Semi-Automatic Calibration for Dereverberation by Spectral Subtraction for Continuous Speech Recognition

Korbinian Riedhammer[1], *Tobias Bocklet*[1], *Juan Rafael Orozco-Arroyave*[1,2], *Elmar Nöth*[1*]

[1] Pattern Recognition Lab, University of Erlangen-Nuremberg, Martensstr. 3, 91058 Erlangen, Germany
[2] GITA Research Group, Universidad de Antioquia, Medellín, calle 67 # 53-108, Colombia
*Corresponding author: noeth@cs.fau.de
Web: www5.cs.fau.de

Abstract

In this article, we describe a semi-automatic calibration algorithm for dereverberation by spectral subtraction. We verify the method by a comparison to a manual calibration derived from measured room impulse responses (RIR). We conduct extensive experiments to understand the effect of all involved parameters and to verify values suggested in the literature. The experiments are performed on a text read by 31 speakers and recorded by a headset and three far-field microphones. Results are measured in terms of automatic speech recognition (ASR) performance using a 1-gram model to emphasize acoustic recognition performance. To accommodate for the acoustic change by dereverberation we apply supervised MAP adaptation to the hidden Markov model output probabilities. The combination of dereverberation and adaptation yields a relative improvement of about 35% in terms of word error rate (WER) compared to the original signal.

1 Introduction

Current state-of-the-art automatic speech recognition (ASR) systems work with remarkable accuracy. However, this usually only holds if close talking microphones like headsets are used as audio source. The use of far-field microphones, e.g., microphones mounted to a table, wall or ceiling, severely degrades the recognition performance as the observed acoustic signal is significantly different to the one captured by close talking microphones. Beside the so-called direct sound, i.e., the speech signal uttered and directly received at the microphone, far-field microphones also capture ambient noise and reverberation effects caused by reflections of the acoustic signal from walls and objects.

Though human speech perception is almost not affected by reverberation effects, already little reverberation within a small room, often not even noticed by humans, can cause ASR systems to fail. Given the common design of an ASR system, this can easily be explained. Most systems are based on spectral analysis and on the change of the spectrum over the time given a certain spoken word or utterance. The system is then trained on recorded speech data which is usually recordings by close talking microphones. In the presence of reverberation, the observed signal is significantly altered, thus spectral analysis and the resulting features computed on clean and reverberated speech signals lead to different results, depending on the actual room acoustics.

To ease the effects of reverberation, ASR systems can be modified at three stages. Top-down, dereverberation can be integrated into the acoustic decoding of the feature sequence as in [1], where the authors use spectral features in combination with hidden Markov models (HMM) for continuous digit recognition. Using HMMs trained on clean speech and a statistical model for reverberation, the feature sequence is dereverberated using recognition scores computed in the decoding process, leading to a significant improvement in terms of recognition performance.

In [2], the acoustic front-end processing is augmented in a way that a voiced/unvoiced detector is used to modify the observed spectrum according to that decision. For voiced segments, the spectrum is re-synthesized according to the estimated fundamental frequency, for unvoiced segments, the lower part of the spectrum is faded out.

Finally, one can try to remove reverberation effects before the feature extraction by filtering the audio data, leaving the ASR system as such unmodified. The idea of spectral subtraction has been around for quite some time, however, mainly used for the denoising of acoustic signals. In [3], the authors propose a simplified reverberation model that can be integrated with spectral subtraction. The idea is to split the acoustic process in two parts, *early* and *late* reverberation. Assuming the effects of the early reverberation to be non-critical to intelligibility as the direct sound is very prominent, the late reverberation is estimated and subtracted from the spectrum.

In this work, we describe in detail how to automatically estimate the decay coefficient from the acquired signal containing clapping events and verify the automatically determined parameters with parameters extracted from RIRs measured in the target room. The adaptation of the HMM output probabilities is performed here to match the resulting new acoustic condition without over-fitting to the spoken text. A systematic parameter exploration and optimization is performed here in terms of ASR for an ambient living assistant scenario and evaluate on real data recorded as part of a large scale usability study. The acoustic conditions like room reverberation time T_{60}, speaker microphone distance (SMD) and speaking direction are both unknown.

