We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

5

Dereverberation and Denoising Techniques for ASR Applications

Fernando Santana Pacheco and Rui Seara
Federal University of Santa Catarina
Brazil

1. Introduction

Over the last few years, advances in automatic speech recognition (ASR) have motivated the development of several commercial applications. Automatic dictation systems and voice dialing applications, for instance, are becoming ever more common. Despite significant advances, one is still far from the goal of unlimited speech recognition, i.e., recognition of any word, spoken by any person, in any place, and by using any acquisition and transmission system. In real applications, the speech signal can be contaminated by different sources of distortion. In hands-free devices, for instance, effects of reverberation and background noise are significantly intensified with the larger distance between the speaker and the microphone. If such distortions are uncompensated, the accuracy of ASR systems is severely hampered (Droppo & Acero, 2008). In the open literature, several research works have been proposed aiming to cope with the harmful effects of reverberation and noise in ASR applications (de la Torre et al., 2007; Huang et al., 2008). Summarizing, current approaches focusing on ASR robustness to reverberation and noise can be classified as model adaptation, robust parameterization, and speech enhancement.

The goal of this chapter is to provide the reader with an overview of the current state of the art about ASR robustness to reverberation and noise, as well as to discuss the use of a particular speech enhancement approach trying to circumvent this problem. For such, we choose to use spectral subtraction, which has been proposed in the literature to enhance speech degraded by reverberation and noise (Boll, 1979; Lebart & Boucher, 1998; Habets, 2004). Moreover, taking into consideration that ASR systems share similar concerns about this problem, such an approach has also been applied successfully as a preprocessing stage in these applications.

This chapter is organized as follows. Section 2 characterizes the reverberation and noise effects over speech parameters. An overview of methods to compensate reverberation and noise in ASR systems is briefly discussed in Section 3, including classification and comparison between different approaches. A discussion of spectral subtraction applied to reverberation reduction is presented in Section 4. In that section we examine how to adjust the parameters of the algorithm; we also analyze the sensitivity to estimation errors and changes in the room response. The combined effect of reverberation and noise is also assessed. Finally, concluding remarks are presented in Section 5.

Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert,
ISBN 978-953-7619-29-9, pp. 550, November 2008, I-Tech, Vienna, Austria

2. Reverberation and noise

Speech communication is so natural to humans that we usually do not perceive some effects. Before reaching a microphone or the listener's ears, speech signals may be modified by the medium in which they are propagating (enclosure). In an ideal anechoic chamber, the signal follows only one path from the source to the receiver. But in typical rooms, surfaces (walls and furniture) reflect the emitted sound; the microphone receives a stream of reflected signals from multiple propagation paths. The whole set of reflections is termed reverberation. Although in this chapter we shall discuss methods to reduce this effect, reverberation is not detrimental at all times. It may give the listener the spatial impression of the enclosure (Everest, 2001); it also increases both the "liveness" and "warmth" of the room, especially important in music. On the other hand, reverberation in excess causes loss of intelligibility and clarity, harming communication or musical performance.

The effect of reverberation can be modeled as the processing of a signal by a linear time invariant system. This operation is represented by the convolution between the room impulse response (RIR) and the original signal, expressed as

$$y(n) = x(n) * h(n) \tag{1}$$

where $y(n)$ represents the degraded speech signal, $x(n)$, the original (without degradation) speech signal, $h(n)$ denotes the room impulse response, and $*$ characterizes the linear convolution operation.

In this approach, room reverberation is completely characterized by the RIR. Fig. 1 shows a typical impulse response measured in a room. A RIR can be usually separated into three parts: the direct response, initial reflections, and late reverberation. The amount of energy and delay of each reflection causes different psychoacoustic effects. Initial reflections (or early reverberation) are acoustically integrated by the ears, reinforcing the direct sound. Since initial reflections do not present a flat spectrum, a coloration of the speech spectrum occurs (Huang et al., 2008). Late reverberation (or reverberation tail) causes a different effect called overlap masking. Speech signals exhibit a natural dynamics with regions presenting noticeably different energy levels, as occurs between vowels and consonants. Reverberation tail reduces this dynamics, smearing the energy over a large interval and masking lower energy sounds.

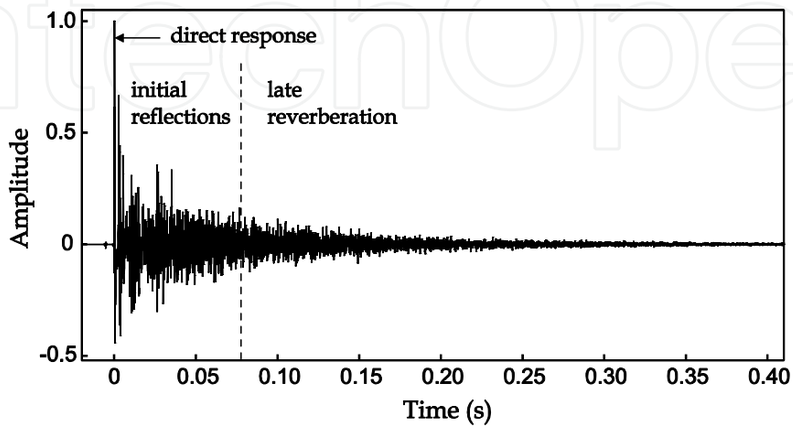


Fig. 1. A typical room impulse response.

It is worth to examine here a real-world example. Fig. 2(a) and (b) illustrate, respectively, the speech signal corresponding to the utterance "enter fifty one" and the associated spectrogram. Notice in these figures the mentioned dynamics in time and the clear harmonic structure with speech resonances marked by darker lines in the spectrogram [see Fig. 2(b)]. Reverberation is artificially incorporated to the original speech signal, by convolving this speech segment with the RIR displayed in Fig. 1. Fig. 2(c) and (d) show, respectively, the reverberant version and the corresponding spectrogram. Now, observe in Fig. 2(c) that the signal is smeared in time, with virtually no gap between phonemes. In addition, notice the difficulty to identify the resonances in Fig. 2(d).

So, how to measure the level of reverberation or how to assess the acoustic quality of a room? Much research has been carried out to define objective parameters correlated with the overall quality and subjective impression exhibited by a room. In this chapter, we present two important parameters used to measure the level of reverberation of an enclosure: reverberation time and early to late energy ratio.

