Assessment of Dereverberation Algorithms for Large Vocabulary Speech Recognition Systems

Koen Eneman, Jacques Duchateau, Marc Moonen, Dirk Van Compernolle, Hugo Van hamme

Katholieke Universiteit Leuven - ESAT
Kasteelpark Arenberg 10
B-3001 Heverlee, Belgium

E-mail: {Koen.Eneman,Jacques.Duchateau}@esat.kuleuven.ac.be

Abstract

The performance of large vocabulary recognition systems, for instance in a dictation application, typically deteriorates severely when used in a reverberant environment. This can be partially avoided by adding a dereverberation algorithm as a speech signal preprocessing step. The purpose of this paper is to compare the effect of different speech dereverberation algorithms on the performance of a recognition system. Experiments were conducted on the Wall Street Journal dictation benchmark. Reverberation was added to the clean acoustic data in the benchmark both by simulation and by re-recording the data in a reverberant room. Moreover additive noise was added to investigate its effect on the dereverberation algorithms. We found that dereverberation based on a delay-and-sum beamforming algorithm has the best performance of the investigated algorithms.

1. Introduction

Automatic speech recognition systems are typically trained under more or less anechoic conditions. Recognition rates therefore drop considerably when signals are applied that are recorded in a moderately or strongly reverberant environment. In the literature, several solutions to this problem are proposed, e.g. in [1, 2, 3, 4]. We can distinguish two types of solutions: (1) a dereverberation algorithm is applied as a speech signal preprocessing step and the recognizer itself is considered as a fixed, black box and (2) robustness is added to the recognizer's feature extraction and (acoustic) modeling. The latter is typically more difficult as it requires access to the core of the recognizer and/or to the necessary training databases.

In this paper, we compare several solutions of the first type in various environmental conditions (amount of reverberation and noise, real recordings). This kind of comparison is rarely found in the literature. An example is [4], but in this paper a poor baseline is used (59% accuracy on clean data for a dictation task), and the behavior of the algorithms is only evaluated on simulated additional reverberation.

The outline of the paper is as follows. In section 2, the investigated dereverberation algorithms are briefly described. The large vocabulary recognizer used in the experiments, and the recognition task are proposed in section 3. Next in section 4, the experiments are described and the results are given and discussed. Finally some conclusions are given in section 5.

2. Dereverberation algorithms

This section gives an overview of the investigated dereverberation algorithms. A general $\mathcal{M}$-channel speech dereverberation system is shown in figure 1. An unknown signal s is filtered by unknown acoustic impulse responses h_1 ... $h_{\mathcal{M}}$, resulting in $\mathcal{M}$ microphone signals y_1 ... $y_{\mathcal{M}}$. Dereverberation deals with finding the appropriate compensator such that the output $\hat{s}$ is as close as possible to the unknown signal s.

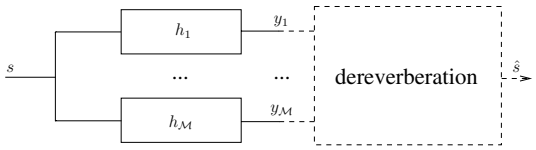

Figure 1: Setup for multi-channel dereverberation

More specifically, the following 4 dereverberation algorithms were compared.

2.1. Delay-and-sum beamforming

Beamforming algorithms [5, 6] exploit the spatial diversity that is present in the different microphone channels. By appropriately filtering and combining the microphone signals spatially dependent amplification can be obtained. In this way the algorithm is able to *zoom in* on the desired signal source and will suppress undesired background disturbances. Although in the first place, beamforming algorithms are used for noise suppression they can be applied to the dereverberation problem as well. As the beamformer focuses on the signal source of interest, only those acoustic waves are amplified that impinge on the array from the same direction as the direct path signal. Waves coming from other directions are suppressed. In this way the amount of reverberation is reduced.

A basic, but nevertheless very popular beamforming scheme is the delay-and-sum beamformer. In this technique the different microphone signals are appropriately delayed and summed together. Referring to figure 1 the output of the delay-and-sum beamformer is given by

$$\hat{s}[k] = \sum_{m=1}^{\mathcal{M}} y_m[k - \delta_m]. \quad (1)$$

For our experiments, we chose $\delta_m = 0$ as the desired signal source was located in front of the (linear) microphone array in the broadside direction (making an angle of 90° with the array).

