ON THE APPLICATION OF REVERBERATION SUPPRESSION TO ROBUST SPEECH RECOGNITION

Roland Maas[1], *Emanuël A.P. Habets*[2], *Armin Sehr*[1], *Walter Kellermann*[1]

[1]Multimedia Communications and Signal Processing,
University of Erlangen-Nuremberg,
Erlangen, Germany
{maas,sehr,wk}@LNT.de

[2]International Audio Laboratories Erlangen†,
Erlangen, Germany
emanuel.habets@audiolabs-erlangen.de

ABSTRACT

In this paper, we study the effect of the design parameters of a single-channel reverberation suppression algorithm on reverberation-robust speech recognition. At the same time, reverberation compensation at the speech recognizer is investigated. The analysis reveals that it is highly beneficial to attenuate only the reverberation tail after approximately $50\,\mathrm{ms}$ while coping with the early reflections and residual late-reverberation by training the recognizer on moderately reverberant data. It will be shown that the overall system at its optimum configuration yields a very promising recognition performance even in strongly reverberant environments. Since the reverberation suppression algorithm is evidenced to significantly reduce the dependency on the training data, it allows for a very efficient training of acoustic models that are suitable for a wide range of reverberation conditions. Finally, experiments with an "ideal" reverberation suppression algorithm are carried out to cross-check the inferred guidelines.

Index Terms— Signal enhancement, reverberation suppression, reverberation robustness, automatic speech recognition

1. INTRODUCTION

For many years, automatic speech recognition (ASR) has been successfully deployed in everyday-life applications, such as dictation systems and telephone hotlines. The main restriction so far is the necessity of close-talking microphones in order to achieve reliable recognition performance. There are, however, numerous scenarios where the employment of distant-talking microphones being installed at fixed positions in the environment would be much more convenient. Since in such scenarios, the speaker is in general several meters away from the microphone, the received signal is distorted by additive noise and reverberation. These effects significantly reduce the ASR performance if no countermeasures are taken.

Focusing on increasing the robustness of an ASR system to reverberation, three classes of algorithms can be distinguished. Firstly, signal enhancement techniques like beamforming and dereverberation can be applied to the input signals. The latter ones can further be categorized into reverberation cancellation and reverberation suppression approaches [1]. A second class of methods address reverberation only in the feature domain, either by compensating [2] or by explicitly modeling reverberation [3]. Finally, one could aim for directly training the acoustic model of the recognizer to reverberant data.

So far, dereverberation and ASR are often considered and optimized separately. Only little effort has been devoted to improving the integration of both components into one system. Combining signal enhancement and ASR can, however, be a very promising way for robust ASR in reverberant environments, as shown in [4].

In [5], the effect of ideal late-reverberation suppression on ASR performance was analyzed by varying the shape of room impulse responses (RIRs) based on two design parameters T and A. These parameters described that the reflections arriving with a delay greater than T (seconds) after the direct sound are attenuated by a factor A.

In this study, we extend [5] by investigating the blind reverberation suppression algorithm proposed in [6], which has similar design parameters, in order to derive an optimum adjustment to a connected-digit recognition task. Moreover, different ways of coping with reverberation at the ASR back-end are considered. We will exemplify that the best results are achieved if only the late reverberation after about $50\,\mathrm{ms}$ is attenuated by the investigated algorithm while the early part is compensated by cepstral mean normalization (CMN) [7] in combination with a recognizer trained on moderately reverberant data. The established results are verified with an "ideal" reverberation suppression algorithm that operates on the RIRs before their convolution with the digit data.

This paper is structured as follows: In Section 2, the considered blind and "ideal" reverberation suppression algorithms are described. Section 3 presents the experimental setup while the experimental results are discussed in Section 4. Finally, conclusions are drawn in Section 5.

2. REVERBERATION SUPPRESSION

In this section, we introduce the blind reverberation suppression algorithm as well as its "ideal" counterpart. In both cases, the design parameters we will focus on are - similarly to [5] - T and A. The specific role of these parameters in the context of their algorithms will be outlined in the following.

