ON SPEECH RECOGNITION PERFORMANCE UNDER NON-STATIONARY ECHO CANCELLATION

Mahdi Triki
Philips Research Laboratories, Eindhoven, The Netherlands

Keywords: Acoustic echo cancellation, Acoustic echo suppression, Non-stationary environment, Automatic speech recognition, Speech enhancement.

Abstract: During the last decades, performance of speech recognizers significantly increased for large vocabulary tasks and adverse environments. To reduce interference, acoustic echo cancellation has been proposed and extensively investigated. Particular attention was paid to the convergence proprieties and the capability to handle double talk. However, in time-varying environment, the echo canceller has the additional task to track the variations of the propagation channel. With this respect, it has been established that algorithms that exhibit fast convergence do not provide necessarily good tracking performances. In such an environment, performance assessment is also challenging and the 'experiment' design is crucial to provide consistent and interpretable results. In the present paper, we reproduce time-varying artifacts by altering the surrounding acoustic environment (using a moving person/robot). The movement characteristics (discrete/continuous) and location (line-of-sight/background) emphasizes different room/algorithms characteristics and provides deeper insights on the system behavior.

1 INTRODUCTION

During the last three decades, performance of speech recognizers significantly increased even for large vocabulary tasks (X. Huang and Hon, 2001). The upper-bound performances of recognizers are generally achieved when a close-talk microphone is recording the speech signal, i.e., when no competing speaker, noise sources and/or reverberation affect the original clean speech signal. Many desired settings may require the speaker to be either far from the microphones or surrounded by one or many noise sources, or both. Indeed, for many applications, close-talk recordings (headset solutions) are not desired for aesthetic and/or convenience reasons; while in various environments (e.g. in living-room, car, hospital), surrounding noise cannot be neglected. In these situations, speech recognizers dramatically fail to reach the minimal threshold of performance that the usability is requiring, even with a small vocabulary size.

Often, prior information about the nuisance sources (e.g. radio, background music) is available. These information could be exploited to alleviate noise, enhance the desired source, and increase the recognition accuracy. Typically, the interference is predicted (using an appropriate adaptive processing), then subtracted from the received signal (as illustrated in Figure 1).

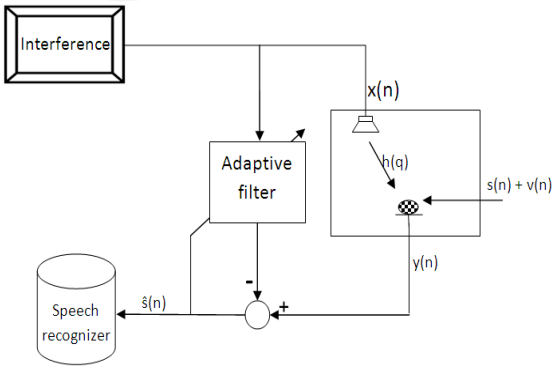

Figure 1: AEC: problem statement.

The enhancement scheme is referred to as Acoustic Echo Cancellation (AEC). AEC was extensively investigated for both enhancement (J. Benesty and Gay, 2001) and recognition (J. Picone and Hartwell, 1988) applications. The reported results assume generally the coupling between the interfering source and the received echo (the propagation channel) stationary. In reality however, this coupling may be time-varying due to the movement of the desired speaker,

Triki M..
ON SPEECH RECOGNITION PERFORMANCE UNDER NON-STATIONARY ECHO CANCELLATION.
DOI: 10.5220/0003182903160321
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing* (BIOSIGNALS-2011), pages 316-321
ISBN: 978-989-8425-35-5

other persons present in the same room, or due to the variations of physical parameters (e.g. temperature). In such a case, the adaptive processing will have the additional task of tracking the variation of the propagation channel. With this respect, it has been established that adaptive algorithms that exhibit good convergence properties in stationary environments do not necessarily provide good tracking performance in a non-stationary environment; because the convergence behavior of an adaptive filter is a transient phenomenon, whereas the tracking behavior is a steady-state property (Haykin, 2002; Triki, 2009).

In this paper, we address some issues related to the performance evaluation of echo-cancellation in time-varying environments. Generally, experimental evaluation should produce meaningful, consistent and interpretable results. In time-varying environments, the assessment is particularly challenging and the 'experiment' design is crucial. On one hand, mimicking the user experience (moving user/capturing-device) is difficult to reproduce and to interpret. On the other hand, simulating the impulse responses offers flexibility and reproducibility, but gives results that are difficult to interpolate to real-world environment. Alternatively, we reproduce the time-varying artifacts by altering the surrounding acoustic environment (while keeping the source and capturing devices fixed). These alterations are introduced by moving person/robot. The movement characteristics (discrete/continuous) and location (line-of-sight/background) represent degrees of freedom that emphasis various room/algorithms characteristics and provide deeper insights on the system behavior.

