Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 46 (2017) 574–584

www.elsevier.com/locate/csl

Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures

A.H. Moore*, P. Peso Parada[1], P.A. Naylor

Department of Electrical and Electronic Engineering, Imperial College London, Exhibition Road, London, United Kingdom

Received 8 April 2016; received in revised form 28 September 2016; accepted 25 November 2016
Available online 8 December 2016

Abstract

Automatic speech recognition in everyday environments must be robust to significant levels of reverberation and noise. One strategy to achieve such robustness is multi-microphone speech enhancement. In this study, we present results of an evaluation of different speech enhancement pipelines using a state-of-the-art ASR system for a wide range of reverberation and noise conditions. The evaluation exploits the recently released ACE Challenge database which includes measured multichannel acoustic impulse responses from 7 different rooms with reverberation times ranging from 0.33 to 1.34 s. The reverberant speech is mixed with ambient, fan and babble noise recordings made with the same microphone setups in each of the rooms. In the first experiment, performance of the ASR without speech processing is evaluated. Results clearly indicate the deleterious effect of both noise and reverberation. In the second experiment, different speech enhancement pipelines are evaluated with relative word error rate reductions of up to 82%. Finally, the ability of selected instrumental metrics to predict ASR performance improvement is assessed. The best performing metric, Short-Time Objective Intelligibility Measure, is shown to have a Pearson correlation coefficient of 0.79, suggesting that it is a useful predictor of algorithm performance in these tests.

Keywords: Automatic speech recognition; Speech enhancement; Dereverberation; Beamforming; Realistic environments; Microphone array signal processing

1. Introduction

Real-world applications of Automatic Speech Recognition (ASR), such as meeting transcription and human-robot interaction, demand that the speaker be some distance from the sound capture device. This leads to degradation of the desired signal by both additive noise and reverberation, both of which have a significant degrading effect on ASR accuracy. It has been proposed that robustness in distant-talking ASR be achieved through three approaches, namely enhancement of the audio signal, front-end based approaches which enhance the signal in the feature domain and back-end methods (Haeb-Umbach and Krueger, 2012). Recent challenges such as the REVERB challenge (Kinoshita et al., 2013) and CHiME3 (Barker et al., 2015) have demonstrated the effectiveness of all three approaches.

* Corresponding author.
E-mail address: alastair.h.moore@imperial.ac.uk (A.H. Moore).

[1] Present address: Cirrus Logic, Marble Arch House, 66 Seymour St., 1st Floor, London W1H 5BT, United Kingdom.

http://dx.doi.org/10.1016/j.csl.2016.11.003

Our work focusses on speech enhancement algorithms and particularly on the case when the ASR system itself cannot be altered, or it is impractical to alter it. This is relevant in many real-world applications where an ASR engine is supplied by a third-party, either as embedded code or as a cloud-based system. It is also relevant during algorithm development where many alternative parameters or algorithm variations are to be assessed. It is, in general, not practical to process the training data with every algorithm and retrain the ASR system for each. This is particularly true when the range of acoustic conditions that need to be considered is wide, as in everyday environments where a broad range of Reverberation Times (RTs), Direct-to-Reverberant Ratios (DRRs), noise types and Signal-to-Noise Ratios (SNRs) may be encountered.

Speech enhancement for human listening has a long history with a great many algorithms being proposed and refined over the decades which attempt to mitigate both additive and convolutive noise (Naylor and Gaubitch, 2010). By exploiting multiple microphones, the most successful algorithms can improve the perceived audio quality and/or intelligibility. Conducting listening tests using human listeners is time consuming and expensive and so instrumental measures have been developed which attempt to model human responses in order to predict the performance of new algorithms (Taal et al., 2011; ITU-T, 2001).

The ultimate performance measure of speech enhancement algorithms for ASR will always be the Word Error Rate (WER) achieved over a particular set of test data. However, we propose that just as instrumental measures can be useful for modelling human listening, they can also be helpful for evaluating speech enhancement for ASR. To test this assertion, the current study is split into three parts. The first establishes the performance bounds of an “off-the-shelf” recogniser for a broad range of acoustic conditions. The second determines the performance improvements obtained by a range of speech enhancement pipelines for a representative subset of acoustic conditions. The third compares the performance measured using WER and relative WER reduction (rWERR) to that predicted by instrumental metrics.

To summarise, the novel contributions of the current study are (i) results for the CHiME3 (Barker et al., 2015) baseline system under more diverse acoustic conditions than have hitherto been considered; (ii) independent verification of the efficacy of linear prediction-based dereverberation, particularly when combined with beamforming (Yoshioka and Nakatani, 2012; Delcroix et al., 2014, 2015); (iii) evaluation of multi-channel speech enhancement for ASR under reverberant conditions with high levels of noise, and especially babble noise; (iv) a comparison showing strong correlation between ASR performance and instrumental metrics. It is hoped that the results of (i) and (iii), which make use of the Acoustic Characterization of Environments (ACE) Challenge database of Acoustic Impulse Responses (AIRs) and noise, will serve as a baseline for future robust ASR systems and speech enhancement algorithms.

In Section 2, we describe the setup of the experimental framework and then evaluate the ASR system for 210 different acoustic conditions with unprocessed audio. In Section 3, we select a subset of 54 acoustic conditions, each of which is enhanced using 6 different enhancement strategies consisting of different combinations of noise reduction, dereverberation and beamforming. In Section 4, the results from Section 3 are compared to the results of instrumental metrics. The significance of the results is discussed in Section 5 and our conclusions presented in Section 6.

2. Experiment 1: effect of acoustic conditions without speech enhancement

The aim of the first experiment is to characterise the performance of a standard state-of-the-art ASR system for a broad range of acoustic conditions using only the received signal. In this way, the effect of reverberation, noise type and SNR can be individually analysed. The results of this experiment also inform the selection of a reduced subset of acoustic conditions for which speech enhancement is used in the second experiment.