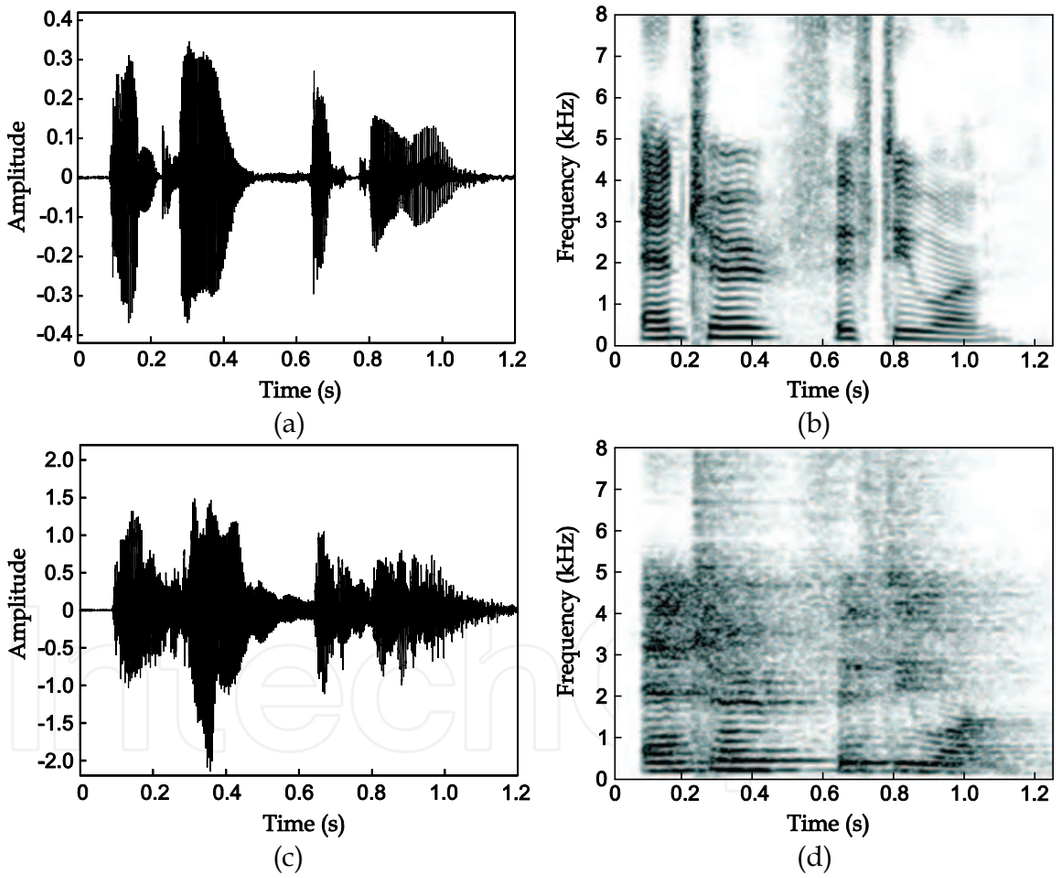


Fig. 2. Reverberation effect over a speech signal. (a) Original speech signal corresponding to the utterance "enter fifty one" and (b) associated spectrogram. (c) Reverberated version of the same previous signal and (d) corresponding spectrogram.

Reverberation time (T_{60}, RT_{60} or $RT60$) is defined as the time interval required for the reverberation to decay 60 dB from the level of a reference sound. It is physically associated with the room dimensions as well as with the acoustic properties of wall materials. The

measurement of the reverberation time is computed through the decay curve obtained from the RIR energy (Everest, 2001). The result can be expressed in terms of either a broadband measure or a set of values corresponding to frequency-dependent reverberation times (for example, RT_{500} corresponds to the reverberation time at the frequency band centered in 500 Hz). To give the reader an idea of typical values, office rooms present T_{60} between 200 ms and 600 ms while large churches can exhibit T_{60} in the order of 3 s (Everest, 2001).

Another objective indicator of speech intelligibility or music clarity is called early to late energy ratio (speech) or clarity index (music) (Chesnokov & SooHoo, 1998), which is defined as

$$C_T = 10\log_{10}\frac{\int_0^T p^2(t)}{\int_T^\infty p^2(t)} \tag{2}$$

where $p(t)$ denotes the instantaneous acoustic pressure and T is the time instant considered as the threshold between early and late reverberation. For speech intelligibility evaluation, it is usual to consider C_{50} ($T = 50$ ms), while C_{80} is a measure for music clarity (Chesnokov & SooHoo, 1998).

Now, considering the separation between early and late reverberation, $h(n)$ can be expressed as

$$h(n) = \begin{cases} 0, & n < 0 \\ h_{\mathrm{d}}(n), & 0 \le n \le N_{\mathrm{d}} \\ h_{\mathrm{r}}(n), & n > N_{\mathrm{d}} \end{cases} \tag{3}$$

where $h_{\mathrm{d}}(n)$ denotes the part of the impulse response corresponding to the direct response plus early reverberation, $h_{\mathrm{r}}(n)$, the other part of the response relating to the late reverberation, $N_{\mathrm{d}} = f_{\mathrm{s}} T$ is the number of samples of the response $h_{\mathrm{d}}(n)$, and f_{s}, the considered sampling rate.

Besides reverberation, sound is also subjected to degradation by additive noise. Noise sources, such as fans, motors, among others, may compete with the signal source in an enclosure. Thus, we should also include the noise effect over the degraded signal model $y(n)$, rewriting (1) now as

$$y(n) = x(n) * h(n) + v(n) \tag{4}$$

where $v(n)$ represents additive noise.

3. Overview of dereverberation and denoising methods

Before introducing methods to tackle reverberation and noise, we present a brief overview of current state-of-the-art ASR technology. The current generation of ASR systems is based on a statistical framework (Rabiner & Juang, 2008). It means that, before system deployment (or test), a first phase of training is mandatory. During training, a set of models is estimated

considering text-labeled utterances (speech corpus and associated transcriptions). Thereby, each model represents a reference pattern of each base unit (word or phoneme, for instance). In order to recognize a given speech signal, the system evaluates the similarity (likelihood score) between the signal and each previously trained model. The most likely word sequence is obtained as an outcome of the process.

During training, models also incorporate acoustic characteristics from the recording, such as reverberation and noise levels. If the system is deployed under similar conditions, one says that training and test are matched and high recognition rate may be expected. Unfortunately, these conditions can differ from the training phase to the effective use, leading to an important acoustic mismatch, and impairing the ASR performance. Considering a real case, if models are trained with clean speech, recorded in a studio, and the system is used for dictation at a noisy office room, the recognition rate may be degraded. Therefore, to improve the system robustness, the mismatch problem between training and test must be tackled. Over the last few decades, a considerable research effort has been directed for reducing the mismatch caused by reverberation and additive noise.

There are several ways to classify existing approaches to the reverberation and noise problems in ASR systems. In this chapter, we choose to group methods based on their location in the speech recognition chain. ASR processing can be roughly separated into two parts: front-end and back-end. Speech parameters are extracted at the front-end module, whereas the likelihood between the input signal and acoustic models is computed at the back-end (or decoder). Considering this classification, the mismatch caused by reverberation and noise can be reduced either before the front-end, or during the front-end processing or even at the back-end module, as shown in Fig. 3. Therefore, methods are grouped into the following classes: speech enhancement, robust parameterization, and model adaptation. Each group is discussed in the following.

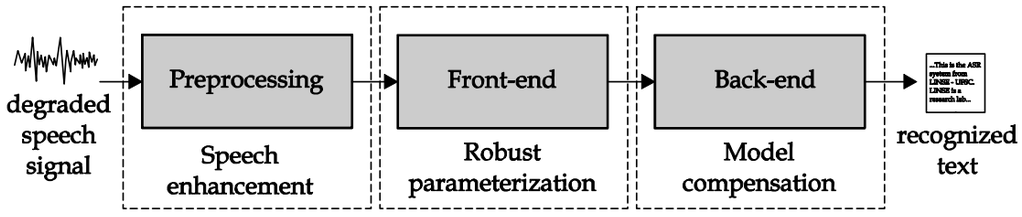


Fig. 3. Classification of methods used in speech recognition applications for reverberation and noise compensation.

3.1 Speech enhancement

Speech enhancement methods attempt to cope with reverberation and noise problems before the signal reaches the front-end. They work as a preprocessing stage in ASR systems.

Methods in this category can be broadly classified by the number of microphones they need to operate, leading to two classes: single and multi-microphone methods (the latter is also termed microphone array).

We briefly describe here some current techniques: beamforming, inverse filtering, kurtosis-based adaptive filtering, harmonicity-based dereverberation, and spectral subtraction.

Beamforming is a classical microphone array approach (Darren et al., 2001). Signals from each microphone are accurately delayed and combined (by a simple sum or a filtering algorithm). As a consequence, the involved algorithm directs the array to the speech source,

reinforcing the speech signal and reducing reverberation and noise from other directions. Although an increase in recognition rate is achieved for noisy speech, the same good effect is not attained for reverberant speech, because conventional microphone array algorithms assume that the target and the undesired signals are uncorrelated (not true for reverberation). In a recent approach, called likelihood maximizing beamforming (LIMABEAM) (Seltzer et al., 2004), the beamforming algorithm is driven by the speech recognition engine. This approach has demonstrated a potential advantage over standard beamforming techniques for ASR applications.

Methods based on inverse filtering have two processing stages: estimation of the impulse responses between the source and each microphone and application of a deconvolution operation. Among other approaches, estimation can be carried out by cepstral techniques (Bees et al., 1991) or a grid of zeros (Pacheco & Seara, 2005); however, some practical difficulties have been noted in real applications, impairing the correct working of these techniques. Regarding the inversion of the RIR, an efficient approach is proposed by Radlović & Kennedy (2000), which overcomes the drawbacks due to nonminimum phase characteristics present in real-world responses.