The remainder of this paper is organized as follows. In section 2, the experimental setup used for the data acquisition and performance analysis is described. Acoustic echo cancellation and noise suppression building blocks are investigated in sections 3 and 4 respectively. Finally, a discussion and concluding remarks are provided in section 5.

2 EXPERIMENTAL SETUP

Throughout this paper, we evaluate different speech preprocessing schemes in order to isolate their performance impact on the recognition rate and motivate further refinements. In the following, we present our experimental setup to assist this progressive unfolding of the speech preprocessing design. Namely, we will describe the data collection procedure, and specify the characteristics of our data recording space (reproducing a living-room environment).

2.1 Data Collection

Defining a formal data collection process is necessary, as it ensures that gathered data is both defined and accurate and that subsequent findings and decisions are valid. The aim of the present work is to investigate the effect of the extrinsic variabilities (noise, reverberation, interference) on the recognition rate. Thus, the data should be collected such to reduce the effect of intrinsic variabilities (that may bias the final conclusions). Specifically, particular attention was paid to:

- Linguistic accent: we have chosen North American native speakers (American or Canadian). The choice was motivated by the fact that our recognition system (that we use for the evaluation) was trained (optimized) for this particular accent.
- Speech rate changes: the variation of the speech rate was alleviated with a two step simulation approach: first we collect the input data, next the various tasks are reproduced using a dummy-head.
- Additive noise: the data collection was performed in a noise-free and low-reverberent environment.

North-American native speakers (4 males, 1 female) were asked to participate in the data collection process. Two dictionaries were defined:

- Controls dictionary, e.g., 'switch on', 'is there any sport program tonight'.
- Artist names dictionary, e.g., 'Madonna', 'Tokio Hotel', 'Laura Pausini'...

The recordings were performed in a noise-free and low-reverberant room (see Figure 2). The speakers were seated in a comfortable chair while they read aloud one-by-one a list of items. The items were displayed using a PowerPoint presentation at constant speed (12 items per minute). The speech signal was captured at 48 kHz.

2.2 Data Recording

We have investigated the recognition accuracy in a living-room environment. The recordings were carried out in a four-by-six meters demonstration room. (see Figure 3 and Figure 5 for schematic representation). The room reverberation time is $T_{60} \approx 300$ *ms*.

In order to account for speech rate variabilities, the control/search commands (recorded during the data collection phase) were reproduced by a KEMAR (Knowles Electronics Manikin for Acoustic Research). The KEMAR was placed at 3 meters distance from the TV set. The audio signal was captured

Figure 2: Data collection room.

Figure 3: Recording room.

by an omnidirectional microphone. The signal was recorded at 48 kHz, then downsampled to 8 kHz (to meet the specifications of the speech recognition engine). The omnidirectional microphone was placed at 30 cm distance from the KEMAR loudspeaker. During the recordings, CNN channel was turned on (the average Signal-to-Interference-Ratio (SIR) was about 10 dB). The North-American accent of CNN speakers makes the TV interference further challenging.

3 ACOUSTIC ECHO CANCELLATION FOR ASR

Acoustic echo arises when an interfering sound (here produced by a TV) is picked up by a microphone, together with its sound wave reflection into the surrounding walls and objects. Usually, the received signal is decomposed into direct sound, reflections that arrives shortly after the direct sound (commonly called early reflections), and reflections that arrive after the early reverberation (called late reverberation and often approximated as white, diffuse, exponentially decaying additive noise) (Habets, 2007). Several adaptive schemes were proposed to estimate the room reverberation and compensate for the interfering TV echo. Among them, the class of the Recursive Least-Squares (RLS) algorithms (Haykin, 2002) (and their frequency domain implementation (Shynk, 1992)) have shown to exhibit a fast convergence, and reduced sensitivity to the color of the input signal. Motivated by the application requirements (fast convergence) and the input characteristics (speech content), an RLS real-time solution was implemented to update the echo-canceller scheme, and used to reduce the interference (TV signal) prior to recognition.

