason behind this degradation is that the algorithm could find the sets of bases representing the class of music signals well, whereas it performed poorly in explaining the sets of bases for the speech signals [15].

Binary time-frequency masking is another method that is proposed to be combined with ICA in literature. The method alleviates the strict constraints of having to know the number of sources in advance and the number of sources being equal to or less than the number of microphones. Experiments conducted by Pedersen *et al.* showed that it is possible to separate mixtures with six sources in nonreverberant conditions using this method [16]. It is also shown that a modified version of this algorithm using two microphones and correlation between the envelopes of the signals can be used for the separation of speech signals in reverberant environments [17,18].

Independent subspace analysis (ISA) is a different type of ICA that increases the dimensionality of the observation from 1 to N. Applying ICA on the transformed signal results in N independent bases, and then these bases are grouped together to represent the different sources in the mixture [13].

2.2.4. NMF

Non-negative matrix factorization (NMF) is based on the decomposition of a non-negative matrix representation of a mixed signal such as its magnitude or power STFT, into the product of two low rank, non-negative matrices: A = BC. In this composition, the columns of B are the basis vectors which together define the structure of the spectro-temporal representation of the separate sources in the mixture. In a similar manner, the rows of the matrix C represent the weights by which the separate sources are active in the mixture.

The sparse NMF method separates a mixture by projecting the mixed feature vector onto the joint subspaces of the sources and calculating the results of the projection in each subspace [13]. A theorem by Laurbeg states that a matrix will have a unique NMF if the row vectors in B is boundary close and the column vectors in C are sufficiently spread [19]. Various NMF algorithms have been proposed in literature for learning B and C from A. Some of these methods rely on favoring temporal continuity constraints and favoring components with slowly varying and sparse gains resulting in learning the sparse representations of the data [20, 21]. Benaroya *et al.* present an approach to single channel source separation where they use an extension of Wiener filtering to non-stationary processes by using GMMs [22]. Schmidt et.al. present a general method that alleviates prior knowledge in NMF based on Gaussian priors [23].

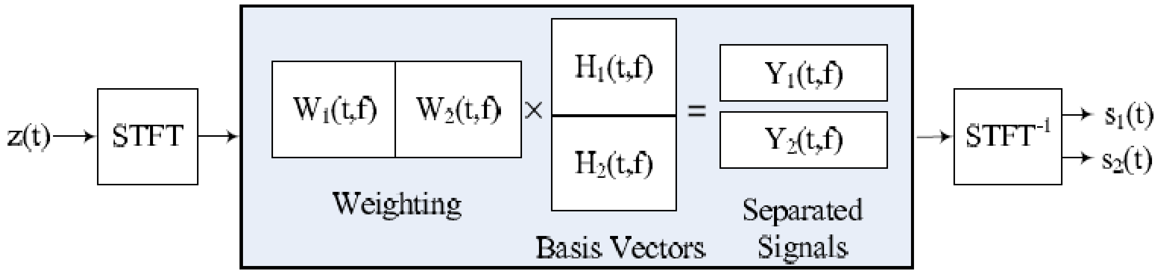


Figure 2.6. Sparse factorization of a spectrum using non-negative matrix factorization [13].

The only limitation in NMF may be described as the requirement of specifying the number of basis vectors. The NMF methods do not perform as expected for speech mixtures where the separate sources overlap extensively in the spectro-temporal domain. Figure 2.6 shows the way NMF decomposes a mixture signal's spectrum into the spectra belonging to the two speakers that contributed to the mixture. As can be seen from the figure, the basis vectors, (H_1(t, f) and H_2(t, f)), and the weights corresponding to these basis vectors, (W_1(t, f) and W_2(t, f)), are calculated and used to recover the spectro-temporal representation of the speakers using an inverse Fourier transform.

NMF can be employed in single channel source separation in either a supervised or an unsupervised way. Unsupervised NMF does the factorization of the mixed signal into the component signals without using any knowledge or training data about the sources, whereas supervised NMF first learns a set of bases by using the training data and uses these bases in the process of factorization [13]. King *et al.* propose a new method called "copy-to-train" for choosing the bases. The advantage of this method is that it overcomes the complication of choosing an optimal number of bases for the training phase [24]. Conventional algorithms for NMF excludes the use of phase information but Parry *et al.* state that the mixture spectrogram depends on the phase of the source STFTs. Authors examined the approach where phase information was not used in the factorization stage and its effects on the source separation, later leveraging a probabilistic representation of phase to improve the results of the source separation [25].

2.2.5. Exemplar Based Sparse Representation

Sparse representations are representations of a signal conveying all or most of the information in the signal using linear combinations of just a small number of basis signals called atoms and the concept has recently gained a lot of attention in the field of signal processing. The atoms used in the representation collectively form what is so called the dictionary. Sparse representations have also been used in separation of audio sources. Using this method, the mixture can be represented by using a dictionary for each of the underlying sources that make up the mixture [26]. In order to find the sparse representation of a signal, it is necessary to find the sparsest linear combination that represents the signal. Such methods exist in the context of NMF where the non-negative basis vectors that are learned are sparse combinations to generate expressiveness in the reconstructions [27]. Another field with such methods is compressed sensing where a discrete-time signal depends on a number of degrees of freedom that is much smaller than its original length [28]. The underlying source can be estimated using parts of the dictionary belonging to only a single source. Sparse representations are also used in the field of pattern recognition by linking the atoms with class labels and expressing the class of the observed signal as weights of the atoms. Use of this method has resulted in state-of-the-art classification algorithms to be developed in the fields of face recognition and phone classification [26].

When speech is considered as the signal of interest, the dictionary atoms are either chosen to include the conventional basis functions such as Fourier coefficients or wavelets in unsupervised cases [20]. In supervised cases, basis functions are learned from the training set in order to constitute the dictionary [29]. In exemplar based sparse representation, however, signals are modeled as sparse linear combinations of "examples" of that signal [26]. These example speech segments may also be called "exemplars" or "episodes" and original speech segments are modeled as a weighted linear combination of these exemplars. It has been shown that exemplar based methods in isolated digit recognition can outperform model-based methods [30]. Gemmeke *et al.* use clean speech exemplars in order to approximate speech features in noise [31]. The use of exemplars has its roots in the rather traditional methods used in speech recognition since it can be considered as template based recognition which is used in dynamic time warping [32].

One of the advantages of using exemplars as atoms in the case of speech recognition is that the dictionary constructed by using such atoms is easier to build since the examples can be taken directly from the speech database. Another advantage is that very sparse representations can be obtained in the case where the speech segment of interest closely follows the speech in the dictionary. Time frames in the exemplars are labeled with an HMM state, obtained by the use of forced alignment of the transcription in the training data so that using exemplar based sparse representations bring the advantage of easy mapping between the atoms and the speech classes [26].

2.2.6. Source-Driven Methods - CASA

In source-driven methods, the underlying speech signals are separated from the mixture without the use of any a priori knowledge about the speakers. Computational auditory scene analysis (CASA) is a well-known method among the source driven methods along with blind source separation (BSS). A CASA based algorithm looks for discriminative features in the mixture signal and works by extracting psychoacoustic cues from the mixture. The first stage is segmentation where the mixture is decomposed into spectro-temporal cells after STFT is performed on the signal which are dominated by either the target or the masker. After segmentation, cues extracted in the segmentation step like common onset-offset, harmonicity and periodicity that are believed to belong to a certain speaker are grouped together in the grouping stage [33].

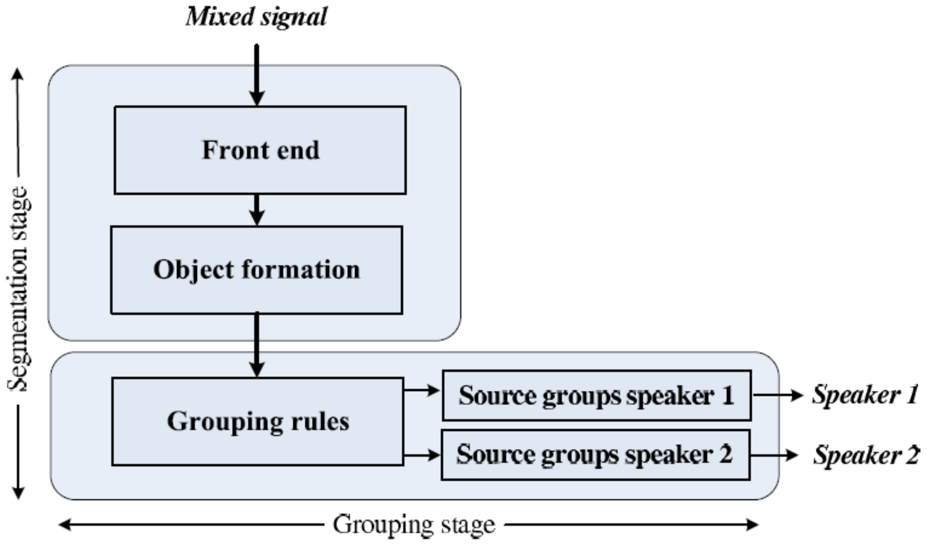


Figure 2.7. Block of CASA-based monaural speech separation [13].

2.2.7. Model-Driven Methods

A model-driven monaural source separation method completely depends on the a priori knowledge about the underlying speakers the mixture. Models for the underlying speakers are used to express the constraints associated with the feature vectors of individual speakers. These include well-known machine learning methods like VQ, Gaussian mixture models (GMM) and hidden Markov models (HMM) [33,13].

MAX-VQ is a model-driven method by Roweis that borrows concepts from both machine learning and speech processing. The author of the paper presents the refiltering approach to separation and denoising for deriving the masker signals based. Factorial-max vector quantization model is used to model clean speech signals from each speaker [34].

Figure 2.8 depicts the way the magnitude STFT of a mixed signal can be represented by the entries of the two speaker models belonging to the speakers in the mixture. These models can be considered as dictionaries represented as vector quantizer codebooks. These codebooks are trained to get the range of the short-time spectral patterns of the speech of a certain person. For the separation of the two speech signals, a finite search is performed over all the codevectors in both of the codebooks for each time frame in order to find the pair that maximizes the likelihood of the combined codevectors to match the real spectrum. In the example shown in Figure 2.8, the first speaker is represented by a codebook that consists of 3 prototypes and the codebook for the second speaker consists of 4 prototypes.

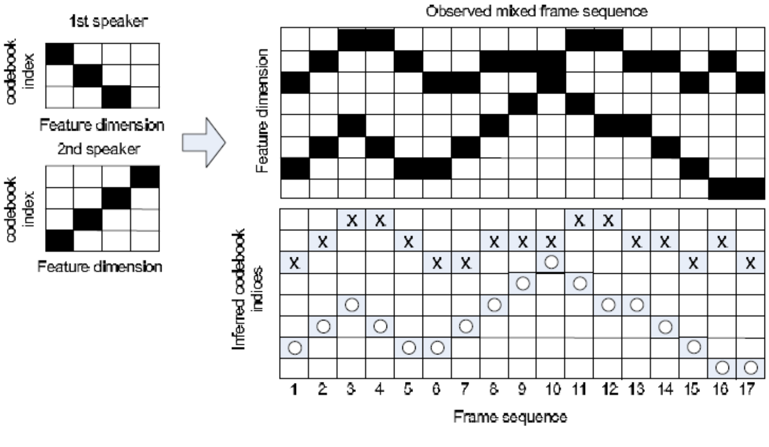


Figure 2.8. An example of two speaker codebooks defining a given mixture [15].

2.2.8. Summary

Up until now, numerous systems have been proposed to solve the problem of extracting or isolating speech in noise. Many of these systems were designed for a particular type of noise and would significantly underperform when the real input varied from the expected signal model, although this is not the case for systems that model the interaural parameters instead of the signals directly, which is not a subject related to the single channel case. Many of the separation methods are not robust to reverberation [13].