Another interesting approach is presented by Gillespie et al. (2001), in which characteristics of the speech signal are used for improving the dereverberation process. There, the authors have demonstrated that the residue from a linear prediction analysis of clean speech exhibits peaks at each glottal pulse while those ones are dispersed in reverberant speech. An adaptive filter can be used for minimizing this dispersion (measured by kurtosis), reducing the reverberation effect. The same algorithm is also used as a first stage of processing by Wu & Wang (2006), showing satisfactory results for reducing reverberation effects when T_{60} is between 0.2 and 0.4 s.

The harmonic structure of the speech signal can be used in harmonicity based dereveberation (HERB) (Nakatani et al., 2007). In this approach, it is assumed that the original signal is preserved at multiples of the fundamental frequency, and so an estimate of room response can be obtained. The main drawback of this technique is the amount of data needed for achieving a good estimate.

Spectral subtraction is another speech enhancement technique, which will be discussed in details in Section 4.

3.2 Robust acoustic features

In this class of techniques, the central idea is to represent the signal with parameters less sensitive to changes in acoustic conditions as reverberation and noise.

A very simple and widespread approach is called cepstral mean normalization (CMN). In this technique, a mean of the parameter vectors is initially obtained. The resulting mean vector is subtracted from each parameter vector. Therefore, the normalized parameters present a long-term average equal to zero. It is possible to demonstrate that this approach improves the robustness with respect to the linear filtering effect introduced by microphones and transmission channels over the speech signal (Droppo & Acero, 2008). It has also been verified experimentally that the CMN reduces the additive noise effect, even though it is ineffective for dereverberation (Droppo & Acero, 2008).

Regarding the reverberation problem, some authors suggest that it cannot be identified within short frames of analysis (in the order of 25 ms), since RIRs usually exhibit large lengths. Two approaches attempt to overcome this problem: relative spectra (RASTA) (Hermansky & Morgan, 1994) and modulation spectrogram (MSG) (Kingsbury, 1998). They

consider slow variations in spectrum, which is verified when a frame size in the order of 200 ms is used. ASR assessments have shown that such approaches improve the recognition accuracy for moderately reverberant conditions (Kingsbury, 1998).
An alternative parameterization technique, named missing feature approach (Palomäki et al., 2004; Raj & Stern, 2005), suggests representing the input signal in a time-frequency grid. Unreliable or missing cells (due to degradation) are identified and discarded or even replaced by an estimate of the clean signal. In the case of reverberation, reliable cells are those in which the direct signal and initial reflections are stronger. Training is carried out with clean speech and there is no need to keep retraining acoustic models for each kind of degradation. So, the identification of unreliable cells is performed only during recognition. A considerable improvement in the recognition rate may be attained; however, to obtain such identification of cells is a very hard task in practice.

3.3 Model adaptation

The main objective of model adaptation approaches is to minimize the mismatch between training and test phases by applying some kind of compensation over the reference model.
The first approach is to include reverberation and noise during training, i.e., contaminating the training material with the same kind of degradation expected for deployment. Reverberation and noise can be recorded during the acquisition of speech corpora or even to be artificially included. International projects have recorded training material in different conditions, such as inside cars in the SpeechDat-Car project (Moreno et al., 2000) or in different environment in the SpeeCon project (Iskra et al., 2002).
Artificial inclusion of reverberation allows generating models with different levels of reverberation (Couvreur & Couvreur, 2004), permitting thus to select the best model match during deployment.
As an alternative to retrain models for each noise condition, the parallel model combination (PMC) technique can be applied. This approach attempts to estimate a noisy speech model from two other models: a previously trained one, based on clean speech, and a noise model, obtained by an on-line estimate from noise segments (Gales & Young, 1995). Promising adaptation results can be achieved by using a small amount of data, whereas the main drawback of the PMC approach is a large computational burden.
A better adjustment can also be accomplished with a set of adaptation data in a maximum *a posteriori* estimation approach (Omologo et al., 1998). A significant increase in recognition rate is achieved, even though a single microphone is used for signal acquisition; however, the robustness to changes in the environmental conditions is still a challenging issue (Omologo et al., 1998).

4. Spectral subtraction

Spectral subtraction is a well-known speech enhancement technique, which is part of the class of short-time spectral amplitude (STSA) methods (Kondoz, 2004). What makes spectral subtraction attractive is its simplicity and low computational complexity, being advantageous for platforms with limited resources (Droppo & Acero, 2008).

4.1 Algorithm

Before introducing spectral subtraction as a dereverberation approach, we shall review its original formulation as a noise reduction technique. Disregarding the effect of reverberation, a noisy signal in (4) can be expressed in frequency domain as

$$Y(k) = X(k) + V(k) \tag{5}$$

where $Y(k)$, $X(k)$, and $V(k)$ denote the short-time discrete Fourier transform (DFT) of $y(n)$, $x(n)$ and $v(n)$, respectively. The central idea of spectral subtraction is to recover $x(n)$ modifying only the magnitude of $Y(k)$. The process can be described as a spectral filtering operation

$$\left|\hat{X}(k)\right|^{\nu} = G(k)\left|Y(k)\right|^{\nu} \tag{6}$$

where ν denotes the spectral order, $\hat{X}(k)$ is the DFT of the enhanced signal $\hat{x}(n)$, and $G(k)$ is a gain function.

Fig. 3 shows a block diagram of a general procedure of spectral subtraction. The noisy signal $y(n)$ is windowed and its DFT is computed. The gain function is then estimated by using the current noisy magnitude samples, the previous enhanced magnitude signal and the noise statistics. Note that the phase of $Y(k)$ [represented by $\angle Y(k)$] remains unchanged, being an input to the inverse DFT (IDFT) block. The enhanced signal is obtained associating the enhanced magnitude and the phase of $Y(k)$, processing them by the IDFT block along with an overlap-and-add operation; the latter to compensate for the windowing.

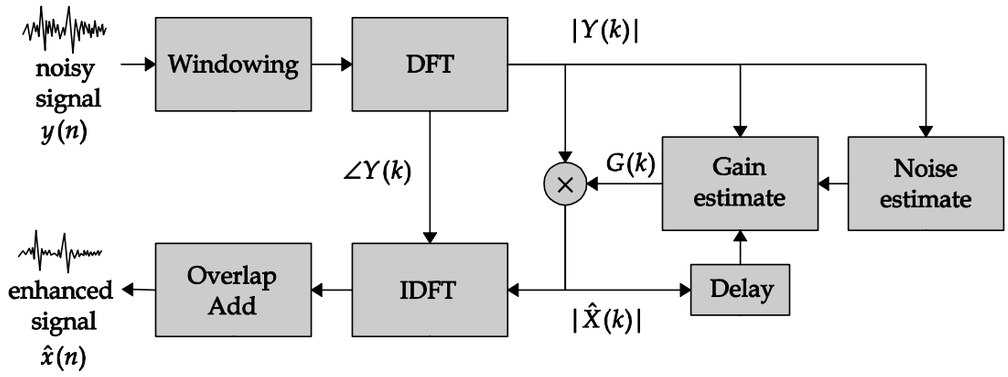


Fig. 3. Block diagram of a general procedure of spectral subtraction.