2.2. Cepstrum based dereverberation

Cepstrum-based dereverberation techniques are another well-known standard for speech dereverberation and rely on the separability of speech and the acoustics in the cepstral domain. The

algorithm that was used in our experiments is based on [7]. It factors the microphone signals into a minimum-phase and an all-pass component. It appears that the minimum-phase component is less affected by the reverberation than the all-pass component. Hence, the minimum-phase cepstra of the different microphone signals are averaged and the resulting minimum-phase component is further enhanced with a low-pass *lifter*. On the all-pass component a spatial filtering or beamforming operation is performed. The beamformer reduces the effect of the reverberation, which acts as uncorrelated additive noise on the all-pass components of the different microphone signals.

2.3. Matched filtering

Another standard procedure for noise suppression and dereverberation is matched filtering. On the assumption that the transmission paths h_m are known (see figure 1), an enhanced system output can be obtained as

$$\hat{s}[k] = \sum_{m=1}^{\mathcal{M}} h_m[-k] \star y_m[k]. \quad (2)$$

In order to reduce complexity the reverse filter $h_m[-k]$ is truncated and the l_e most significant (i.e. last l_e) coefficients of $h_m[-k]$ are retained to obtain e_m such that

$$\hat{s}[k] = \sum_{m=1}^{\mathcal{M}} e_m[k] \star y_m[k]. \quad (3)$$

A disadvantage of this technique is that the transmission paths h_m need to be known in advance. However it is known that matched filtering techniques are quite robust against wrong transmission path estimates. During our research we provided the true impulse responses h_m to the algorithm as an extra input. In the case of experiments with real-life data the impulse responses were estimated with an NLMS adaptive filter based on white noise data.

2.4. Matched filtering subspace dereverberation in the frequency domain

We used a matched filtering-based dereverberation algorithm that relies on 1-dimensional frequency-domain subspace estimation (see section IIc of [8]). An LMS type updating algorithm for this approach was also proposed in this paper.

A key assumption in the derivation of the algorithm in [8] is that the norm of the transfer function matrix $\boldsymbol{\beta}(f) = \|H_1(f)...H_{\mathcal{M}}(f)\|$ (with $H_m(f)$ the frequency-domain representation of $h_m[k]$, see figure 1) needs to be known in advance, which is the weakness of this approach. We can get around this by measuring parameter $\boldsymbol{\beta}$ beforehand. This is however unpractical, hence an alternative is to fix $\boldsymbol{\beta}$ to an environment-independent constant, e.g. $\boldsymbol{\beta} = \mathbf{1}$.

3. Recognizer and database

3.1. Recognition system

For the recognition experiments, the speaker-independent large vocabulary continuous speech recognition system was used that has been developed at the ESAT-PSI speech group of the K.U.Leuven. A detailed overview of this system can be found in [9, 10] (concerning the acoustic modeling) and in [11, 12] (mainly concerning the search engine).

In the recognizer, the acoustic features are extracted from the speech signal as follows. Every 10 ms a power spectrum is calculated on a 30 ms window of the pre-emphasized 16 kHz data. Next, a non-linear mel-scaled triangular filterbank is applied and the resulting mel spectrum with 24 coefficients is transformed into the log domain. Then these coefficients are mean normalized (subtracting the average) in order to add robustness against differences in the recording channel. Next, the first and second order time derivatives of the 24 coefficients are added, resulting in a feature vector with 72 features. Finally, the dimension of this feature vector is reduced to 39 using the MIDA algorithm (an improved LDA algorithm [13]) and these features are decorrelated (see [14]) to fit to the diagonal covariance Gaussian distributions used in the acoustic modeling.

The acoustic modeling, estimated on the SI-284 (WSJ1) training data with 69 hours of clean speech (Sennheiser close-talking microphone), is gender independent and based on a phone set with 45 phones, without specific function word modeling. A global phonetic decision tree defines the 6559 tied states in the cross-word context-dependent and position-dependent models. Each state is modeled as a mixture of tied Gaussian distributions, the total number of Gaussians being 65417. The benchmark trigram language model was estimated on 38.9 million words of WSJ text. With this recognition system, a word error rate (WER) of 1.9% was found on the benchmark test set described below with real time recognition on a 2.0 GHz Pentium 4 processor.

It is important to note that in this baseline recognition system, no specific robustness for (additive) noise or for reverberation is integrated, nor in the feature extraction nor in the acoustic modeling. So if robustness for noise or reverberation is observed in the experiments, it is the result of the additional signal preprocessing step based on the dereverberation algorithm.