2.3. Reverberation

2.3.1. Introduction

Reverberation can be described as the sound that persists after the original sound source ceases to produce the sound. Reverberation or reverb occurs when a sound is produced in an enclosed space due to the large number of echoes reflecting off the surroundings and then gradually decreasing in amplitude as the sound is absorbed by the walls and air. It can be easily identified when the original sound from the source ceases to exist but its reflections continue to be heard as they decay and finally can no longer be heard. The duration of the decay of these reflections is associated with what is called the reverberation time and it receives special attention in the architectural design of large chambers since this acoustic feature of the room will directly affect the acoustic quality in terms of the activity of interest. Reverberation can be considered as the many thousands of echoes arriving in very quick succession, roughly with a time interval of 0.01 – 1 ms. between echoes [35].

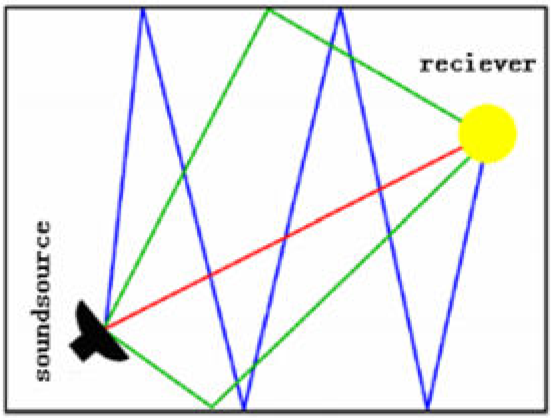


Figure 2.9. Representation of the different reverb types [36].

The red path in the Figure 2.9 is indicative of the direct sound coming from the source and reaching the microphone directly and the green paths indicate the early echoes, whereas the blue path represents the late reverberation [36].

Table 2.2. Comparison of the three components of the RIR [12].

	Direct-path	**Early echoes**	**Late Reverberation**
Relative timing	0 ms.	1-100 ms.	>100 ms.
Information about	source location	room geometry	room size, materials
Change with distance	r^{-2}	constant	constant
Change with motion	moderate	slow	details change rapidly
Effects on intelligibility	improve	improve	diminish
Interaural parameters	main trend	perturb mean	increase variance

2.3.2. Direct-path

The direct path sound is the sound coming from the source and directly reaching the microphone. It is the shortest path from the source to the listener so that direct path sound arrives at the microphone first. Its energy is inversely proportional with the square of the direct distance between the source and the listener. In a non-reverberant environment, this would be the only path that sound takes from source to listener [12].

2.3.3. Early Echoes

Early echoes are the sound waves that arrive immediately after the arrival of the direct-path sound. Sounds that are reflected off from large, regular surfaces are called specular. Early echoes can be considered “specular”. Although they can be treated as separate sources on their own, human perception groups these together with the direct path sound because they arrive to the listener in a time interval short enough after the direct path. Typically, early reflections arrive within 1 to 100 ms. of the direct-path. Early echoes convey information about the geometry of the enclosing such as its volume or the number and orientation of walls [12].

Early echoes also improve the intelligibility of speech due to the increase in energy arriving at the listener. The speech intelligibility tests conducted by Bradley *et al*. confirm the importance of early reflections in terms of the good conditions for speech intelligibility in a room. These experiments show that early reflections helped increase the SNR ratio and thus the speech intelligibility scores [37].

Image methods can be used for the analysis and simulation of the acoustic properties of enclosures. This method has been described in a paper by Allen and Berkeley. The authors chose a simple rectangular enclosure for investigating the method. The resulting impulse response obtained via the image method simulates the acoustic properties of the room reverberation when convolved with any desired speech signal [38]. Since reverberation is a wave phenomenon, another approach for simulation is called “ray tracing” where the computational simulation is based on ray approximation [39].

2.3.4. Late Reverberation

The echoes that arrive after the early reflections are called “late reverberation”. It consists of a very large number of higher order reflections as a result of sound waves scattering off of walls and objects in the room. The nature of reflections in late reverberation is mostly diffuse rather than specular. Late reverberation is useful for the characterization of the size of the room and the materials used in the walls because the walls exhibit a frequency dependent reflectance.

Although late reverberation may include specular reflections, it mostly includes diffuse reverberant energy that causes remarkable degradation in the intelligibility of speech especially for foreign speakers and people with hearing related impairment [12].

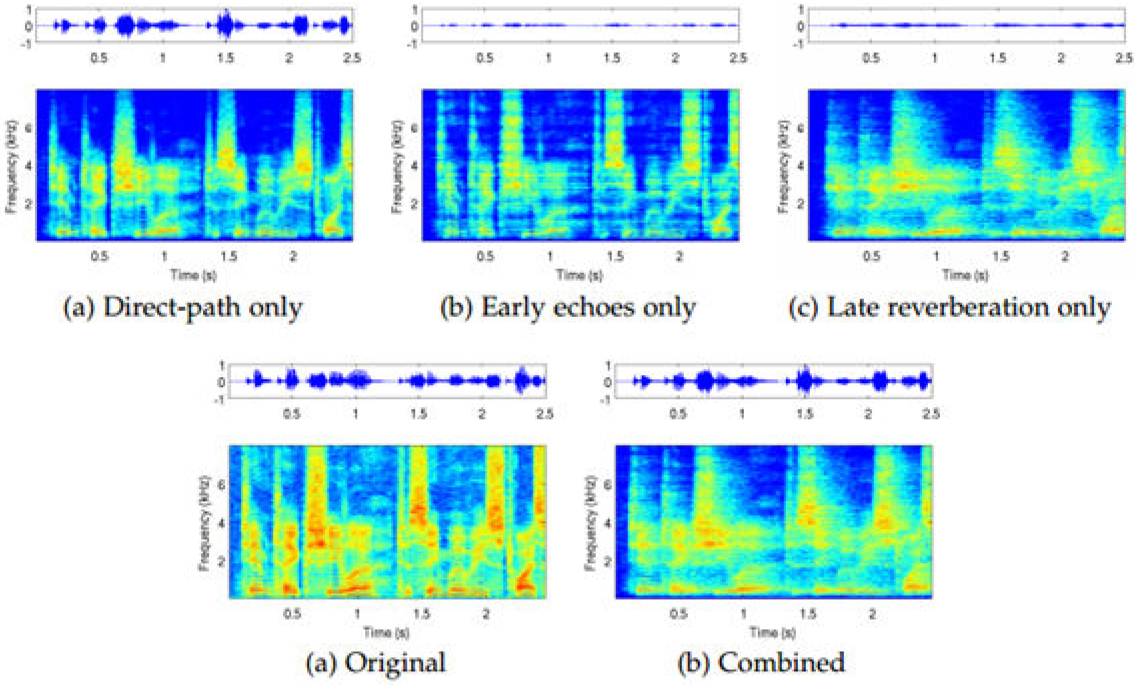


Figure 2.10. An example utterance convolved with different parts of a RIR [12].

2.3.5. Reverberation Time

Reverberation time, also denoted as RT_{60}, is the time required for reflections to decay by 60 dB below the level of the direct sound. 60 dB of decrease corresponds to the sound pressure to decrease to 1/1000 of its initial value, as shown in Figure 2.11.

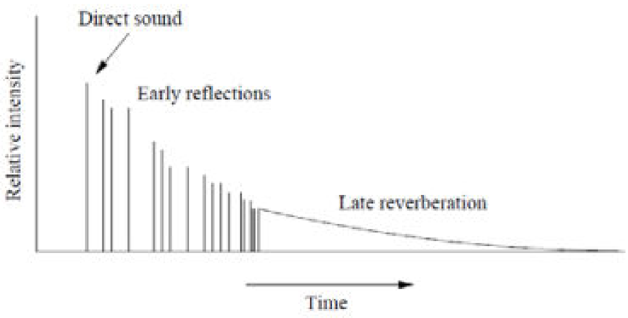
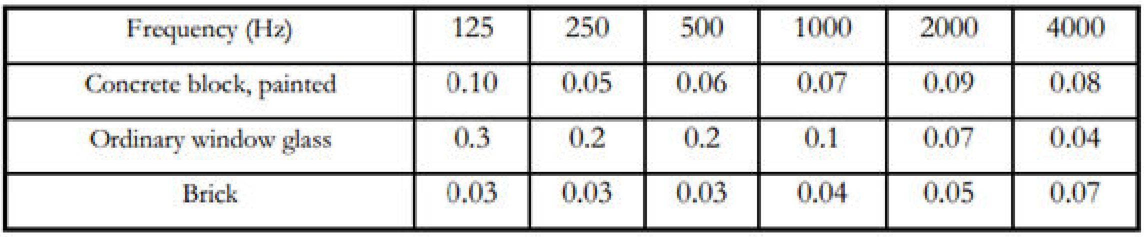


Figure 2.11. Decrease of the sound pressure with time [36].

Speech signals are produced in the 0-20kHz frequency range. Due to the frequency dependency of the absorption coefficients of the walls in the room, the reverberation time RT_{60} measured will also be frequency dependent.

Table 2.3. “Absorption parameter of materials.” [36]



Frequency (Hz)	125	250	500	1000	2000	4000
Concrete block, painted	0.10	0.05	0.06	0.07	0.09	0.08
Ordinary window glass	0.3	0.2	0.2	0.1	0.07	0.04
Brick	0.03	0.03	0.03	0.04	0.05	0.07

Absorption parameters of three common materials found in the walls are shown in Table 2.3. It can be seen from the table that the value of the absorption parameter changes as the frequency changes. Air also absorbs speech energy especially present in the high frequency components of the sound signal, causing the reverberation time to have larger values at a low frequencies and smaller values at high frequencies [36].

2.3.6. Effects of Reverberation on Speech Signals

Reverberation causes numerous destructive impacts on the spectrotemporal characteristics of speech signals. These impacts include temporal smearing, filling dips and gaps in the temporal envelope. It also makes the energy present at low frequencies more prominent and flattens the formant transitions. Impacts of reverberation may be classified as either self-masking or overlap-masking [40]. The self-masking effects are caused by early reflections that arrive at the receiver within 100 ms. after the direct sound whereas the overlap-masking effects are due to late reverberation. Overlap masking smears the direct sound over time and masks the following sounds. Overlap-masking effects of reverberation are the primary contributors to the degradation of speech recognition performance in both human listeners and ASRs [41].

Early studies in room acoustics show that multiple reflections of sound in a room can ideally be expressed as an exponential decay of the sound energy. This causes the room impulse response to have an exponentially decaying shape [42], defined as,

$$h^2(t) \sim e^{-\frac{6\ln(n)}{RT_{60}}t} \tag{2.8}$$

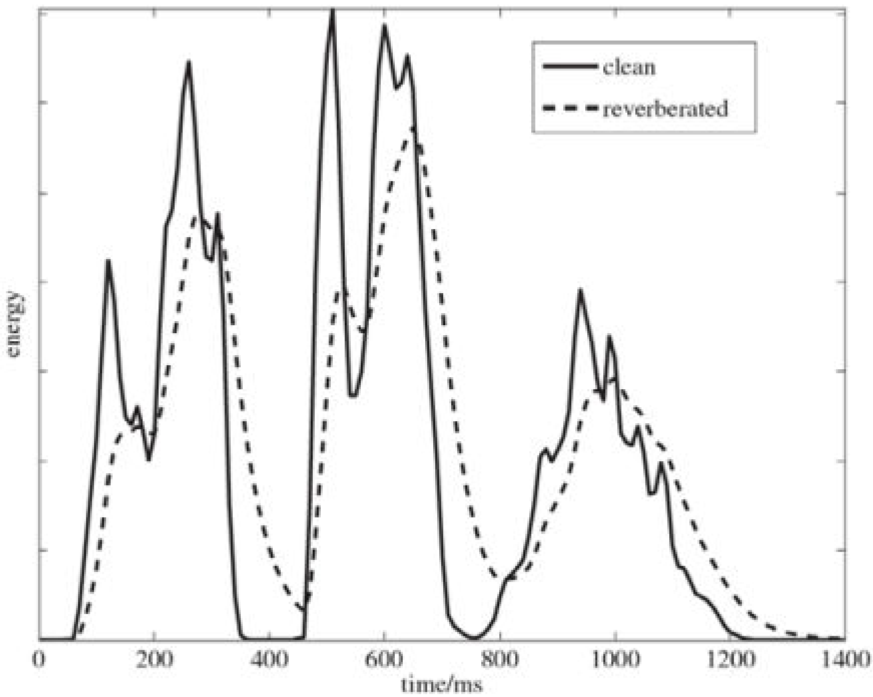


Figure 2.12. Energy contours of clean and reverberated speech signals [42].