The blocks of gain and noise estimates are the most critical part in the process and the success of this technique is strongly dependent on determining adequate gains. In the following, we shall discuss this approach considering a power spectral subtraction example. Processing signals in the power spectral domain, i.e., $\nu = 2$, and assuming that signal and noise are uncorrelated, we have

$$\left|\hat{X}(k)\right|^2 = \left|Y(k)\right|^2 - \left|\hat{V}(k)\right|^2 \tag{7}$$

or even

$$\left|\hat{X}(k)\right|^2 = G(k)\left|Y(k)\right|^2 \tag{8}$$

for which the most simple estimate of the gain $G(k)$ is given by

$$G(k)=\begin{cases}1-\dfrac{1}{\mathrm{SNR}(k)}, & \mathrm{SNR}(k)>1\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

with

$$\mathrm{SNR}(k)=\frac{|Y(k)|^2}{|\hat{V}(k)|^2} \tag{10}$$

where $\mathrm{SNR}(k)$ is the *a posteriori* signal-to-noise ratio and $\hat{V}(k)$ is the noise estimate. Although necessary to prevent $|\hat{X}(k)|$ from being negative, the clamping introduced by the conditions in (9) causes some drawbacks. Note that gains are estimated for every frame and at each frequency index independently. Observing the distribution of these gains in a time-frequency grid, one notes that neighbor cells may display varying levels of attenuation. This irregularity over the gain gives rise to tones at random frequencies that appear and disappear rapidly (Droppo & Acero, 2008), leading to an annoying effect called musical noise.
More elaborate estimates for $G(k)$ are proposed in the literature, aiming to reduce musical noise. An improved approach to estimate the required gain is introduced by Berouti et al. (1979), which is given by

$$G(k)=\max\left\{\left[1-\alpha\left(\frac{1}{\mathrm{SNR}(k)}\right)^{\frac{\nu}{2}}\right]^{\frac{1}{\nu}},\ \beta\right\} \tag{11}$$

where α and β are, respectively, the oversubtraction and spectral floor factors. The oversubtraction factor controls the reduction of residual noise. Lower levels of noise are attained with higher α; however, if α is too large, the speech signal will be distorted (Kondoz, 2004). The spectral floor factor works to reduce the musical noise, smearing it over a wider frequency band (Kondoz, 2004). A trade-off in β choice is also required. If β is too large, other undesired artifacts become more evident.
It is important to point out that speech distortion and residual noise cannot be reduced simultaneously. Moreover, parameter adjustment is dependent on the application. It has been determined experimentally that a good trade-off between noise reduction and speech quality is achieved with power spectral subtraction $(\nu=2)$ by using α between 4 and 8, and $\beta=0.1$ (Kondoz, 2004). This set-up is considered adequate for human listeners, since, as a general rule, human beings can tolerate some distortion, but are sensitive to fatigue caused by noise. We shall show in Section 4.4 that ASR systems usually are more susceptible to speech distortion, and so $\alpha<1$ could be a better choice for reducing the recognition error rate.

4.2 Application of spectral subtraction for dereverberation

An adaptation of spectral subtraction has been recently proposed to enhance speech degraded by reverberation (Lebart & Boucher, 1998; Habets, 2004). It will be discussed in details later on.

In order to tackle room reverberation by using spectral subtraction, some fundamental relations must be established. Firstly, the autocorrelation $r_y(\ell)$ of the reverberant signal is defined. Therefore, disregarding the additive noise effect, we get

$$r_y(\ell) \equiv r_y(n, n+\ell) = \mathrm{E}[y(n)y(n+\ell)] = \mathrm{E}[\sum_{k=-\infty}^{n} x(k)h(n-k) \sum_{m=-\infty}^{n+\ell} x(m)h(n+\ell-m)]. \tag{12}$$

Given the nature of the speech signal and of the RIR, one can consider $x(n)$ and $h(n)$ as independent statistical processes. Thus,

$$r_y(\ell) = \sum_{k=-\infty}^{n} \sum_{m=-\infty}^{n+\ell} \mathrm{E}[x(k)x(m)]\mathrm{E}[h(n-k)h(n+\ell-m)]. \tag{13}$$

Considering a RIR modeled by modulating a zero-mean random sequence with a decaying exponential (Lebart & Boucher, 1998), one can write

$$h(n) = w(n)\,\mathrm{e}^{-\tau n}u(n) \tag{14}$$

where $w(n)$ represents a white zero-mean Gaussian noise with variance σ_w^2, $u(n)$ denotes the unit step function, and τ is a damping constant related to the reverberation time, which is expressed as (Lebart & Boucher, 1998)

$$\tau = \frac{3\ln 10}{T_{60}}. \tag{15}$$

Thus, the second r.h.s. term in (13) is written as

$$\mathrm{E}[h(n-k)h(n+\ell-m)] = \mathrm{e}^{-2\tau n}\,\sigma_w^2\,\mathrm{e}^{\tau(k+m-\ell)}\,\delta(k-m+\ell) \tag{16}$$

where $\delta(n)$ represents the unit sample sequence.
Then, substituting (16) into (13), we obtain

$$r_y(\ell) = \mathrm{e}^{-2\tau n} \sum_{k=-\infty}^{n} \mathrm{E}[x(k)x(k+\ell)]\sigma_w^2\,\mathrm{e}^{2\tau k}. \tag{17}$$

Now, considering the threshold N_d, defined in (3), one can split the summation in (17) into two parts. Thereby,

$$r_y(\ell) = \mathrm{e}^{-2\tau n} \sum_{k=-\infty}^{n-N_\mathrm{d}} \mathrm{E}[x(k)x(k+\ell)]\sigma_w^2\,\mathrm{e}^{2\tau k} + \mathrm{e}^{-2\tau n} \sum_{k=n-N_\mathrm{d}+1}^{n} \mathrm{E}[x(k)x(k+\ell)]\sigma_w^2\,\mathrm{e}^{2\tau k}. \tag{18}$$

In addition, the autocorrelation of the $y(n)$ signal, computed between the samples $n-N_\mathrm{d}$ and $n-N_\mathrm{d}+\ell$, can be written as

$$r_y(n-N_\mathrm{d}, n-N_\mathrm{d}+\ell) = \mathrm{e}^{-2\tau(n-N_\mathrm{d})} \sum_{k=-\infty}^{n-N_\mathrm{d}} \mathrm{E}[x(k)x(k+\ell)]\sigma_w^2\,\mathrm{e}^{2\tau k}. \tag{19}$$

Then, from (18), the autocorrelation between the samples n and $n+\ell$ is given by

$$r_y(n,n+\ell) = r_{y_\mathrm{r}}(n,n+\ell) + r_{y_\mathrm{d}}(n,n+\ell) \tag{20}$$

with

$$r_{y_\mathrm{r}}(n,n+\ell) = \mathrm{e}^{-2\tau N_\mathrm{d}}\, r_y(n-N_\mathrm{d}, n-N_\mathrm{d}+\ell) \tag{21}$$

and

$$r_{y_\mathrm{d}}(n,n+\ell) = \mathrm{e}^{-2\tau n} \sum_{k=n-N_\mathrm{d}+1}^{n} \mathrm{E}[x(k)x(k+\ell)]\sigma_w^2\, \mathrm{e}^{2\tau k} \tag{22}$$

where $r_{y_\mathrm{r}}(n,n+\ell)$ and $r_{y_\mathrm{d}}(n,n+\ell)$ are the autocorrelation functions associated with the signals $y_\mathrm{r}(n)$ and $y_\mathrm{d}(n)$, respectively. Signal $y_\mathrm{r}(n)$ is related to the late reverberation, as a result of the convolution of $h_\mathrm{r}(n)$ and $x(n)$. Variable $y_\mathrm{d}(n)$ is associated with the direct signal and initial reflections, being obtained through the convolution of $h_\mathrm{d}(n)$ and $x(n)$.
Now, from (20), the short-time power spectral density (PSD) of the degraded signal $S_y(n,k)$ is expressed as

$$S_y(n,k) = S_{y_\mathrm{r}}(n,k) + S_{y_\mathrm{d}}(n,k) \tag{23}$$

where $S_{y_\mathrm{r}}(n,k)$ and $S_{y_\mathrm{d}}(n,k)$ are the PSDs corresponding to the signals $y_\mathrm{r}(n)$ and $y_\mathrm{d}(n)$, respectively. From (21), the estimated value $S_{y_\mathrm{r}}(n,k)$ is obtained by weighting and delaying the PSD of the degraded speech signal. Thus,

$$S_{y_\mathrm{r}}(n,k) = \mathrm{e}^{-2\tau N_\mathrm{d}}\, S_y(n-N_\mathrm{d},k). \tag{24}$$

Then, assuming that $y_\mathrm{d}(n)$ and $y_\mathrm{r}(n)$ are uncorrelated, the late reverberant signal can be treated as an additive noise, and the direct signal can be recovered through spectral subtraction.

