ON THE USE OF EMPIRICALLY DETERMINED IMPULSE RESPONSES FOR IMPROVING DISTANT TALKING SPEECH RECOGNITION

Thomas Plötz and Gernot A. Fink

Intelligent Systems Group, Robotics Research Institute
Dortmund University of Technology, Dortmund, Germany
Email: {Thomas.Ploetz,Gernot.Fink}@udo.edu

ABSTRACT

Recognition rates of distant talking speech recognition applications substantially decrease if the acoustic environment contains reverberation. Although standard approaches for compensating such distortions, e.g. cepstral mean subtraction (CMS), are quite effective, they are not appropriate for dynamic human machine interaction. When only short portions of speech are uttered by speakers at different positions, compensation methods fail that require several seconds of speech. For this kind of applications we present a dereverberation approach utilizing empirically determined impulse responses. Prior to speaking users are asked to produce some impulse-like signal (clapping their hands, or snipping the fingers) which is used for compensation. By means of an experimental evaluation on the German Verbmobil corpus we demonstrate the promising potential of the approach.

Index Terms— De-reverberation, distant talking speech recognition, impulse responses, cepstral mean subtraction

1. INTRODUCTION

For intuitive human-machine interaction (HMI) applications speech represents an indispensable modality which allows for truly natural communication with technical systems. Automatic speech recognition (ASR) for HMI applications often needs to deal with speech recorded by distant microphones. In contrast to the (rather intrusive, hence inappropriate) use of head-sets, in this setting the quality of the recorded utterances is limited. Reasons for this are additionally recorded noise and distortions of the speech signal caused by the recording environment (reverberations).

Due to such distortions signal enhancement represents a pre-requisite for distant talking ASR. In fact, numerous techniques for the enhancement of acoustic conditions have been described. These techniques are either directly related to the actual ASR process (e.g. speech enhancement by room effect compensation at the level of feature extraction), or to acoustic signal processing in general. The majority of related approaches address de-reverberation by exploiting filtering techniques based on linear systems theory. By the analysis of impulse responses effective compensation can be realized.

Most filtering based speech enhancement techniques rely on the detection and analysis of ideal impulse responses. Furthermore, often rather large portions of speech data uttered by a speaker at a certain (fixed) position are required for the successful compensation of acoustic distortions. Considering non-artifical settings for dynamic HMI applications it becomes clear that techniques of this kind are likely to fail at the desired compensation of room effects or noise. In such scenarios speakers are usually wandering around while uttering short statements, which violates the pre-requisites.

In our research we focus on practical solutions to the aforementioned problem. We are aiming at the improvement of a distant talking ASR system used for voice control of the electrical installation (lights, sun-blinds etc.) of a smart house – the FINCA [1]. One of the fundamental challenges in this setting is the existence of long-term reverberations severely distorting acoustic signals. We empirically determined a reverberation time $T_{60} = 519$ms which clearly indicates the highly reverberant character of the acoustic environment our ASR system needs to deal with.

Within this context we investigated the use of empirically determined impulse responses for the enhancement of short spoken instructions, which is described in this paper. In the next section the relevant related work is briefly reviewed. Our novel approach for improving distant talking speech recognition is described in section 3. The results obtained in an experimental evaluation based on the German Verbmobil corpus are presented in section 4 and the advantages and limitations of the approach are discussed.

2. RELATED WORK

Especially addressing the compensation of reverberations in acoustic environments numerous approaches have been developed. Basically, (blind) deconvolution techniques can be applied (cf. e.g. [2]), which, technically, corresponds to filtering based on linear systems theory.

State-of-the-art speech recognition systems, however, use an approximation of the aforementioned deconvolution technique, namely *cepstral mean subtraction* (CMS) (cf. e.g. [3, ch. 10]) as an integral part of the feature extraction stage. The goal is to compensate for convolutive speech distortions caused by different recording equipments. However, as CMS is employed within the analysis windows used for feature extraction (usually 10 to 20ms long) distortions caused by reverberations that operate on substantially longer time-scales can hardly be compensated.

Therefore, in [4] and later in [5, 6] the use of *long-term log-spectral mean subtraction* (LLSMS) was proposed to counteract the effects of reverberation on speech recorded in distant talking settings. The method principally works like CMS. The most notable difference is that the short-term log-spectral estimate is computed over windows of 1 to 4s length and averaged over up to 12s of speech. Additionally, after normalization speech is re-synthesized by an overlap-add technique and normal feature extraction is applied afterwards (as for clean speech). Though speech normalized such exhibits substantial audible distortions considerable improvements in recognition quality are reported for data recorded with a distant microphone in a meeting room.