2.1. ASR system

The ASR employed in this work corresponds to the updated Kaldi CHiME3 baseline recipe available in the Git Kaldi repository.[2] This system is an improvement of the original baseline which includes a feature-space maximum likelihood linear regression transformation on the DNN features and output hypothesis rescoring with a Kneser−Ney smoothed 5-gram model and an RNN language model (Hori et al., 2015). By choosing a system that can be freely downloaded and used, new algorithms can be objectively compared to the baselines presented here.

[2] https://github.com/kaldi-asr/kaldi/tree/master/egs/chime3/s5 commit id: 7e44bc74c3388cd134485672fd1c9687c68073b9.

Table 1
AIRs from ACE Challenge database used in Experiment 1 with associated acoustic parameters for the first microphone channel. AIRs marked with an asterisk are used in Experiment 2.

AIR	RT [s]	DRR [s]	C_{50} [dB]
Building Lobby 1	0.68	8.1	14.1
Building Lobby 2	0.75	5.1	11.8
Lecture Room 1 1	0.68	3.8	10.9
Lecture Room 1 2	0.61	6.4	13.1
Lecture Room 2 1*	1.34	3.8	7.3
Lecture Room 2 2*	1.28	3.6	7.1
Meeting Room 1 1	0.47	10.6	16.6
Meeting Room 1 2	0.47	7.0	13.4
Meeting Room 2 1*	0.38	9.8	17.6
Meeting Room 2 2*	0.39	8.7	18.0
Office 1 1	0.36	10.5	19.7
Office 1 2	0.33	4.3	15.3
Office 2 1*	0.41	13	19.4
Office 2 2*	0.38	9.5	14.6

The system is trained using the real and simulated noisy scenarios provided in the CHiME3 Challenge, which are based on the Wall Street Journal (WSJ0)(Garofalo et al., 2007) 5 k corpus and contains utterances read from the Wall Street Journal newspaper. The real scenarios (1600 utterances) were recorded in 4 environments (on a bus, in a cafe, in a pedestrian area and at a street junction) by 4 different speakers using a 6 channel microphone array. The simulated scenarios (7138 utterances) were produced by mixing the clean data from the WSJ0 5 k training set with background noise recorded in the same 4 environment types as the the real data. See (Barker et al., 2015) for full details of how this mixing was performed. The average SNR of the recordings is reported to be 5 dB. Only the fifth channel of each recording is employed in the training, i.e., no beamforming or speech enhancement of any sort is applied.

In the testing stage, the language model weight is set to the value that provides the lowest WER in the development set, which in this case is 6.

Evaluation is based on the “SI ET 05” simulation test set. This consists of 330 read speech utterances from the “no verbal punctuation” (NVP) part of the WSJ0 speaker-independent medium vocabulary (5 k) evaluation set. Since we are specifically interested in the performance variations caused by different acoustic conditions, the clean test set was used to generate the simulated multi-channel noisy reverberant speech from the clean signals as described in Section 2.2.

2.2. Evaluation data generation

Evaluation data were generated to simulate a 4-element linear microphone array using the recently-released ACE Challenge database (Eaton et al., 2015, 2016a, b). This database is uniquely relevant in that it contains AIRs for a large number of rooms with two source distances per room, for a variety of microphone arrays. Table 1 lists the relevant acoustic parameters associated with the first element of the microphone array in each case. In addition to the AIRs, the ACE Challenge database contains noise recordings made using room-microphone setups identical to the setups used to measure the corresponding AIRs. In this way, the noise and reverberation are entirely consistent.

Multichannel noisy reverberant speech was simulated according to the signal model

$$x_m(t) = \sum_{\tau=0}^{L-1} h_m(\tau) s(t-\tau) + \nu_m(t), \tag{1}$$

where $x_m(t)$ is the signal received at the mth microphone, $h_m(t)$ is the L-tap, time-invariant impulse response between the source and the mth microphone, $s(t)$ is the desired speech signal and $\nu_m(t)$ is the additive noise at mth microphone. The microphone signals are therefore simulated by convolving each of the 330 speech files from the CHiME3 simulation evaluation set “SI ET 05” with each of the AIRs from the ACE Challenge database (7 rooms × 2 source-array distances per room), as described above. The A-weighted speech power of the reverberant speech at the first microphone