This article is structured as follows. After a detailed description of the data in Section 2, the dereverberation algorithm following [3] and its required parameters are briefly introduced in Section 3. Section 4 explains the automatic calibration procedure used to determine the required decay coefficient. Section 5 and 6 describe the speech recognition system and analyze the experiments using different dereverberation parameters and adaptation techniques. Section 7 presents the conclusions of this work.

2 Data

The ASR system is trained on a subset of the German VERBMOBIL [4] corpus (11,714 utterances, 257,810 words, about 25 hours); the speakers were aged around 27 ± 8 years. For the evaluation of the dereverberation algorithm, a subset of the FAU IISAH Corpus [5] was used. 31 speakers (19m, 12f) aged 61 to 78 (65 ± 5) read the German version of "North Wind and the Sun", a phonetically rich fable from Aesop, resulting in about 22 minutes of speech. The data was synchronously acquired using a head mounted SHURE WH20XLR and three T.BONE GZ400 far-field microphones mounted to the wall behind, opposite and to the far right of the speakers in approx. 1.5 m height. Further details of the data used in this paper can be found in [6]

3 Dereverberation Algorithm

The algorithm is based on modeling the RIR $h(t)$ as an exponentially decaying Gaussian white noise $b(t)$ [3]

$$h(t) = b(t)e^{-\rho t}, t \geq 0 \tag{1}$$

where ρ is the decay coefficient linked to the room reverberation time as

$$\rho = \frac{3 \ln 10}{T_{60}} \quad . \tag{2}$$

ISBN 978-3-8007-3640-9

The reverberated signal $x(t)$ can now be obtained by a convolution of the ideal anechoic signal $s(t)$ with the RIR $h(t)$ as

$$x(t) = e^{-\rho t} \int_{-\infty}^{t} s(\nu) b(t-\nu) e^{\rho \nu} \mathrm{d}\nu \quad . \tag{3}$$

In [3], Lebart *et al.*split the reverberation process in an early part containing mainly the direct signal and a few reflections, and a late part containing mainly the reverberation by introducing a T_{mix} as transition point. This late part is estimated and subtracted from the spectrum, leading to an approximation of the original signal. The subtraction is in practice realized by a short-term spectral attenuation as

$$\hat{S}(m,k) = G(m,k) X(m,k) \tag{4}$$

where $\hat{S}$ is the estimated amplitude spectrum of the dereverberated signal using m as time index and k as frequency index, X the spectrum of the reverberant signal, and G is the gain defined as

$$G(m,k) = \begin{cases} 1 - \frac{1}{\sqrt{\text{SNR}_{\text{pri}}(m,k)}} & \text{if } \hat{S}(m,k) \geq \lambda \sqrt{\hat{\gamma}_{\text{rr}}(m,k)} \\ \frac{\lambda \sqrt{\hat{\gamma}_{\text{rr}}(m,k)}}{|X(m,k)|} & \text{otherwise} \end{cases} \tag{5}$$

where λ is a weighting factor controlling the spectral floor, ensuring that there are no negative estimates in $\hat{S}(m,k)$. The reverberation power spectral density (PSD) $\hat{\gamma}_{\text{rr}}$ can be estimated from the PSD of the past reverberated signal as

$$\hat{\gamma}_{\text{rr}}(m,k) = e^{-2\rho T_{\text{mix}}} \hat{\gamma}_{\text{xx}}(m - T_{\text{mix}}, k) \quad . \tag{6}$$

Both SNR_{pri} and $\hat{\gamma}_{\text{xx}}$ are estimated as a running average using the update weight β in

$$\begin{aligned} \text{SNR}_{\text{pri}}(m,k) \quad = \quad & \beta \ \text{SNR}_{\text{pri}}(m-1,k) \\ & + (1-\beta) \max[0, \frac{X(m,k)^2}{\hat{\gamma}_{\text{rr}}} - 1] \end{aligned} \tag{7}$$

and

$$\hat{\gamma}_{\text{xx}}(m,k) = \beta \, \hat{\gamma}_{\text{xx}}(m-1,k) + (1-\beta) X(m,k)^2 \quad . \tag{8}$$

