CLAP YOUR HANDS! CALIBRATING SPECTRAL SUBTRACTION FOR DEREVERBERATION

Uwe Zäh, Korbinian Riedhammer, Tobias Bocklet, Elmar Nöth

Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg
Martensstraße 3, 91058 Erlangen, GERMANY
korbinian.riedhammer@informatik.uni-erlangen.de

ABSTRACT

Reverberation effects as observed by room microphones severely degrade the performance of automatic speech recognition systems. We investigate the use of dereverberation by spectral subtraction as proposed by Lebart and Boucher and introduce a simple approach to estimate the required decay parameter by clapping hands. Experiments on small vocabulary continuous speech recognition task on read speech show that using the calibrated dereverberation improves WER from 73.2 to 54.7 for the best microphone. In combination with system adaptation, the WER could be reduced to 28.2, which is only a 16% relative loss of performance comparison to using a headset instead of a room microphone.

Index Terms— speech recognition, robustness

1. INTRODUCTION

Current automatic speech recognition (ASR) systems work with impressive speed and accuracy when a close talking microphone, usually a headset, is used. However, this performance drops severely if the data is acquired using a far-field microphone, usually mounted on a table, wall or ceiling. Beside the so-called direct sound which should be about the same as received by a close talking microphone, the far-field microphone also captures ambient noise and, indoors, effects of reverberation, i.e., reflections of the acoustic signal from walls and objects mixed with the direct sound.

Although often not even perceived by humans, already little reverberation causes severe trouble for ASR systems as the observed signal is completely altered, as illustrated on the top and middle of Figure 1. In the transition from voiced to unvoiced, the energy observed in the lower spectrum is smeared into the following unvoiced segment. Vice versa, the energy observed in the upper spectrum is smeared into the voiced segment. As most ASR systems use some sort of spectral analysis, the problem is imminent: The features extracted from close talking microphones show strongly different frequency components than features extracted from far-field microphones.

To ease the effects of reverberation, the recognition process can be modified at three stages. Beginning top-down, reverberation can be integrated into the actual decoding of the feature sequence. In [1], the authors use spectral features in combination with hidden Markov models (HMM) for continuous digit recognition. By modeling the reverberation as an additive component to the mean values of the output distributions, the authors could show a significant improvement in terms of recognition performance.

Instead of changing the acoustic model, one can take reverberation into account when extracting the features. In [2], the authors use a voiced/unvoiced detector to modify the observed spectrum according to its decision: For voiced segments, the upper part of the spectrum is cleared and vice versa. By integrating this into their feature extraction, the authors could show a significant improvement for isolated (command) word recognition in an embedded setup.

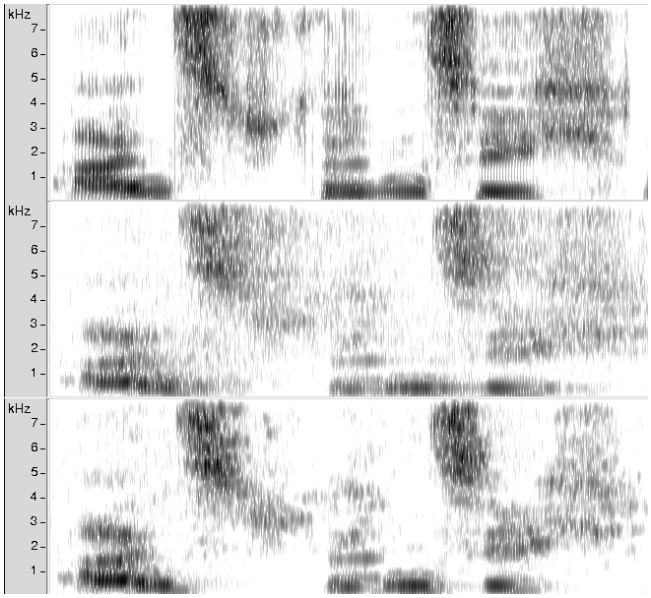

Fig. 1. Example spectrum of /aInst/ /StrI|t@n/ /zIC/ acquired by headset (top), room microphone (middle) showing the smear effect in the voiced/unvoiced transition; the bottom spectrum shows the segment after dereverberation, reducing that effect.

