IMPULSE RESPONSE ESTIMATION FOR ROBUST SPEECH RECOGNITION IN A REVERBERANT ENVIRONMENT

Mirco Ravanelli, Alessandro Sosi, Piergiorgio Svaizer, Maurizio Omologo

Fondazione Bruno Kessler-Irst
via Sommarive 18, 38123 Trento, Italy
{mravanelli|alesosi|svaizer|omologo}@fbk.eu

ABSTRACT

This paper refers to a voice-enabled smart-home scenario, for which contaminated speech is produced to train a distant-speech recognition system. The impulse response measurement process is investigated, with a specific focus on its impact on speech recognition performance. Experimental results, related to a phone-loop and to a word-loop task, show that a significant change of performance is obtained when using different techniques for impulse response estimation. In particular, the best performance is obtained when an exponential sine sweep excitation sequence is used, with a proper choice of its length and of the energy with which it is propagated in the environment.

Index Terms— Smart home, Distant-speech recognition, Room impulse response, Speech contamination, Auralization.

1. INTRODUCTION

Speech interaction with distant microphones is an important step towards the development of easy-to-use voice interfaces in the automated home context. Nowadays, the most common home controlling interfaces are still based on touch screens, keyboards, PDAs, tablet PCs, or other similar devices. Major trends are towards a diffused use of handheld devices in home automation. Nevertheless, the evolution of the human-machine interfaces, in parallel with other technologies for the future digital home, is in the direction of achieving a more natural interaction. Speech recognition, thanks to the high degree of maturity achieved over the last years, is being progressively introduced in this application field [1], generally based on the use of a headset, of a telephone like input device, or of so-called open-air microphones although at this moment applied quite rarely and with unsatisfactory performance. One of the biggest challenges for a massive introduction of ASR technologies in home automation systems is the increase of robustness against spontaneous speech and uncontrolled acoustic conditions. In this regard, the ASR input variability related to microphone location is one of the critical issues which often determines a significant loss of performance. Although most of the users could simply try to speak close to the microphone, and in a rather controlled way, the expectation is that in the future users would require to be able to interact at four-five meters from microphones in a crowded room, with music playing, and other possible active sound sources.

The EC DIRHA project (*http://dirha.fbk.eu*) has the purpose of developing basic technologies that enable distant speech interaction in a home environment based on a distributed microphone network. Multiple microphone devices will be installed in different rooms in order to monitor selectively acoustic and speech activities observable inside any space of the household.

To tackle ASR robustness to environmental noise and reverberation, one of the most effective approaches is based on training the speech recognizer with contaminated speech that is characterized by the acoustic properties of the given enclosure. The main target of this work is to investigate on the possible influence on recognition performance of different parameter settings and choices in the acoustic impulse response estimation with indirect methods. In the remainder of the paper, we will show that type, length and energy of the excitation signal diffused in the environment play a crucial role to derive impulse responses of high quality, which allow one to increase significantly the robustness of the recognition system. Past works had not analyzed the variability in recognition performance due to different methods for impulse response measurement, although differences have been evidenced at perceptual level.

The paper is organized as follows: Section 2 introduces the methods for impulse response estimation here explored for speech contamination purposes. Section 3 describes the experimental setup and the data sets used for training and testing of the speech recognizer. Section 4 provides details about the speech recognition system and the related investigated tasks, while Section 5 reports on experimental results obtained with different techniques and settings. Finally, Section 6 gives some concluding remarks and outlines possible future activities.

The research leading to these results has partially received funding from the European Union's 7th Framework Programme (FP7/2007-2013) under grant agreement n° 288121 - DIRHA. M. Ravanelli contributed to the work thanks to funding provided by DomoticArea s.r.l. and UniHospital s.r.l., under FESR project "Tecnologie di riconoscimento vocale a distanza dai microfoni in sanita' e domotica" - call n. 2/2010 of PAT. IR estimation was investigated by A. Sosi, student of Univ. of Trento, during his master thesis.