In summary, there are 5 parameters that need to be set. ρ is linked to T_{60} and describes the room acoustics. T_{mix} is the offset distinguishing between the early and late reverberation, where the latter part is estimated. λ controls the spectral floor by controlling the comparison between $\hat{S}$ and $\sqrt{\hat{\gamma}_{\text{rr}}}$. β effects the running average computations (a β close to one results in little variance but long averaging duration). Last, the decimator controls how fine-grained the steps of m are. In [3], the authors recommend to set $\lambda = 0.1$ and $T_{\text{mix}} = 0.050$ s but give no hints on the other parameters involved.

4 Decay Parameter Estimation

The decay coefficient is directly linked to the target room acoustics. However, the best way to estimate ρ is to extract it from at least one measured RIR, as it requires a careful setup and defined hardware.

The original calibration that was proposed in [3] alongside the dereverberation algorithm was designed to work online using continuous speech input, constantly measuring the decay and accepting a measure if the decay duration exceeded a certain length. However the description is rather vague about the implementation details, plus, the number of estimates and the resulting average accuracy seem to be hard to control.

Furthermore, a static estimate of the dereverberation parameters is desired as the ASR system will be adapted once to the resulting signal quality. For an always-on system like an ambient living assistance system, a short initialization phase to determine the parameters seems also more reasonable than a continuous estimate, as the amount of speech input will be rather small compared to silence and other ambient noises.

If one plans to apply the algorithm in off-the-shelf consumer products, the required parameter estimation needs to be fast and easy. Traditionally, RIR measures are done using some sort of clicking devices that can produce a somewhat unique impulse in the signal. However, as the decay coefficient seems not very sensitive to different RIR instances, anything producing a reasonable sharp impulse, like clapping hands, should be sufficient.

We propose a somewhat semi-automatic approach. After initializing the system, the user moves freely in the target room and claps his hands for several times with short times of silence in between until notified by the system. In contrast to traditional procedures, the clapping events are segmented automatically and the decay coefficient is extracted. To consider false or bad estimates, the median value is used to finalize the system setup. Once a certain amount of estimates has been acquired, the system notifies the user that the setup is complete. In detail, the algorithm can be described in a six steps:

1. Initialize the system.
2. Identify the time t_c of the next clapping event by determining the next strong peak in the raw input signal x. x is acquired at the microphone, has no unit and is normalized to $[-1,1]$. To avoid the ill-posed problem of finding the derivative within the signal, we compute a regularized derivation using a Gaussian kernel $g(t)$ in

$$\begin{aligned} x'(t) \quad & \approx \textstyle\int_u x'(t-u) g(u) \mathrm{d}u \\ & = \textstyle\int_u x(u) g'(t-u) \mathrm{d}u \end{aligned} \tag{9}$$

   which is valid for small $\sigma > 0$ where σ is the standard deviation for the Gaussian kernel. Thus, the problem of finding a relative maximum in x' can be reduced to finding a relative maximum in

$$\begin{aligned} \bar{x}(t) \quad & = \alpha \textstyle\int_u x(u) g'(t-u) \mathrm{d}u \\ & = \textstyle\int_u x(u) \alpha g'(t-u) \mathrm{d}u \end{aligned} \tag{10}$$

   using suitable $\sigma > 0$. We set the normalization constant $\alpha = \frac{\sigma\sqrt{\pi}}{\sqrt{2}}$ to normalize the derived Gaussian kernel.
3. Accept candidate maximum at t_c if it exceeds a certain threshold θ, i.e., $\bar{x}(t_c) > \theta$, otherwise repeat (2).
4. Fit a line $y \mapsto 2\rho t + 2b$ to $\log(x)$ on the interval $[t_c + T_{\text{mix}}, t_c + T_{\text{mix}} + T_{\text{max}}]$ where T_{max} is the length of the time interval to consider for the least square fit and T_{mix} is the delay after the observed impulse.
5. Save the estimated ρ, repeat from (2) until enough estimates.
6. Use the median of the saved estimates to initialize the dereverberation algorithm; notify user.