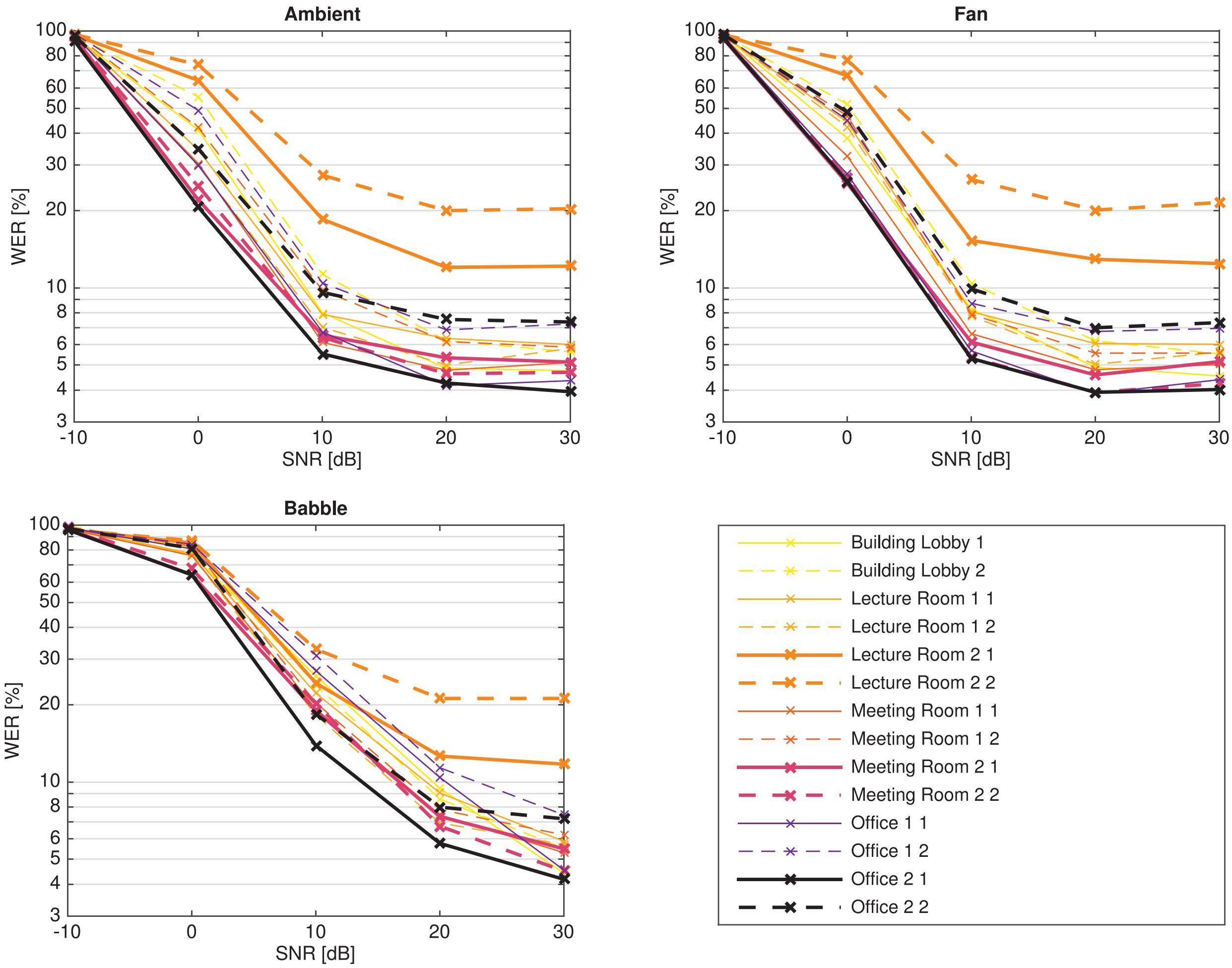

Fig. 1. WER as a function of SNR for all AIRs with Ambient, Fan and Babble noise. Bold lines indicate AIRs selected for use in Experiment 2.

(i.e., the clean speech convolved with the measured impulse response for the first channel) was determined according to ITU-P56 (ITU-T, 1993). For each utterance a randomly selected segment of multichannel noise from the appropriate noise recording was added, adjusting the level to obtain A-weighted SNRs of −10, 0, 10, 20 and 30 dB. This process was repeated for each available noise type in the ACE Challenge database, namely Ambient, Fan and Babble. Ambient noise consisted of the background noise in each room and so, depending on the ventilation system, equipment in the room and any pedestrian traffic could be quite variable in its nature. Fan noise was produced by a portable fan heater with an approximately stationary noise characteristic. Babble noise was recorded live by between 4 and 7 simultaneously active talkers, reading from a variety of source material. For both Fan and Babble noise the Ambient noise is inevitably also present in the background but the level of the Fan or Babble is dominant in those cases. A Matlab script to generate the evaluation data as described is available.[3]

2.3. Results

Fig. 1 shows the WER achieved for each AIR and noise type as a function of SNR. It can be seen that, for all noise types, noise is the dominant source of error at low SNRs, with WER tending to 100% at −10 dB SNR. It should be noted that 100% WER is not an upper bound since there is no limit to the number of insertions which could occur,

[3] http://www.commsp.ee.ic.ac.uk/~sap/resources/csl-ace-asr.

but for this system and stimuli the insertions peaked at 0 dB SNR and dropped again at −10 dB SNR. At 30 dB SNR the WER varies between 3.9 and 21.5% due to the differing reverberation, independently of the noise type. At intermediate SNRs there are substantial differences according to the AIR, noise type and SNR. The results for Fan and Ambient noise are very similar, reflecting the fact that, for most rooms, the ambient noise was due to stationary noise sources. For these noise types the performance plateaus at 20 dB SNR. With Babble noise, as is well known, the performance is considerably worse for a given SNR and so the performance continues to improve as the SNR is increased to 30 dB. The substantial differences in WER for the different AIRs and for each noise type and at any particular SNR demonstrates the huge influence of the diverse reverberation properties of the ACE Challenge database.

It is interesting that certain AIRs result in consistently better (or worse) performance compared to the other AIRs, regardless of the noise type and SNR. We therefore select a subset of the noisy reverberant speech for speech enhancement and subsequent analysis in Section 3. The 6 selected AIRs are taken from rooms “Lecture Room 2”, “Meeting Room 2” and “Office 2” since together they span the range of achieved WERs with the AIRs representing best and worst WERs being “Office 2 1” and “Lecture Room 2 2”, respectively. We note that these also represent the extreme values of the DRR and C_{50}. We choose to use both AIRs from each room so that the effect of RT, which is constant in each room, can be separated from that of DRR, which varies with the source-microphone distance.

3. Experiment 2: effect of speech enhancement algorithms

The aim of the second experiment is to compare the effectiveness of different realisations of speech enhancement with the specific goal of reducing the WER of a standard ASR system. In particular, variations in performance due to a diverse range of acoustic conditions are considered.

Based on the results of Section 2, the data for rooms “Lecture Room 2”, “Meeting room 2” and “Office 2” are selected for speech enhancement as they represent the full range of performance observed in Experiment 1. To reduce the total number of test conditions only SNRs of 0, 10 and 20 dB are considered as these represent the region of greatest variation in the results of Experiment 1.

3.1. Enhancement algorithms

Speech enhancement algorithms of three types are tested in the current study—non-linear speech enhancement, spatial filtering and linear dereverberation—and these will be discussed below. As well as testing each algorithm individually, a number of combinations of algorithms are additionally considered.