2. IMPULSE RESPONSE ESTIMATION

The Impulse Response (IR) is one of the most representative features characterizing an acoustic space. In the case of indoor reverberant room, if one assumes to deal with a linear time-invariant acoustical transmission system, IR provides a complete description of the changes a sound signal undergoes when it travels from one point in space to another [2].

IR estimation is a topic which has been widely discussed in the literature of the last two decades. The early proposed methods were referred to as *Direct*, i.e. methods based on diffusing in the environment a signal of impulsive nature, as a gun shot or a bursting balloon. These methods were then replaced by *Indirect* ones, characterized by using excitation signals different from the Dirac function, primarily due to the advantage of providing a higher Signal-to-Noise Ratio (SNR) which was not guaranteed by the former one. In the case of indirect methods, a known excitation signal is reproduced at a given point (for instance through a loudspeaker), and the corresponding signal is observed by a microphone placed at the other point in space. In general, the related observation is affected by environmental noise and non-linearities that may be introduced in the measurement chain by instrumentation.

In the category of indirect methods, some of the most commonly used state-of-the-art techniques are *Maximum Length Sequence* (MLS), *Linear Sine Sweep* (LSS), *Exponential Sine Sweep* (ESS) [3].

2.1. Maximum Length Sequence

Originally proposed by Schroeder [4], the MLS technique provides an indirect measurement of IR based on a finite-length pseudo-random sequence of pulses, which has spectral properties almost equivalent to a pure white noise [5]. Based on this technique, the impulse response is derived by cross-correlation between the MLS sequence (i.e., input signal) and the microphone signal (i.e., output signal). As mentioned above, if compared to direct methods MLS offers a better SNR; however, it is sensitive to non-linearities introduced by the measurement system, as shown in Figure 1(a).

2.2. Linear Sine Sweep

LSS is an indirect technique [3] characterized by an excitation input signal consisting in a sine whose frequency sweeps linearly with time (also referred to as *chirp*). Denoting with ω_1 and ω_2 the initial and final angular frequencies of the sweep and with L its length, it can be defined as follows:

$$x\left(t\right) = \sin\left(\omega_{\mathit{1}}\, t + \frac{\left(\omega_{\mathit{2}} - \omega_{\mathit{1}}\right)}{L}\frac{t^2}{2}\right) \tag{1}$$

As in the case of MLS, the impulse response derives from a cross-correlation between input and output signals. Besides a better SNR than in the MLS case, LSS introduces the advantage of a better (although not perfect) processing of the non-linearities.

2.3. Exponential Sine Sweep

The ESS technique, introduced by Farina [3], is based on an exponential time-growing frequency sweep, as described by the following relationship:

$$x(t) = sin\left[\frac{\omega_1 \cdot L}{ln\left(\frac{\omega_2}{\omega_1}\right)}\left(e^{\frac{t}{L}\cdot ln\left(\frac{\omega_2}{\omega_1}\right)} - 1\right)\right] \tag{2}$$

An advantage offered by ESS is the immunity against harmonic distortions. As shown by Figure 1(c), a perfect separation can be observed between the contributions due to non-linearities (i.e., harmonic distortion), appearing in the left part of the estimated IR, and contributions related to the linear impulse response (i.e., reverberation), observable in its right part. On the other hand, in the case of MLS this separation can not be found, which introduces noticeable artifacts in the estimated IR. In the case of LSS, although contributions due to harmonic distortion are mainly in the left part of estimated IR, the linear part of the IR can be affected at low frequencies. Another advantage offered by ESS is that its excitation signal spectrum is pink (note that it's white-like for both MLS and LSS), which ensures to have a better SNR at lower frequencies. This is a desirable feature both at perceptual level and for speech recognition purposes, for which a mel-filter bank is used with higher resolution in the lower part of the frequency axis. Due to this coloration in frequency, pre-equalizing signals is necessary, before applying cross-correlation between input and output signals, which consists in a filter with 3dB/octave gain.