4.3 Estimation of reverberation time

In order to implement the previously presented procedure, one initially must obtain the parameter τ, since it is used for estimating the power spectral density of the late reverberant signal (24). Given that τ is related to the reverberation time, one estimates T_{60} from the captured signal.
Some approaches have been proposed recently for blind estimation of the reverberation time. In this case, blind means that only the captured signal is available. Maximum-likelihood (ML) approaches are proposed for T_{60} estimation by Ratnam et al. (2003) and Couvreur & Couvreur (2004). The main difficulty to estimate T_{60} is the requirement of silence regions between spoken words. Particularly in short utterances, this condition may not be fulfilled, leading to a considerable error in the T_{60} estimate.

In this chapter, instead of evaluating a specific algorithm we opt to assess the sensitivity of an ASR system to errors in the estimate of T_{60}. Experimental results showing the performance of spectral subtraction algorithm under such errors are presented in the next section.

4.4 Performance assessment in ASR systems

We have used the spectral subtraction approach as a preprocessing stage in an ASR system. The chosen task here consists of recognizing digit strings representing phone numbers in the Brazilian Portuguese language. Speech data recorded through telephone and sampled at 8 kHz are used as the original signal. In this experiment, we use 250 recordings, taken with several speakers. The reverberant speech signal is generated through a linear convolution between the original speech data and a given RIR. We have considered three different impulse responses, which are obtained by using the well-known image method to model acoustic room responses (Allen & Berkley, 1979). Room configurations used in the simulation experiments are given in Table 1.

In the spectral subtraction stage, the degraded signal is segmented into 25 ms frames, with an overlapping of 15 ms, and weighted by a Hamming window. The threshold T is fixed in 40 ms. We have considered magnitude subtraction ($\nu = 1$), since previous research works have obtained very good results with this configuration (Habets, 2004). From the modified magnitude spectrum and (original) phase signal, the enhanced signal is recovered by an overlap-and-add algorithm.

Parameter		Room #1	Room #2	Room #3
Dimensions (m)		7×7×3.5	6×8×3	9×8×3
Speaker position		(2.5, 3.8, 1.3)	(2.0, 3.0, 1.5)	(4.5, 5.0, 1.0)
Microphone position		(3.3, 3.0, 0.7)	(3.0, 3.5, 0.6)	(5.5, 6.5, 0.5)
Reflection coefficients	Walls	0.9	0.9	0.9
	Floor and ceiling	0.6	0.6	0.9
Resulting T_{60}(s)		0.68	0.73	0.83

Table 1. Parameters used for obtaining the room impulse responses.

Assessments have been carried out by using a speaker-independent HMM-based speech recognition system. Experiments are performed with word-based models, one for each of 11 digits (0 to 9 in Brazilian Portuguese plus the word "meia"[1]).

Acoustic features are extracted by a mel-cepstrum front-end developed for distributed speech recognition (DSR) (ETSI, 2002). This front-end includes a preprocessing stage of noise reduction using a Wiener filter (ETSI, 2002). Feature extraction is also carried out at each 25 ms frame, with an overlapping of 15 ms.

From each segment, 12 mel-frequency cepstral coefficients (MFCC) and the energy are computed, along with the first- and second-order derivatives. Thus, the final parameter vector is composed of 39 elements.

Recognition is performed by a Viterbi decoder with beam searching and word-end pruning (Young et al., 2002).

[1] In Brazilian Portuguese, it is common to speak "meia" for representing the number six. It is short for "meia dúzia" (half a dozen).

The results of the speech recognition task are presented in terms of the sentence error rate (SER), defined as

$$\mathrm{SER}(\%) = \frac{N_\mathrm{e}}{N_\mathrm{s}} 100 \tag{25}$$

where N_e is the number of sentences incorrectly recognized, and N_s is the total number of sentences in the test (250 in this evaluation). We have decided to use SER since for digit string recognition (phone numbers, in our case) an error in a single digit renders ineffective the result for the whole string. Note that SER is always greater than or equal to the word error rate (WER).

For the original speech data, SER is equal to 4%. For the reverberant data, obtained by the convolution of the original speech with the RIRs, SER increases to 64.4%, 77.6%, and 93.6% for Room #1, Room #2 and Room #3, respectively. This result reinforces the importance of coping with reverberation effects in ASR systems.

In order to evaluate spectral subtraction applied to reducing reverberation in ASR systems, we present the following simulation experiments:

i) Selection of oversubtraction factor α and spectral floor factor β. Here, we verify the best combination of parameters considering a speech recognition application.

ii) Sensitivity to errors in the estimate of T_{60}. Since an exact estimation of reverberation time could be difficult, we assess here the sensitivity of ASR to such errors.

iii) Effect of RIR variation. We evaluate the effect of speaker movement, which implies changes in the RIR.

iv) Effect of both reverberation and noise over ASR performance. In real enclosures, reverberation is usually associated with additive noise. We also assess this effect here.

4.4.1 Selection of oversubtraction factor and spectral floor

The first parameter we have evaluated is the oversubtraction factor α. Previous research works (Lebart & Boucher, 1998; Habets, 2004) assume α equal to 1. In contrast to them, we use the general formulation given by (11). We have evaluated α for different values of β and here we show the best results obtained using β = 0.2 and the particular value of T_{60} for each room (see Table 1). Fig. 4 shows the SER as a function of the oversubtraction factor between 0.4 and 1.3.

For Room #1 and Room #2, the best result is obtained with α = 0.7, which corresponds to an undersubtraction level. For Room #3, the best result is also obtained for α <1.

These particular results for reverberation reduction are in accordance with those obtained in studies about noise reduction discussed by Virag (1999) and Chen et al. (2006). Virag (1999) has verified that the oversubtraction parameter should be lower in ASR systems than for human listeners. Chen et al. (2006) have used a Wiener filter for denoising considering the factor α less than unity, leading to a satisfactory reduction in the distortion level over the resulting signal.

The influence of the spectral floor factor β, parameter that controls the masking level of musical noise, is shown in Fig. 5. For the three assessed room responses, the best result is obtained for β = 0.2, i.e., suggesting that it is important to maintain a certain level of masking noise. Note also that by not using any spectral flooring (β = 0) the SER increases. These results point out that ASR systems tolerate better residual noise than the inherent distortion provoked by the spectral subtraction processing, provided the noise level is not too high.

4.4.2 Sensitivity to errors in the estimation of the reverberation time

As discussed in Section 4.2, reverberation time must be estimated by the spectral subtraction algorithm. Since this estimate is subject to errors, it is important to evaluate the effect of such errors over ASR performance. The sensitivity to errors in the estimation of T_{60} has been assessed at the operating point $\alpha = 0.7$ and $\beta = 0.2$. We use the same set of RIRs as in Table 1. In the spectral subtraction algorithm, errors over T_{60} are introduced by varying the