Figure 2.12 illustrates the effects of reverberation on the speech signal. Energy is estimated as short-term energy in frames of about 25 ms. It can be seen from the figure that reverberation causes an extension to the overall durations of sounds in the time domain. This extension along with the exponential decay is referred to as the reverberation tail [42]. The same effect will be seen in the frequency domain, too [36].

2.3.7. Effects of Reverberation on Speech Recognition

State-of-the-art ASR systems have a performance comparable to that of humans in environments where clean speech can be obtained. But real-life conditions introduce noise and reverberation into the environment. While humans have a robust speech recognition performance in such adverse conditions with noise and reverberation, automatic speech recognition systems are not robust to these environmental conditions. This is a major limitation in the deployment of speech recognition technologies. Thus, robustness against such adverse environmental conditions received considerable attention lately.

Although there has been some progress on the development of algorithms robust to noise, robustness to reverberation has remained to be a tough challenge. These algorithms show an increase in performance in the presence of stationary noise such but perform poorly in the presence of more realistic degradations such as background music, background speech or reverberation. It should be noted that humans show a relatively good performance in speech recognition in the presence of such challenging conditions [43].

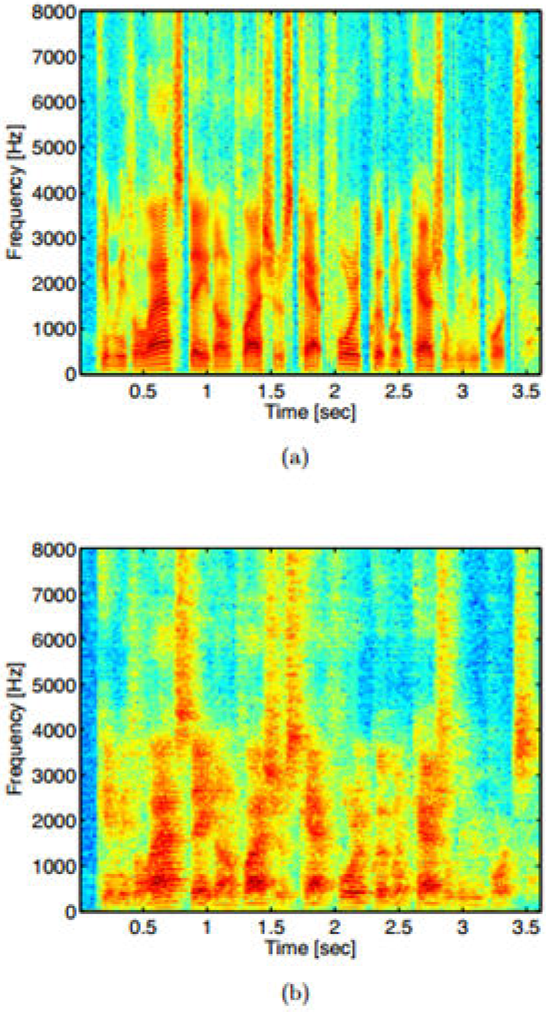
2.3.8. Effects on Human Auditory System

Nabelek and Robinson conducted a study investigating the intelligibility of words in various reverberant conditions using subjects of various ages. The results show that reverberation causes degradation in the intelligibility of speech for subjects of all ages. As the reverberation time increased, intelligibility got worse. In the case of monaural listening, word recognition accuracy was 99.7% for anechoic speech, but recognition accuracy dropped to 97.0%, 92.5%, and 87.7% for reverberation times of 0.4 s., 0.8 s, and 1.2 s, respectively. This same study also revealed that binaural listening instead of the monaural improved speech intelligibility by nearly 5-25% [12].

Reverberation time directly affects the human auditory perception of speech. .For smaller values of reverberation time, the sound and thus the perception is enhanced. On the other hand, large reverberation time values causes sounds to extend in time, interfering with the succeeding sounds and degrading the speech recognition [43].

2.3.9. Impact of Reverberation on ASR

Since reverberation causes sound to remain in the environment even after the original sound ceases to exist, it causes spectro-temporal smearing and thus distortion of the sound signal. Such a distortion results in a remarkable degradation in recognition performance of the ASR systems since these systems basically work by matching the spectral features of the signal with learned patterns. There will be a mismatch between the clean spectral patterns expected by the ASR system and the real distorted spectral patterns.



(a)

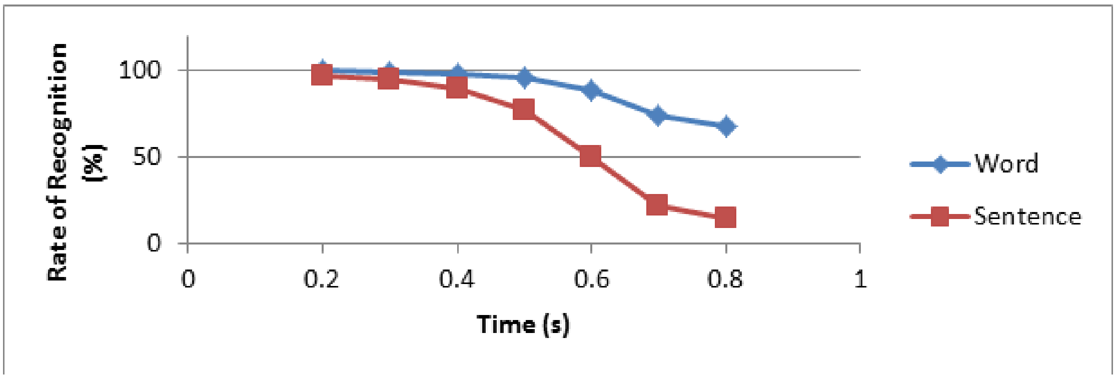
(b)

Figure 2.13. Spectrograms of clean and reverberated speech signals [43].

Impact of reverberation on ASR may also be seen in Figure 2.13. In Figure 2.13a, the spectrogram corresponding to a typical clean speech signal is plotted and in Figure

2.13b the spectrogram corresponding to reverberated speech with an RIR at RT_{60} of 300 ms can be seen. The mismatch between these spectrograms is clear and is detrimental regarding the accuracy of automatic speech recognition systems [43].

Increase in reverberation time is inversely proportional with the recognition performance of the ASR systems. Figure 2.18 shows an experiment of speech recognition conducted under reverberant conditions. The experiment involves recognition of digits from zero to nine in German.



	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Word	99.89	99.47	98.49	96.29	89.1	73.58	68.22
Sentence	97.3	94.61	89.71	76.96	49.51	21.57	13.97

Figure 2.14. Recognition results under reverberant conditions [36].

As can be seen from Figure 2.14, recognition accuracy decreases as the reverberation time increases. This result holds true for both word recognition and sentence recognition accuracies, though sentence recognition seems to be more sensitive to an increase in reverberation time [36].

Figure 2.15 shows the word error rates (WER) for the DARPA RM Database in the presence of reverberation. ASR performance degradation is shown by the WER having a value of 6.7% for clean speech and a value of 51% for reverberation time of 300 ms.

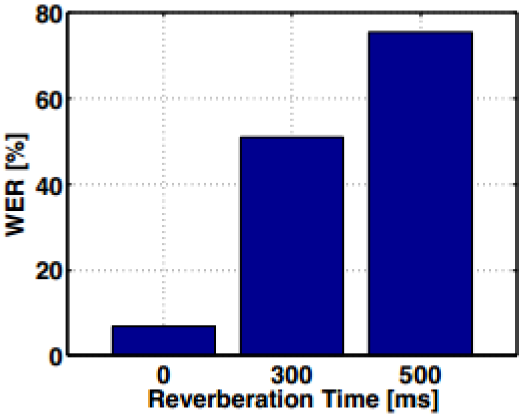


Figure 2.15. Baseline WER in reverberant conditions [43].

Figure 2.15 is an example there is a mismatch between the test and the training set. ASR system is trained on clean speech and testing is carried out using various reverberation conditions. Using oracle knowledge of the test environment it would also be possible to train the system with the data from the test environment, thus performing a matched training and testing. Even matched training is not enough to decrease the word error rate to reasonable levels though. In the case of matched training, the WER has a value of 20% for reverberation time of 500 ms compared to word error rate of 6.7% for clean conditions. The spectro-temporal smearing caused by reverberation leads to interference between the neighboring sounds. But since speech includes different sound units, the effect of reverberation on a particular sound unit also depends on the previous sound units along with the effect of the room impulse response [43].

2.4. Previous Work

Mandel *et al*. investigated the performance of several source separation systems with respect to the increase in the recognition of the ASR systems and human listeners. It has been shown that while ASR performance increases for non-reverberant mixtures in the non-oracle systems, improvement is much restricted for reverberant mixtures. The experiments performed in this paper have a fixed reverberation time of approximately 550 ms [3]. This study is mostly related to finding a consistent evaluation metric for the source separation algorithms in reverberant environments.

Another paper by Park *et al.* proposed a robust interaural time difference (ITD) extraction method for binaural separation of target speech in reverberant environments. Signals are separated by comparing extracted ITDs with the ITD corresponding to the location of target speech [44].

Roman and Wang propose a method for increasing the SNR values by using a two-stage monaural speech separation system. The system takes advantage of both inverse filtering of the RIR and a pitch-based speech segregation method. At the end of the inverse filtering stage, the harmonicity target signal is partially restored while masking signals are further smeared. The system is tested against different levels of reverberation, starting from anechoic mixtures to reverberation time of 0.35 seconds. The mixing includes input SNR of -5 dB, 0 dB and 5 dB and leads to considerable improvement in the output SNR values. This study also includes different noise types and masker locations [45].

Koutras *et al.* present an online BSS method to separate convolutive speech signals in the presence of moving speakers and reverberation. Separation of convolutive speech mixtures is accomplished in the time domain without any prior information using the maximum likelihood estimation principle. The proposed method improves the recognition accuracy of the ASR system by more than 10% in all adverse mixing situations so that it can be used as a front-end to separate simultaneous speech of moving speakers in the presence of reverberation [46].

Another paper by Koutras *et al.* investigate the overdetermined blind speech separation problem and try to improve the speech recognition accuracy of simultaneous speakers in real room environments using a number of microphones larger than the number of sound sources. This method makes use of NxN BSS networks that process all combinations of the mixture signals in the frequency domain. Experiments are conducted using an array of two to ten microphones with two simultaneous speakers. The results of the experiments show that the speech separation performance is improved when the number of microphones exceeds two and the recognition accuracy of an HMM based ASR increases by over 6%. This result indicates that increasing the number of microphones also increase the recognition accuracy obtained [47].

3. METHODOLOGY

3.1. Corpus

The corpus used in the experiments is the GRID corpus. This same corpus is used for both single-talker and two-talker scenarios in order to maintain consistency throughout the research.

GRID is a large sentence corpus including both audio and video data. It is made available for research in the field of multi-talker speech recognition. The corpus consists of 1000 utterances spoken by 34 talkers in total, with 18 being male and the remaining 16 being female. Only the audios supplied in the corpus are used in this research [48].

The sentences from the speakers in the Grid corpus consists of sentences like “place red at F 1 again”, i.e. they are of the form:

<command:4> <color:4> <preposition:4> <letter:25> <digit:10> <adverb:4>

The number of different choices to choose from for each component in the sentence is placed right next to the component in the form shown above. There are 34 talkers in the corpus and the number of sentences per talker is 500, giving a total corpus size of 17000 sentences [49].