Non-linear speech enhancement (SS). We include in this study a version of non-linear speech enhancement based on Spectral Subtraction (SS). Based on the work of Xiong et al. (2014), we have implemented a single channel algorithm that includes cepstral smoothing to reduce the musical noise commonly associated with this type of non-linear processing in the Short Time Fourier Transform (STFT) domain. The noise Power Spectral Density (PSD) is first estimated using minimum statistics (Martin, 2001), the late reverberation PSD is estimated according to (Lebart et al., 2001) which uses an estimate of the RT from (Eaton et al., 2013). After estimating the speech PSD the Minimum Mean Squared Error (MMSE) gain is found using (Breithaupt et al., 2008).

Spatial filtering (DSB). Spatial filtering, or beamforming, attempts to enhance the signal based on the spatial distribution of the sound sources, interference, reverberation and noise. Many beamformers have been proposed which are optimal under different assumptions about the spatial properties of the noise and reverberant soundfields (see, for example, Brandstein and Ward, 2001; Benesty et al., 2008). The simplest and arguably the most robust is the classic Delay-and-Sum Beamformer (DSB) and we include this beamformer in our evaluation. In this case, only the Time-Differences-of-Arrival (TDOAs) of the desired sound at each microphone must be estimated. We obtain such an estimate using Generalised Cross-Correlation with Phase Transformation (GCC-PHAT) (Knapp and Carter, 1976) applied to the Linear Predictive Coding (LPC) residuals, where a single set of LP coefficients are jointly estimated using all 4 channels (Gaubitch et al., 2006).

Linear dereverberation (WPE). Multichannel linear filtering has the potential to offer high performance dereverberation, assuming the AIRs are known exactly (Miyoshi and Kaneda, 1988). However, in practice these are difficult to estimate with sufficient accuracy. Using an autoregressive model of the reverberation process, linear prediction and inverse filtering can be applied in subbands to achieve dereverberation (Nakatani et al., 2010). In the Weighted Prediction Error (WPE) approach (Yoshioka et al., 2010), an estimate of the desired signal energy is used to weight the prediction error to take account of the non-stationary nature of speech. To avoid over-whitening the speech signal, no attempt is made to dereverberate the early reflections. A set of dereverberation filters are designed for each channel using an iterative approach to alternately estimate the desired signal energy and the linear prediction filter.

The WPE algorithm is inherently multiple-input, multiple-output. When applied in isolation, without being combined with beamforming, the ASR uses only the first channel.

3.2. Enhancement pipelines

As well as evaluating the individual enhancement methods, we have additionally considered the following combinations of methods, which we refer to here as speech enhancement pipelines.

DSB + SS. In this combination of the above two methods, the output of the DSB is processed using single channel non-linear speech enhancement (SS).

WPE + DSB. The four output channels from the WPE dereverberation algorithm are time aligned using the same TDOAs as calculated for the DSB in isolation, such that the performance of the beamforming when combined with WPE is not impacted (whether positively or negatively) by the accuracy of the TDOAs estimation.

WPE + DSB + SS. In this combination of speech enhancement processing, the single channel output of WPE + DSB is further processed using the above non-linear speech enhancement (SS).

None. Baseline performance is achieved as in Experiment 1, where the first channel of noisy reverberant speech is input to the ASR without any front-end enhancement.

3.3. Results

Fig. 2 shows the WER obtained with each of the enhancement pipelines, arranged in order of increasing performance. The differences between the means of all the pipelines are statistically significant ($p < 0.005$) other than those indicated by braces. We can therefore conclude that, in isolation, both the multichannel algorithms (i.e., WPE and DSB) are successful at reducing the WER and that the best performance is obtained when the two algorithms are combined into a single pipeline (i.e., WPE + DSB). On the other hand, the single channel algorithm (SS) does not lead to an overall improvement under the conditions tested. Moreover, when SS is employed on the output of either DSB or WPE + DSB the performance of the combination is actually degraded.

To gain more insight into the conditions under which each of the enhancement pipelines are effective, Fig. 3 shows the WER for each pipeline according to the noise type and SNR. For clarity each plot only shows the results for the “Lecture Room 2 2”, “Office 2 1” and “Office 2 2” AIRs as these represent the range of results and results for the other AIRs followed a similar pattern. Comparing the variation in performance due to different AIRs, the overall trends observed in Experiment 1 are also evident in the enhanced signals. In particular, performance in Babble noise at 0 dB SNR is dominated by noise as indicated by the similarity of the WER for the different AIRs. The similarity between the performance in Fan and Ambient noise at each SNR suggests that the statistical properties of these two noise types have a similar impact on the enhancement pipelines.

The pipelines are arranged according to their overall performance, as in Fig. 2, such that deviations from the mean behaviour are characterized by non-monotonically decreasing lines. The WPE algorithm is noteworthy in this regard. At both 10 and 20 dB SNR it performs notably better than the rank order neighbouring pipelines. This suggests that its low average performance seen in Fig. 2 is due to the high WERs obtained at 0 dB SNR.

At the lowest SNR the relative difference between the AIRs can be seen to be broadly consistent. As the SNR increases the WERs start to converge for the better performing pipelines. This suggest that at higher SNRs, the better performing pipelines have a relatively large improvement on the less favourable reverberation conditions. These