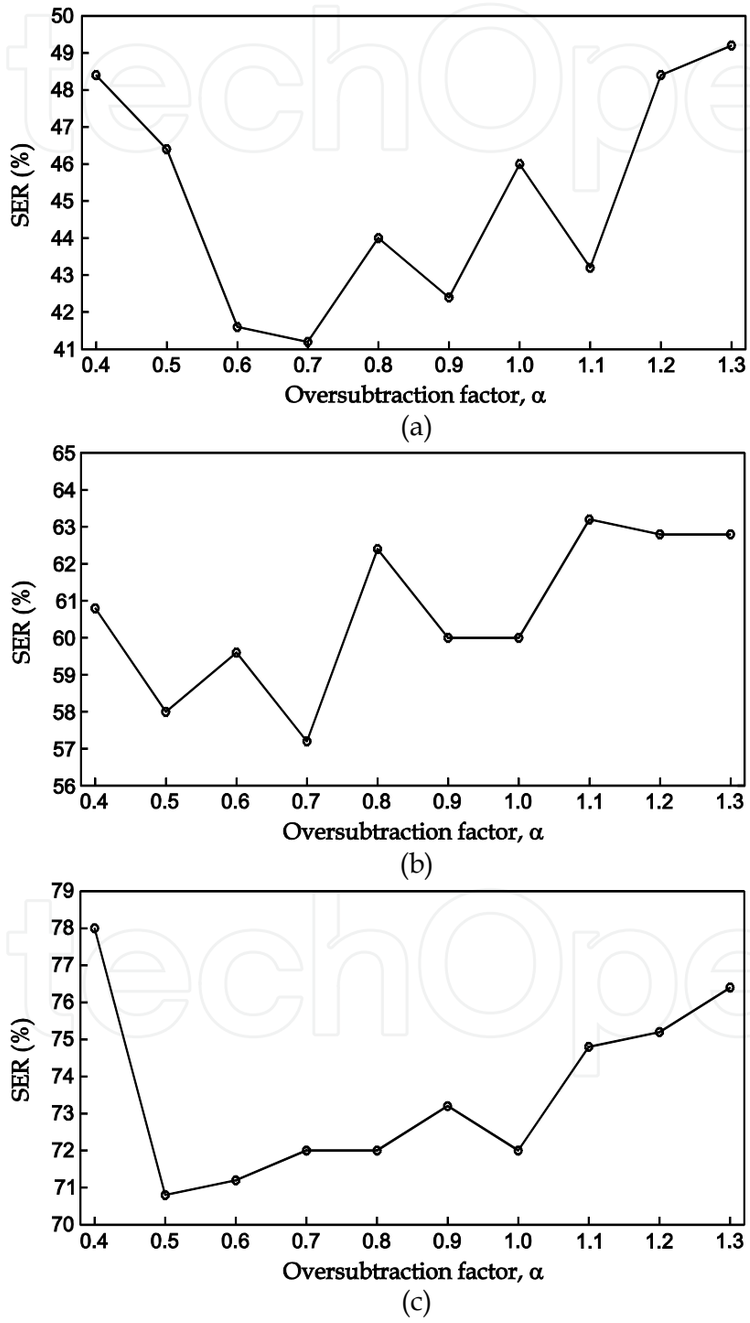


Fig. 4. Variation in SER as a function of α for $\beta = 0.2$ and the corresponding T_{60}.
(a) Room #1. (b) Room #2. (c) Room #3.

value from 0.3 to 1.3 s using steps of 0.2 s. Fig. 6 presents the SER in terms of such a variation. Ideally, the method should be less sensitive to errors in the estimation of T_{60}, since a blind estimate is very cost demanding in practice. Achieved results point out that even for an inaccurate estimate of T_{60}, the performance degradation is still tolerable.

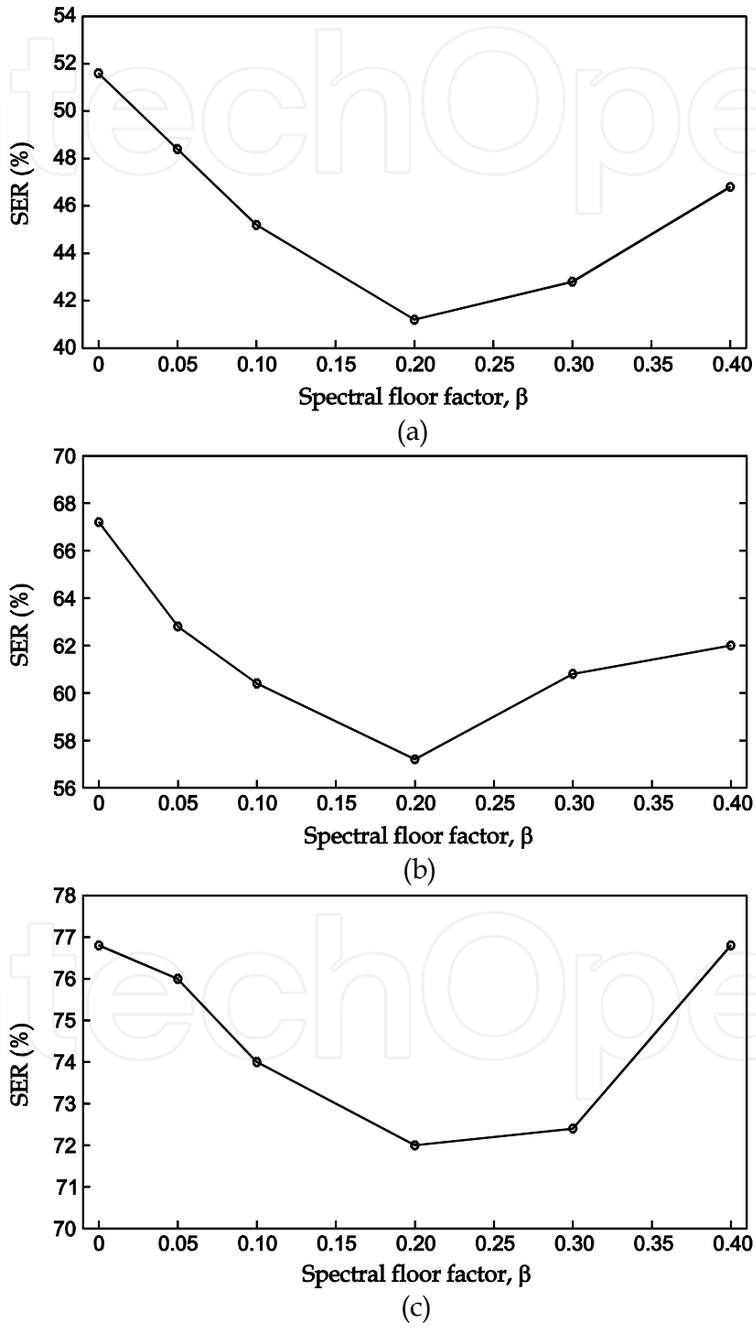
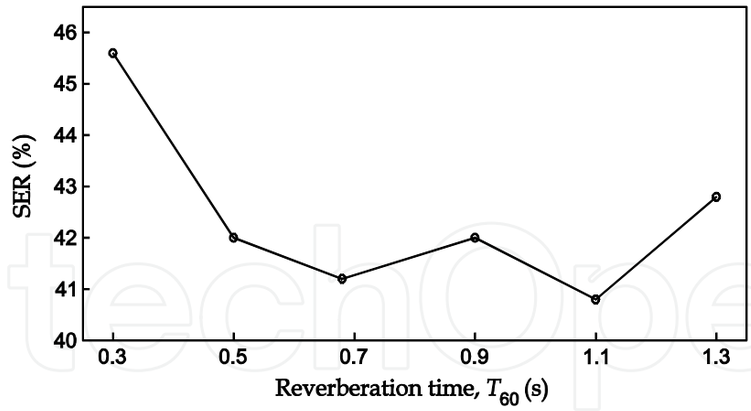
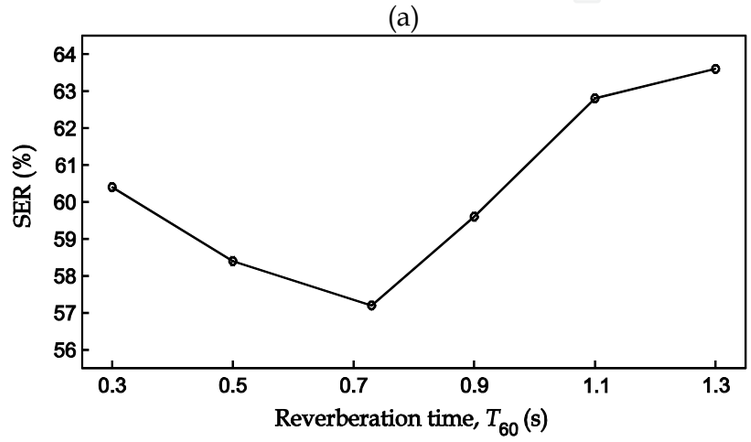