3.2. Reverberation

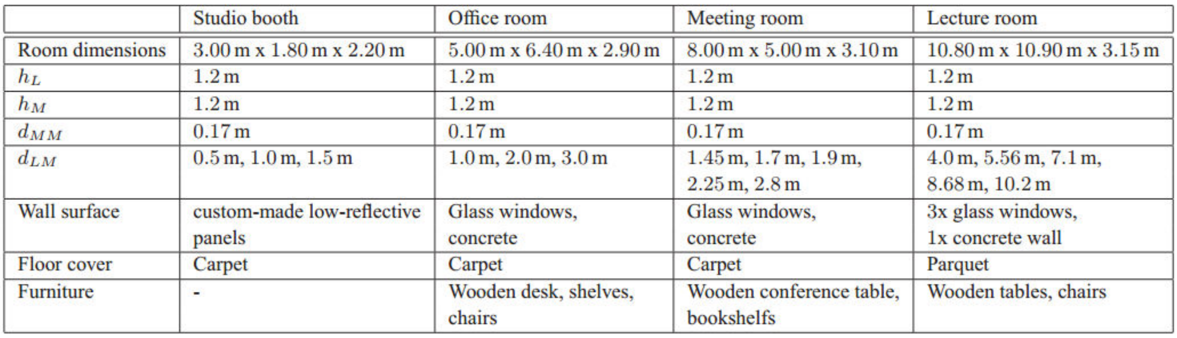
The reverberated utterances are obtained by convolving the speech signals with real room impulse responses (RIRs). Assuming that the overall system is a linear time-invariant system, the RIR would be enough to completely describe the acoustic properties like sound propagation and reflections of the sound that is characteristic of the room it represents. With $h_j(k)$ being the room impulse responses, (where j = 1, ..,M and where M is the number of microphones used in the configuration), s(k) being the non-reverberant speech signal, reverberant signals can be represented by

$$x_j(k) = s(k) * h_j(k) \tag{2.9}$$

where ∗ stands for the convolution operation.

The room impulse responses used in this research are obtained using The Aachen Impulse Response (AIR) database. The Aachen Impulse Response (AIR) database is a set of impulse responses that were measured in a wide variety of rooms. AIR database was assembled using measurements from real places so that signal processing algorithms being developed for reverberant environments can take advantage of these measurements to allow for realistic studies of signal processing algorithms in reverberant life RIRs. The places that these RIRs were measured include a studio booth, an office room, a meeting room and a lecture room. Different speaker-to-microphone distances helped production of room impulse responses with different reverberation times using the same places.

Table 3.1. Properties of different rooms in the AIR database [50].



	Studio booth	Office room	Meeting room	Lecture room
Room dimensions	3.00 m x 1.80 m x 2.20 m	5.00 m x 6.40 m x 2.90 m	8.00 m x 5.00 m x 3.10 m	10.80 m x 10.90 m x 3.15 m
h_L	1.2 m	1.2 m	1.2 m	1.2 m
h_M	1.2 m	1.2 m	1.2 m	1.2 m
d_{MM}	0.17 m	0.17 m	0.17 m	0.17 m
d_{LM}	0.5 m, 1.0 m, 1.5 m	1.0 m, 2.0 m, 3.0 m	1.45 m, 1.7 m, 1.9 m, 2.25 m, 2.8 m	4.0 m, 5.56 m, 7.1 m, 8.68 m, 10.2 m
Wall surface	custom-made low-reflective panels	Glass windows, concrete	Glass windows, concrete	3x glass windows, 1x concrete wall
Floor cover	Carpet	Carpet	Carpet	Parquet
Furniture	-	Wooden desk, shelves, chairs	Wooden conference table, bookshelfs	Wooden tables, chairs

- Low-Reverberant Studio Booth

  It should be noted that one of the most important features of the low-reverberant studio booth is the very low reverberation time of the room and this RT_{60} value is almost constant over the frequency range.

- Office room

  The second room is a typical office room including standard office furniture.

- Meeting room

The meeting room is chosen taking into account the fact that it constitutes a realistic example in terms of representing the acoustic environment where different speakers may be talking simultaneously in the case of a meeting.

- Lecture room

Among the rooms where the RIRs are measured, the largest room in the AIR database is a lecture room with chairs and desks. The loudspeaker is placed in front of the lecturer and the recording system is placed at different rows which are at various lengths from the lecturer. This place is useful for depicting a typical lecture. [50,51]

Table 3.2. RT_{60} of different places in the AIR database [50].

	RT_{60} in s
Studio booth ($\overline{RT}_{60}$= 0.12 s)	**0.08**
	0.11
	0.18
Office room ($\overline{RT}_{60}$= 0.43 s)	**0.37**
	0.44
	0.48
Meeting room ($\overline{RT}_{60}$ = 0.23 s)	0.21
	0.22
	0.21
	0.24
	0.25
Lecture room ($\overline{RT}_{60}$ = 0.78 s)	**0.70**
	0.72
	0.79
	0.80
	0.81
	0.83

Table 3.2 shows the reverberation time for each room on every measuring position.

A mean value RT_{60} for each room is calculated as the average over all positions. The bold values indicate the RT_{60} values that the RIR's chosen from the database had. The reason behind the choice of these reverberation time values is to examine the effects of reverberation as its most important, characteristic feature, namely the reverberation time, varies in a wide range, with RIRs including only the early reflections, as well as ones including late reverberation.

The following figures include the plots for these different room impulse responses from the AIR database that are used for the experiments:

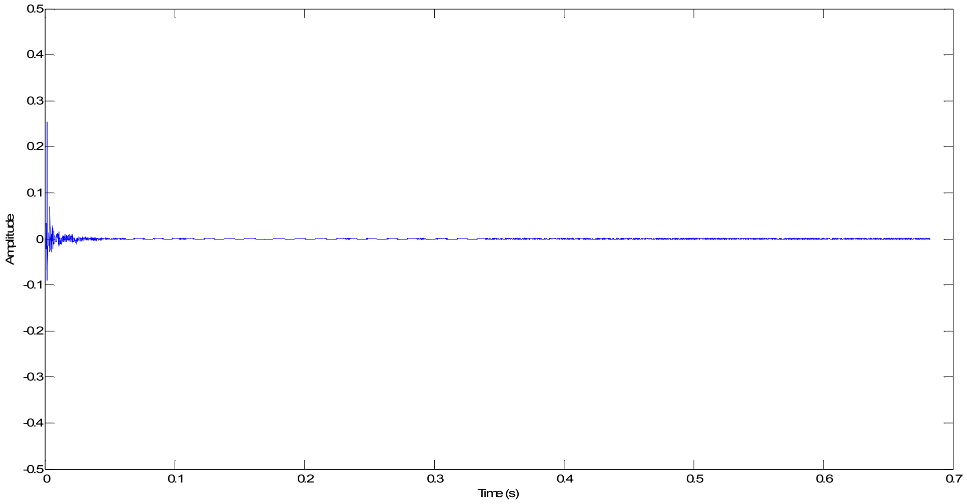


Figure 3.1. RIR with RT_{60}=0.08 seconds.

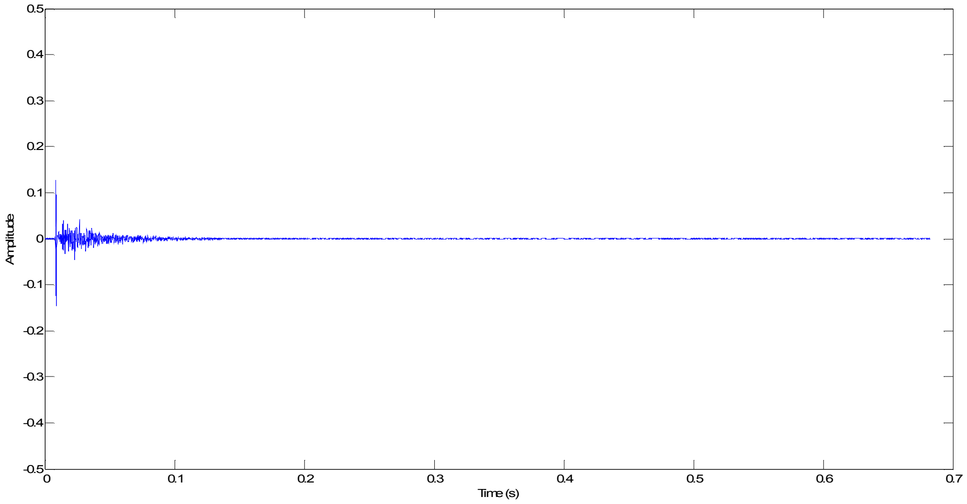


Figure 3.2. RIR with RT_{60}=0.25 seconds.

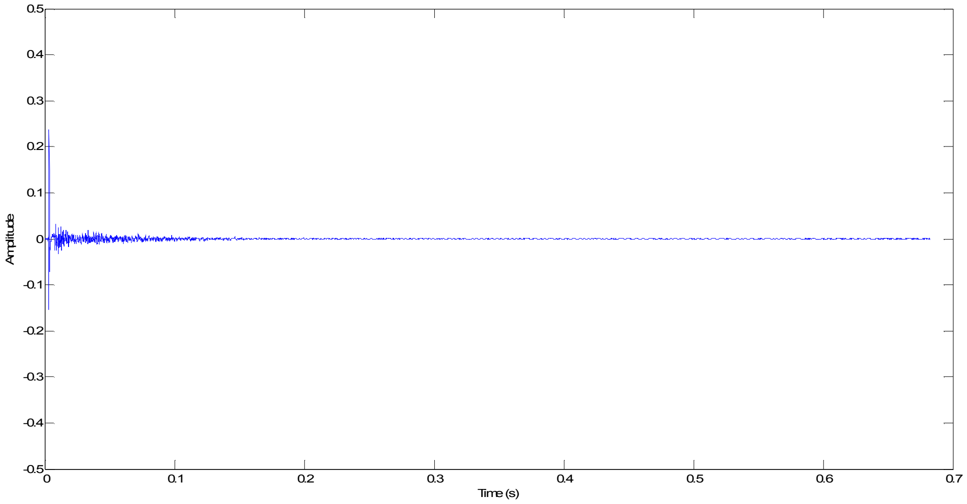


Figure 3.3. RIR with RT_{60}=0.37 seconds.

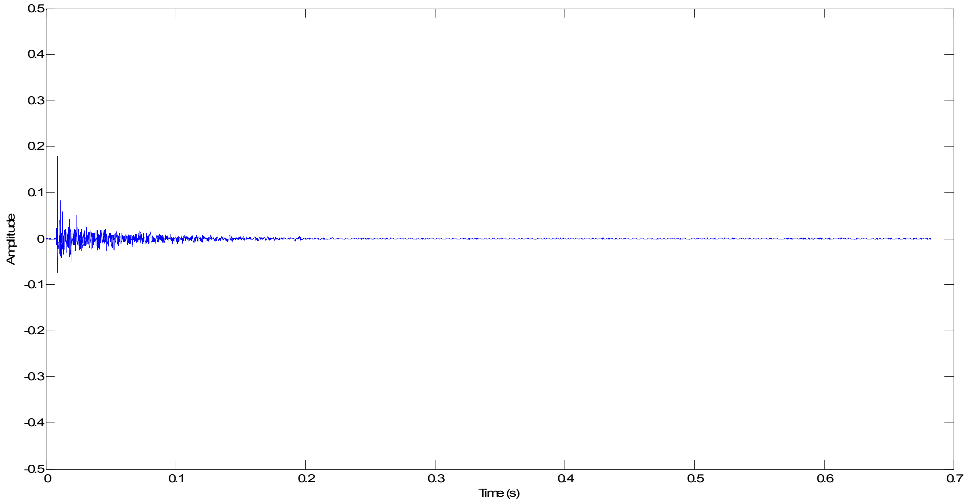


Figure 3.4. RIR with RT$_{60}$=0.48 seconds.