For the experiments in this article, we set $\sigma = 0.008$, $\theta = 0.5$ and $T_{\text{max}} = 300$ ms.

5 Recognition System and Adaptation

The ASR system used for this work is based on semi-continuous HMMs sharing 500 Gaussian densities with full covariance matrices. The acoustic models are polyphones, i.e., phones with variable sized context. The first 12 mel-frequency cepstral coefficients and their first order derivatives were used as features. For both training and experiments, cepstral mean subtraction was applied. After training the system on the VERBMOBIL data, the vocabulary was replaced by the words of the German version of "North Wind and the Sun" (108 words, 71 disjoint) and supplemented with a few common reading errors. For the latter decoding, we trained 1-gram and 2-gram models, however we will focus on 1-gram recognition results to emphasize the performance of the acoustic decoding.

For acoustic adaptation, the speakers were split into two groups. Supervised adaptation of the HMM output probabilities using MAP [7] was performed on one group and the resulting models

	original		*best derev*	
mic	1-gram	2-gram	1-gram	2-gram
CT-0	18.0	3.6	—	
CT	21.7	6.2	—	
R1	73.4	53.8	53.0	23.4
R2	78.9	64.3	59.8	35.4
R3	77.1	62.3	54.9	24.8

Table 1: Baseline recognition results (WER) for the control (CT-0) and target group (CT) close talking recordings and the far-field microphones R1-3 using 1-gram and 2-gram language models.

#	R1	R2	R3
Measure 1	12.40	12.29	12.65
Measure 2	11.97	12.51	12.41
Median Est.	12.27	10.97	12.23

Table 2: RIR measured from different microphone positions [6]

were tested on the other group and vice versa. The effect of adapting to a certain word sequence may still be present but is rather unlikely as all HMMs share the same Gaussian densities and the transition probabilities and mixture weights are not adapted.

6 Experiments

6.1 Baseline

Table 1 shows the results for a baseline experiment applying the ASR system to the data set of the control group CT-0 that roughly matches the training speaker age. The rather high word error rate (WER) of 18.0% using the 1-gram language model and the more reasonable WER of 3.6% using the 2-gram language model confirms the well-known fact that ASR performance strongly depends on a proper language model. For this article, we use a 1-gram language model, to see how much the dereverberation actually contributed to the recognition process. In the following, WER results refer to the use of the 1-gram language model if not otherwise stated.

The results on the CT data shows that due to the strong age mismatch of training and test speakers, the WER drops about 4 percentage points to 21.7%. This similarly holds using the 2-gram language model, however the language model can compensate for some losses.

Applying the ASR system to the original far-field microphone data R1-3 yields WER above 70%. Comparing the recognition scores with the location of the microphones confirms the intuition that a microphone placed opposite of the speaker produces best results while placed right behind him is probably the worst location.

6.2 Estimated vs. Measured RIR

Before dereverberating the acoustic signal, we analyze the performance of the automatic calibration, that is the automatic estimation of the decay parameter ρ. Two RIR were measured at the two speaker positions using the MLS technique [8]. For the automatic estimation, the notification step was skipped but a test person his clapped hands for 20 times, resulting in up to 18 estimates. Beside for R2 which results in many obviously wrong estimates, the extracted median is close to the values extracted from the measured RIR:

6.3 Parameter Exploration

Motivated by the observation that even "ideal" decay coefficients vary and to verify the heuristically defined remaining parameters, we conduct an extensive parameter sampling. Starting from the

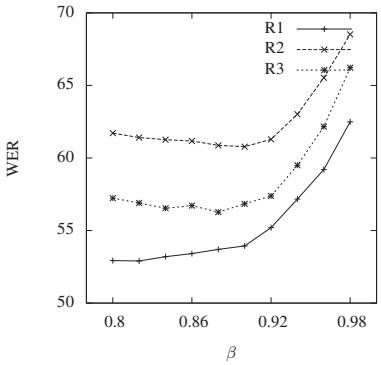

Figure 1: Recognition rates (WER) on dereverberated audio data for variable β using 1-gram language models.