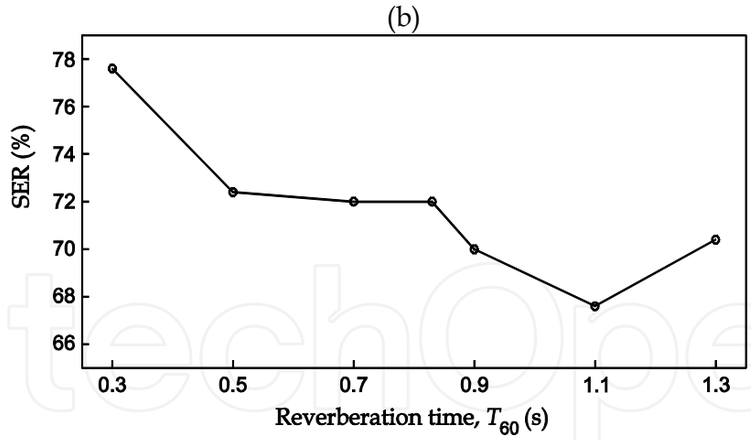
Fig. 5. Variation in SER as a function of β, keeping $\alpha = 0.7$ and the corresponding T_{60}. (a) Room #1. (b) Room #2. (c) Room #3.



(a)



(b)



(c)

Fig. 6. Variation of SER as a function of T_{60} using $\alpha = 0.7$ and $\beta = 0.2$. (a) Room #1. (b) Room #2. (c) Room #3.

4.4.3 Effect of room impulse response variation caused by a moving speaker

The variation of the RIR in a particular enclosure is analyzed considering Room #1 (Table 1) and a moving speaker. The reference position of the speaker shown in Table 1 is shifted by 0.5 m (with a 0.25 m step) in both dimensions (length and width). Fig. 8 shows the ground plan of the enclosure, marking the positions of microphone and (moving) speaker. By using

this configuration, eight different RIRs are obtained. A set of reverberated audio signals is determined convolving each room response with the input signals from the test set. The spectral subtraction algorithm is configured with $\alpha = 0.7$, $\beta = 0.2$, and $T_{60} = 0.68$ s.

Results are presented in Table 2. Regarding the column "without processing", we observe that even small changes in the speaker position affect the ASR performance.

Test condition		SER (%)	
		Without processing	Spectral subtraction
Reference response		64.4	41.2
Speaker position shifted in the x axis by	−0.50 m	64.4	44.8
	−0.25 m	60.8	45.6
	+0.25 m	47.2	26.8
	+0.50 m	40.0	29.6
Speaker position shifted in the y axis by	−0.50 m	46.8	29.6
	−0.25 m	52.0	33.6
	+0.25 m	65.6	48.8
	+0.50 m	69.2	51.2

Table 2. SER as a function of room impulse response changes.

Although some performance reduction was expected, the effect of changes in the speaker position over the recognition rate is still considerable. In general, we verify that the larger the distance between speaker and microphone, the larger the error rate. These results confirm the need for making use of robust dereverberation techniques to cope with impulse response changes.

Spectral subtraction improves the recognition rate for all considered conditions. Error rates are reduced between 10 and 20 percentage points with respect to the standard front-end. Ideally, error rates should be less than or equal to the reference error rate (see Table 2). Although this is not verified, no instability is observed in the technique discussed here in contrast to some approaches presented in the open literature (Bees et al., 1991).

4.4.4 Combined effect of reverberation and noise

The combined effect of reverberation and additive noise is evaluated considering the addition of noise to the reverberant audio signals of Room #1 (Table 1). Samples of noise are obtained from the set available in Hansen & Arslan, (1995). We have considered two types of noise: the first one is named large city noise (LCI) and the other is white Gaussian noise (WGN), with three signal-to-noise ratio (SNR) levels: 5, 10, and 15 dB.

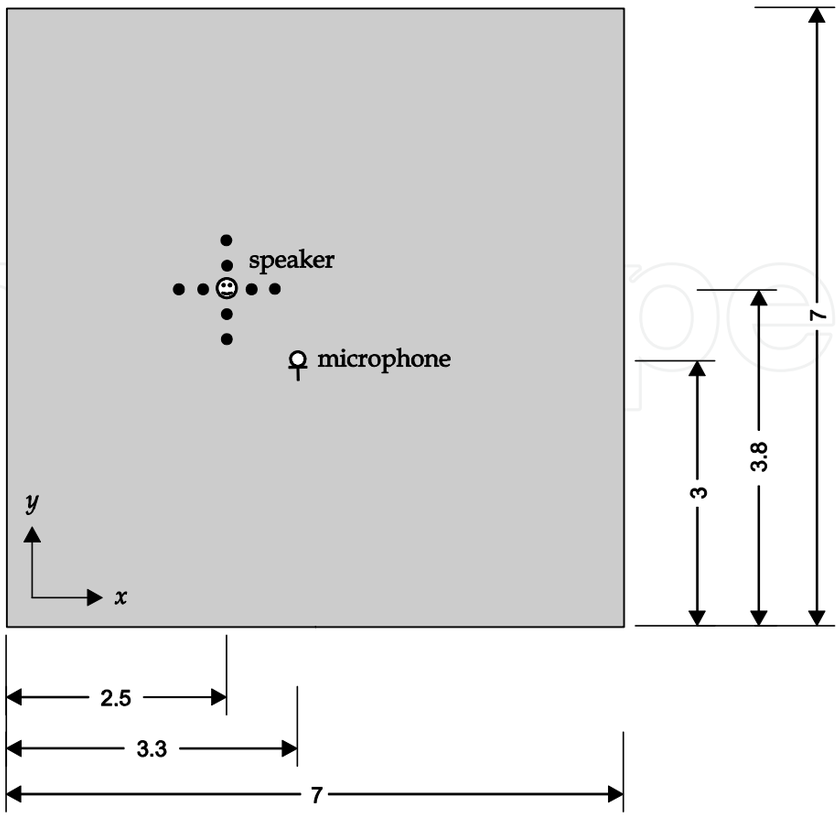


Fig. 7. Ground plan of the room showing speaker and microphone positions (dimensions in m). Speaker position is shifted with a 0.25 m step.

Test condition	SER (%)	
	Without processing	Spectral subtraction
Only reverberation	64.4	41.2
Reverberation + large city noise at SNR 15 dB	75.6	59.6
Reverberation + large city noise at SNR 10 dB	85.6	75.2
Reverberation + large city noise at SNR 5 dB	97.6	92.4
Reverberation + white Gaussian noise at SNR 15 dB	84.0	66.8
Reverberation + white Gaussian noise at SNR 10 dB	91.6	80.0
Reverberation + white Gaussian noise at SNR 5 dB	98.4	95.6

Table 3. Combined effects of reverberation and noise.

Table 3 shows the SER values. Column "without processing" presents the deleterious effect of reverberation and noise over the speech recognition performance. Error rate increases significantly as SNR decreases.
With spectral subtraction, the error is reduced for all considered situations, although it is still high for the worst noise settings. Apart from that, we do not observe any kind of instability, seen in some other approaches.

5. Concluding remarks

This chapter has characterized effects of reverberation and noise over ASR system performance. We have shown the importance of coping with such degradations in order to improve ASR performance in real applications. A brief overview of current dereverberation and denoising approaches has been addressed, classifying methods according to the point of operation in the speech recognition chain. The use of spectral subtraction applied to dereverberation and denoising in ASR systems has been discussed, giving rise to a consistent formulation to treat this impacting problem. We assessed the used approach considering the sentence error rate over a digit string recognition task, showing that the recognition rate can be significantly improved by using spectral subtraction. The impact on the choice of algorithm parameters has been assessed under different environmental conditions for performance. Finally, it is important to mention that reverberation and noise problems in ASR systems continue to be a challenging subject for the signal processing community.

6. References

ETSI (2002). Speech processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms, European Telecommunications Standards Institute (ETSI) Std. ES 202 050 V.1.1.1, Oct. 2002.

Allen, J. B. & Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, Vol. 65, No. 4, Apr. 1979, pp. 943-950.