The reverberant signals that will be considered in the ASR experiments result from the convolution of anechoic speech signals and causal room impulse responses. In the short-time Fourier transform (STFT) domain, the reverberant spectral coefficients $Z(m,k)$, where m denotes the time-frame index and k the discrete frequency index, can be written as [6, 8]

$$Z(m,k) = Z_{\mathrm{e}}(m,k) + Z_{\ell}(m,k),$$

where

$$Z_{\mathrm{e}}(m,k) = \sum_{m'=0}^{N_{\mathrm{e}}-1} H(m',k)\, S(m-m',k)$$

The authors would like to thank the Deutsche Forschungsgemeinschaft (DFG) for supporting this work (contract number KE 890/4-1).

†A joint institution of the University of Erlangen-Nuremberg and Fraunhofer IIS.

denotes the early spectral speech component,

$$Z_\ell(m,k) = \sum_{m'=N_\mathrm{e}}^{\infty} H(m',k)\, S(m-m',k)$$

denotes the late reverberant spectral speech component, and $H(m,k)$ represents the RIR in the STFT domain. The parameter N_e, $N_\mathrm{e} \geq 1$, defines which portion of the RIR is considered as late reverberation. Specifically, it is the time (measured relative to the arrival time of the direct sound) at which we assume that the late reverberation starts. Denoting the number of samples between successive analysis frames by R and the sampling frequency in Hz by f_s, we have

$$T = N_\mathrm{e}\, R \,/\, f_s.$$

2.1. Blind reverberation suppression

Although $S(m,k)$ and $H(m,k)$ are unknown in practice, an estimate of $Z_\mathrm{e}(m,k)$ can be obtained using spectral enhancement methods given an estimate of the late reverberant spectral variance $\lambda_\ell(m,k) = \mathcal{E}\left\{|Z_\ell(m,k)|^2\right\}$, where $\mathcal{E}\{\cdot\}$ denotes the mathematical expectation. Using a statistical reverberation model that depends on the direct-to-reverberation ratio (DRR) and the reverberation time of the RIR, we can obtain an estimate of $\lambda_\ell(m,k)$ using the fact that [6]

$$\lambda_\ell(m,k) = \mathrm{e}^{-2\alpha R (N_\mathrm{e}-1)}\, \lambda_\mathrm{r}(m - N_\mathrm{e} + 1, k),$$

where

$$\begin{aligned}\lambda_\mathrm{r}(m,k) &= [1-\kappa]\; \mathrm{e}^{-2\alpha R}\, \lambda_\mathrm{r}(m-1,k) \\ &\quad + \kappa\, \mathrm{e}^{-2\alpha R}\, \lambda_z(m-1,k), \\ \lambda_z(m,k) &= \mathcal{E}\{|Z(m,k)|^2\},\end{aligned}$$

and κ, $0 < \kappa \leq 1$, is a smoothing parameter related to the DRR as shown in [6]. Furthermore,

$$\alpha = \frac{\ln(10)}{\mathrm{T}_{60}\, f_s}$$

is related to the reverberation time T_{60}.

In this study, we have used the log-spectral amplitude estimator [9] to estimate the early speech component $Z_\mathrm{e}(m,k)$. The corresponding gain function is given by

$$G(m,k) = \frac{\xi(m,k)}{1+\xi(m,k)}\, \exp\left(\frac{1}{2} \int_{\zeta(m,k)}^{\infty} \frac{e^{-t}}{t}\, \mathrm{d}t \right),$$

where

$$\xi(m,k) = \frac{\lambda_\mathrm{z}(m,k)}{\lambda_\ell(m,k)},$$

denotes the *a priori* signal-to-reverberation ratio and

$$\zeta(m,k) = \frac{\xi(m,k)}{1+\xi(m,k)}\, \frac{|X(m,k)|^2}{\lambda_\ell(m,k)}.$$

Using A^{-1} as a lower bound for the gain $G(m,k)$ alleviates speech distortions but also limits the amount of late reverberation that can be reduced. An estimate of the early spectral speech component $Z_\mathrm{e}(m,k)$ can now be obtained by applying the constrained gain function to the reverberant spectral coefficient $Z(m,k)$, i.e.,

$$\hat{Z}_\mathrm{e}(m,k) = \max\left[G(m,k), A^{-1}\right]\; Z(m,k).$$

Finally, given the estimated spectral component $\hat{Z}_\mathrm{e}(m,k)$ the early speech component can be obtained using the inverse STFT.