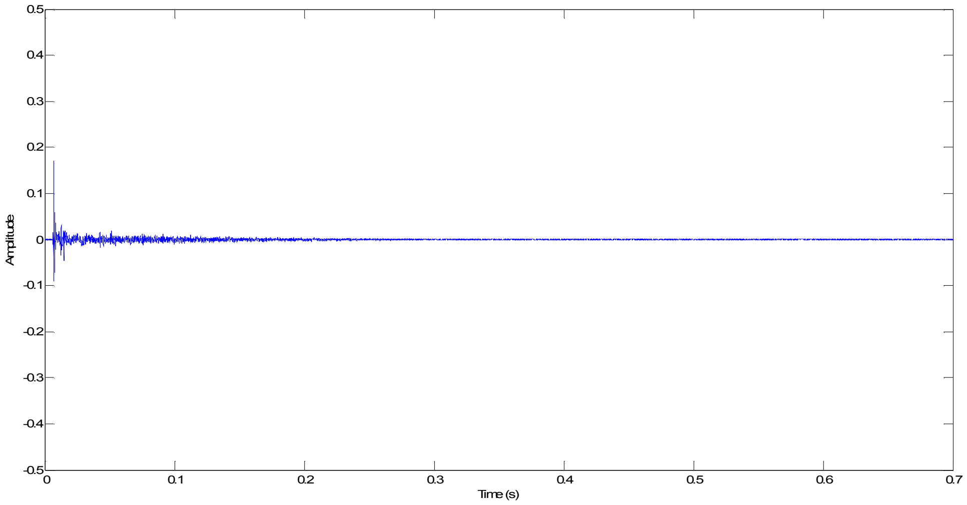


Figure 3.5. RIR with RT$_{60}$=0.70 seconds.

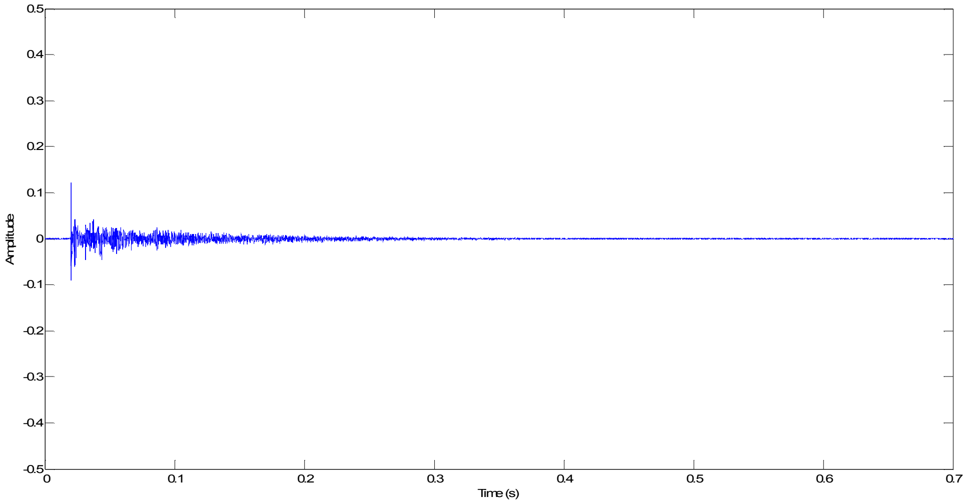


Figure 3.6. RIR with RT_{60}=0.80 seconds.

3.3. Automatic Speech Recognition

HTK is used as the automatic speech recognizer in the experiments. 39-dimensional mel frequency cepstral coefficients (MFCCs) derived from FFT-based log spectra are used for the acoustic feature vectors, with 12 mel-cepstral coefficient, the logarithmic frame energy and the corresponding delta and acceleration coefficients. The corresponding configuration in the HTK is “MFCC_E_D_A”. Whole-word HMM models are preferred due to the small vocabulary size. A left-to-right model topology is chosen with 7 Gaussian mixtures per state with diagonal covariance matrices. Each phoneme in the word is represented by two states, resulting in words with different number of phonemes to be represented with different number of states:

- words with 4 states: at, by, in, a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, x, y, z, one, two, three, eight.
- words with 6 states: bin, lay, place, set, blue, green, red, white, with, four, five, six, nine, now, please, soon.
- words with 8 states: again, zero.
- words with 10 states: seven

After the training and creation of speaker-independent HMMs, speaker dependent HMMs are estimated using these speaker-independent HMMs and performing 4 more iterations of expectation-maximization training using the 500 training utterances for each speaker. The grammar used in the recognition is as shown in Figure 3.7:

```
$command=bin|lay|place|set;
$colour=blue|green|red|white;
$preposition=at|by|in|with;
$letter=a|b|c|d|e|f|g|h|i|j|k|l|m|n|o|p|q|r|s|t|u|v|x|y|z;
$number=zero|one|two|three|four|five|six|seven|eight|nine;
$last=again|now|please|soon;
($command sp $colour sp $ preposition sp $letter sp
$number sp $last)
```

Figure 3.7. ASR grammar.

3.4. Scoring

Each utterance receives a score based on how many of the letter – digit keywords are recognized correctly. If none of the keywords are recognized, the score is 0. If either the letter or the digit is recognized correctly, the score is 1 and if both are correct, score of 2 is given. Afterwards, average is taken over these scores to give the final recognition result. The final score is the average over all the utterances in the test set and it is expressed as a percentage.

The recognized utterances are also recorded in a file as they are recognized. The first item on the line is the name of the file being processed which is indicative of the true utterance transcription e.g. s4_pgad7n. The rest of the line contains the letter and the digit as they are recognized [12,52,53]. Sample lines are shown in Figure 3.8.

```
s1 bgaa5a a 5
s2 bgwi2a y 2
```

Figure 3.8. ASR file name samples.

4. EXPERIMENTS AND RESULTS

4.1. One Talker Scenario

4.1.1. Reverberation

The first step is to obtain the reverberated utterances which are going to form the test set that will be used in the single-talker scenario. A test set of 600 utterances randomly selected from the corpus with each speaker having nearly the same number of utterances is created. This set is used as the basis set for all single-speaker test cases.

The test set with clean utterances are convolved with the selected room impulse responses in a batch process to create the reverberated audios with the reverberation times of interest. These reverberated utterances are put in folders named “0.08”, “0.25”, “0.37”, “0.48”, “0.70”, and “0.80” for further processing in the ASR end. Each folder contained the same 600 utterances reverberated with the specified reverberation levels.

4.1.2. Automatic Speech Recognition

600 utterances at each of the six predefined reverberation levels are tested against the clean and reverberated models. The clean models are produced by training the ASR system with the 17000 clean utterances. For the creation of the reverberated models, a reverberated copy of the Grid corpus training set was prepared by convolving the 17,000 utterance Grid training set with a room impulse response with a reverberation time of 300 ms. This set was used to train the recogniser in order to get the reverberated speech models. The selection of RT_{60}=300 ms for training is due to the fact that typical average room reverberation time is about 300 ms as stated by Sawada *et al.* [54].

Thus, it is important to see the effects of training on the ASR system according to this typical reverberation setting.

4.1.3. Results

Table 4.1. ASR results using clean models.

Reverberation Time (RT_{60}) (s)	% Accuracy
0.08	99.8
0.25	86.9
0.37	76.9
0.48	44.6
0.7	48.4
0.8	24.3

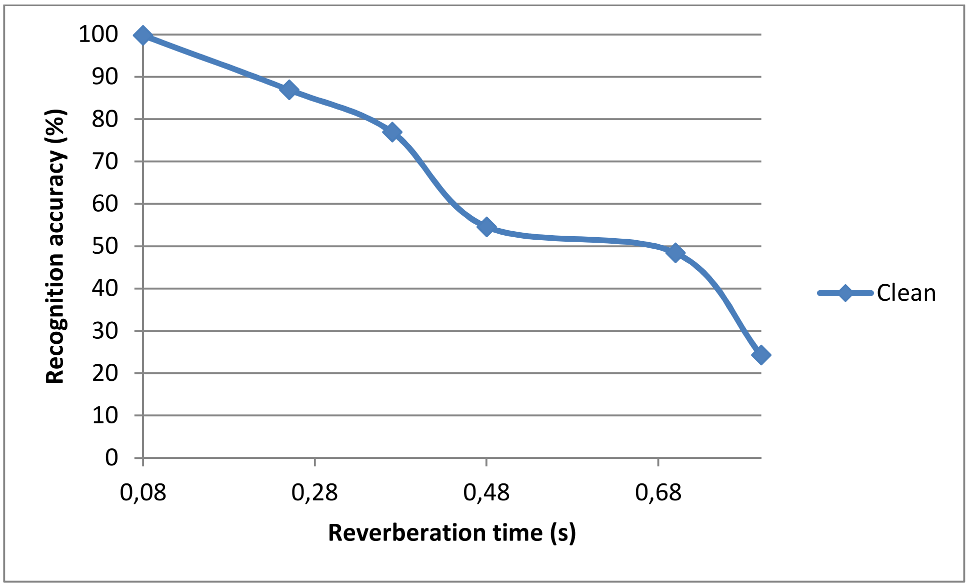


Figure 4.1. Plot of ASR results using clean models in single-talker scenario.

Results of ASR in single-talker case using the clean models show that early reflections do not cause a noticeable change in recognition performance while recognition is severely degraded with an increase in the reverberation time. Recognition performance plateaus in the 0.5-0.7 s range and then starts to rapidly decline as the RT$_{60}$ value approaches 0.80.

Table 4.2. ASR results using reverberated models.

Reverberation Time (RT60) (s)	% Accuracy
0.08	84.2
0.25	96.0
0.37	94.8
0.48	88.7
0.7	84.4
0.8	64.7

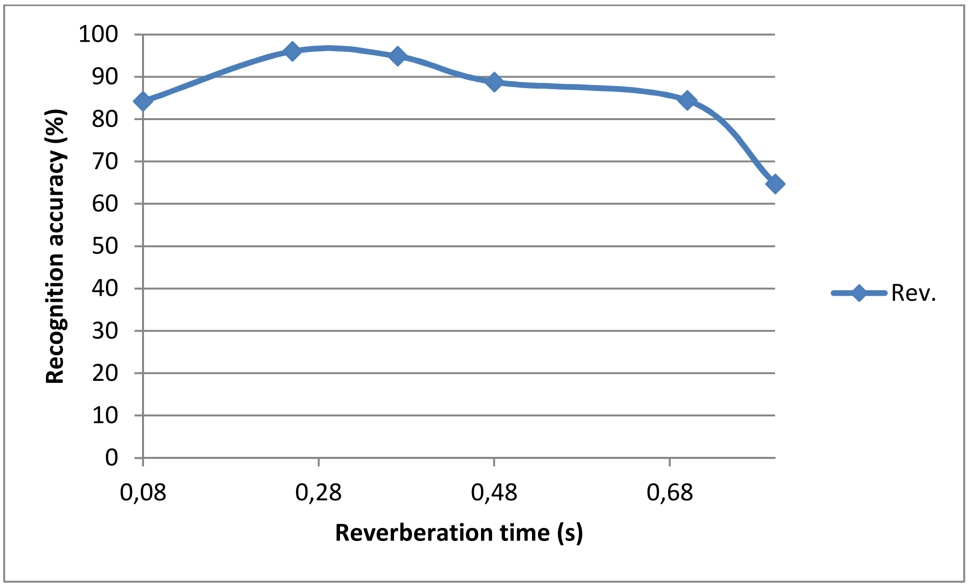


Figure 4.2. Plot of ASR results using reverberated models in single-talker scenario.

It can be observed from Figure 4.2 that the best result is obtained where the test setting matches the training setting, where reverberation time is around 300 ms.

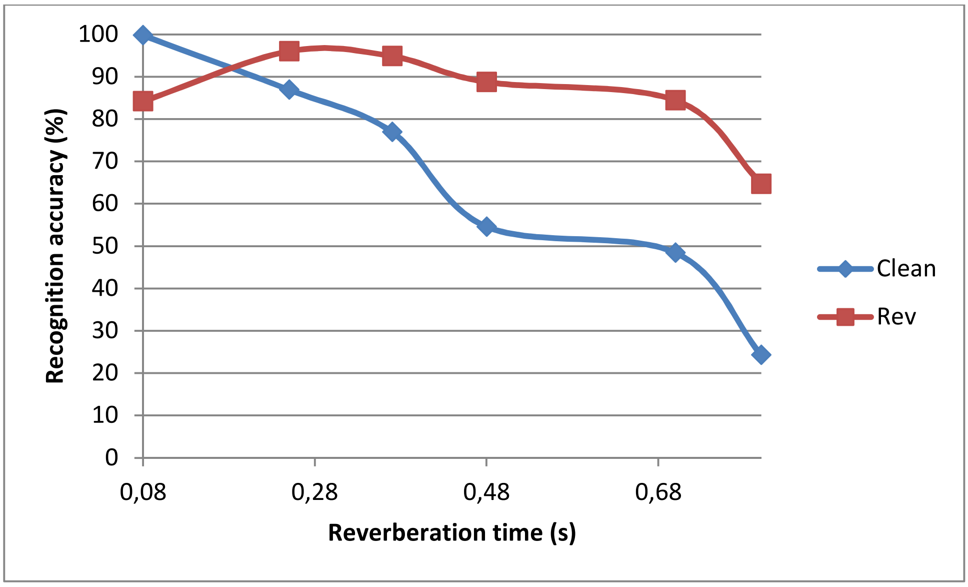


Figure 4.3. Comparison of clean and reverberated training in single-talker scenario.