Finally, one can try to remove reverberation effects before feature extraction by preprocessing the audio data thus leaving the recognition system as such untouched. Spectral subtraction has been around for quite some time, however mainly used to remove noise from the signal. In [3], the authors propose a simplified reverberation model that can be integrated with spectral subtraction. The core idea is to split the process in two parts, *early* and *late* reverberation. Assuming the early part of reverberation as non-critical to intelligibility, the late part is estimated and subtracted from the spectrum.

Most work based on [3] deals with subjective or objective acoustic quality assessment (e.g. [4]). We investigate the use of the model for automatic continuous speech recognition in contrast to connected digits or command word recognition.

Although the methods mentioned above help reducing the effects of reverberation, they not only require extensive modifications of the speech recognition system or time-consuming (re-)training using modified training data but also knowledge about the target room and conditions where the system is employed. Thus, there is a lot of overhead work to do for each installation. Thinking of an off-the-shelf application to work with room microphones like an ambient living assistant system or meeting assistant, the adaptation to the previously unknown acoustic characteristics needs to be easy, fast and robust, similar as is speaker adaptation for commercially available dictation systems – it must be reliable and doable in a few minutes.

This article is structured as follows. After a short description of the data in Section 2, the dereverberation model of Lebart and Boucher [3] and its required parameters are introduced in Section 3. In Section 4, an easy, fast and robust approach to estimate the critical dereverberation parameter is described. In Section 5, the used recognition system and adaptation techniques are briefly described. Section 6 provides setup and analysis of a series of experiments showing the use of the proposed algorithms on genuine reverberated data. Section 7 summarizes the work and concludes with an outlook.

2. DATA

For the training of the ASR system, a subset of the German VERBMOBIL [5] corpus (11,714 utterances, 257,810 words, about 25 hours) was used; the speakers were aged around 27 ± 8 years.

For the evaluation of the dereverberation algorithm, 34 speakers (22 male, 12 female; age 65 ± 5) read the text "The North Wind and the Sun", a phonetically rich text from Aesop, resulting in about 25 minutes of speech. The acquisition was done using 4 high quality SHURE microphones in a demo room of an ambient living assistant system installed at our lab: a headset (*ct*) and 3 room microphones (*R1, R2, R3*) mounted on different positions in the room as depicted in Figure 2. This data set will be referenced as TEST.

Large age differences among the speakers in training and test will strongly affect the recognition performance [6]. In order to evaluate the system performance on a data set by speakers of similar age to the the training set, we use close talk recordings of that same text (see above) read by 38 speakers (31 male, 7 female; aged 30 ± 9) resulting in about 30 minutes of speech to get an idea of the expected performance loss due to elderly speech. This data set will be referenced as CONTROL.

All speech data was acquired at a sampling rate of 16 kHz and quantized using 16 bit.

3. DEREVERBERATION BY SPECTRAL SUBTRACTION

In the following, dereverberation by spectral subtraction is briefly introduced; for a more detailed explanation, we refer to [3].

What actually reaches the ear is the produced acoustic signal convolved with an acoustic impulse response (of the room), modeled as

$$f(t) = (h \star s)(t) \tag{1}$$

where $f(t)$ is the observed signal, $s(t)$ the original clean signal and $h(t)$ is the acoustic transfer function, for closed rooms usually assumed to be room impulse response (RIR). If the RIR were known, the dereverberation could be done by de-convolution with the signal. However, the RIR is dependent on the position in the room and requires very sophisticated measurements, usually done by a sound engineer.

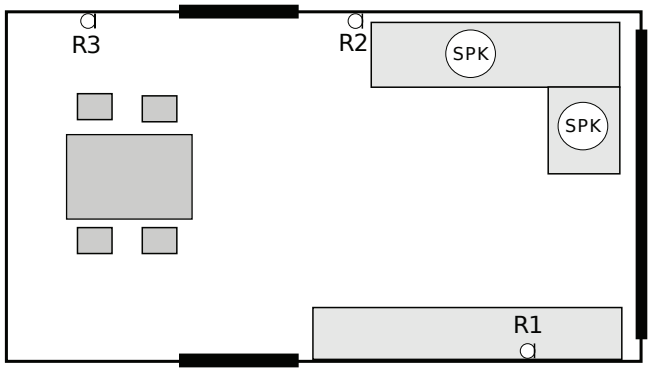

Fig. 2. Sketch of the room microphones mounted on the walls in approx. $1.5\mathrm{m}$ height; the speaker (SPK) is seated on the sofa (either the long or short part) and equipped with a headset (approx. room dimensions: $3.5\mathrm{m} \times 6.5\mathrm{m}$).