We first consider a stationary (time-invariant) environment. In such configuration, the echo-paths do not change, and only the steady-state convergence of the AEC is focal. For recognition performance analysis, we distinguish the substitution, insertion and deletion error rates, defined as:

$$\text{substitute} = \frac{\#\text{substituted commands}}{\#\text{total commands}}$$

$$\text{inserte} = \frac{\#\text{inserted commands}}{\#\text{total commands}}$$

$$\text{delete} = \frac{\#\text{deleted commands}}{\#\text{total commands}}$$

where # denotes the cardinality operator. Intuitively, insertion and deletion errors refer respectively to false positive and false negative detection errors, while a substitution occurs when a command is well detected but misrecognized. The recognition was performed using a Philips Speech Recognition system. The models used by the engine were trained with US-English speech data.

Figure 4 illustrates the recognition accuracy as a function of the order(length) of the FIR echo canceller. We observe that without echo-cancellation ($L = 0$), the recognition system do not reach the usability threshold. Moreover, the recognition performance increases with the AEC length: the longer the AEC, the better the echo-path modeling, and the further the echo is suppressed. However for AEC length ($L > 1024$), only minor additional gain was observed (particularly for 'substitute' and 'delete' measures): in this region, modeling errors are small compared to estimation and adaptation errors.

Next, we investigate non-stationary (time-varying) scenarios. We have defined and compared the recognition accuracy in four settings:

- No-mvt: no interferent person (stationary propagation)
- Back-mvt: an interferent person continuously moving on the background region (Fig 5.(a)). In such scenario, the direct sound and early reverberation are still time-invariant, only late reverberation varies.

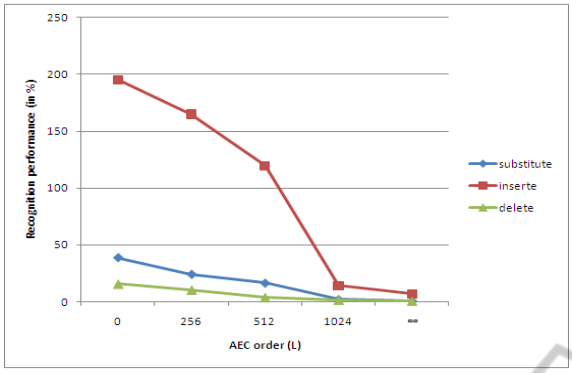

Figure 4: Recognition performances (in %) for processed (echo-cancelled) signal function of the AEC order (L).

- Kont-mvt: an interferent person is continuously moving on line-of-sight region (Fig 5.(b)). Thus, all the room reverberation components (direct, early, and late) are varying.
- Disk-mvt: an interferent person is moving on the line-of-side region in a discrete fashion (step - 'immobile' 5 seconds - step ...).

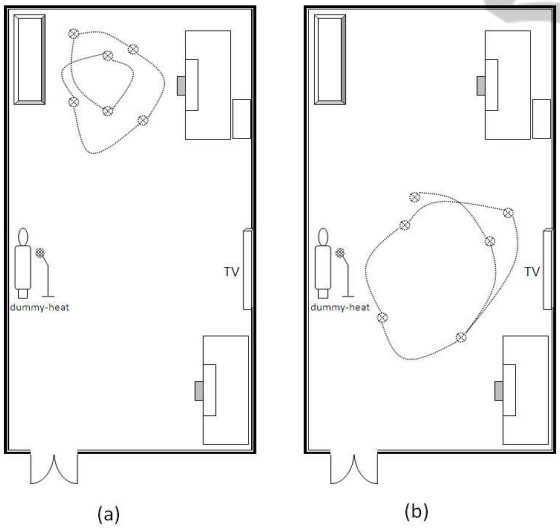

Figure 5: Recording scenarios for AEC tracking analysis.

In these scenarios,the echo-paths are altered by the presence of an external disturbance (moving persons). However, equivalent artifacts may occur when the position of the source (desired speaker) or the capturing microphone varies. The predefined scenarios have a double advantage. First, they are *simple* to simulate and reproduce. Second, they *decorrelate* the effects of the variation of early vs. late reverberation, as well as the convergence vs. tracking capabilities of the AEC solutions.

As the insertion errors could be handled to some extend, for instance, by a 'press-to-speak' button, we limit our attention to the substitution and the deletion errors and we define the recognition error rate as:

$$RER = \frac{\#\text{substituted + deleted commands}}{\#\text{total commands}} = substitute + delete$$

Figure 6 compares the recognition error rate gain, i.e,

$$RER_{Gain} = RER_{\text{noisy signal}} - RER_{\text{with AEC}}$$

computed for the four previously described scenarios.