2.2. Ideal reverberation suppression

In one part of our analysis, we will simulate "ideal" suppression of the late reverberant component $Z_\ell(m,k)$, where "ideal" implies that the speech signal is not distorted. Consequently, an upper bound for the speech recognition performance that can be achieved with dereverberation approaches attenuating only the late reverberation. Given $S(m,k)$ and $H(m,k)$, we can compute the "ideally" dereverberated spectral coefficients using

$$Z_\mathrm{i}(m,k) = Z_\mathrm{e}(m,k) + A^{-1} \cdot Z_\ell(m,k),$$

where A, $0 \leq A \leq 1$, is the above-mentioned real-valued parameter that allows us to control the amount of reverberation reduction. The time-domain signal is then obtained by computing the inverse STFT of $Z_\mathrm{i}(m,k)$. It is worthwhile noting that such an "ideal" reverberation suppression algorithm cannot be implemented in practice since $S(m,k)$ and $H(m,k)$ are not available.

3. ANALYSIS CONDITIONS

ASR experiments with a connected-digit recognition task are carried out to find guidelines for setting the design parameters T and A along with choosing a suitable speech recognizer. This task is chosen for evaluation since the probability of the current digit can be assumed to be independent of the preceding digits so that the recognition rate is solely determined by the degree the processed data match the recognizer's acoustic model.

For recognition, we employed the ASR toolkit HTK [10] with word-level HMMs using three Gaussian densities per state. From the input signals, features consisting of 13 mel-frequency cepstral coefficients (MFCCs), including the 0'th, as well as 13 delta and 13 acceleration coefficients are derived. Furthermore, CMN is applied. The sampling rate is $20\,\mathrm{kHz}$, and the analysis frame length used for the feature extraction is set to $25\,\mathrm{ms}$ at a frame shift of $10\,\mathrm{ms}$.

To obtain the reverberant test data, the clean-speech TI digits data are convolved with different RIRs measured at different loudspeaker and microphone positions in four rooms with the characteristics given in Table 1. A strict separation of training and test data is maintained in all experiments both for speech and RIRs. Each test utterance is convolved with an RIR selected randomly from a number of measured RIRs in order to simulate changes of the RIR during the test. Unless stated otherwise, the test data has been processed with the blind reverberation suppression algorithm described in Section 2.1, where the averaged T_{60} values from Table 1 are assumed to be known.

In the following, we distinguish three differently trained recognizers:

1.) The "clean recognizer" is trained on clean, i.e., anechoic, data.

2.) The "moderately reverberant recognizer" is trained on clean data that have been convolved with RIRs from the least reverberant room R1.

3.) The "dereverberated recognizer" is trained on clean data that have been convolved with RIRs from the least reverberant room R1 and preprocessed by the blind reverberation suppression algorithm according to Section 2.1 with $(A,T) = (5\,\mathrm{dB}, 64\,\mathrm{ms})$.

Room	Type	T_{60}	d	DRR
R1	lab	300 ms	2.0 m	+ 4 dB
R2	conf. room	600 ms	2.0 m	+ 0.5 dB
R3	conf. room	700 ms	2.0 m	- 0.5 dB
R4	lecture room	900 ms	4.0 m	- 4 dB

Table 1. Summary of room characteristics: d denotes the distance between speaker and microphone.

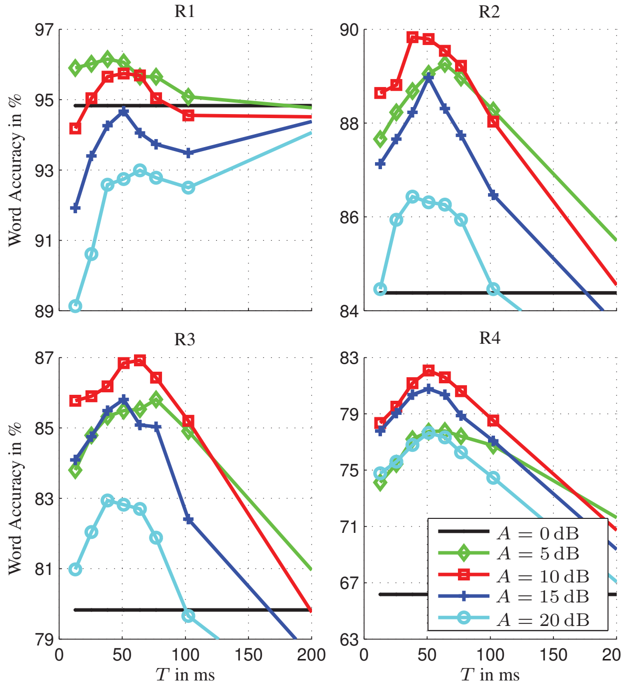

Fig. 1. Word accuracy for dereverberated test data for all rooms and the "clean recognizer".

4. ANALYSIS RESULTS

4.1. The proper aggressiveness

Fig. 1 shows the word accuracy over T with A as parameter when applying the blind reverberation suppression algorithm to the test data while using the "clean recognizer". For all rooms, the curves indicate the same trend: The word accuracy has a pronounced global maximum around $T \approx 50\,\mathrm{ms}$. Moreover, an attenuation of about $A = 10\,\mathrm{dB}$ seems to be most effective. We deduce from these results that a too strong suppression A and values of $T \ll 50\,\mathrm{ms}$ distort the early speech components. As we will see in Section 4.3, those restrictions can be very well compensated by the ASR back-end. On the other hand, considering the optimum configuration $(A, T) \approx (10\,\mathrm{dB}, 50\,\mathrm{ms})$, the algorithm shows a very promising performance, e.g., a relative reduction in word error rate of 47% compared to the baseline in the most reverberant room R4.

4.2. The role of processing artifacts

To obtain deeper insights into the behavior of this algorithm, we employed the "dereverberated recognizer". As can be seen in Fig. 2, this procedure of adapting the recognizer to the processing characteristics strongly improves the recognition results. Consider, e.g., room R4: The word error rate achieved at $(A, T) = (10\,\mathrm{dB}, 50\,\mathrm{ms})$ is reduced by 77% compared to the same configuration in Fig. 1. Considering that the "dereverberated recognizer" is trained on processed data of a single room, namely room R1, the consistent improvement over all rooms indicates that the blind reverberation suppression algorithm significantly reduces the room dependency from the training data. Thus, the reverberation suppression allows for a very efficient training of acoustic models that are suitable for a wide range of reverberation conditions.

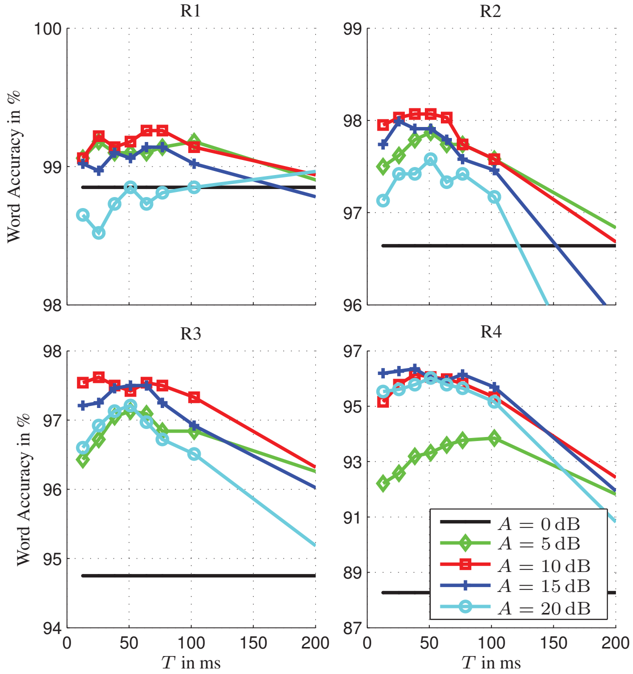

Fig. 2. Word accuracy for dereverberated test data for all rooms and the "dereverberated recognizer".

To investigate the question how strongly the processed data are affected in terms of artifacts, we tested the processed data on the "moderately reverberant recognizer". The results in Fig. 3 show an almost identical behavior for $A \leq 15\,\mathrm{dB}$ and $T \approx 50\,\mathrm{ms}$ compared to Fig. 2, which leads to the conclusion that the algorithm does not introduce considerable artifacts for this configuration. This has also been confirmed by informal listening tests.

Note that Fig. 3 shows slightly better results than Fig. 2 for the more reverberant test cases $A \leq 5\,\mathrm{dB}$ since the data used to train the "moderately reverberant recognizer" in Fig. 3 are more reverberant than the processed training data used for the "dereverberated recognizer" in Fig. 2. For $A \geq 10\,\mathrm{dB}$, the test data are sufficiently well dereverberated to better fit the acoustic model underlying Fig. 2.

4.3. The proper balance between preprocessing and back-end

The results above suggest to divide the task of reverberation compensation between preprocessing and back-end. Preprocessing appears to be most effective when "focusing" on the late reverberation starting at $T \approx 50\,\mathrm{ms}$. At the same time, the ASR back-end should be trained on slightly reverberant data since the early reflections can be very well modeled by HMMs in combination with CMN.

To cross-check this hypothesis, we applied the "ideal" reverberation suppression algorithm to the test data of the most reverberant room R4 for different T and A. Fig. 4 shows the results for a) the "clean" and b) the "moderately reverberant recognizer". Especially for low attenuation of $A \leq 20\,\mathrm{dB}$, it is obviously of significant advantage to keep the early reflections, i.e., $T \approx 50\,\mathrm{ms}$, and use a recognizer trained on slightly reverberant data. In contrast, attenuating even the early reflections while employing the “clean recognizer” is less effective.

4.4. The role of the analysis frame length

For all the above experiments, the frame length used for feature extraction has been set to $25\,\mathrm{ms}$ with a frame shift of 40%. In the final experiment, we varied the frame length for the training and test data from $15\,\mathrm{ms}$ to $35\,\mathrm{ms}$ while keeping the frame shift at 40% of the corresponding frame length. The results for room R4 are depicted in Fig. 5 for a) the “clean” b) the “moderately reverberant recognizer”. For the case of the “clean recognizer”, a longer analysis frame brings a slight increase in word accuracy, whereas a shorter frame remarkably degrades the performance. This might be explained by the fact that CMN as an intra-frame method can better compensate for reverberation with increasing frame length. Considering the “moderately reverberant recognizer”, we see that the commonly used frame length of $25\,\mathrm{ms}$ is well justified. The claim that the early reflections before some $T \approx 50\,\mathrm{ms}$ should be kept seems to hold independently of the frame length.

5. SUMMARY AND CONCLUSIONS

The optimum configuration of the blind reverberation suppression algorithm proposed in [6] was investigated in this contribution for the recognition of reverberated speech signals. The design parameters T and A describe that the reflections arriving with a delay greater than T after the direct sound are attenuated by a maximum level of A. The results show that it is preferable to approximately set $(A, T) = (10\,\mathrm{dB}, 50\,\mathrm{ms})$, independently of the analysis frame length used for feature extraction. Experiments with both the blind and the “ideal” reverberation suppression algorithm confirm that it is most effective to compensate for the early reflections by employing a recognizer trained on moderately reverberant or processed data.

6. REFERENCES

[1] E.A.P. Habets, *Single- and multi-microphone speech dereverberation using spectral enhancement*, Ph.D. thesis, 2007.

[2] H.-G. Hirsch and H. Finster, “A New HMM Adaptation Approach for the Case of a Hands-free Speech Input in Reverberant Rooms,” in *Proc. Interspeech*, 2006, pp. 781–784.

[3] A. Sehr, R. Maas, and W. Kellermann, “Reverberation Model-Based Decoding in the Logmelspec Domain for Robust Distant-Talking Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 18, no. 7, pp. 1676–1691, 2010.

[4] M. Delcroix, T. Nakatani, and S. Watanabe, “Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 324–334, Feb. 2009.

[5] A. Sehr, E.A.P. Habets, R. Maas, and W. Kellermann, “Towards a better understanding of the effect of reverberation on speech recognition performance,” in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.

[6] E. A. P. Habets, S. Gannot, and I. Cohen, “Late reverberant spectral variance estimation based on a statistical model,” *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–774, Sept. 2009.

[7] B. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification.,” *Journal of the Acoustical Society of America*, 1974.

[8] K. Lebart, J. M. Boucher, and P. N. Denbigh, “A new method based on spectral subtraction for speech dereverberation,” *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[9] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, University of Cambridge, 2002.

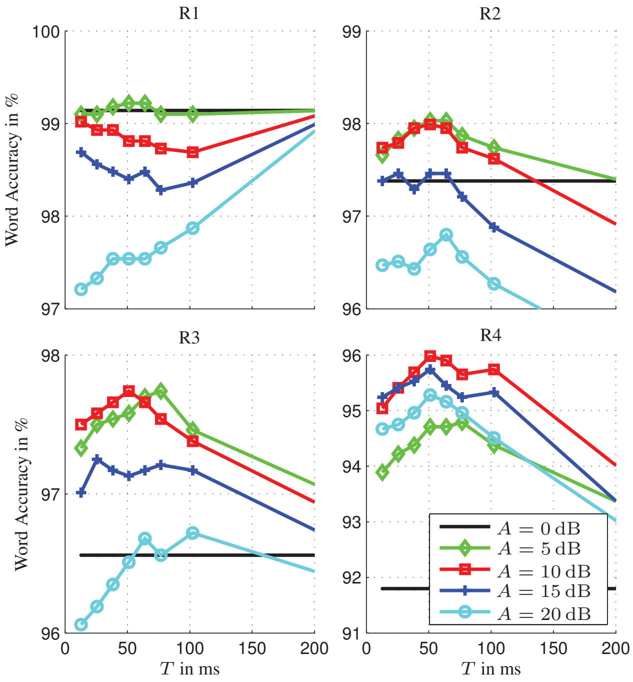

Fig. 3. Word accuracy for dereverberated test data for all rooms and the “moderately reverberant recognizer”.

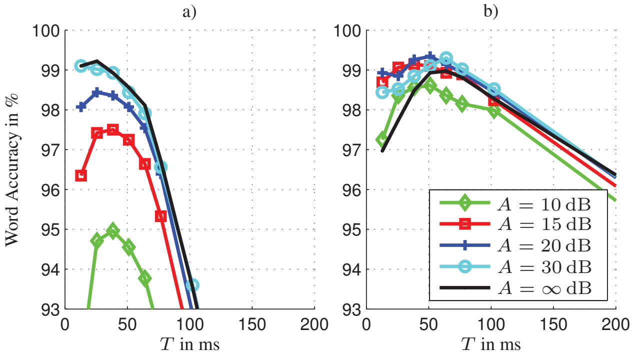

Fig. 4. Word accuracy for “ideally” dereverberated test data for room R4 and a) the “clean” b) the “moderately reverberant recognizer”.

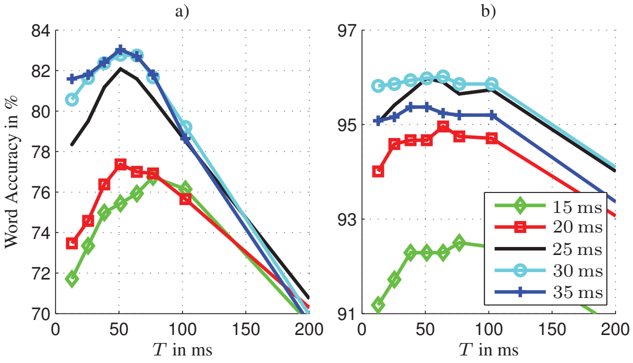

Fig. 5. Word accuracy for dereverberated test data at $A = 10\,\mathrm{dB}$ for room R4 and a) the “clean” b) the “moderately reverberant recognizer”. The frame length is varied from $15\,\mathrm{ms}$ to $35\,\mathrm{ms}$.