2.4. From Impulse Response to Contaminated Speech

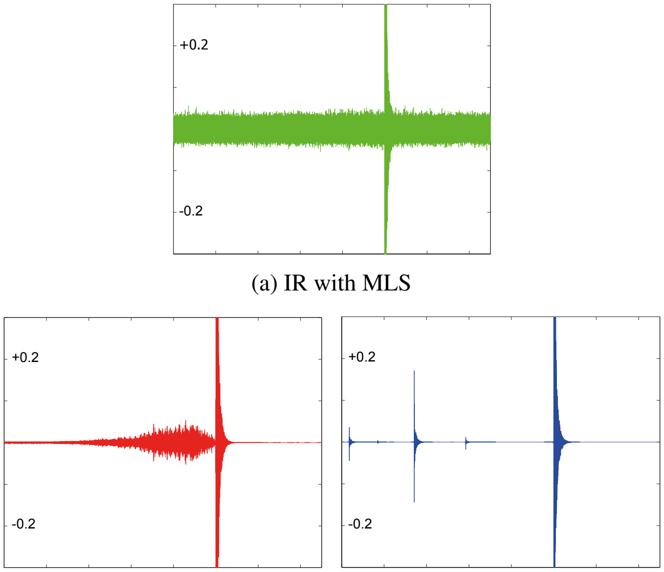
In the literature, transforming a clean (i.e., dry or anechoic) signal into a reverberated one is commonly referred to as auralization process, often applied in the case of binaural rendering [6]. For speech recognition purposes, this operation is often referred to as contamination process [7], as shown in the following relationship:

$$y_{rev}(t) = s_{clean}(t) * h(t) + n(t) \tag{3}$$

where one can note that the clean signal is convoluted with the given impulse response $h(t)$, and then further processed by adding a noise signal $n(t)$ that may consist in environmental background noise signals of variable amplitude, to account for different SNRs to simulate.

3. EXPERIMENTAL SETUP AND CORPORA

The following of this paper reports on experimental results of distant-speech recognition which were obtained using contaminated data (for training purposes) and real data collected in an apartment available in the context of DIRHA project (for evaluation purposes). Different recording sessions were performed in the given environment, to collect both speech material and audio signals useful to estimate impulse responses.



(a) IR with MLS

(b) IR with LSS

(c) IR with ESS

Fig. 1. Impulse responses provided by the three given techniques, when the measurement process has been affected by a harmonic distortion introduced by the amplifier-loudspeaker chain while diffusing the excitation signal in the environment.

3.1. Experimental Setup

The above-mentioned apartment comprises different rooms which were equipped with a network of microphone arrays for preliminary experiments under DIRHA project. The complete experimental set-up is based on the use of 30 omnidirectional microphones *Shure MX391*, a harmonic-nested array consisting of 13 electret microphones, 6 multi-channel audio card *RME Octamic II*, and a *Focusrite Saffire Pro 40* clocked one another. All the input/output signals were characterized by 48 kHz sampling frequency and 16 bit accuracy.

The remainder of this work will focus on the system behaviour for a speaker and a microphone both located in the living-room. The sound source is frontal to the microphone, located at a distance of about 4 m. The estimated reverberation time T_{60} is about 600 ms.

For impulse response measurement purposes, different categories of loudspeakers were considered. In particular, the following results are based on the use of a professional studio monitor, i.e., *Genelec 8030*, and of an inexpensive consumer speaker for computer, both active but characterized by a quite different frequency response and nonlinearities introduced when transducing sound.

In order to estimate impulse responses, MLS, LSS, and ESS excitation sequences were diffused by each loudspeaker in the environment, with varying length and amplifying settings. Each excitation signal was preceded and followed by fade-in and fade-out sequences of 50 ms duration, in order to avoid the introduction of any possible numerical clicks. Speech data collection was then conducted with all the real speakers located in the same position where the loudspeaker was previously placed. For comparison purposes, the speech uttered by each speaker was also recorded with a professional close-talking *Countryman E6* microphone.