A closer look at squared power value of the RIRs measured at different positions in the room suggests that, in logarithmic scale, the decay of the RIR can be modeled by the *decay line* $t \mapsto 2\rho t + 2b$ which can be determined by a least square fit. ρ is then considered to be the *decay parameter*. For example, the decay parameter estimation of the Aachen Lecture Room using six measured RIR [7] yields $\rho = 7.8 \pm 0.09$. Similar results on other recorded RIRs suggest ρ as an invariant for the room acoustics, which will be exploited in the following.

In [3], the reverberation process is split into an early ($t = [0, T_{\text{mix}}[$) and late ($t = [T_{\text{mix}}, \infty[$) part, each modeled by exponentially decaying Gaussian white noise $B(t)$ utilizing the decay parameter ρ:

$$h_{\text{LB}}(t) = \sigma e^{-\rho t} B(t) 1_{[0,T_{\text{mix}}[}(t) + \sigma e^{-\rho t} B(t) 1_{[T_{\text{mix}},\infty[}(t) \tag{2}$$

The early part is considered to carry most (undistorted) information, thus only the late part is removed.

Recalling the original idea of additive noise n in the frequency domain, the spectral subtraction is given (without derivation) as

$$\text{FT}[s](t,\xi) = \sqrt{|\text{FT}[f](t,\xi)|^2 - E(|\text{FT}[n](t,\xi)|^2)} \cdot e^{i\phi(\cdot)} \tag{3}$$

with FT denoting the (short time) Fourier transform and $\phi(\cdot)$ the angle of the polar representation of $\text{FT}[f](t,\xi)$. In words, the clean and observed signal share the phase but the amplitude is changed according to the expectation of the noise. For the dereverberation, the noise is considered to be the late part of the reverberation and can be estimated as

$$\begin{aligned} E(|\text{FT}[n](t,\xi)|^2) &= e^{-2\rho T_{\text{mix}}} \cdot E(|\text{FT}[f](t - T_{\text{mix}},\xi)|^2) \\ &= e^{-2\rho T_{\text{mix}}} \gamma(t - T_{\text{mix}},\xi) \end{aligned} \tag{4}$$

where $\gamma(t,\xi)$ is approximated in a discrete manner as

$$\gamma(t,\xi) \approx r \cdot \sum_{j=0}^{l-1} \beta^j |\text{FT}[f][l-j,k]|^2 \tag{5}$$

for a suitable normalization parameter r and $0 < \beta < 1$, with a recommended $\beta = 0.9$. Thus for the actual dereverberation, only the parameters T_{mix} and ρ are required. While [3] suggests to set

$T_{\text{mix}} = 50$ (ms), the decay parameter ρ was not part of their research and remains to be estimated. Figure 1 shows a segment of clean headset speech, corresponding room microphone signal and the dereverberated signal using the decay parameter estimated as described in the following section.

4. ESTIMATING THE DECAY PARAMETER

As mentioned above, the exponential decay parameter ρ can be estimated from the squared power value of the RIR. Instead of properly measuring the RIR, we propose a rather simple calibration method: In a short training phase, the user moves around the target room and generates impulses. In our case these were generated by clapping hands and rapidly closing a large book. The recorded impulses can be automatically segmented and result in multiple estimations of ρ. To account for possibly bad estimates, the mean value is chosen.

Experiments show pretty stable values; for our ambient assistant living room, we estimated $\rho \approx 11.9 \pm 0.3$ using 10 instances, a rather robust estimate given the presence of slight background noise like a nearby larger street and some construction work in front of the window. As extensive experiments on different data and RIRs [8] indicate that choosing a slightly larger ρ than the actually measured one leads to better results, we chose to add a constant and set $\rho = 12.5$.