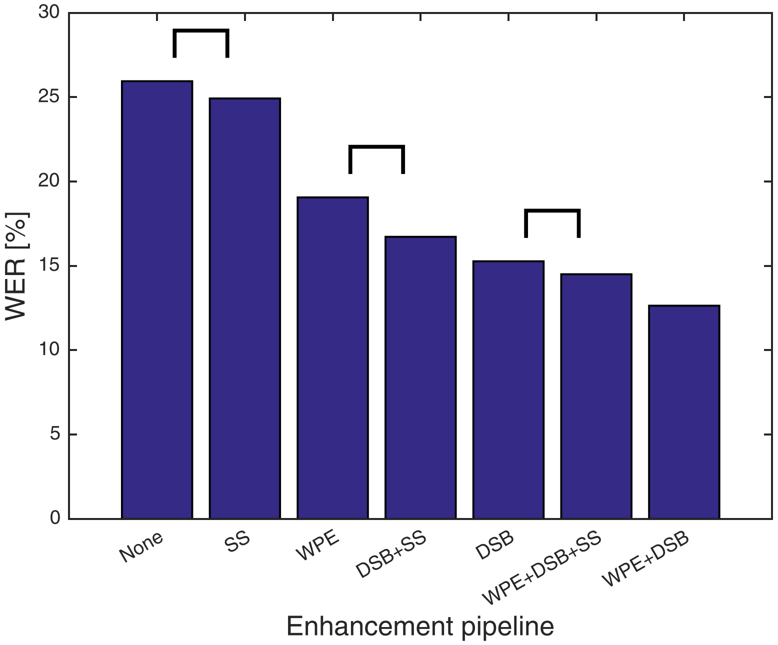

Fig. 2. WER for each enhancement pipeline averaged over 6 selected AIRs, 3 noise types and 3 SNRs. Pipelines are ordered on merit and braces above bars indicate pairs of pipelines which are not significantly different at $p < 0.005$ level.

relationships can also be observed in Fig. 4 which shows the rWERR of each pipeline for each AIR with respect to the corresponding noisy reverberant speech. It is remarkable that for the most challenging AIR the best case pipeline, WPE + DSB, achieves an rWERR of up to 82%. In contrast there are a number of datapoints indicating negative rWERRs. Almost all of these have SS in the pipeline, suggesting that the poor performance may be due to the introduction of artefacts due to the non-linear processing.

4. Evaluation using instrumental metrics

To evaluate the performance of each speech enhancement pipeline over the full range of acoustic conditions employed in the second experiment is very time consuming. During the development of new algorithms it is therefore desirable to predict WER scores using more easily computed instrumental measures, where the clean non-reverberant speech reference is available. In this study, we consider 3 possible instrumental measures. The first, Normalised Signal-to-Reverberation Ratio (NSRR), is based on the segmental SNR but includes normalisation of the noisy reverberant signal such that the direct path power is the same as in the reference signal (Naylor et al., 2010). The second, Perceptual Evaluation of Speech Quality (PESQ), is designed to model the perceived audio quality of narrowband speech (ITU-T, 2001). The third, Short-Time Objective Intelligibility Measure (STOI), attempts to model human speech intelligibility (Taal et al., 2011).

Fig. 5 shows the WER for every combination of acoustic condition and enhancement pipeline plotted against the corresponding instrumental measures. Whilst WER does not seem to vary systematically with NSRR there is a clear relationship with PESQ and STOI, especially for Ambient and Fan noise.

To be useful in predicting the performance benefit of speech enhancement algorithms, the improvement in WER should be correlated with improvements in the instrumental metrics. Fig. 6 shows these correlations and the corresponding Pearson correlation coefficient, ρ, for each metric. The correlations are statistically significant in all cases but the size of the correlation coefficient for STOI, 0.79, is much larger than for NSRR, 0.46, and PESQ, 0.55. For each point in Fig. 6, the colour represents the value of the metric for the “None” pipeline. The gradient in colours for both PESQ and STOI suggests there is a tendency for more severely degraded signals to benefit more in terms of rWERR for a particular relative improvement in the instrumental metric. This is consistent with the non-linear shape of the corresponding curves in Fig. 5.