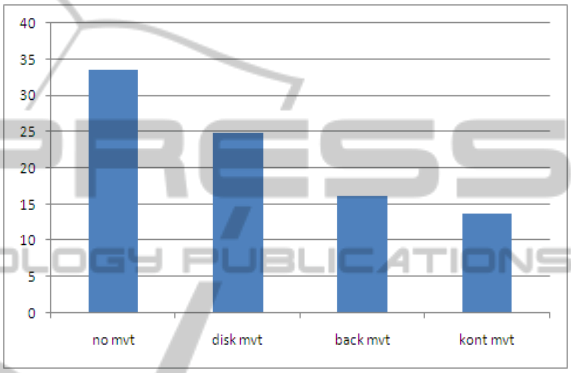

Figure 6: RER_{Gain} (in %) for the different non-stationary scenarios.

One may notice that despite the RLS algorithm has a relatively good steady-state performance (scenario 'No-mvt') and rapid convergence (scenario 'Disk-mvt'), its tracking capabilities (scenarios 'Back-mvt' and 'Kont-mvt') is not sufficient to capture continuous variations of the propagation channel. To alleviate this problem, spectral-based post-processing is proposed and investigated in the following section.

4 ACOUSTIC ECHO AND NOISE SUPPRESSION FOR ASR

We have observed that using solely adaptive FIR filters to perform echo cancellation would require a large number of coefficients. This results in large memory requirements and large convergence time. Moreover, perfect tracking of the non-stationarities in the propagation channel is problematic. Thus, additional measures have to be taken to guarantee robustness. In communication systems, spectral post-processing has been proposed at the AEC output. The basic idea is to estimate the amplitude spectrum of the desired signal and combine it with the phase available

from the degraded signal for reconstructing the enhanced signal. In practice, a time-varying gain filter is designed to reconstruct the desired signal. A number of well-known gain functions $G_n(f)$ can be formulated as (Etter and Moschytz, 1994; Tashev, 2006):

$$G_n(f) = \begin{cases} \left[1-\gamma\left(\frac{|R_n(f)|}{|Y_n(f)|}\right)^{\alpha}\right]^{\beta} & \text{if } \gamma\left(\frac{|R_n(f)|}{|Y_n(f)|}\right)^{\alpha} < 1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where Y_n and R_n denote the amplitude spectrum of the received and the remaining noise signals. The remaining noise signal originates both from the remaining interference (after echo cancellation) and the ambient noise. These approaches have been relatively successful due to their implementation simplicity and their robustness against non-ideal circumstances. They were extensively investigated and optimized for communication systems. Particularly, it has been shown that oversubtraction ($\gamma > 1$), smoothing the gain factor G_n, and constraining the gain minimum (i.e. $G_n(f) = \min(G_{min}, G_n(f))$ enhances the audio quality and reduces the musical noise (M. Berouti and Makhoul, 1979).

However, it is well established that increasing the audio quality does not lead necessarily to a better recognition rate. Indeed, recognizers hinge critically (only) on spectral information. Any processing leading to spectral distortion (especially time-varying coloration) may seriously affect their performance. Moreover as features extraction is performed in the log spectral domain, computational stability issues may arise (e.g. $\log(x) \;/\; x \to 0$), which is not always well handled with commercial recognition engines. Thus, the oversubtraction factor γ set a tradeoff between noise/interference reduction and stationary spectral distortion, while the gain dynamics (via the choice of G_{min}) leads to a compromise between the noise/interference tracking capability and dynamic spectral distortion.

We have implemented three post-processing methods:

1) spectral magnitude subtraction ($\alpha = 1, \beta = 1$).

2) minimum mean square error (MMSE) estimation ($\alpha = 2, \beta = 1$).

3) MMSE estimation in the log-spectral domain (Ephraim and Malah, 1985).

We have compared the recognition performances after the post-processing for the four tracking scenarios described in Section 3 ('No mvt', 'Back mvt', 'Kont mvt', 'Disk mvt'). None of the post-processing schemes consistently outperforms the others. In the following, we will limit our attention to the spectral magnitude subtraction technique as it is easier to implement and to interpret.