The two cases, including recognition with clean and reverberated speech models, are superimposed in Figure 4.3 for comparison purposes. Recognition performance declined for RT_{60}<150 ms due to the mismatch between the actual setting and the reverberant model of the training due to the negative effects of reverberation in this range using the clean model being smaller than the effects of the mismatch when using the reverberated model. But the negative effects of reverberation are remedied as reverberation time is past around 150 ms where these negative effects would start to severely degrade the recognition accuracy if the clean model were to be used. Although the mismatch between the trained model and the actual case again starts to increase as reverberation time increases beyond the 300 ms value, performance degradation in the reverberated model case is not as steep as in the clean model case. Even at RT_{60}=0.80 s, recognition accuracy remains well over 60%, whereas in the clean model case it approaches 20%. Hence, it can be stated that training the ASR system with a typical reverberation level of 300 ms positively affects the performance for settings where reverberation time is greater than 200 ms. but causes about 15% average decrease in performance for settings with early reflections only.

4.2. Two-Talker Scenario

4.2.1. Preparation of the Datasets

The first step in the separation process is to prepare the test set that is going to be used as the basis for the reverberation, source separation and recognition processes respectively. The basis test set for the two-talker scenario is created by mixing the utterances of the 34 different speakers used in the single talker scenario test set. Two-by-two mixing is performed starting from the first speaker's first utterance by mixing the current speaker's utterance with the first utterance of the next speaker. Then the second utterance of the first speaker is mixed with the first utterance of the third speaker and this is continued until there remained no utterance belonging to the first speaker that was not used for the mixing. The same process is repeated for the next speaker using the utterances of the speaker subsequent to the current one excluding the ones already used and a test set of 292 utterances is created which will serve as the clean reference in the separation step. The goal behind using such an algorithm for mixing is having a balanced mixing between the utterances of the different speakers. Also with such a mixing algorithm, no mixed sample containing the two utterances of the same speaker are included in the test set, resulting in a cross-mixing among the speakers. The audio samples from each speaker are normalized before mixing so that the signal-to-signal ratio (SSR) used in the two-talker scenario can be assumed to be 0 dB.

The next step is the reverberation process that will create the copies of the two-talker scenario clean test set with the reverberation levels of interest. The same reverberation time values are used in the two-talker scenario as the ones used in the single-talker scenario for comparison purposes. The test set with clean utterances are convolved with the room impulse responses in the AIR database having reverberation time values of "0.08", "0.25", "0.37", "0.48", "0.70", and "0.80" in a batch process to create the reverberated audios with the reverberation times of interest. These reverberated utterances are put in folders named "0.08", "0.25", "0.37", "0.48", "0.70", and "0.80" for further processing in the separation and ASR end. Each folder contained the same 292 utterances reverberated with the specified reverberation levels. Thus, 7 folders containing the clean and reverberated utterances are made ready for subsequent processing.

4.2.2. Speech Separation

Supervised NMF is used for the separation of the mixed speeches. The script used for the separation of the mixed speeches is based on the software provided by Virtanen et.al. [55]. First, a set of spectral atoms are learned for the first and the second speaker by using the single-speaker utterances. Clean single-talker utterances are used for the training in the clean cases and reverberated single-talker utterances are used in the reverberated cases. After the spectral atoms are learned, the mixture signals are represented as linear combinations of these atoms [56].

The sampling frequency used is f_s = 16000 Hz as it is the sampling frequency of the speech samples contained in the GRID corpus. The window size used is 60 ms and the number of adjacent frames used in factorization is 1 because increasing this parameter would slow down the separation process although not giving a remarkable difference in the separation results.

The number of atoms used for training both the first and the second speaker models is 1000 and random sampling is chosen as the method of sampling the atoms from the training data. Identities of the talkers are known during the separation process and this information is used for speaker dependent training of the separation system so that speaker identification is not required as an additional step. After the training, speech dictionaries are formed using these atoms for both of the speakers, the separation process of the mixed signals takes place and each separated mixture is decomposed and saved as two separate audio files labeled with the speaker identity and the abbreviation for the true transcription of the utterance as such information will be necessary in the automatic speech recognition phase. This separation process results in the creation of 584 audio files to be created for the clean and the reverberated cases.

4.2.3. Results

Since a separation system is added before the ASR system in the two-talker scenario, the recognition performance of the overall system suffers a remarkable decrease due to error encountered in the separation process even without the introduction of reverberation in the environment. Recognition accuracy of the separated clean mixtures using the clean models is 38.49%. This accuracy is important for setting a reference point in the comparison of the single-talker and the two-talker scenarios, as well as also being a reference point in the two-talker clean versus reverberated recognition results.

584 separated utterances at each of the six predefined reverberation levels are tested against the clean and reverberated models. For the training of the models for the ASR system, a similar route is followed as the one used in the single-talker scenario since the mixed utterances are separated beforehand. 17000 clean utterances in the GRID corpus are used in the training of the clean models. In the training of the reverberated models, the same set that is created for the single-talker case is used which is formed by convolving the 17,000 utterance Grid training set with a room impulse response with reverberation time of 300 ms.

Table 4.3. ASR results using clean models in two-talker scenario.

Reverberation Time (RT60) (s)	% Accuracy
0.08	40.4
0.25	18.2
0.37	15.4
0.48	10.8
0.7	10.3
0.8	9.8

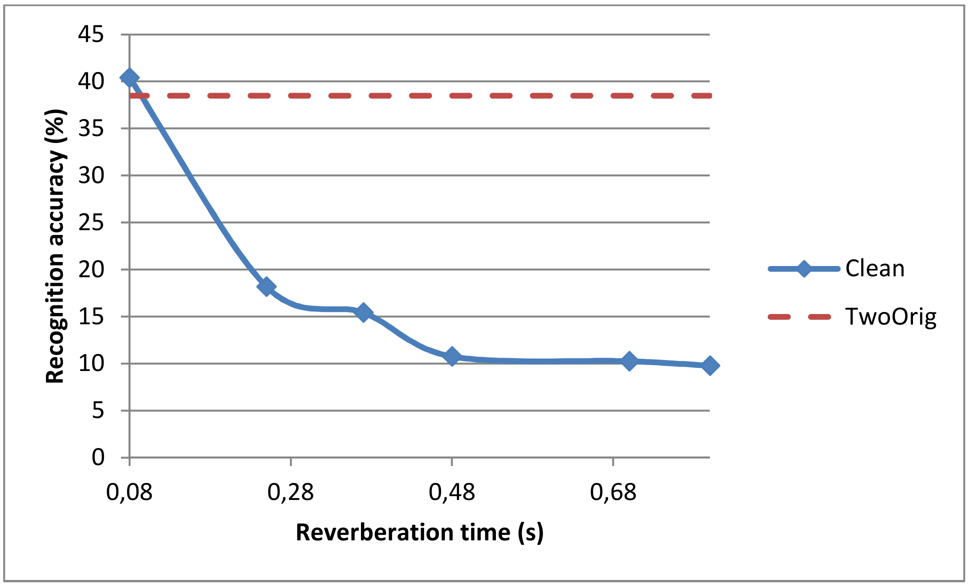


Figure 4.4. Plot of results for clean training in two-talker scenario.

Results of ASR in the two-talker case using the clean models show that utterances with only early reflections are recognized with an accuracy even slightly higher than the accuracy obtained with clean utterances. Early reflections seem to have enhanced the speech features in favor of the speech recognition performance causing such a phenomenon. Introduction of greater reverberation time values results in a rapid decrease in recognition and the recognition accuracy stabilizes around 10% as the reverberation time is around 0.8 s.

Table 4.4. ASR results using a moderately reverberated model in two-talker scenario.

Reverberation Time (RT60) (s)	% Accuracy
0.08	16.8
0.25	22.7
0.37	18.4
0.48	15.3
0.7	13.5
0.8	12.2

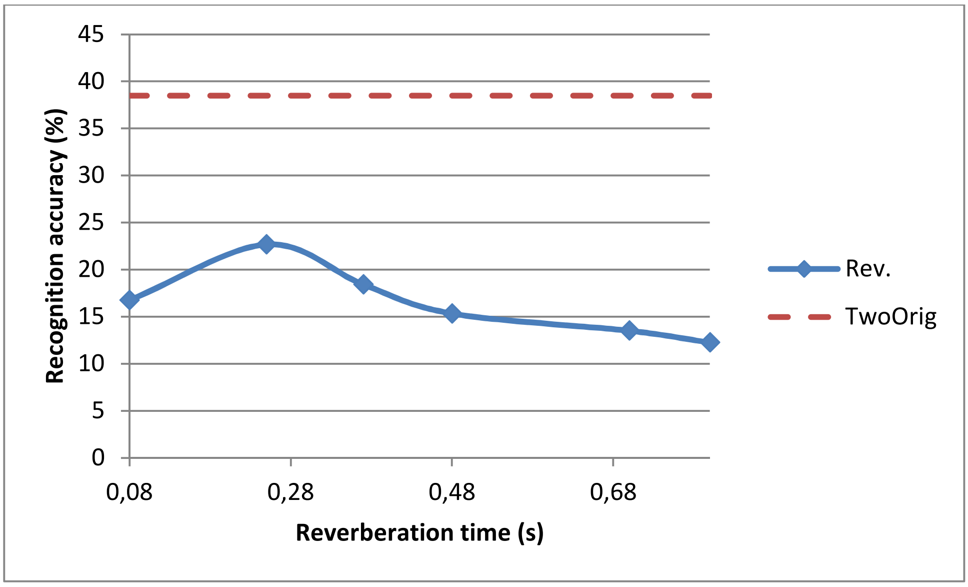


Figure 4.5. Plot of results for reverberated training in two-talker scenario.

In the case where reverberated utterances are tested against the reverberated models, recognition accuracy is mainly determined by the proximity of the real conditions to the reverberation time used in training the models. It can be observed from Figure 4.5 that the best result is obtained where the test setting matches the training setting, where reverberation time is around 300 ms but it is still about 15% below the accuracy of the non-reverberated utterances, namely the reference point. The worst case is when RT_{60}=0.8 s where accuracy drops nearly 25% below the reference point.

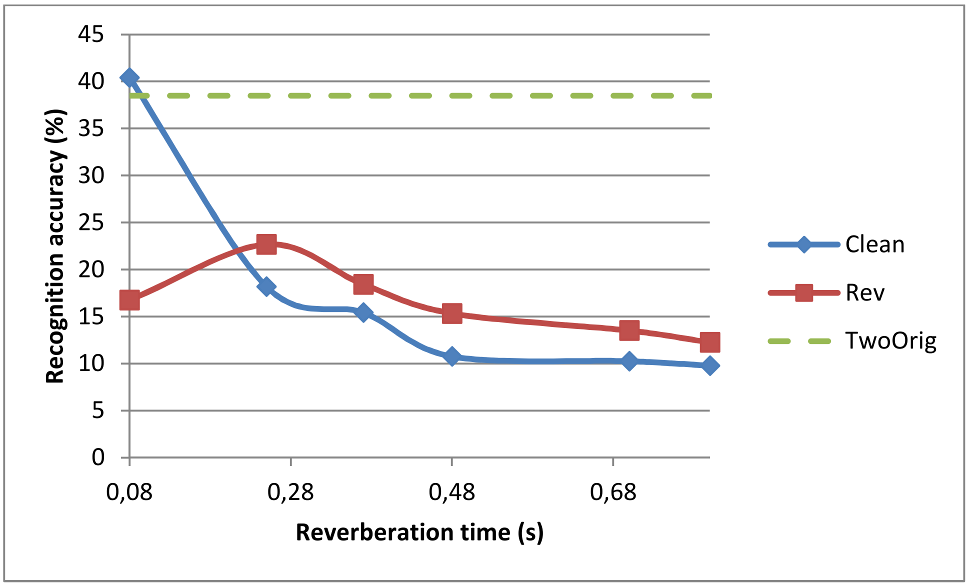


Figure 4.6. Comparison of clean and reverberated training in two-talker scenario.