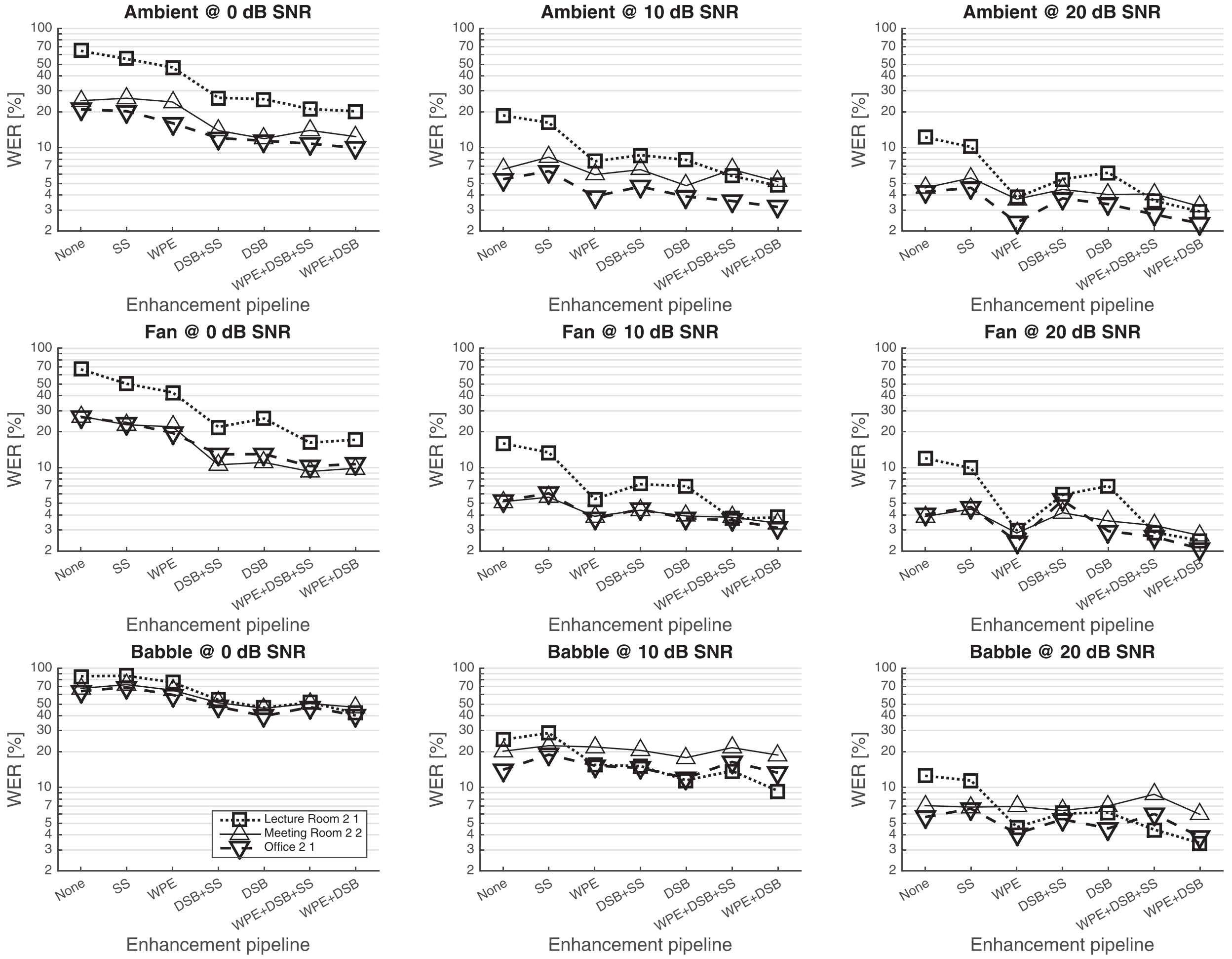

Fig. 3. WER for each algorithm and AIR. Each plot shows results for a particular SNR and noise type.

5. Discussion

In this paper, we have evaluated the CHiME3 baseline ASR system in a wide range of acoustic conditions that significantly extends the breadth of acoustic conditions employed in the CHiME3 challenge itself. The multichannel noisy reverberant speech signals were simulated using 14 AIRs measured in 7 different rooms with RTs varying from 0.33 to 1.34 s. Noise of three different types, recorded for the same microphone arrangements as the AIR measurements was mixed at SNRs ranging from −10 to 30 dB.

In the first experiment, where the noisy reverberant signals were not enhanced, the WERs ranged from 3.9 to 99.3% indicating that in favourable acoustic conditions the ASR system is capable of excellent performance and that in the most severe cases the degradation of the signal due to reverberation and noise was complete. Even at the highest SNR the difference in reverberation properties between the AIRs caused the WER to vary from 3.9 to 20%. This clearly demonstrates the deleterious effect of reverberation on the ASR system.

In the second experiment, a representative subset of the signals from the first experiment, including those corresponding to the best and worst case AIRs, were enhanced using 6 different processing pipelines. At 20 dB SNR the best performing pipeline achieved WERs between 2.1 and 7.1%, clearly demonstrating the effectiveness of the speech enhancement.

Of the different algorithms tested it was found that single channel non-linear speech enhancement used both in isolation and as part of a pipeline had a negative impact on the WER. In contrast both beamforming, using a DSB, and dereverberation, using WPE, contribute positively and combining the two led to the best results. Our comparison

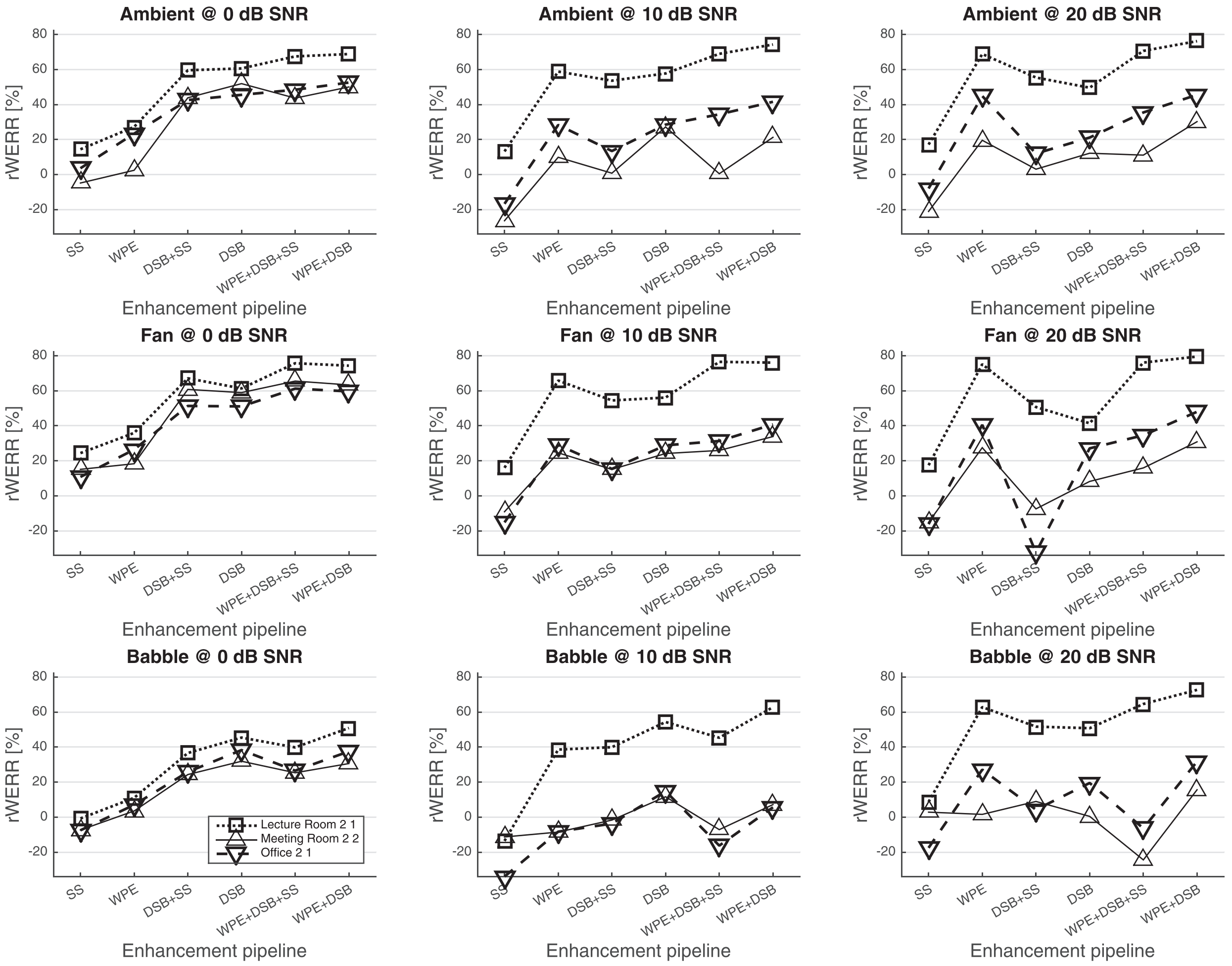

Fig. 4. rWERR for each algorithm and AIR. Each plot shows results for a particular SNR and noise type.