Next, we focus on the effect of the gain dynamic on the recognition accuracy. In communication systems, it has been noticed that noise suppressors suffer from the rapid fluctuation of the SNR both in time and frequency domains. It has been shown that reducing the gain dynamic (by introducing a minimum gain constraint, i.e., $G_n(f) = \min(G_{min}, G_n(f))$) reduces auditory artifacts. For speech recognition, our experiments show that imposing a minimum gain constraint is also required. The recognition error rate and the insertion error rate function of the minimum gain G_{min} are plotted ('kont mvt' scenario) in Figure 7. No oversubtraction was performed (i.e. $\gamma = 1$)

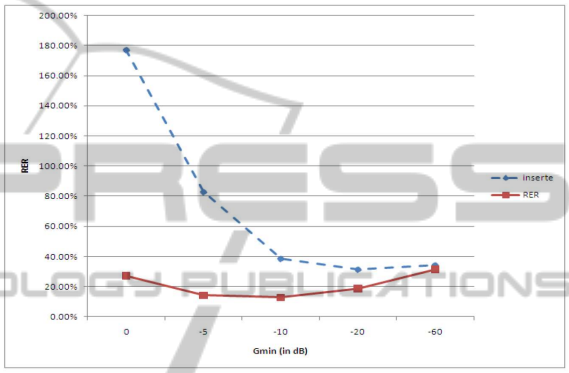

Figure 7: Insertion and Recognition Error Rate function of the minimum gain G_{min} (in dB), for the 'kont mvt' scenario.

Finally, we investigate the effect of the subtraction factor γ. In communication systems, it was shown that oversubtraction $\gamma > 1$ improves the audio quality. We

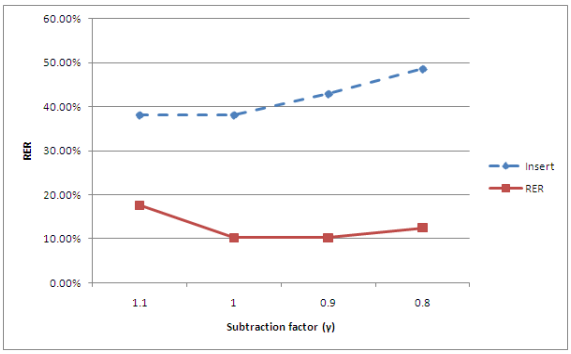

Figure 8: Insertion and Recognition Error Rate function of the subtraction factor γ, for the 'kont mvt' scenario.

observe (in Figure 8) that oversubtraction improves the insertion at the expense of recognition error rates, as it allows for further noise subtraction. On the detection region, undersubtraction seems advantageous as it allows for less spectral distortions.

5 CONCLUDING REMARKS

In the present paper, we have investigated the performance of echo-cancellation for voice control devices operating in non-stationary propagation conditions. Four distinct scenarios have been defined and analyzed. In addition to be simple to simulate and to reproduce, these scenarios decouple the effect of early vs. late reverberation as well as the convergence vs. tracking capabilities of the AEC solutions. This provides additional insights on reverberation artifacts/effects, and allows better design of the adaptive schemes.

Our experimental investigation has confirmed that AEC systems that exhibit good convergence properties in stationary environment do not necessarily provide good tracking performance in non-stationary environment. We have also shown that spectral subtraction based post-processing may alleviate non-stationary reverberations. Moreover, particular attention should be paid to limit the gain dynamics and to the subtraction factor selection.

REFERENCES

Ephraim, Y. and Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. on Acoustic, Speech and Signal Processing*.

Etter, W. and Moschytz, G. (1994). Noise reduction by noise-adaptive spectral magnitude expansion. *Journal of the Audio Engineering Society*.

Habets, E. (2007). *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*. PhD thesis, Technische Universiteit Eindhoven.

Haykin, S. (2002). *Adaptive Filter Theory*. Prentice Hall.

J. Benesty, T. Gansler, D. M. M. S. and Gay, S. (2001). *Advances in Network and Acoustic Echo Cancellation*. Springer.

J. Picone, M. J. and Hartwell, W. (1988). Enhancing the performance of speech recognition with echo cancellation. In *IEEE Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*.

M. Berouti, R. S. and Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In *IEEE Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*, volume 4, pages 208–211.

Shynk, J. (1992). Frequency-domain and multirate adaptive filtering. *IEEE Signal Processing Magazine*.

Tashev, I. (2006). *Defeating Ambient Noise: Practical Approaches for Noise Reduction and Suppression*. Tutorial at ICASSP.

Triki, M. (2009). Performance issues in recursive least-squares adaptive gsc for speech enhancement. In *IEEE Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP)*.

X. Huang, A. A. and Hon, H. (2001). *Spoken Language Processing*. Carnegie Mellon University.