Type	Use	Spks	Sent.	Words	Hours
Cont. APASCI	Train	176	3.9 k	33 k	∼6
Phon. Rich	Test	11	1.7 k	12 k	∼2
Commands	Test	11	1.3 k	7 k	∼1.5

Table 1. Characteristics of speech material used for training, and for testing of the system on phone and word loop tasks, respectively.

3.2. Data Set Description

The speech material used to train the distant-speech recognizer consists in contaminated versions (one for each IR measurements settings) of APASCI [8], an italian corpus of phonetically rich sentences, whose main characteristics are summarized in the Table 1.

In order to evaluate speech recognition performance, the utterances pronounced by real speakers (which were not in the APASCI corpus) in the above-mentioned apartment were organized in two tasks. The first one is a phone task related to a set of phonetically rich sentences, while the second one regards a list of sentences typical of a possible command-and-control home application. In the former case, different lists of sentences (about 150 per speaker) were used, while in the latter case the same list of 125 commands was shared by all the speakers. Main features of these two tasks are also summarized in Table 1 and reprised in the following section.

4. SPEECH RECOGNITION SYSTEM AND TASKS

The speech recognition system investigated in this work is based on a standard front-end processing consisting of a pre-emphasis step followed by feature extraction. The pre-emphasized signal is blocked and Hamming windowed into frames of 20 ms duration (with 50% overlapping). For each frame, 12 Mel-frequency Cepstral Coefficients (MFCCs) and the log-energy are extracted. MFCCs are normalized by subtracting the means, while the log-energy is normalized with respect to the maximum value on the whole utterance. The resulting normalized MFCCs and log-energy, together with their first and second order derivatives, are then arranged into a single observation vector of 39 components. Acoustic modeling operates at context-independent phone-like unit level, and is derived by applying the *Baum-Welch* algorithm, while the recognition step is accomplished by using *Viterbi* algorithm.

4.1. Phone Loop task

In this task, a reduced set of 26 phone-like units of the Italian language was chosen, which consists of 5 vowels, 5 fricatives, 4 affricates. 6 occlusives, 3 nasals, and 3 liquids. An acoustic model of silence/background noise is also used. Each phone-like unit is modeled with a three-state left-to-right continuous density HMM, with diagonal covariance matrix and with output probability distributions represented by means of mixtures of 55 Gaussian components.

4.2. Word Loop

In the Word Loop (WL) task, we adopted the same acoustic modeling realized for the PL task. The vocabulary consists of 233 words used to create any of the above-mentioned command-and-control sentences. Although the introduction of a grammar or language modeling would increase the system performance, in this work a word loop task was prefered to better emphasize any experimental evidence at acoustic level, which would otherwise be partially missed. It is worth noting that the vocabulary includes a quite large number of short and confusable words.

5. EXPERIMENTAL RESULTS

5.1. Baseline Results

In our past studies [7], contaminated speech-acoustic models were derived by using impulse responses obtained diffusing in the environments LSS signals of less than 5 s length. Moreover, dynamics was limited in order to avoid any possible artifacts due to harmonic distortion. On the other hand, in this work length (L) and dynamics (VOL) of the excitation sequence are objects of study. Table 2 shows the close-talking baseline results which represent a set of lower bound error rates to use as reference for the following experiments. Acoustic models were trained on the non-contaminated (clean) APASCI speech material, while test is performed on the close-talk microphone of the 11 speakers.

Training	Test	PER (%)	WER (%)
Clean APASCI	Close-talk	29.8	19.5

Table 2. Close-talking speech recognition performance expressed in terms of Phone Error Rate (PER) and Word Error Rate (WER), for the PL and the WL tasks, respectively.

5.2. Influence of the Excitation Length

A first set of experiments was conducted to investigate on the impact that excitation length has on recognition performance. In this case, the impulse responses were derived based on diffusing in the environment of MLS, LSS, ESS excitation signals with a low amplifying level. This experiment, in which SNR highly depends on the length of the emitted signal, mainly concerns the analysis of system sensitivity to noise introduced by the different proposed techniques rather than on harmonic distortion. For each IR measurement settings, a contaminated version of APASCI were generated to train the acoustic model of the ASR system.