of WPE, DSB and WPE + DSB is similar to the results of the comparison by Delcroix et al. (2015) between WPE, Minimum Variance Distortionless Response (MVDR) and WPE + MVDR. Note that we chose to use a DSB beamformer rather than an MVDR beamforer because, unlike (Delcroix et al., 2015), our tests include acoustic conditions with non-stationary noise and inaccuracies in estimating the noise statistics can negatively affect the performance of MVDR beamforming. Furthermore, our results include rooms with up to 1.3 s, compared to 0.7 s in the REVERB challenge. Since the specific beamformers and acoustic conditions vary between the two studies, a direct comparison of the WERs achieved in each case is not meaningful. Nevertheless, it is interesting that in their study the relative

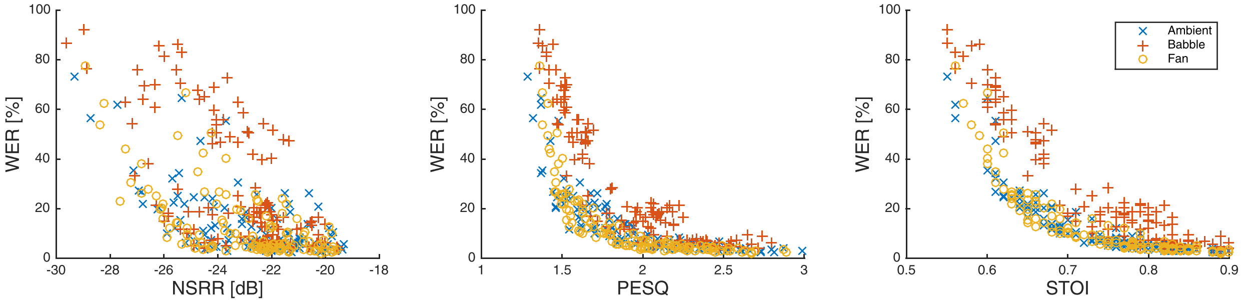

Fig. 5. WER as a function of instrumental metrics NSRR, PESQ and STOI for all acoustic conditions and enhancement algorithms.

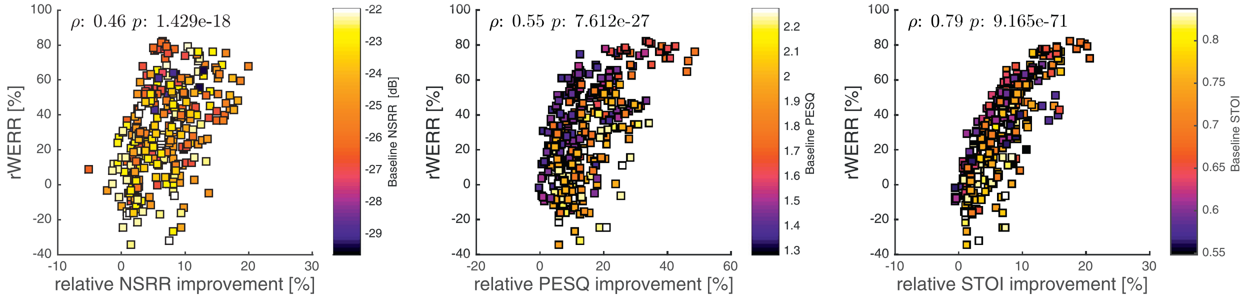

Fig. 6. rWERR as a function of relative improvement in instrumental metrics NSRR, PESQ and STOI for all acoustic conditions and enhancement algorithms. Text insets show Pearson correlation, ρ, and p-value. The colour of each point indicates the value of the metric in the baseline case.

performance of MVDR and WPE were reversed for Simulated and Real data. Our results agree with their Real data result suggesting that the naturalness of the noise recordings increased the realism in our simulations.

In the present study, room reverberation was simulated by convolving dry speech with acoustic impulse responses measured in real rooms. The filtering effect is therefore more realistic than the common approach where the AIRs themselves are generated by simple acoustic simulation, e.g., using the image source method (Allen and Berkley, 1979). However, this method of simulation does not include the temporal variations of the acoustic channel from the desired source to the microphones which occur if either is moving. Since all speech enhancement methods tested here used offline implementations, any change in the spatial properties of the sound field over the observed duration of the received signals due to changes in the acoustic channel would likely degrade performance. Online enhancement algorithms, which update their parameters in response to changes in the signal characteristics, can mitigate such degradation but depend on the ability to track acoustic changes. An analysis of algorithm behaviour under non-stationary conditions is increasing in importance, as demonstrated by the CHiME3 challenge, but as yet there is no database of recorded speech under dynamic conditions in which the SNR can be controlled or with such a diverse range of reverberation conditions as was used in the current study.

The similarity of the results obtained with Fan noise and Ambient noise are perhaps not surprising given the stationary nature of building ventilation and computer fans, which was the primary source of noise in most of the simulated conditions. This raises the question of whether in future work it is sufficient to evaluate new algorithms on either one or the other.