5. SYSTEM DESCRIPTION

5.1. Baseline System

The ASR system used for this work is based on semi-continuous Hidden Markov Models sharing 500 Gaussian densities with full covariances. The acoustic models are on polyphones, i.e., phones with variable sized context. The computed features are the short time energy anlongside 12 mel-frequency cepstral coefficients, and their first order derivatives. After training the system on the VERBMOBIL data, the vocabulary was replaced by the words of the read text "The North Wind and the Sun" (108 words, 71 disjoint). For the latter decoding, only a uni-gram language model was used in order to put more emphasis on the acoustic properties of the data. For details of the training refer to [9].

5.2. Adaptation

To recover from the performance loss due to the age difference of the speakers in training and test, the ASR system is adapted. First to match the age and in a second step, to the (de)reverberated data. This is done to account for (remaining) reverberation effects and dereverberation artifacts like "musical noise" which sounds somewhat similar to talking into an empty tin. As we have a limited amount of data, we do 3-fold cross-validation experiments (two thirds adaptation, one third test) and give average results. The adaptation was achieved using MLLR [10] followed by MAP [11] and iterating that procedure for 10 times. The resulting 7 types of recognizers are listed in Table 1.

6. EXPERIMENTS

6.1. Baseline

Table 2 shows results applying the unadapted recognition system to the acquired recordings of the TEST data. Using the headset recordings of the CONTROL data which roughly matches the age distribution of the training set, a baseline word error rate (WER) of 18.9 was achieved. The rather high value is due to the fact that only a uni-gram language model is used in order to put emphasis on the acoustic properties of the data. As expected, the WER for the elderly speakers is increased to 24.3, which is a relative change of 29% confirming earlier findings in [6].

ID	adaptation data
rec-ct	*ct*, original
rec-R1-o	*R1-o*, original
rec-R2-o	*R2-o*, original
rec-R3-o	*R3-o*, original
rec-R1-d	*R1-d*, dereverberated
rec-R2-d	*R2-d*, dereverberated
rec-R3-d	*R3-d*, dereverberated

Table 1. List of recognizers resulting from the different adaptation approaches: ID (left) and adaptation data (right).

	ct	R1	R2	R3
original	24.3	73.2	78.2	76.1
dereverberated	—	54.7	60.5	57.0

Table 2. Word error rate (WER) applying the unadapted recognition system to the headset (*ct*) and room microphone (*R1, R2, R3*) recordings of the TEST data and the respective dereverberated instances.

The performance severely degrades using the room microphones according to their position. *R1* catches most of the direct sound and thus yields the far best recognition rates, however, the WER relatively increases by more than 200% compared to the headset recordings. *R2* is mounted behind the speakers and results in the worst performance as it is considered to be in the acoustically worst position: nearly no direct sound and reverberation generated from all surrounding walls and objects. *R3* is the most distant microphone showing a WER somewhat between *R1* and *R2* – it seems to catch less or more homogeneous reverberation.

After dereverberation, the tremendously high WER of the room microphones could be greatly enhanced. For the best microphone *R1*, the WER could be reduced by roughly a third to 54.7. Similarly, the performance is increased for the other room microphones. However, WER in that range are unacceptable for any use. In the next section, we investigate additional age and acoustic adaptation to further improve the performance.

6.2. Age Adaptation

Results above showed that if speaker age in training and test differs, recognition performance decreases. Therefore, we adapt the ASR system to elderly speech as described in Section 5 using the headset TEST data. Column "ct" in Table 3 shows the measured improvements for both headset and room microphone data; after dereverberating the room microphone data, the improvement in WER as observed in the baseline experiment still holds. However, the achieved WER of 34.7 using R1 is still unacceptable, thus adaptation to the acoustic properties of (de)reverberated speech seems necessary.

test data \ adaptation data	ct	orig.	derev.
ct	**12.4**	—	—
R1-o	57.5	**38.2**	45.5
R2-o	66.8	**50.5**	59.7
R3-o	64.0	**44.0**	50.6
R1-d	34.7	44.8	**28.2**
R2-d	44.7	51.9	**41.0**
R3-d	39.0	47.8	**32.3**

Table 3. Results on the TEST data in WER using different data for adaptation (columns) and test (rows); the baseline experiment for *ct* data without any adaptation is WER = 24.3.