In Figure 2, one can note that an increase of L corresponds to an improved performance for all the investigated techniques, thanks to the improved SNR in the IR measurement process. Results also show that ESS provides the best performance at any excitation signal length. This fact can be due to a better SNR at lower frequencies (pink spectrum), e.g. below 2-3 kHz, typically more critical in speech recognition. MLS and LSS, characterized by a white-like spectrum, don't have this interesting property. The experimental results also show that MLS has a higher sensitivity to noise than the other two techniques, however in general the difference in performance tends to decrease when L increases. Although not reported here, it is worth noting that the PL task provides a quite similar experimental trend.

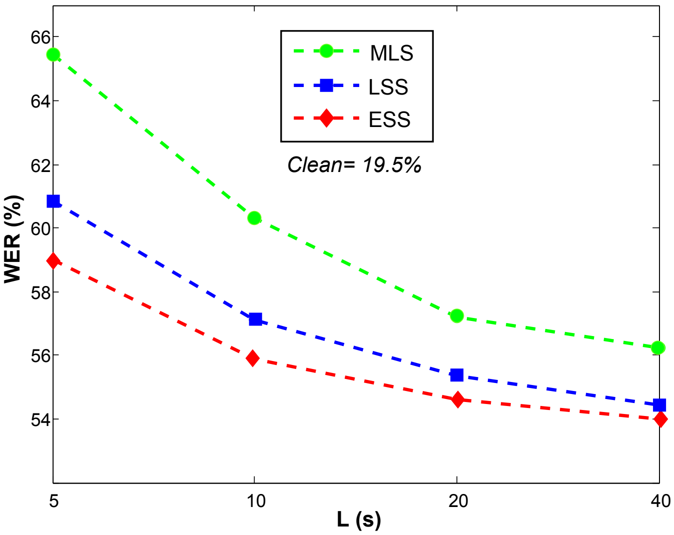


Fig. 2. Performance in terms of WER (%) obtained on the WL task when varying the length L of the excitation signal, given the minimum output level $VOL = V_1$ with professional loudspeaker Genelec 8030.

5.3. Influence of the Output Level

A second set of experiments regarded the analysis of recognition performance when impulse response measurements are realized with different dynamics at loudspeaker output level. When the monitor emits the excitation signal with a larger amplifier output level, together with an increase of SNR in the IR measurement process, one can introduce harmonic distortions. Results reported in Figure 3 refer to the case of L=*5 s*, which corresponds to a situation of SNR highly depending on the sound energy diffused in the environment.

Experimental results show that ESS outperforms the other two techniques. As previously, for lower dynamics (V_1,V_2), this is due to a better management of SNR. On the other hand, at higher levels of dynamics (V_4,V_5) the best performance provided by ESS is mainly due to a better management of harmonic distortions.

While for MLS and LSS, one should choose a trade-off setting in order to have a satisfactory SNR in IR measurement without introducing artifacts due to non-linearities, it is clear that ESS overco-

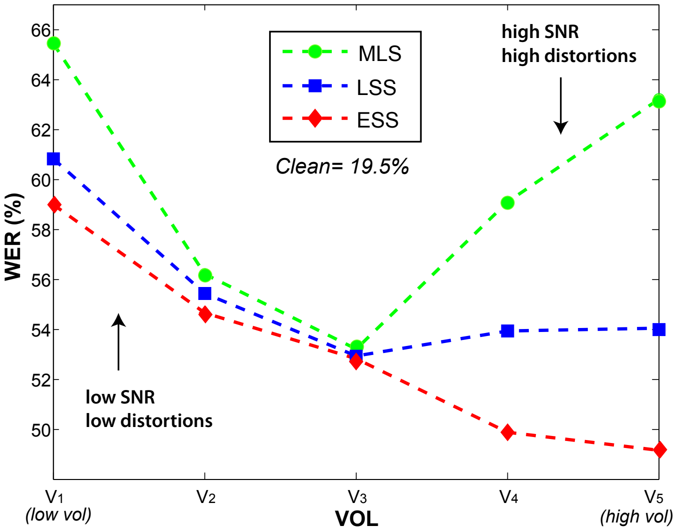


Fig. 3. Performance in terms of WER (%) obtained on the WL task when varying the loudspeaker output level with professional monitor Genelec 8030.

mes this trade-off ensuring to have better performance also when the output level of the active monitor increases.