Comparison between the two cases, where recognition of reverberated utterances is performed using clean and reverberated speech models can be made by investigating Figure 4.6 where both graphs are superimposed. Due to the mismatch between the actual setting and the reverberant model of the training being more detrimental than the reverberation itself for RT_{60}<200 ms, recognition performance is below the clean model case for this range of reverberation time values. The gap is at its widest for RT_{60}=0.08 s as the mismatch causes a degradation of about 25%. For RT_{60}>300 ms, recognition accuracy shows a similar trend for both the clean and reverberated training cases, with the reverberated training case having an average performance boost of 4% over the clean training case.

Hence, it can be stated that training the ASR system with a typical reverberation level of 300 ms positively affects the performance for settings where reverberation time is greater than 200 ms. but causes about 25% average decrease in performance for settings with early reflections only.

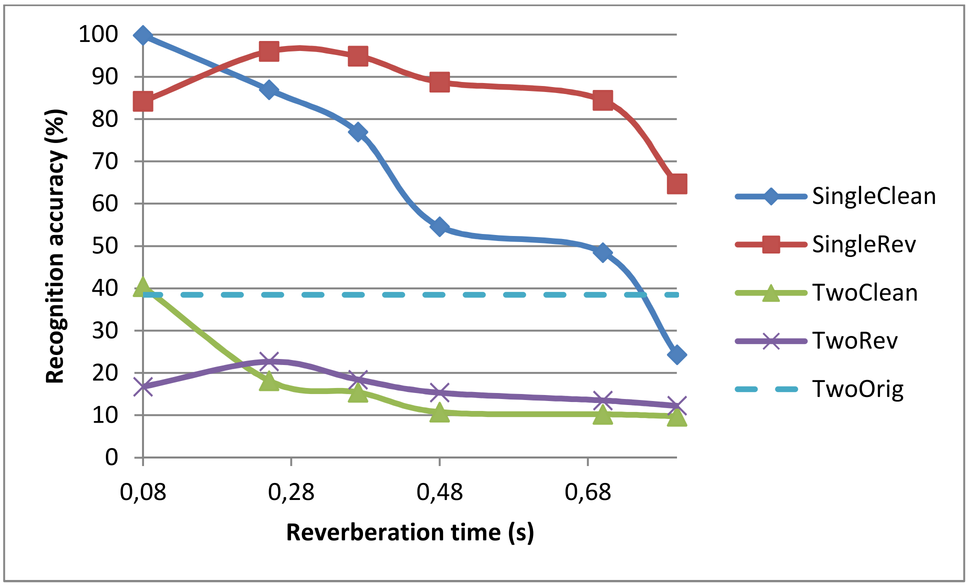


Figure 4.7. Comparison of clean and rev. training in one- and two-talker scenarios.

For RT_{60}=80 ms where there is only early reflections present, mismatched training dramatically decreases the recognition performance to half the reference point in the two-talker scenario, whereas there is a smaller decrease in the single-talker scenario. But in both scenarios, there is a decrease in performance until the RT_{60} value of 200 ms because until this value, destructive effects of training mismatch is greater than that of reverberation. Positive effects of training with RT_{60}=300 ms can be seen after the RT_{60}=200 ms point.

Recognition performance decrease at RT_{60}=0.4 s in the clean case is very steep for the single-talker scenario. Although training with reverberated models follow a similar trend in both scenarios, because there is no such abrupt decrease in the clean case of the two-talker scenario, performance boost of obtained by the reverberated model remains to be more restricted compared to the performance boost in the single-talker scenario.

In addition, reference point in the two-talker scenario is about 40% as opposed to the 100% performance in the single-talker case. If the error were to decrease for the clean model case, a performance increase with a similar ratio would be expected in the reverberated training case. Hence, having a separation system that performs better would also have a corrective effect in the performance boost obtained by training the system with reverberated utterances.

The system is also trained with each reverberation level that is present in the reverberated test sets for comparing the results obtained by using a moderate level of reverberation with the matched training cases and to investigate the effects of mismatch in case of using these reverberation levels as predefined constant values. For each reverberation level, both the separation system and the ASR system are trained with the test speech set that has been reverberated with the RIRs having reverberation time matching the RT_{60} value of the speech signals to be tested. Resulting recognition rates are tabulated in Table 4.5.

Table 4.5. ASR results using models with predefined RT_{60} values in two-talker scenario.

Training RT_{60} (s) **Test set RT_{60} (s)**	**0.08**	**0.25**	**0.37**	**0.48**	**0.7**	**0.8**
0.08	**41.1%**	17.0%	13.5%	12.1%	9.3%	10.0%
0.25	22.1%	**24.6%**	22.4%	20.9%	18.6%	15.2%
0.37	19.8%	15.7%	**24.7%**	22.0%	17.5%	15.4%
0.48	13.3%	12.4%	16.1%	**23.8%**	14.6%	16.0%
0.7	12.5%	10.7%	13.4%	14.5%	**22.6%**	20.6%
0.8	9.4%	7.4%	11.4%	12.4%	14.5%	21.2%

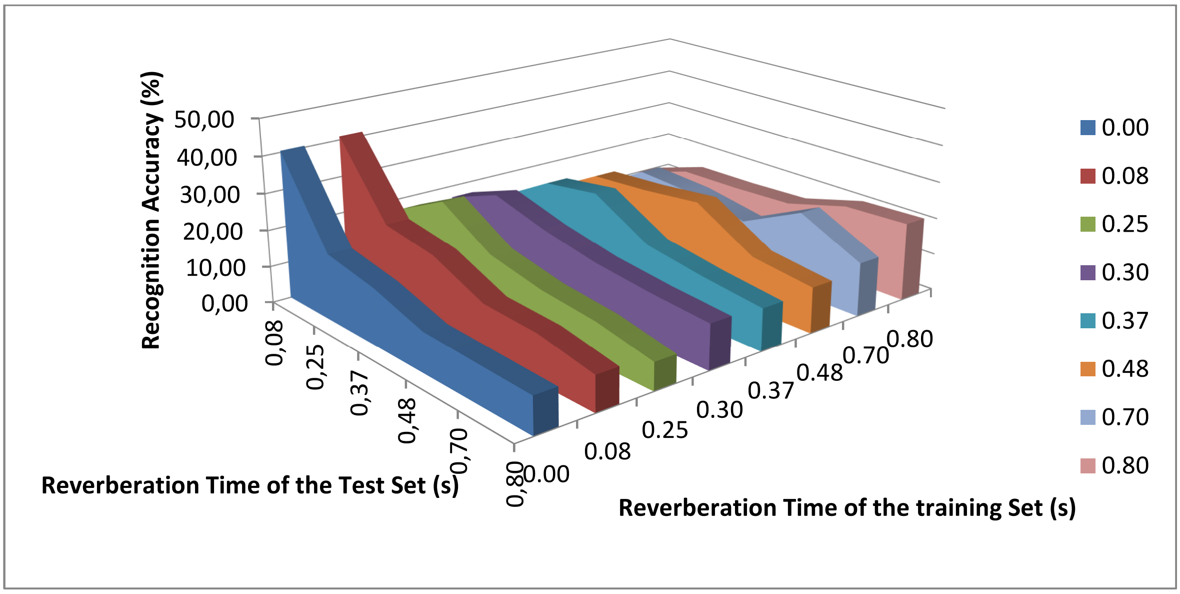


Figure 4.8. Comparison of training with different rev. levels in two-talker scenario.

Figure 4.8 shows the superimposed graph of recognition accuracies obtained by training the system with a moderate level of reverberation and with constant reverberation values corresponding to the ones present in the test set for the two-talker scenario. As can be deduced from the figure, the best performance is obtained where the true reverberation time matches the reverberation time of the training set. The maximum recognition accuracy obtained is around 25% for each case.

5. CONCLUSION

ASR systems perform reasonably well in environments where the speech signals are clean. But real life conditions introduce noise and reverberation. While human speech recognition is robust to noise and reverberation, ASR systems cannot usually cope well with these detrimental effects. Reverberation may cause severe spectro-temporal destructive effects on the speech signal. The performance of an ASR system trained with clean speech degrades remarkably when tested under such conditions.

The two-talker case where the speech of one person overlaps with the other is another type of noise that significantly degrades the performance of ASR systems. Adding a preprocessing stage of source separation should solve the problem if the speech separation were to be done without error but error in the speech separation system also causes a decrease in the overall recognition performance. This speech separation system is also affected by real-life conditions such as noise and reverberation and this means additional margins of error for the overall system.

The objective of this research was to examine the effects of different levels of reverberation on the automatic recognition of concurrent, overlapping speeches in the two-talker monaural scenario and training the system with reverberated samples in order to increase the overall recognition performance of the ASR system. Performance degradation due to the effects of reverberation in single-talker scenario was also examined for comparison purposes. The experiments were carried out using clean and reverberated utterances in training in order to figure out the performance increase that training on reverberated samples would yield.

First, a reference database of clean one-talker utterances was assembled from the GRID corpus. Then its reverberated copies with different reverberation time values were produced. The ASR system was trained with moderately reverberated copies of the utterances with a fixed reverberation time of RT_{60}=300 ms and the results of the ASR system's recognition rate were recorded. Then for the two-talker scenario, utterances were mixed together two by two and the resulting two-talker speech samples were fed into the separation and recognition system. Results were obtained for both cases where the system components were trained with clean speech signals from the speakers and with moderately reverberated copies of the clean signals with RT_{60}=300 ms.

Results of ASR in single-talker case using the clean models showed that early reflections do not cause a noticeable change in recognition performance while recognition is severely degraded with an increase in the reverberation time. For training with reverberated signals, it was observed that the best result is obtained where the test setting matches the training setting, where reverberation time is around 300 ms.

Results of ASR in the two-talker case using the clean models show that utterances with only early reflections are recognized with an accuracy even slightly higher than the accuracy obtained with clean utterances. Introduction of greater reverberation time values results in a rapid decrease in recognition and the recognition accuracy. By testing the reverberated utterances against the reverberated models, it was observed that recognition accuracy is mainly determined by the proximity of the real conditions to the reverberation time used in training the models and the best result is obtained where the test setting matches the training setting, as also was the case in the one-talker scenario, where reverberation time is around 300.

It has been experimentally shown that training the ASR system with a typical reverberation level of 300 ms positively affects the performance for settings where reverberation time is greater than 200 ms. but causes about 25% average decrease in performance for settings with early reflections only. Reverberated training case showed an average performance boost of 4% over the clean training case although this performance it remains below the increase in the single-talker scenario.

In addition, reference point in the two-talker scenario is about 40% as opposed to the 100% performance in the single-talker case. If the error were to decrease for the clean model case, a performance increase with a similar ratio would be expected in the reverberated training case. Hence, having a separation system that performs better would also have a corrective effect in the performance boost obtained by training the system with reverberated utterances.

Lastly, the system is trained for each reverberation time value present in the test set. In each case, the best performance is obtained where the true reverberation time matches the reverberation time of the training set. The maximum recognition accuracy obtained is around 25% for each case. This result indicates that matched training yields the highest recognition rates, thus making use of such an adaptive method would yield the best overall performance in the case where online detection of the present reverberation value of the speech is possible or the reverberation time is constant and known in advance. Otherwise, in the presence of variable reverberation where the current reverberation time value is not continuously calculated, a good choice of reverberation time for training would be to use the moderate reverberation time of 300 ms.