6.3. Combined Acoustics and Age Adaptation

In a second step, we adapted the ASR system to (de)reverberated and elderly speech at the same time using the original room microphone data of the TEST data (labeled *R[1,2,3]-o*) to see if only adaptation without dereverberation could do the job. Note that we used data of R1 to test on R1, R2 to test on R2 and so on.

As shown in column "orig.", the WER is reduced from 57.5 to 38.2 (a relative reduction of 34%) for the matching acoustic condition. Testing with the dereverberated signals which still show some reverberation effects (see bottom of Figure 1), the WER could not outperform the age-only adapted system, suggesting that the dereverberated signal is acoustically closer to the headset signal than to the reverberated speech.

In a last step, we adapted the ASR system to the dereverberated elderly speech using the *R[1,2,3]-o* data dereverberated with the method described in this work, resulting in *R[1,2,3]-d*. Column "derev." shows that the combination of dereverberation and adaptation yields WER between 28.2 and 41.0 depending on the microphone position. With respect to the original unadapted system applied to headset recordings, this is a relative increase of 16% in terms of WER for R1.

7. CONCLUSION

Using room microphones is a very convenient way to acquire speech for ambient assistant living systems, meeting assistants, automatic translation devices or similar speech enabled devices. However, it usually comes with ambient noise or, indoors, with reverberation effects. In Section 1, we listed several approaches how to reduce these effects but all require extensive system modifications or re-training with knowledge of the target room characteristics (e.g. a carefully measured RIR).

In this work, we showed that the combination of dereverberation and adaptation yields impressive results in terms of speech recognition. For the best microphone, the word error rate could be lowered from 73.2 to 28.2, a relative improvement of about 61.5%. Given the WER using headset microphones is 24.3 without any preprocessing or adaptation, the combined method reduces the relative increase of WER due to age and reverberation from over 200% to 16%.

Though the experiments were conducted using a small vocabulary and read speech to study the effects of acoustic improvement for speech recognition, we expect similar findings for spontaneous speech with larger vocabulary as one would apply a proper language model which tremendously boosts the performance. This may also give insight whether or not the adaptation resulted in overfitting, as the same text was read in every recording thus certain phone and word sequences were favored.

Though not new in its idea, we proposed a method which is easy, fast and robust and can be integrated into off-the-shelf products, similar to speaker adaptation in commercially available dictation systems. To get going after installation, the user would first clap his hands for a few times to automatically estimate the dereverberation parameter and then read a predefined text, similar to dictation systems. Using this data, the system can be adapted to the new acoustic and speaker characteristics.

For future work, we plan to confirm these results in different acoustic settings and to compare the estimated reverberation parameters with ones estimated from carefully measured RIRs. Having the focus on human-machine-interaction, recognition performance on spontaneous reverberated speech would be of special interest – we therefore plan to confirm our findings on data like the ICSI or AMI meeting corpus.

8. REFERENCES

[1] A. Sehr and W. Kellermann, "Model-based dereverberation of speech on 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. EUSIPCO 2008*, 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] E.A.P. Habets, "Single-channel speech dereverberation based on spectral subtraction," in *Proc. 15th Annual Workshop on Circuits, Systems and Signal Processing (ProRISC 2004)*, 2004, pp. 250–254.

[5] W. Wahlster, Ed., *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, Berlin, 2000.

[6] J.G. Wilpon and C. N. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. ICASSP*, 1996, pp. 349–352.

[7] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. Digital Signal Processing (DSP)*, 2009.

[8] U. Zäh, "Dereverberation of speech signals – a case study," Diploma Thesis at the Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen-Nürnberg, 2009.

[9] G. Stemmer, *Modeling Variability in Speech Recognition*, vol. 19 of *Studien zur Mustererkennung*, Logos Verlag, Berlin, Germany, 2005.

[10] C. Legetter and P. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, pp. 249–264.

[11] J.L. Gauvain and C.H. 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.