As the transfer function in a distant talking scenario also varies with the speaker's position, in [7] (and certain related publications of its authors) the position dependent normalization of speech by applying CMS was proposed. The setting described is comparable to the one investigated in our work. Especially for the enhancement of short utterances its capabilities seem, however, rather limited. Furthermore, it is not clear to what extent the experimental setup really suffers from reverberation since the particular ambient reverberation time has not been given.

3. EMPIRICALLY DETERMINED IMPULSE RESPONSES FOR DISTANT TALKING ASR

In order to allow for an effective improvement of the distant talking speech recognition system in the highly reverberant environment of the FINCA we use the aforementioned LLSMS approach (cf. previous section) as the basis for our developments. The original approach relies on long speech durations (optimal estimation window 12s) which is, unfortunately, not feasible for our and related Ambient Intelligence scenarios. Since the speech interaction with the smart house is reduced to certain control commands uttered on (almost) constantly changing locations it is unrealistic to rely on spoken utterances of this length.

Therefore, we perform *log-spectral mean subtraction* (LSMS) by using a "spontaneous" estimate of the log-spectral density. Ideally, this estimate would reflect the transfer function of the room observed in the distant talking setting. A rather good estimate of the desired log-spectral density can be obtained by passing an appropriate test signal (usually a sine *sweep*) through the acoustic pipeline. The estimate could, however, also be approximated by signals that exhibit impulse like characteristics (most importantly an approximately flat frequency distribution) while still being "natural" in the setting of HMI.

In these premises, we derive coarse approximations of log-spectral estimates of the ideal impulse response from sounds of hand clapping or snipping the fingers. Obviously, these sounds differ largely from an ideal impulse and, therefore, compensations based on these estimates are still suffering from severe distortions. Consequently, the recognition quality can not be expected to reach the same level as in the case of compensating with a log-spectrum based on the measured response. Since it is, however, impracticable to rely on ideal impulses for the desired ASR improvement in HMI applications the use of such coarse approximations seems more promising compared to no compensation at all.

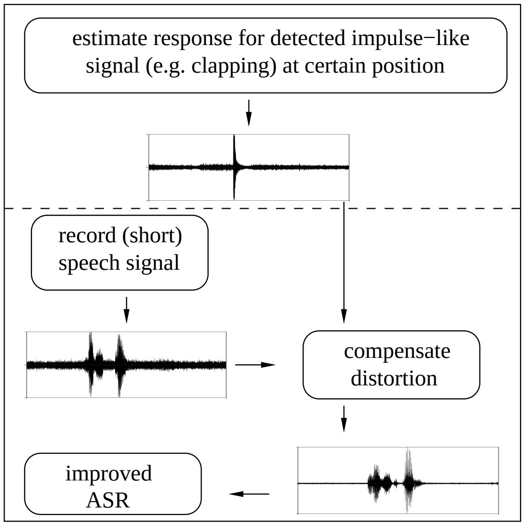
estimate response for detected impulse−like signal (e.g. clapping) at certain position
record (short) speech signal
compensate distortion
improved ASR

Fig. 1. Empirically estimated impulse responses for improved distant talking ASR – system overview

In figure 1 an overview of the system for distant talking ASR improvement based on empirically determined impulse responses is given. Prior to some spoken command the particular speaker produces some impulse like signal – he claps his hands or snips with his fingers (upper part of the figure) – at the position where he, afterwards, utters the command. By means of the impulse response estimated on the "initiation" signal deconvolution, i.e. alleviation of room reverberations, is performed (lower, right part) and the improved signal is passed to distant talking ASR.

In order to more formally describe our approach let us first look at the principle idea behind LLSMS. As in our approach this technique assumes that clean speech s is passed through an acoustic channel c which causes convolutive distortions resulting in an observed signal $\tilde{s} = s * c$. Additive noise is not taken into account in this model. When transforming this relation into the log-spectral domain – and considering the power-spectral density only for simplicity – the convolutive distortion c becomes an additive one:

$$\log|\tilde{S}| = \log|S| + \log|C|$$

Computing the long-term average, i.e. the expectation of $\log|\tilde{S}|$, which is in fact the conditional expectation given the excitation signal s, one obtains:

$$\mathcal{E}\{\log|\tilde{S}|\}|_s = \mathcal{E}\{\log|S|\}|_s + \mathcal{E}\{\log|C|\}|_s = \text{LLSM}$$

The resulting estimate of the long-term log-spectral mean (LLSM) consists of a portion accounting for the channel, which, however, depends on the speech data itself, and the average log-spectrum of speech.