Bees, D.; Blostein, M. & Kabal, P. (1991). Reverberant speech enhancement using cepstral processing. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'91)*, Vol. 2, pp. 977–980, Toronto, Canada, Apr. 1991.

Berouti, M.; Schwartz, R. & Makhoul, J. Enhancement of speech corrupted by acoustic noise. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'79)*, Vol. 4, pp. 208-211, Washington, USA, Apr. 1979.

Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 27, No. 2, Apr. 1979, pp. 113-120.

Chen, J.; Benesty, J.; Huang, Y. & Doclo, S. (2006). New insights into the noise reduction Wiener filter. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 4, July 2006, pp. 1218–1234.

Chesnokov, A. & SooHoo, L. (1998). Influence of early to late energy ratios on subjective estimates of small room acoustics. *Proceedings of the 105th AES Convention*, pp. 1–18, San Francisco, USA, Sept. 1998.

Couvreur, L. & Couvreur, C. (2004). Blind model selection for automatic speech recognition in reverberant environments. *Journal of VLSI Signal Processing*, Vol. 36, No. 2-3, Feb./Mar. 2004, pp. 189-203.

de la Torre, A.; Segura, J. C.; Benitez, C.; Ramirez, J.; Garcia, L. & Rubio, A. J. (2007). Speech recognition under noise conditions: Compensation methods. In: *Speech Recognition and Understanding,* Grimm, M. & Kroschel, K. (Eds.), pp. 439-460, I-Tech, ISBN 978-3-902-61308-0. Vienna, Austria.

Droppo, J. & Acero, A. (2008). Environmental robustness. In: *Springer Handbook of Speech Processing,* Benesty, J.; Sondhi, M. M. & Huang, Y. (Eds.), pp. 653-679, Springer, ISBN 978-3-540-49125-5, Berlin, Germany.

Everest, F. A. (2001). *The Master Handbook of Acoustics.* 4 ed., McGraw-Hill, ISBN 978-0-071-36097-5, New York, USA.

Gales, M. J. F. & Young, S. J. (1995). A fast and flexible implementation of parallel model combination. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95)*, Vol. 1, pp. 133–136. Detroit, USA, May 1995.

Gillespie, B. W.; Malvar, H. S. & Florêncio, D. A. F. (2001). Speech dereverberation via maximum-kurtosis subband adaptive filtering. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, Vol. 6, pp. 3701–3704. Salt Lake City, USA, May 2001.

Habets, E. A. P. (2004). Single-channel speech dereverberation based on spectral subtraction. *Proceedings of the 15th Annual Workshop on Circuits, Systems and Signal Processing (ProRISC'04)*, pp. 250–254, Veldhoven, Netherlands, Nov. 2004.

Hansen, J. H. L. & Arslan, L. (1995). Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit-card corpus. *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 3, May 1995, pp. 169–184.

Hermansky, H. & Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 4, Oct. 1994, pp. 578–589.

Huang, Y.; Benesty, J. & Chen, J. (2008). Dereverberation. In: *Springer Handbook of Speech Processing,* Benesty, J.; Sondhi, M. M. & Huang, Y. (Eds.), pp. 929-943, Springer, ISBN 978-3-540-49125-5, Berlin, Germany.

Iskra, D.; Grosskopf, B.; Marasek, K.; van den Heuvel, H.; Diehl, F. & Kiessling, A. (2002). SPEECON - Speech databases for consumer devices: Database specification and validation. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'2002)*, pp. 329-333, Las Palmas, Spain, May 2002.

Kingsbury, B. E. D. (1998). *Perceptually Inspired Signal-processing Strategies for Robust Speech Recognition in Reverberant Environments*. PhD Thesis, University of California, Berkeley.

Kondoz, A. M. (2004). *Digital Speech: Coding for Low Bit Rate Communication Systems*. 2 ed, Wiley, ISBN 978-0-470-87008-2, Chichester, UK.

Lebart, K. & Boucher, J. M. (1998). A new method based on spectral subtraction for the suppression of late reverberation from speech signals. *Proceedings of the 105th AES Convention*, pp. 1–13, San Francisco, USA, Sept. 1998.

Moreno, A.; Lindberg, B.; Draxler, C.; Richard, G.; Choukri, K.; Euler, S. & Allen, J. (2000). SPEECHDAT-CAR. A large speech database for automotive environments. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000)*, Vol. 2, pp. 895–900, Athens, Greece, May/June 2000.

Nakatani, T.; Kinoshita, K. & Miyoshi, M. (2007). Harmonicity-based blind dereverberation for single-channel speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 1, Jan. 2007, pp. 80–95.

Omologo, M.; Svaizer, P. & Matassoni, M. (1998). Environmental conditions and acoustic transduction in hands-free speech recognition. *Speech Communication*, Vol. 25, No. 1-3, Aug. 1998, pp. 75–95.

Pacheco, F. S. & Seara, R. (2005). A single-microphone approach for speech signal dereverberation. *Proceedings of the European Signal Processing Conference (EUSIPCO'05)*, pp. 1–4, Antalya, Turkey, Sept. 2005.

Palomäki, K. J.; Brown, G. J. & Barker, J. P. (2004). Techniques for handling convolutional distortion with 'missing data' automatic speech recognition. *Speech Communication*, Vol. 43, No. 1-2, June 2004, pp. 123–142.

Rabiner, L. & Juang, B.-H. (2008). Historical perspectives of the field of ASR/NLU. In: *Springer Handbook of Speech Processing,* Benesty, J.; Sondhi, M. M. & Huang, Y. (Eds.), pp. 521-537, Springer, ISBN 978-3-540-49125-5, Berlin, Germany.

Radlović, B. D. & Kennedy, R. A. (2000). Nonminimum-phase equalization and its subjective importance in room acoustics. *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 6, Nov. 2000, pp. 728–737.

Raj, B. & Stern, R. M. (2005). Missing-feature approaches in speech recognition. *IEEE Signal Processing Magazine*, Vol. 22, No. 5, Sept. 2005, pp. 101–116.

Ratnam, R.; Jones, D. L.; Wheeler, B. C.; O'Brien Jr., W. D.; Lansing, C. R. & Feng, A. S. (2003). Blind estimation of reverberation time. *Journal of the Acoustical Society of America*, Vol. 114, No. 5, Nov. 2003, pp. 2877–2892.

Seltzer, M. L.; Raj, B. & Stern, R. M. (2004). Likelihood-maximizing beamforming for robust hands-free speech recognition. *IEEE Transactions on Speech and Audio Processing*, Vol. 12, No. 5, Sept. 2004, pp. 489–498.

Virag, N. (1999). Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 2, Mar. 1999, pp. 126-137.

Ward, D. B.; Kennedy, R. A. & Williamson, R. C. (2001). Constant directivity beamforming. In: *Microphone Arrays: Signal Processing Techniques and Applications,* Brandstein, M. & Ward, D. (Eds.), pp. 3-17, Springer, ISBN 978-3-540-41953-2, Berlin, Germany.

Wu, M. & Wang, D. (2006). A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 3, May 2006, pp. 774–784.

Young, S.; Kershaw, D.; Odell, J.; Ollason, D.; Valtchev, V. & Woodland, P. (2002). *The HTK Book (for HTK Version 3.2)*. Cambridge University, Cambridge, UK.

Speech Recognition

Edited by France Mihelic and Janez Zibert

ISBN 978-953-7619-29-9

Hard cover, 550 pages

Publisher InTech

Published online 01, November, 2008

Published in print edition November, 2008

Chapters in the first part of the book cover all the essential speech processing techniques for building robust, automatic speech recognition systems: the representation for speech signals and the methods for speech-features extraction, acoustic and language modeling, efficient algorithms for searching the hypothesis space, and multimodal approaches to speech recognition. The last part of the book is devoted to other speech processing applications that can use the information from automatic speech recognition for speaker identification and tracking, for prosody modeling in emotion-detection systems and in other speech processing applications that are able to operate in real-world environments, like mobile communication services and smart homes.

InTech Europe

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen