# An Investigation into the Effectiveness of Enhancement in ASR Training and Test for Chime-5 Dinner Party Transcription


## Abstract

Despite the strong modeling power of neural network acoustic models, speech enhancement has been shown to deliver additional word error rate improvements if multi-channel data is available. However, there has been a longstanding debate whether enhancement should also be carried out on the ASR training data. In an extensive experimental evaluation on the acoustically very challenging CHiME-5 dinner party data we show that: (i) cleaning up the training data can lead to substantial error rate reductions, and (ii) enhancement in training is advisable as long as enhancement in test is at least as strong as in training. This approach stands in contrast and delivers larger gains than the common strategy reported in the literature to augment the training database with additional artificially degraded speech. Together with an acoustic model topology consisting of initial CNN layers followed by factorized TDNN layers we achieve with 41.6 % and 43.2 % WER on the DEV and EVAL test sets, respectively, a new single-system state-of-the-art result on the CHiME-5 data. This is a 8 % relative improvement compared to the best word error rate published so far for a speech recognizer without system combination.


## Introduction

Neural networks have outperformed earlier GMM based acoustic models in terms of modeling power and increased robustness to acoustic distortions. Despite that, speech enhancement has been shown to deliver additional WER improvements, if multi-channel data is available. This is due to their ability to exploit spatial information, which is reflected by phase differences of microphone channels in the STFT domain. This information is not accessible by the ASR system, at least not if it operates on the common log mel spectral or cepstral feature sets. Also, dereverberation algorithms have been shown to

consistently improve ASR results, since the temporal dispersion of the signal caused by reverberation is difficult to capture by an ASR acoustic model BIBREF0.

However, there has been a long debate whether it is advisable to apply speech enhancement on data used for ASR training, because it is generally agreed upon that the recognizer should be exposed to as much acoustic variability as possible during training, as long as this variability matches the test scenario BIBREF1, BIBREF2, BIBREF3. Multi-channel speech enhancement, such as acoustic BF or source separation, would not only reduce the acoustic variability, it would also result in a reduction of the amount of training data by a factor of $M$, where $M$ is the number of microphones BIBREF4. Previous studies have shown the benefit of training an ASR on matching enhanced speech BIBREF5, BIBREF6 or on jointly training the enhancement and the acoustic model BIBREF7. Alternatively, the training data is often artificially increased by adding even more degraded speech to it. For instance, Ko et al. BIBREF8 found that adding simulated reverberated speech improves accuracy significantly on several large vocabulary tasks. Similarly, Manohar et al. BIBREF9 improved the WER of the baseline CHiME-5 system by relative 5.5% by augmenting the training data with approx. 160hrs of simulated reverberated speech. However, not only can the generation of new training data be costly and time consuming, the training process itself is also prolonged if the amount of data is increased.

In this contribution we advocate for the opposite approach. Although we still believe in the argument that ASR training should see sufficient variability, instead of adding degraded speech to the training data, we clean up the training data. We make, however, sure that the remaining acoustic variability is at least as large as on the test data. By applying a beamformer to the multi-channel input, we even reduce the amount of training data significantly. Consequently, this leads to cheaper and faster acoustic model training.

We perform experiments using data from the CHiME-5 challenge which focuses on distant

multi-microphone conversational ASR in real home environments BIBREF10. The CHiME-5 data is heavily degraded by reverberation and overlapped speech. As much as 23% of the time more than one speaker is active at the same time BIBREF11. The challenge's baseline system poor performance (about 80% WER) is an indication that ASR training did not work well. Recently, GSS enhancement on the test data was shown to significantly improve the performance of an acoustic model, which had been trained with a large amount of unprocessed and simulated noisy data BIBREF12. GSS is a spatial mixture model based blind source separation approach which exploits the annotation given in the CHiME-5 database for initialization and, in this way, avoids the frequency permutation problem BIBREF13.

We conjectured that cleaning up the training data would enable a more effective acoustic model training for the CHiME-5 scenario. We have therefore experimented with enhancement algorithms of various strengths, from relatively simple beamforming over single-array GSS to a quite sophisticated multi-array GSS approach, and tested all combinations of training and test data enhancement methods. Furthermore, compared to the initial GSS approach in BIBREF13, we describe here some modifications, which led to improved performance. We also propose an improved neural acoustic modeling structure compared to the CHiME-5 baseline system described in BIBREF9. It consists of initial CNN layers followed by TDNN-F layers, instead of a homogeneous TDNN-F architecture.

Using a single acoustic model trained with 308hrs of training data, which resulted after applying multi-array GSS data cleaning and a three-fold speed perturbation, we achieved a WER of 41.6% on the development (DEV) and 43.2% on the evaluation (EVAL) test set of CHiME-5, if the test data is also enhanced with multi-array GSS. This compares very favorably with the recently published top-line in BIBREF12, where the single-system best result, i.e., the WER without system combination, was 45.1% and 47.3% on DEV and EVAL, respectively, using an augmented training data set of 4500hrs total.

The rest of this paper is structured as follows. Section SECREF2 describes the CHiME-5 corpus, Section

SECREF3 briefly presents the guided source separation enhancement method, Section SECREF4 shows the ASR experiments and the results, followed by a discussion in Section SECREF5. Finally, the paper is concluded in Section SECREF6.

## CHiME-5 corpus description

The CHiME-5 corpus comprises twenty dinner party recordings (sessions) lasting for approximately 2hrs each. A session contains the conversation among the four dinner party participants. Recordings were made in kitchen, dining and living room areas with each phase lasting for a minimum of 30mins. 16 dinner parties were used for training, 2 were used for development, and 2 were used for evaluation.

There were two types of recording devices collecting CHiME-5 data: distant 4-channels (linear) Microsoft Kinect arrays (referred to as units or `U') and in-ear Soundman OKM II Classic Studio binaural microphones (referred to as worn microphones or `W'). Six Kinect arrays were used in total and they were placed such that at least two units were able to capture the acoustic environment in each recording area. Each dinner party participant wore in-ear microphones which were subsequently used to facilitate human audio transcription of the data. The devices were not time synchronized during recording. Therefore, the W and the U signals had to be aligned afterwards using a correlation based approach provided by the organizers. Depending on how many arrays were available during test time, the challenge had a single (reference) array and a multiple array track. For more details about the corpus, the reader is referred to BIBREF10.

## Guided source separation

GSS enhancement is a blind source separation technique originally proposed in BIBREF13 to alleviate the speaker overlap problem in CHiME-5. Given a mixture of reverberated overlapped speech, GSS aims

to separate the sources using a pure signal processing approach. An EM algorithm estimates the parameters of a spatial mixture model and the posterior probabilities of each speaker being active are used for mask based beamforming.

An overview block diagram of this enhancement by source separation is depicted in fig:enhancementblock. It follows the approach presented in BIBREF12, which was shown to outperform the baseline version. The system operates in the STFT domain and consists of two stages: (1) a dereverberation stage, and (2) a guided source separation stage. For the sake of simplicity, the overall system is referred to as GSS for the rest of the paper. Regarding the first stage, the multiple input multiple output version of the WPE method was used for dereverberation ($M$ inputs and $M$ outputs) BIBREF14, BIBREF15 and, regarding the second stage, it consists of a spatial MM BIBREF16 and a source extraction (SE) component. The model has five mixture components, one representing each speaker, and an additional component representing the noise class.

The role of the MM is to support the source extraction component for estimating the target speech. The class affiliations computed in the E-step of the EM algorithm are employed to estimate spatial covariance matrices of target signals and interferences, from which the coefficients of an MVDR beamformer are computed BIBREF17. The reference channel for the beamformer is estimated based on an SNR criterionBIBREF18. The beamformer is followed by a postfilter to reduce the remaining speech distortions BIBREF19, which in turn is followed by an additional (optional) masking stage to improve crosstalk suppression. Those masks are also given by the mentioned class affiliations. For the single array (CHiME-5) track, simulations have shown that multiplying the beamformer output with the target speaker mask improves the performance on the U data, but the same approach degrades the performance in the multiple array track BIBREF13. This is because the spatial selectivity of a single array is very limited in CHiME-5: the speakers' signals arrive at the array, which is mounted on the wall at some distance, at very similar impinging angles, rendering single array beamforming rather ineffective. Consequently, additional

masking has the potential to improve the beamformer performance. Conversely, the MM estimates are more accurate in the multiple array case since they benefit from a more diverse spatial arrangement of the microphones, and the signal distortions introduced by the additional masking rather degrade the performance. Consequently, for our experiments we have used the masking approach for the single array track, but not for the multiple array one.

GSS exploits the baseline CHiME-5 speaker diarization information available from the transcripts (annotations) to determine when multiple speakers talk simultaneously (see fig:activity). This crosstalk information is then used to guide the parameter estimation of the MM both during EM initialization (posterior masks set to one divided by the number of active speakers for active speakers' frames, and zero for the non-active speakers) and after each E-step (posterior masks are clamped to zero for non-active speakers).

The initialization of the EM for each mixture component is very important for the correct convergence of the algorithm. If the EM initialization is close enough to the final solution, then it is expected that the algorithm will correctly separate the sources and source indices are not permuted across frequency bins. This has a major practical application, since frequency permutation solvers like BIBREF20 become obsolete.

Temporal context also plays an important role in the EM initialization. Simulations have shown that a large context of 15 seconds left and right of the considered segment improves the mixture model estimation performance significantly for CHiME-5 BIBREF13. However, having such a large temporal context may become problematic when the speakers are moving, because the estimated spatial covariance matrix can become outdated due to the movement BIBREF12. Alternatively, one can run the EM first with a larger temporal context until convergence, then drop the context and re-run it for some more iterations. As shown later in the paper, this approach did not improve ASR performance. Therefore,

the temporal context was only used for dereverberation and the mixture model parameter estimation, while for the estimation of covariance matrices for beamforming the context was dropped and only the original segment length was considered BIBREF12.

Another avenue we have explored for further source separation improvement was to refine the baseline CHiME-5 annotations using ASR output (see fig:enhancementblock). A first-pass decoding using an ASR system is used to predict silence intervals. Then this information is used to adjust the time annotations, which are used in the EM algorithm as described above. When the ASR decoder indicates silence for a speaker, the corresponding class posterior in the MM is forced to zero.

Depending on the number of available arrays for CHiME-5, two flavours of GSS enhancement were used in this work. In the single array track, all 4 channels of the array are used as input ($M = 4$), and the system is referred to as GSS1. In the multi array track, all six arrays are stacked to form a 24 channels super-array ($M = 24$), and this system is denoted as GSS6. The baseline time synchronization provided by the challenge organizers was sufficient to align the data for GSS6.

Experiments ::: General configuration

Experiments were performed using the CHiME-5 data. Distant microphone recordings (U data) during training and/or testing were processed using the speech enhancement methods depicted in Table TABREF6. Speech was either left unprocessed, enhanced using a weighted delay-and-sum beamformer (BFIt) BIBREF21 with or without dereverberation (WPE), or processed using the guided source separation (GSS) approach described in Section SECREF3. In Table TABREF6, the strength of the enhancement increases from top to bottom, i.e., GSS6 signals are much cleaner than the unprocessed ones.

The standard CHiME-5 recipes were used to: (i) train GMM-HMM alignment models, (ii) clean up the training data, and (iii) augment the training data using three-fold speed perturbation. The acoustic feature vector consisted of 40-dimensional MFCCs appended with 100-dimensional i-vectors. By default, the acoustic models were trained using the LF-MMI criterion and a 3-gram language model was used for decoding BIBREF10. Discriminative training (DT) BIBREF22 and an additional RNN-based language model (RNN-LM) BIBREF23 were applied to improve recognition accuracy for the best performing systems.

Experiments ::: Acoustic model

The initial baseline system BIBREF10 of the CHiME-5 challenge uses a TDNN AM. However, recently it has been shown that introducing factorized layers into the TDNN architecture facilitates training deeper networks and also improves the ASR performance BIBREF24. This architecture has been employed in the new baseline system for the challenge BIBREF9. The TDNN-F has 15 layers with a hidden dimension of 1536 and a bottleneck dimension of 160; each layer also has a resnet-style bypass-connection from the output of the previous layer, and a “continuous dropout” schedule BIBREF9. In addition to the TDNN-F, the newly released baseline also uses simulated reverberated speech from worn microphone recordings for augmenting the training set, it employes front-end speech dereverberation and beamforming (WPE+BFIt), as well as robust i-vector extraction using 2-stage decoding.

CNN have been previously shown to improve ASR robustness BIBREF25. Therefore, combining CNN and TDNN-F layers is a promising approach to improve the baseline system of BIBREF9. To test this hypothesis, a CNN-TDNNF AM architecture consisting of 6 CNN layers followed by 9 TDNN-F layers was compared against an AM having 15 TDNN-F layers. All TDNN-F layers have the topology described above.

ASR results are given in Table TABREF10. The first two rows show that replacing the TDNN-F with the CNN-TDNNF AM yielded more than 2% absolute WER reduction. We also trained another CNN-TDNNF model using only a small subset (worn + 100k utterances from arrays) of training data (about 316hrs in total) which has produced slightly better WERs compared with the baseline TDNN-F trained on a much larger dataset (roughly 1416hrs in total). For consistency, 2-stage decoding was used for all results in Table TABREF10. We conclude that the CNN-TDNNF model outperforms the TDNNF model for the CHiME-5 scenario and, therefore, for the remainder of the paper we only report results using the CNN-TDNNF AM.

Experiments ::: Enhancement effectiveness for ASR training and test

An extensive set of experiments was performed to measure the WER impact of enhancement on the CHiME-5 training and test data. We test enhancement methods of varying strengths, as described in Section SECREF5, and the results are depicted in Table TABREF12. In all cases, the (unprocessed) worn dataset was also included for AM training since it was found to improve performance (supporting therefore the argument that data variability helps ASR robustness).

In Table TABREF12, in each row the recognition accuracy improves monotonically from left to right, i.e., as the enhancement strategy on the test data becomes stronger. Reading the table in each column from top to bottom, one observes that accuracy improves with increasing power of the enhancement on the training data, however, only as long as the enhancement on the training data is not stronger than on the test data. Compared with unprocessed training and test data (None-None), GSS6-GSS6 yields roughly 35% (24%) relative WER reduction on the DEV (EVAL) set, and 12% (11%) relative WER reduction when compared with the None-GSS6 scenario. Comparing the amount of training data used to train the acoustic models, we observe that it decreases drastically from no enhancement to the GSS6 enhancement.

Experiments ::: State-of-the-art single-system for CHiME-5

To facilitate comparison with the recently published top-line in BIBREF12 (H/UPB), we have conducted a more focused set of experiments whose results are depicted in Table TABREF14. As explained in Section SECREF16, we opted for BIBREF12 instead of BIBREF13 as baseline because the former system is stronger. The experiments include refining the GSS enhancement using time annotations from ASR output (GSS w/ ASR), performing discriminative training on top of the AMs trained with LF-MMI and performing RNN LM rescoring. All the above helped further improve ASR performance. We report performance of our system on both single and multiple array tracks. To have a fair comparison, the results are compared with the single-system performance reported inBIBREF12.

For the single array track, the proposed system without RNN LM rescoring achieves 16% (11%) relative WER reduction on the DEV (EVAL) set when compared with System8 in BIBREF12 (row one in Table TABREF14). RNN LM rescoring further helps improve the proposed system performance.

For the multi array track, the proposed system without RNN LM rescoring achieved 6% (7%) relative WER reduction on the DEV (EVAL) set when compared with System16 in BIBREF12 (row six in Table TABREF14).

We also performed a test using GSS with the oracle alignments (GSS w/ oracle) to assess the potential of time annotation refinement (gray shade lines in Table TABREF14). It can be seen that there is some, however not much room for improvement.

Finally, cleaning up the training set not only boosted the recognition performance, but managed to do so using a fraction of the training data in BIBREF12, as shown in Table TABREF15. This translates to significantly faster and cheaper training of acoustic models, which is a major advantage in practice.

Discussion ::: Temporal context configuration for GSS

Our experiments have shown that the temporal context of some GSS components has a significant effect on the WER. Two cases are investigated: (i) partially dropping the temporal context for the EM stage, and (ii) dropping the temporal context for beamforming. The evaluation was conducted with an acoustic model trained on unprocessed speech and the enhancement was applied during test only. Results are depicted in Table TABREF17.

The first row corresponds to the GSS configuration in BIBREF13 while the second one corresponds to the GSS configuration in BIBREF12. First two rows show that dropping the temporal context for estimating statistics for beamforming improves ASR accuracy. For the last row, the EM algorithm was run 20 iterations with temporal context, followed by another 10 without context. Since the performance decreased, we concluded that the best configuration for the GSS enhancement in CHiME-5 scenario is using full temporal context for the EM stage and dropping it for the beamforming stage. Consequently, we have chosen system BIBREF12 as baseline in this study since is using the stronger GSS configuration.

Discussion ::: Analysis of speaker overlap effect on WER accuracy

The results presented so far were overall accuracies on the test set of CHiME-5. However, since speaker overlap is a major issue for these data, it is of interest to investigate the methods' performance as a function of the amount of overlapped speech. Employing the original CHiME-5 annotations, the word distribution of overlapped speech was computed for DEV and EVAL sets (silence portions were not filtered out). The five-bin normalized histogram of the data is plotted in Fig. FIGREF19. Interestingly, the percentage of segments with low overlapped speech is significantly higher for the EVAL than for the DEV set, and, conversely, the number of words with high overlapped speech is considerably lower for the EVAL than for the DEV set. This distribution may explain the difference in performance observed between

the DEV and EVAL sets.

Based on the distributions in Fig. FIGREF19, the test data was split. Two cases were considered: (a) same enhancement for training and test data (matched case, Table TABREF20), and (b) unprocessed training data and enhanced test data (mismatched case, Table TABREF21). As expected, the WER increases monotonically as the amount of overlap increases in both scenarios, and the recognition accuracy improves as the enhancement method becomes stronger.

Graphical representations of WER gains (relative to the unprocessed case) in Tables TABREF20 and TABREF21 are given in Figs. FIGREF22 and FIGREF25. The plots show that as the amount of speaker overlap increases, the accuracy gain (relative to the unprocessed case) of the weaker signal enhancement (BFIt) drops. This is an expected result since BFIt is not a source separation algorithm. Conversely, as the amount of speaker overlap increases, the accuracy gain (relative to None) of the stronger GSS enhancement improves quite significantly. A rather small decrease in accuracy is observed in the mismatched case (Fig. FIGREF25) for GSS1 in the lower overlap regions. As already mentioned in Section SECREF3, this is due to the masking stage. It has previously been observed that using masking for speech enhancement without a cross talker decreases ASR recognition performance. We have also included in Fig. FIGREF25 the GSS1 version without masking (GSS w/o Mask), which indeed yields significant accuracy gains on segments with little overlap. However, since the overall accuracy of GSS1 with masking is higher than the overall gain of GSS1 without masking, GSS w/o mask was not included in the previous experiments.

## Conclusions

In this paper we performed an extensive experimental evaluation on the acoustically very challenging CHiME-5 dinner party data showing that: (i) cleaning up training data can lead to substantial word error

rate reduction, and (ii) enhancement in training is advisable as long as enhancement in test is at least as strong as in training. This approach stands in contrast and delivers larger accuracy gains at a fraction of training data than the common data simulation strategy found in the literature. Using a CNN-TDNNF acoustic model topology along with GSS enhancement refined with time annotations from ASR, discriminative training and RNN LM rescoring, we achieved a new single-system state-of-the-art result on CHiME-5, which is 41.6% (43.2%) on the development (evaluation) set, which is a 8% relative improvement of the word error rate over a comparable system reported so far.

Acknowledgments

Parts of computational resources required in this study were provided by the Paderborn Center for Parallel Computing.