For every analysis window a normalized speech spectrum can then be obtained from the normalized power spectrum $|\hat{S}| = \exp(\log|S| - \text{LLSM})$ combined with the original phase information. The complete normalized signal then results from overlap-adding the contributions of all the analysis windows.

In our LSMS approach we use a compensation term that represents a channel estimate only. It is computed *independently* from any speech data that need to be normalized on the basis of a suitable excitation signal e – e.g. the clapping of hands – which is passed through the acoustic channel. From the distorted version $\tilde{e}$ of this excitation an estimate of the LSM normalization term

$$\mathcal{E}\{\log|\tilde{E}|\}|_e = \mathcal{E}\{\log|E|\}|_e + \mathcal{E}\{\log|C|\}|_e = \text{LSM}$$

can be derived. It represents a reasonable estimate of the channel distortion if the excitation used has an approximately flat spectrum.

4. EXPERIMENTAL EVALUATION

Since the (spontaneous) interaction of humans with a smart house is difficult to evaluate in a quantitative and objective manner we decided to conduct recognition experiments based on a standard speech recognition corpus. Together with impulse like signals (sweep, clapping, snipping) it was replayed through a loudspeaker within the FINCA. The effectiveness of ASR improvement based on exploiting empirically determined impulse responses is measured by the changes in word error rates (WER).

4.1. Baseline Recognition System

As the baseline system for our experiments we used a recognizer designed for the 5k spontaneous speech recognition task defined by the 1996 evaluation of the German Verbmobil project. The system was trained on over 27 hours of spontaneously spoken dialogs that were elicited from more than 600 speakers in a fictitious appointment scheduling scenario.

In the feature extraction stage 12 mel-frequency cepstral coefficients and a normalized energy term together with delta and delta-delta coefficients are computed. In order to reduce short term channel distortions we apply causal cepstral mean normalization. The acoustic model is based on semi-continuous Hidden-Markov Models (HMMs) with a shared codebook of 1k Gaussians, linear topology, and a data dependent number of states. On the basis of triphone models, acoustic units that can be trained robustly are built by applying state clustering after the first Baum-Welch reestimation step. Afterwards, another 9 re-estimation steps are applied. During decoding a bi-gram language model with a test-set perplexity of 64.2 is used by a recognizer that uses time-based search-tree copies and operates in a strictly time-synchronous manner. All recognition systems described in this paper were built using the tools and methods provided by the ESMERALDA development environment [8]. As can be seen in table 1 the baseline clean speech recognition system achieves a word error rate of 28.1% on this challenging task.

4.2. Results

For the recognition experiments in the reverberant environment ($T_{60} = 519$ ms) of the FINCA we replayed the data of the 1996 Verbmobil test set through a loudspeaker (Behringer TRUTH B2030A) and recorded the reverberant signals by a artificial-head microphone (Sennheiser KU 100) positioned at a distance of approximately 1.5 meters. All acoustic signals were sampled at 16 kHz and only the left channel of the recorded stereo-data was used.

On the reverberant data the recognizer trained on clean speech achieves only a quite unsatisfactory performance with a word error rate of almost 72%.

As a kind of baseline method for compensating the effects of reverberation we first applied LLSMS[1] to both the clean training and the reverberated test data using different analysis window lengths (1.024s and 4.096s). Both normalization techniques significantly increase the error rate when applied to clean speech data as shown in table 1. Note that the relative increases in error rate are roughly consistent with the results published in [5] for the much easier task of recognizing digit strings where the clean speech baseline achieved a word error rate as low as 1%. Furthermore, also in our experiments the configuration with the longer window size outperforms the 1s version when applied to the reverberant data though both window lengths are longer then the observed reverberation time T_{60} of 519 ms.

Second, we used the recording of a so-called *sweep*, i.e. a sine signal of 10.4 s length with logarithmically increasing frequency and uniform energy distribution, for rather accurately estimating the log-spectral density of the true transfer function. In fact this sweep-based estimate represents the optimal impulse response that can be achieved in this setting. As both LLSMS and LSMS cause quite noticeable acoustic distortions regardless of the normalization applied, we used the system trained on 1s LLSMS data for decoding on the test normalized by the sweep-based log-spectral estimate. In this configuration a rather mediocre performance with an error rate of more than 68% is achieved.