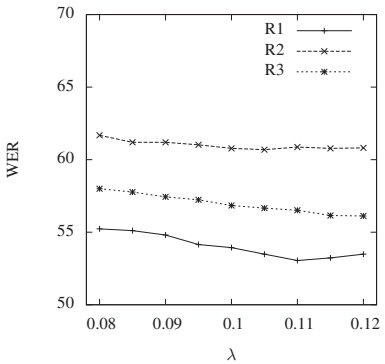

Figure 2: Recognition rates (WER) on dereverberated audio data for variable λ using 1-gram language models.

fixed parameter set $\beta = 0.9$, $\rho = 12.5$, $\lambda = 0.1$, $T_{\text{mix}} = 0.05$ and a decimator (step size Δt) of 64 (resp. 4ms at a sampling rate of 16 kHz), for each parameter, four will be fixed and one sampled to see its effect on the recognition performance.

The update parameter β controls the variance and accuracy of the running averages. A higher value leads to a higher contribution of the following samples. In terms of WER, a value around $\beta = 0.88$ yields best results (cf. Figure 1). Any larger value than $\beta = 0.9$ severely degrades the performance. Surprisingly, the variation of the most intuitive and room dependent parameter ρ modeling the exponential decay of the RIR has only little effect on the recognition performance. However, WER scores converge to an optimum in the area of the expected real decay coefficient.

The parameter λ controls the spectral floor. A lower value leads to less subtraction and a higher value enforces the reduction of reverberation. As illustrated in Figure 2, WER scores converge for values around $\lambda = 0.11$.

The T_{mix} parameter directly controls the dereverberation process in terms of reverberation estimation and thus when potential reverberation is removed. In general, a higher value, i.e., postponing the time where the late reverberation is supposed to start, seems to help (cf. Figure 3). However, the higher the value is, the less reverberation will be captured.

The decimator (or step size Δt) controls how fine-grained the dereverberation process is. While a larger value increases processing speed, it also noticeably affects the recognition performance as shown in Figure 4. For 16 kHz, a decimator value around 40 (or 2.5ms) yields the best WER. As ρ and λ intuitively control the dereverberation process, we fully explored both parameters at the same time to see if they are linked some way. In general, using low values for both parameters yields worst WER, and choosing more appropriate values consistently yields better WER. Although the above parameter sampling did not reveal any new insights on how to choose parameters at the first glance, there

ISBN 978-3-8007-3640-9

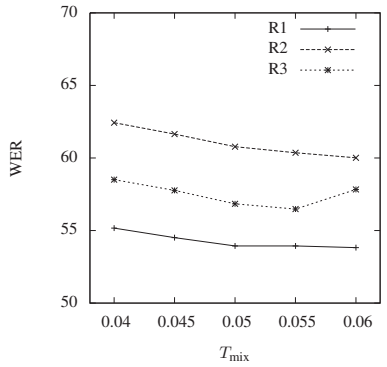

Figure 3: Recognition rates (WER) on dereverberated audio data for variable T_{mix} using 1-gram language models.

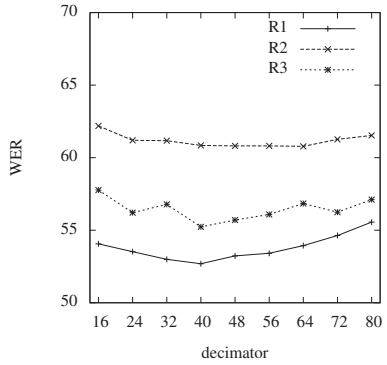

Figure 4: Recognition rates (WER) on dereverberated audio data for variable decimator values using 1-gram language models.

is one interesting fact. Given the numbers, the parameters with the most effect on WER are β, the decimator Δt and T_{mix}. That means on the other hand, if one can come up with any estimates for ρ, λ and possibly T_{mix}, they all will do the job as long as they are in the region of the ideal parameters.