Reverberation time is a frequency dependent value and different RT_{60} values for different frequency ranges may also impact the overall performance of the system, which was a subject not covered in this study. Another factor that can be investigated in terms of its effects on the recognition accuracy is the SSR used when mixing the speech signals in the two-talker case. The speech mixing process in this study constrained the speech signals to come from different speakers. Further studies may be conducted, examining the effects for two speech signals belonging to the same speaker. Gender dependency of the results may also be investigated. Using a database with simultaneous speech from real speakers may also produce more realistic results than using synthetically produced mixtures. Although this research used NMF and exemplar based sparse representation as the method of speech separation, it would be worthwhile extending the research to see how using other techniques would affect the results. The identities of the speakers were given to the algorithm in this study so that speaker identification step was avoided. Effects of reverberation on speaker identification in two-talker monaural case is another topic that can be further studied.

REFERENCES

1. Huang, X., A. Acero and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2001.

2. Cook, M., J.R. Hershey and S.J. Rennie, “Monaural Speech Separation And Recognition Challenge”, *Computer Speech and Language*, Vol. 24, pp. 1-15, 2010.

3. Mandel, M.I., S. Bressler, B.S. Cunningham and D.P.W. Ellis, “Evaluating Source Separation Algorithms With Reverberant Speech”, *IEEE Transactions On Audio, Speech And Language Processing*, Vol. 18, pp. 1872-1883, 2010.

4. Zue V., R. Cole., *Survey of the State of the Art in Human Language Technology (Studies in Natural Language Processing)*, Cambridge University Press, NY, USA, 1998

5. Hermansky, H., “Perceptual Linear Predictive (PLP) Analysis Of Speech”, *Journal of the Acoustical Society of America*, Vol. 87, pp. 1738–1752, 1990.

6. Zue, V., J. Glass, M. Phillips and S. Seneff, “The MIT SUMMIT Speech Recognition System: A Progress Report”, *Proceedings of the Third DARPA Speech and Natural Language Workshop, Hidden Valley, Pennsylvania*, 1990.

7. Jurafsky D. and J.H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition,* Prentice-Hall, Inc, Lebanon, IN, *USA, 2008.*

8. Pallett, D.S., “A Look At NIST's Benchmark ASR Tests: Past, Present, And The Future”, *Proceedings of ASRU*, pp.483-488, Virgin Islands, USA, 2003.

9. Young, S., G. Evermann, M. Gales, T. Hain, D.Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, *The HTK Book (for HTK Version 3.4),* 2009.

10. Haykin, S., Z. Chen, "The Cocktail Party Problem", *Neural Computation*, Vol. 17, pp. 1875-1902, 2006

11. Divenyi, P., B. Cunningham, D. Ellis, D. Wang, "Separating Speech from Speech Noise", 2005, http://labrosa.ee.columbia.edu/projects/speechsep , accessed at January 2013.

12. Mandel, M.I., *Binaural Model-Based Source Separation and Localization*, Ph.D. Thesis, Columbia University, 2010.

13. Mowlaee, P., *New Strategies for Single-channel Speech Separation*, Ph.D. Thesis, Aalborg University, 2010.

14. Hyvarinen, A. and E. Oja, "Independent Component Analysis: Algorithms And Applications", *Journal of Neural Networks*, Vol. 13, pp. 411–430, 2000.

15. Jang, G.J. and T.W. Lee, "A Maximum Likelihood Approach To Single Channel Source Separation", *The Journal of Machine Learning Research*, Vol. 4, pp. 1365–1392, 2004.

16. Pedersen, M.S., D. Wang, J. Larsen and U. Kjems, "Overcomplete Blind Source Separation By Combining ICA And Binary Time-Frequency Masking", *Proceedings of IEEE Workshop on Machine Learning for Signal Processing*, Mystic, CT, USA, 2005.

17. Pedersen, M.S., D. Wang, J. Larsen and U. Kjems, "Separating Underdetermined Convolutive Speech Mixtures", *Independent Component Analysis and Blind Signal Separation*, Vol. 3889, pp. 674–681, 2006.

18. Pedersen, M.S., D. Wang, J. Larsen and U. Kjems, “Two-Microphone Separation Of Speech Mixtures”, *IEEE Transaction on Neural Networks*, Vol. 19, pp. 475 –492, 2008.

19. H. Laurberg, “Uniqueness Of Non-Negative Matrix Factorization”, Proceedings of IEEE 14th Workshop on Statistical Signal Processing, pp. 44–48, Madison, WI, USA, 2007.

20. Virtanen, T., “Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria,”, *IEEE Transaction on Audio, Speech, and Language Processing*, Vol. 15, pp. 1066–1074, 2007.

21. Schmidt, M. and R. Olsson, “Single-Channel Speech Separation Using Sparse Non-Negative Matrix Factorization”, *Proceedings of Interspeech*, pp. 2614–2617, Pittsburgh, PA, USA, 2006.

22. L. Benaroya, F. Bimbot, and R. Gribonval, “Audio Source Separation With A Single Sensor,” *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, pp. 191–199, 2006.

23. M. N. Schmidt and H. Laurberg, “Non-Negative Matrix Factorization With Gaussian Process Priors” *Computational Intelligence and Neuroscience*, Vol. 2008, pp. 152-162, 2008.

24. King, B. and L. Atlas, “Single-Channel Source Separation Using Simplified Training Complex Matrix Factorization”, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4206–4209, Dallas, TX, USA, 2010.

25. Parry, R. and I. Essa, “Incorporating Phase Information For Source Separation Via Spectrogram Factorization”, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 661–664, Honolulu, HI, USA, 2007.

26. Gemmeke, J.F., T. Virtanen and A. Hurmalainen, “Exemplar-Based Sparse Representations for Noise Robust Automatic Speech Recognition”, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, pp. 2067 – 2080, 2011.

27. Lee, D. D. and H. S. Seung, “Algorithms For Non-Negative Matrix Factorization”, *Proceedings of Neural Information Processing Systems*, pp. 556–562, Vancouver, British Columbia, Canada, 2001.

28. Cand´es, E. J., and M. B. Wakin, “An Introduction To Compressive Sampling”, *IEEE Signal Processing Magazine*, Vol. 25, pp. 21–30, 2008.

29. Sivaram, G.S.V.S., S.K. Nemala, M. Elhilali, T.D. Tran and H. Hermansky, “Sparse Coding For Speech Recognition”, *Proceedings of International Conference on Audio, Speech and Signal Processing*, pp. 4346 - 4349, Dallas, TX, USA, 2010.

30. J. F. Gemmeke, L. ten Bosch, L.Boves, and B. Cranen, “Using Sparse Representations For Exemplar Based Continuous Digit Recognition”, *Proceedings of EUSIPCO*, pp. 1755– 1759, Glasgow, Scotland, 2009

31. Gemmeke, J.F., H. Van hamme, B. Cranen, and L. Boves, “Compressive Sensing For Missing Data Imputation In Noise Robust Speech Recognition”, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 4, pp., 272–287, 2010.

32. Wachter, M.D., M. Matton, K. Demuynck, P. Wambacq, R. Cools and D. Van Compernolle, “Template Based Continuous Speech Recognition”, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, pp. 1377–1390, 2007.

33. Radfar, M.H., R.M. Dansereau and A. Sayadiyan, “Monaural Speech Segregation Based On Fusion Of Source-Driven With Model-Driven Techniques” *Speech Communication*, Vol. 49, pp. 464– 476, 2007.

34. Roweis, S., “Factorial Models And Refiltering For Speech Separation And Denoising”, Proceedings of Interspeech, pp. 1009–1012, Geneva, Switzerland, 2003.

35. “Reverberation”, 2012, http://en.wikipedia.org/wiki/Reverberation, accessed at March 2013.

36. Beining, M., *Improving Automatic Speech Recognition for a Speech Input in hands-free Mode Inside Rooms*, MS. Thesis, Niederrhein University, 2010.

37. Bradley, J.S., H. Sato and M. Picard, “On The Importance Of Early Reflections For Speech In Rooms”, *Journal of the Acoustical Society of America*, Vol. 113, pp. 3233–3244, 2003.

38. Allen, J.B. and D.A. Berkley, “Image Method For Efficiently Simulating Small-Room Acoustics*”, Journal of the Acoustical Society of America*, Vol. 65, pp. 943–950, 1979.

39. Schroeder, M.R., “Digital Simulation Of Sound Transmission In Reverberant Spaces”, *Journal of the Acoustical Society of America*, Vol. 47, pp. 424–431, 1970.

40. Nabelek, A.K., T.R. Letowski and F.M. Tucker, “Reverberant Overlap And Self-Masking In Consonant Identification”, *The Journal of the Acoustical Society of America*, Vol. 86, pp. 1259–1265, 1989.

41. Boril, H., O. Sadjadi, J.H.L. Hansen, “A Study On Combined Effects Of Reverberation And Increased Vocal Effort On ASR”, *LISTA‘12 Workshop*, pp. 16–19, Edinburgh, UK, 2012.

42. Hirsch, H.G., *Automatic Speech Recognition in Adverse Acoustic Conditions, in Advances in Digital Speech Transmission,* John Wiley & Sons, Ltd, Chichester, UK, 2008.

43. Kim, C., *Signal Processing For Robust Speech Recognition Motivated By Auditory Processing*, Ph.D. Thesis, Carnegie Mellon University, 2010.

44. Park, J.H. and H.K.Kim, "Interaural Time Difference Estimation for Binaural Speech Separation in Reverberant Multi-source Environments", *Advanced Signal Processing Conference Proceedings*, 2012.

45. Roman, N., D. Wang, "A Pitch-based Model for Separation of Reverberant Speech", Proceedings of Interspeech, pp. 2109-2112, Lisbon, Portugal, 2005.

46. Koutras, A., E. Dermatas, G. Kokkinakis, "Blind Speech Separation Of Moving Speakers In Real Reverberant Environments", *Proceedings of the ICASSP*, Istanbul, Turkey, 2000.

47. Koutras, A., E. Dermatas, G. Kokkinakis, "Improving Simultaneous Speech Recognition in Real Room Environments Using Overdetermined Blind Source Separation", *Proceedings of the Interspeech*, Aalborg, Denmark, 2001.

48. Cook, M., "The GRID Audiovisual Sentence Corpus", 2007, http://spandh.dcs.shef.ac.uk/gridcorpus, accessed at January 2013.

49. Cooke, M., J. Barker, S. Cunningham and X. Shao, "An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition", *The Journal of the Acoustical Society of America*, Vol. 120, pp. 2421-2424, 2006.

50. Jeub, M., M. Schäfer, H. Krüger, C.M. Nelke, C. Beaugeant and P. Vary, "Do We Need Dereverberation for Hand-Held Telephony?", *Proceedings of International Congress on Acoustics*, Sydney, Australia,. 2010.

51. Jeub, M., M. Schäfer and P. Vary, "Binaural Room Impulse Response Database for the Evaluation of Dereverberation Algorithms*", Proceedings of IEEE International Conference on Digital Signal Processing*, pp. 1–4 , Santorini, Greece, July 2009.

52. Ma, N.,Y. Lu and Martin Cooke, “Speech Separation Challenge: Baseline Recogniser And Scoring Scripts”, 2006, http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge/recogniser.pdf, accessed at March 2013.

53. Vincent, E., “The 2nd Chime Challenge Baseline Scoring, Decoding and Training Scripts”, 2012, http://spandh.dcs.shef.ac.uk/chime_challenge/grid/README.pdf, accessed at February 2013.

54. Sawada, H., S. Araki and S. Makino, “Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment”, *IEEE Transactions on Audio, Speech & Language Processing - TASLP*, Vol. 19, pp. 516-527, 2011.

55. Virtanen, T., “Audio processing software”, 2012, http://www.cs.tut.fi/~tuomasv/software.html, accessed at March 2013.

56. Virtanen, T., J.F. Gemmeke and B. Raj, “Active-Set Newton Algorithm for Overcomplete Non-Negative Representations of Audio”, *Transactions on Audio Speech and Language Processing, not yet published.*