3.2. Data set

We evaluated the effect of the different dereverberation algorithms on the recognizer's performance using the well-known speaker-independent Wall Street Journal (WSJ) benchmark recognition task with a 5k word closed (so without out-of-vocabulary words) vocabulary.

Results are given on the November 92 evaluation test set with non-verbalized punctuation. This set consists of 330 sentences, amounting to about 33 minutes of speech, uttered by eight different speakers (which are not in the trainset), both male and female. It is recorded at 16 kHz and contains almost no additive noise, nor reverberation. In the experiments, different levels of reverberation and additive noise will be obtained or by simulation, or by playing back the clean audio and making new recordings with a microphone array.

4. Experiments

This section describes the experiments and gives and discusses the results. The effect of several environmental variables were investigated in separate experiments: the reverberation time, the number of microphones, the amount of additive noise, and the setup in real-life recordings. The reference experiment has a reverberation time of 547 ms (for a microphone distance of 94 cm and a room of 36 m^3), a setup with 6 equidistant microphones, and uses data without additive noise. This setup with a 19.7% WER when no dereverberation algorithm is applied, was chosen to produce possibly *significant* experimental results.

4.1. Reverberation time

First, the effect of the reverberation time on the recognition performance was measured. The reverberation time T_{60} is defined

as the time that the sound pressure level needs to decay to -60 dB of its original value. Typical reverberation times are in the order of hundreds or even thousands of milliseconds. For a typical office room T_{60} is between 100 and 400 ms, for a church T_{60} can be several seconds long.

For the simulation, the recording room is assumed to be rectangular and empty, with all walls having the same reflection coefficient. The reverberation time can then be computed from the reflection coefficient ρ and the room geometry using Eyring's formula [15] :

$$T_{60} = \frac{0.163V}{-S \log \rho}, \quad (4)$$

where S is the total surface of the room and V is the volume of the room.

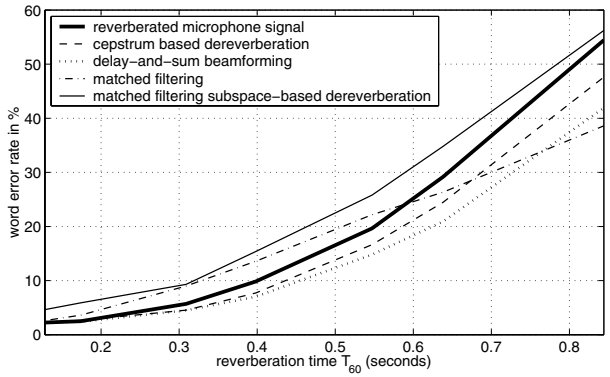

Figure 2: Performance (WER) vs. reverberation time

The results are given in figure 2. As could be expected, the WER increases drastically for a higher reverberation time. The matched filtering algorithms seem to deteriorate the WER, at least for relatively small reverberation times corresponding to an office room. On the other hand the algorithms based on the cepstrum and on delay-and-sum beamforming improve the result for any reverberation time. Delay-and-sum beamforming is the best, a relative improvement of about 25% is found.

4.2. Number of microphones

The microphones are placed on a linear array at a distance of 5 cm of each other. The number of microphones has been lowered from the reference 6 to 2 to detect performance losses.

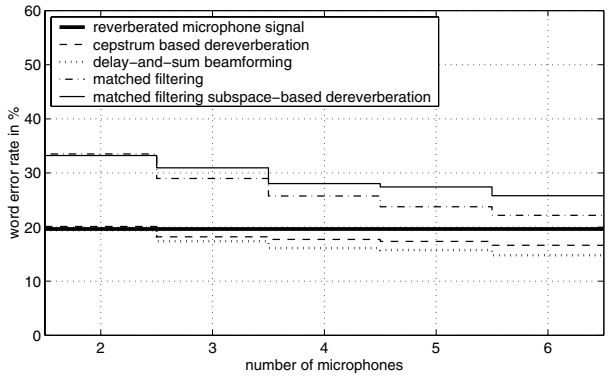

Figure 3: Performance (WER) vs. number of microphones

The results are given in figure 3. It can be observed that if the number of microphones is increased the performance of the algorithms improves gradually. This performance improvement is probably due to the higher number of degrees of freedom and to the increased spatial sampling that is obtained when more microphones are involved.

4.3. Additive noise

In these experiments noise has been added to the multi-channel speech recordings at different (non frequency weighted) signal-to-noise ratios (SNR). The source for spatially correlated noise (simulated or real-life as in section 4.4) makes an angle of about 45° with the microphone array.

In figures 4, 5, and 6, the results are given for 3 types of noise: uncorrelated white noise, spatially correlated white noise, and spatially correlated speech-like noise respectively. As a reference, we also investigated the clean signals with the additive noise but without reverberation.

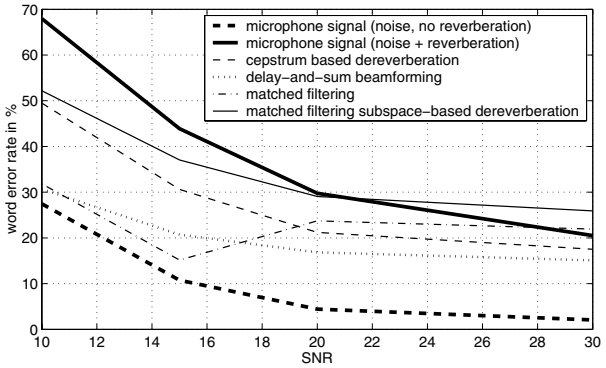

Figure 4: WER vs. SNR for uncorrelated white noise

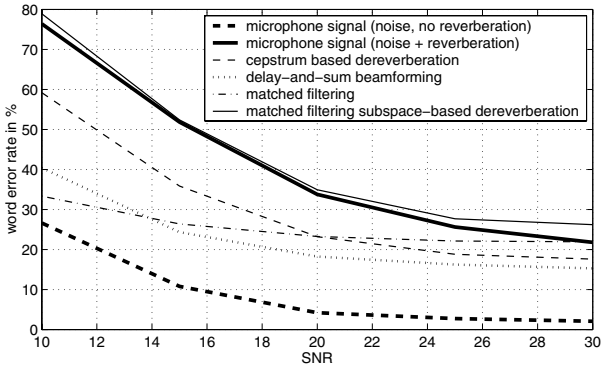

Figure 5: WER vs. SNR for spatially correlated white noise

In general we can see that the recognition system (in which, as said, no additive noise robustness is incorporated) is more robust to speech-like noise than to white noise. Moreover compared to reverberation, additive noise has a smaller negative impact on the performance of the recognizer, for instance in an office environment. We can furthermore conclude that spatially correlated (white) noise has a worse effect on the recognizer than uncorrelated noise. Comparing the algorithms, the delay-and-sum beamformer again seems to outperform the other methods. Note that if higher relative improvements are obtained for low SNR, this may be due to the fact that the differ-

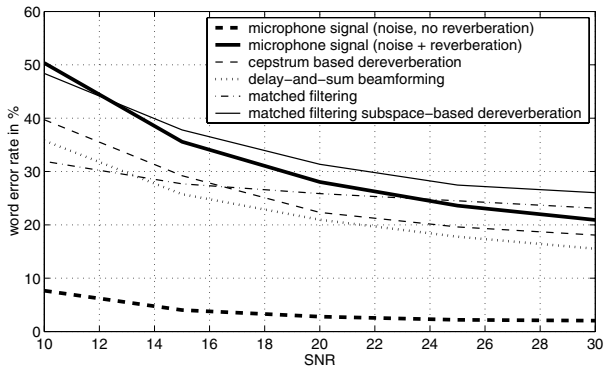

Figure 6: WER vs. SNR for spatially correlated speech-like noise

ent algorithms also incorporate noise reduction abilities (rather than dereverberation capabilities).

4.4. Real-life experiments

For the real-life experiments, recordings were made in the (69 m^3 large) ESAT speech lab, using different room acoustics. The audio was sent through a loudspeaker and recorded with a 6 microphone array. Only in the last (fourth) experiment, there was an extra loudspeaker with spatially correlated speech-like noise, resulting in a 8dB SNR.

Exp. number	exp 1	exp 2	exp 3	exp 4
Mic. distance (m)	1.9	1.9	1.3	1.3
T_{60}	0.12	0.28	0.24	0.29
reverberated signal	6.4%	16.8%	14.1%	50.0%
cepstrum based	6.0%	14.0%	13.6%	42.4%
delay-and-sum	6.2%	15.3%	14.6%	37.0%
matched filtering	/	/	24.9%	44.7%
subspace-based	10.0%	25.4%	21.4%	56.6%

Table 1: Performance (WER) on real-life recordings

The results are given in table 1. We can see from the table that in real-life situations, improvements can only be found for the cepstrum based algorithm and for the delay-and-sum beamformer. Unfortunately, the improvements are also smaller than for simulated data: up to 25% (relative) for experiment 4 with additive noise, and between 5% and 15% for experiments without additive noise.

5. Conclusions and further research

In general, we can conclude that applying dereverberation algorithms in the preprocessing of a recognizer can partly cancel the deterioration due to reverberation. From the investigated algorithms, a simple one (algorithmically) performed the best in most cases: the delay-and-sum beamformer.

In the future, the situation with both reverberation and additive noise should be investigated further by (1) adding algorithms for noise removal (in the preprocessing) or for noise robustness (in the recognizer) and by (2) checking the complementarity of these methods with the dereverberation algorithms evaluated in this paper.

6. Acknowledgments

This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven, in the frame of the Interuniversity Poles of Attraction Programme P5/22 and P5/11, the Concerted Research Action GOA-MEFISTO-666 of the Flemish Government, IWT project 000401: MUSETTE-II and was partially sponsored by Philips-PDSL. The scientific responsibility is assumed by its authors.

7. References

[1] D. Van Compernolle, W. Ma, F. Xie, and M. Van Diest, "Speech recognition in noisy environments with the aid of microphone arrays," *Speech Communication*, vol. 9, no. 5-6, pp. 433–442, December 1990.

[2] D. Giuliani, M. Omologo, and P. Svaizer, "Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation," in *Proc. International Conference on Spoken Language Processing*, vol. III, Philadelphia, U.S.A., October 1996, pp. 1329–1332.

[3] L. Couvreur, C. Couvreur, and C. Ris, "A corpus-based approach for robust ASR in reverberant environments," in *Proc. International Conference on Spoken Language Processing*, vol. I, Beijing, China, October 2000, pp. 397–400.

[4] B. Gillespie and L. Atlas, "Acoustic diversity for improved speech recognition in reverberant environments," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, vol. I, Orlando, U.S.A., May 2002, pp. 557–560.

[5] D. Van Compernolle and S. Van Gerven, "Beamforming with microphone arrays," in *COST 229 : Applications of Digital Signal Processing to Telecommunications*, V. Cappellini and A. Figueiras-Vidal, Eds., 1995, pp. 107–131.

[6] B. Van Veen and K. Buckley, "Beamforming : A versatile approach to spatial filtering," *IEEE Magazine on Acoustics, Speech and Signal Processing*, vol. 36, no. 7, pp. 953–964, July 1988.

[7] Q.-G. Liu, B. Champagne, and P. Kabal, "A microphone array processing technique for speech enhancement in a reverberant space," *Speech Communication*, vol. 18, no. 4, pp. 317–334, June 1996.

[8] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 425–437, September 1997.

[9] J. Duchateau, "Hmm based acoustic modelling in large vocabulary speech recognition," Ph.D. dissertation, K.U.Leuven, ESAT, November 1998, available from http://www.esat.kuleuven.ac.be/~spch.

[10] J. Duchateau, K. Demuynck, and D. Van Compernolle, "Fast and accurate acoustic modelling with semi-continuous HMMs," *Speech Communication*, vol. 24, no. 1, pp. 5–17, April 1998.

[11] K. Demuynck, "Extracting, modelling and combining information in speech recognition," Ph.D. dissertation, K.U.Leuven, ESAT, February 2001, available from http://www.esat.kuleuven.ac.be/~spch.

[12] K. Demuynck, J. Duchateau, D. Van Compernolle, and P. Wambacq, "An efficient search space representation for large vocabulary continuous speech recognition," *Speech Communication*, vol. 30, no. 1, pp. 37–53, January 2000.

[13] J. Duchateau, K. Demuynck, D. Van Compernolle, and P. Wambacq, "Class definition in discriminant feature analysis," in *Proc. European Conference on Speech Communication and Technology*, vol. III, Aalborg, Denmark, September 2001, pp. 1621–1624.

[14] K. Demuynck, J. Duchateau, D. Van Compernolle, and P. Wambacq, "Improved feature decorrelation for HMM-based speech recognition," in *Proc. International Conference on Spoken Language Processing*, vol. VII, Sydney, Australia, December 1998, pp. 2907–2910.

[15] H. Kuttruff, *Room Acoustics*, 2nd ed. Ripple Road, Barking, Essex, England: Applied Science Publishers LTD, 1979.