6. Conclusions

The performance of the CHiME3 baseline ASR system has been evaluated in a diverse range of acoustic conditions using the ACE Challenge database of AIRs and noise, which goes beyond the conditions of the CHiME3 challenge. The benefit of speech enhancement processing has been clearly demonstrated, with rWERRs of up to 82%. The STOI metric has been shown to be well correlated with the rWERR and so can be helpful in evaluating the performance of novel algorithms prior to testing as part of an ASR system.

Acknowledgements

The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement nos. 609465 and ITN-GA-2012-316969.

This work was supported by the Engineering and Physical Sciences Research Council (Grant number EP/M026698/1).

Data Statement

The data used in this study were derived from publicly and commercially available databases which can be accessed at: http://www.ee.ic.ac.uk/naylor/ACEweb and https://catalog.ldc.upenn.edu/ldc93s6a.

References

Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. 65 (4), 943–950.

Barker, J., Marxer, R., Vincent, E., Watanabe, S., 2015. The third CHiME speech separation and recognition challenge: dataset, task and baselines. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 504–511.

Benesty, J., Chen, J., Huang, Y., 2008. Microphone Array Signal Processing. Springer-Verlag, Berlin, Germany.

Brandstein, M.S., Ward, D.B. (Eds.), 2001. Microphone Arrays: Signal Processing Techniques and Applications. Springer-Verlag, Berlin, Germany.

Breithaupt, C., Krawczyk, M., Martin, R., 2008. Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4037–4040. doi: 10.1109/ICASSP.2008.4518540.

Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Ito, N., Kinoshita, K., Espi, M., Araki, S., Hori, T., Nakatani, T., 2015. Strategies for distant speech recognition in reverberant environments. EURASIP J. Adv. Signal Process. 2015 (1), 1–15.

Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Ito, N., Kinoshita, K., Espi, M., Hori, T., Nakatani, T., Nakamura, A., 2014. Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge. In: Proceedings of REVERB Challenge Workshop, vol. 1. Florence, Italy, pp. 1–8.

Eaton, J., Gaubitch, N.D., Moore, A.H., Naylor, P.A., 2015. The ACE challenge – corpus description and performance evaluation. In: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). New Paltz, NY, USA. doi: 10.1109/WASPAA.2015.7336912.

Eaton, J., Gaubitch, N.D., Moore, A.H., Naylor, P.A., 2016a. ACE Challenge Results Technical Report. Technical Report. Imperial College London.

Eaton, J., Gaubitch, N.D., Moore, A.H., Naylor, P.A., 2016b. Estimation of room acoustic parameters: the ACE challenge. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (10), 1681–1693. doi: 10.1109/TASLP.2016.2577502.

Eaton, J., Gaubitch, N.D., Naylor, P.A., 2013. Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, Canada, pp. 161–165. doi: 10.1109/ICASSP.2013.6637629.

Garofalo, J., Graff, D., Paul, D., Pallett, D., 2007. CSR-I (WSJ0) Complete. Corpus LDC93S6A. Linguistic Data Consortium (LDC), Philadelphia.

Gaubitch, N.D., Ward, D.B., Naylor, P.A., 2006. Statistical analysis of the autoregressive modeling of reverberant speech. J. Acoust. Soc. Am. 120 (6), 4031–4039.

Haeb-Umbach, R., Krueger, A., 2012. Reverberant speech recognition. Techniques for Noise Robustness in Automatic Speech Recognition. John Wiley & Sons, pp. 251–281.

Hori, T., Chen, Z., Erdogan, H., Hershey, J.R., Roux, J.L., Mitra, V., Watanabe, S., 2015. The MERL/SRI system for the 3rd chime challenge using beamforming, robust feature extraction, and advanced speech recognition. In: Proceeding of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 475–481.

ITU-T, 1993. Objective Measurement of Active Speech Level. Recommendation P.56. International Telecommunications Union (ITU-T).

ITU-T, 2001. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. Recommendation P.862. International Telecommunications Union (ITU-T).

Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., Maas, R., 2013. The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech. In: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). New Paltz, NY, USA, pp. 1–4. doi: 10.1109/WASPAA.2013.6701894.

Knapp, C., Carter, G., 1976. The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. 24 (4), 320–327.

Lebart, K., Boucher, J.M., Denbigh, P.N., 2001. A new method based on spectral subtraction for speech de-reverberation. Acta Acoust. 87, 359–366.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9, 504–512. doi: 10.1109/89.928915.

Miyoshi, M., Kaneda, Y., 1988. Inverse filtering of room acoustics. IEEE Trans. Acoust. Speech Signal Process. 36 (2), 145–152.

Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B.-H., 2010. Speech dereverberation based on variance-normalized delayed linear prediction. IEEE Trans. Audio Speech Lang. Process. 18 (7), 1717–1731.

Naylor, P.A., Gaubitch, N.D. (Eds.), 2010. Speech Dereverberation. Springer, London.

Naylor, P.A., Gaubitch, N.D., Habets, E.A.P., 2010. Signal-based performance evaluation of dereverberation algorithms. J. Electr. Comput. Eng. 2010, 1–5. doi: 10.1155/2010/127513.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2125–2136. doi: 10.1109/TASL.2011.2114881.

Xiong, F., Moritz, N., Rehr, R., Anemuller, J., Meyer, B., Doclo, T.G.G., Goetze, S., 2014. Robust ASR in reverberant environments using temporal cepstrum smoothing for speech enhancement and an amplitude modulation filterbank for feature extraction. In: Proceedings of REVERB Challenge Workshop, vol. 1. Florence, Italy.

Yoshioka, T., Nakatani, T., 2012. Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening. IEEE Trans. Audio Speech Lang. Process. 20 (10), 2707–2720.

Yoshioka, T., Nakatani, T., Kinoshita, K., Miyoshi, M., 2010. Speech dereverberation and denoising based on time varying speech model and autoregressive reverberation model. In: Cohen, I., Benesty, J., Gannot, S. (Eds.), Speech Processing in Modern Communication. Springer, Berlin, Germany, pp. 151–182.