5.4. Other Results

The absolute best performance were obtained using ESS technique with the maximum investigated length and the maximum output level of the excitation signal, simultaneously. Table 3 shows the results obtained with *L=40 s* and *VOL=* V_5 *(high vol)*, for both PL and WL tasks. Under GEN and PC columns, word error rates are reported for the use of the professional studio monitor (*Genelec 8030*) and of the inexpensive PC monitor, respectively. For both tasks, experiments confirm the advantage of adopting ESS method since, especially with this settings, we have both an high SNR and immunity to harmonic distortions. In particular, it is worth noting that word error rates decrease from 54.5% to 45.5% when the MLS is replaced by the ESS excitation sequence. Results also show that the use of a PC monitor, combined with the ESS method, does not introduce significant drawbacks, with just 0.5% word error rate difference between the two cases.

	MLS		LSS		ESS	
	GEN	PC	GEN	PC	GEN	PC
Phone Loop	51.1	51.7	49.3	49.9	48.8	48.9
Word Loop	54.6	56.5	47.6	48.7	45.5	46.0

Table 3. Recognition performance, expressed in terms of phone/word error rates, obtained with *L=40 s* and *VOL=*V_5 (*high vol*), using two different active monitors.

6. CONCLUSIONS

The results reported in this paper evidence the impact of impulse response estimation in the development of a distant-speech recognition system, when a contamination method is used to simulate noisy and reverberated speech signals for training purposes. In particular, we showed that ESS technique outperforms other IR measurments methods, especially when long and high dynamic excitation signals are emitted in the acoustic environment. This work is mainly focused on the application of the proposed method with a speaker/sound source located frontal to a microphone at four meter distance. Next activities will regard further studies on different sound source positions and orientations in the given home environment, even when there is no direct-path between the source and the microphone. Under the DIRHA project these methods will be extended to the case of a microphone array including beamforming and enhancement techniques. One of the main challenges of the project is to ensure satisfactory performance in a speech interaction based application for smart home, also when a quite limited number of microphones can be installed in the environment. With this regard, the study reported in this paper is of fundamental importance in order to direct the future activities regarding both the design of the microphone network distributed in the household and the development of a proper acoustic modeling for robust recognition purposes. In this case, in fact, the ultimate goal is to find the best trade-off between system performance, cost and invasiveness of the solution.

7. REFERENCES

[1] B. Lecountex, M. Vacher, and F. Portet, “Distant Speech Recognition in a Smart Home: Comparison of Several Multisource ASRs in Realistic Conditions,” *Interspeech*, August 2011.

[2] H. Kuttruff, “Room acoustic-fifth edition,” *Spon Press*, 2009.

[3] A. Farina, “Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique,” *110th AES Convention*, February 2000.

[4] M.R. Schroeder, “Diffuse sound reflection by maximum-length sequences,” *Acoust.Soc.Am*, April 1974.

[5] G.B. Stan, J.J. Embrechts, and D. Archambeau, “Comparison of different impulse response measurement techniques,” *Journal of the Audio Engineering Society*, vol. 50, pp. 249–262, April 2002.

[6] M. Vorlander, “Auralization,” *Springer*, 2007.

[7] M. Matassoni, M. Omologo, D. Giuliani, and P. Svaizer, “Hidden markov model training with contaminated speech material for distant-talking speech recognition,” *Computer Speech and Language*, November 2000.

[8] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, “Speaker independent continuous speech recognition using an acoustic-phonetic italian corpus,” *ICSLP*, September 1994.