The right side of Table 1 shows the recognition performance of using the best parameters found in the experiments above. For both 1-gram and 2-gram language models, WER could be reduced by an absolute value of around 20% and 30% respectively. However, error rates above 50% are still not satisfactory.

6.4 Acoustic Adaptation

Dereverberated signals acquired by a far-field microphone sound different than close-talking recordings altering the spectrum and the extracted features. To accommodate that acoustic change, adaptation of the acoustic models is the next step to do.

Table 3 shows the results of the different adaptation scenarios. First, the system is adapted to elderly close-talking speech using the CT data. The WER is improved by an absolute value of 1.5 percentage points for the CT data and by about 5 percentage points when testing on the reverberated R1-3 data. In a second step, the system is adapted to reverberated data to check if sole adaptation does the job and dereverberation is not required. Note that this is done for each microphone separately. This closely links to training on reverberated speech which has shown to improve recognition rates. However, that only works if the acoustic conditions match. Compared to the unadapted system, the WER for the R1-3 data could be improved by about 8 percentage points showing that acoustic codebook adaptation indeed helps. In a last step, the system is adapted to the data that is dereverberated with the best parameter combination. The WER is improved by about 5 percentage points for each microphone compared to the performance of the unaltered system. This confirms, that acoustic adaptation to dereverberated data yields the best results.

	adaptation data		
mic	CT	rev	derev
CT	20.2	—	
R1	67.2	62.0	47.6
R2	75.5	70.6	54.8
R3	72.2	66.9	49.8

Table 3: Recognition results (WER) using 1-gram language models and the close talking, reverberated (far-field acquired) and dereverberated audio data for acoustic model adaptation.

7 Conclusion

According to the experiments conducted in this paper the decay coefficient required for the dereverberation algorithm can be estimated in a fast, easy and robust way using the proposed algorithm. The dereverberation algorithm helps speech recognition performance, even if the parameters are not ideal. Unfortunately, parameters that cannot be directly extracted from the RIR have more effect on the performance than the parameter that can be estimated. Acoustic model adaptation to accommodate for the differences between the training signal acoustics and the dereverberated data is required and yields good results. Though the combination of dereverberation and adaptation leads to a relative improvement of roughly 35% for far-distant microphones, the current WER are still not satisfactory. The use of further technologies seems required.

8 Acknowledgment

Juan Rafael Orozco-Arroyave is under grants of COLCIENCIAS through the call Nº 528.

References

[1] A. Sehr and W. Kellermann, “Model-based dereverberation of speech in the mel-spectral domain,” in *Proc. Asilomar Conference on Signal, Systems, and Computers*, 2008.

[2] R. Petrick, K. Lohdeand, M. Lorenz, and R. Hoffmann, “A new feature analysis method for robust asr in reverberant environments based on the harmonic structure of speech,” in *Proc. European Signal Processing Conference (EUSIPCO)*, 2008.

[3] K. Lebart and J. M. Boucher, “A new method based on spectral subtraction for speech dereverberation,” *ACOUSTICA*, vol. 87, pp. 359–366, 2001.

[4] W. Wahlster, ed., *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin, Germany: Springer, 2000.

[5] W. Spiegl, K. Riedhammer, S. Steidl, and E. N"oth, “FAU IISAH Corpus – A German Speech Database Consisting of Human-Machine and Human-Human Interaction Acquired by Close-Talking and Far-Distance Microphones,” in *Proc. Language Resources and Evaluation (LREC)*, 2010.

[6] U. Zäh, K. Riedhammer, T. Bocklet, and E. Nöth, “Clap Your Hands! Calibrating Spectral Substraction for Dereverberation,” in *Proc. IEEE Int’l Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

[7] J. Gauvain and C. Lee, “Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[8] D. Rife and J. Vanderkooy, “Transfer-function measurement with maximum-length sequences,” *Journal of the Audio Engineering Society*, vol. 37, pp. 419–444, 1989.