[1] We used the implementation of the LLSMS method supplied by Gelbart & Morgan on the web page accompanying their paper [5] `http://www.icsi.berkeley.edu/Speech/papers/asru01-meansub-corr.html`.

condition	train / test	window length	WER [%]	Δ WER [%]
baseline	clean / clean	–	28.1	–
LLSMS	clean / clean	1s	32.5	15.7
		4s	37.1	32.0
baseline	clean / reverb.	–	71.7	155.2
LLSMS	clean / reverb.	1s	65.1	131.7
		4s	**60.3**	114.6

Table 1. Results of baseline recognizer without any compensation and LLSMS approach (deviations of more than 1.2% are significant at a level of 95%)

LSMS estimate	length of excitation	smoothing	WER [%]	Δ WER [%]
sweep	10.4	–	68.4	13.4
		median	65.1	8.0
snipping	3.2	median	62.1	3.0
clapping	4.1	median	**61.7**	2.3

Table 2. Results for the proposed LSMS approach with different normalization of the reverberant *test* data (training condition: LLSMS, 1 s window, Δ WER wrt. LLSMS, 4 s)

Finally, we derived coarse estimates of the true impulse's log-spectral density from sounds of hand clapping and finger snipping. When using those raw estimates directly for LSMS the resulting speech data is too heavily distorted to be useful. Therefore, we smoothed the raw log-spectral estimates by applying a median filter with a window size of 60 frequency bins. The results obtained are shown in the lower part of table 2. For purposes of comparison we also applied the same smoothing to the sweep-based estimate, which resulted in a significant reduction of the word error rate. Interestingly, the best results – given the complexity of the task – with error rates around 62% are achieved when using the smoothed LMS estimates of hand clapping and finger snipping. With respect to the best-performing LLSMS approach this corresponds to a relative increase in word error rate of less than 3%.

5. SUMMARY

In this first study on using spontaneous estimates of impulse responses for compensating effects of reverberation in distant talking speech we successfully proved the principal effectiveness of the proposed approach. Not requiring several seconds of speech signals for the compensation allows for rapid dereverberation. In fact no speech at all is required for the estimation of the room transfer function since we could demonstrate that the use of impulse-like signals like a clap of the hands or a snip with the fingers is sufficient.

The proposed approach is especially relevant for the application domain of dynamic human machine interaction where short statements are uttered by speakers with varying positions. In order to improve distant talking ASR users are simply asked to perform certain impulse-like signals prior to talking. In an experimental evaluation based on replaying (via loudspeaker) the test set of the German Verbmobil corpus in our heavily reverberant smart house the effectiveness of the approach could be demonstrated.

So far we did not yet investigate systematically the effect of varying speaker positions on the process. In informal experiments we found, however, that in our setting the reverberation characteristics varies slowly with speakers' positions. Therefore, a rather coarse sampling of the interaction space will probably be sufficient which can then be integrated with the automatic tracking of speech sources. In future work we will also focus on the improvement of the absolute recognition rates e.g. by applying acoustic adaptation techniques. In fact in this paper we concentrated on the general proof-of-concept of the proposed approach and did not focus on more sophisticated speech recognition issues.

6. ACKNOWLEDGMENTS

We gratefully acknowledge the valuable support of Marius Hennecke in developing the various software modules making up our environment for experimental evaluations.

7. REFERENCES

[1] Thomas Plötz, "The FINCA: A Flexible, Intelligent eNvironment with Computational Augmentation," www.finca.irf.de, 2007.

[2] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments," *Signal Processing*, vol. 86, no. 6, pp. 1260–1277, 2006.

[3] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, Englewood Cliffs, New Jersey, 2001.

[4] Carlos Avendano, Sangita Tibrewala, and Hynek Hermansky, "Multiresolution channel normalization for ASR in reverberant environments," in *Proc. European Conf. on Speech Communication and Technology*, Rhodes, Greece, 1997, pp. 1107–1110.

[5] David Gelbart and Nelson Morgan, "Evaluating long-term spectral subtraction for reverberant ASR," in *Proc. Workshop on Automatic Speech Recognition and Understanding*, Madonna di Campiglio, Italy, 2001, pp. 103–106.

[6] David Gelbart and Nelson Morgan, "Double the trouble: Handling noise and reverberation in far-field automatic speech recognition," in *Proc. Int. Conf. on Spoken Language Processing*, Denver, 2002, pp. 2185–2188.

[7] Longbiao Wang et al., "Robust distant speech recognition based on position dependent CMN," in *Proc. Int. Conf. on Spoken Language Processing*, Jeju, Korea, 2004, pp. 2409–2052.

[8] Gernot A. Fink and Thomas Plötz, "ESMERALDA: A development environment for HMM-based pattern recognition systems," http://sourceforge.net/projects/esmeralda, 2007.