THE SECOND 'CHIME' SPEECH SEPARATION AND RECOGNITION CHALLENGE: DATASETS, TASKS AND BASELINES

Emmanuel Vincent
Inria
Nancy, France
emmanuel.vincent@inria.fr

Jon Barker
University of Sheffield
Sheffield, UK
j.barker@dcs.shef.ac.uk

Shinji Watanabe
Jonathan Le Roux
MERL
Cambridge, MA, USA

Francesco Nesta[*]
Marco Matassoni
FBK-Irst
Trento, Italy

ABSTRACT

Distant-microphone automatic speech recognition (ASR) remains a challenging goal in everyday environments involving multiple background sources and reverberation. This paper is intended to be a reference on the 2nd 'CHiME' Challenge, an initiative designed to analyze and evaluate the performance of ASR systems in a real-world domestic environment. Two separate tracks have been proposed: a small-vocabulary task with small speaker movements and a medium-vocabulary task without speaker movements. We discuss the rationale for the challenge and provide a detailed description of the datasets, tasks and baseline performance results for each track.

Index Terms— Noise-robust ASR, 'CHiME' Challenge

1. INTRODUCTION

Despite tremendous progress in close-microphone ASR for broadcast news, telephone speech or meeting speech, robust distant-microphone ASR in everyday environments remains a challenging goal. In parallel to research in the Speech and Language (SL) community, new techniques have emerged in the Audio and Acoustic Signal Processing (AASP) and Machine Learning for Signal Processing (MLSP) communities which are currently changing the face of robust ASR [1–3].

The 1st 'CHiME' Challenge [4] held in 2011 was the first concerted evaluation of ASR systems in a real-world domestic environment involving both reverberation and highly dynamic background noise made up of multiple sound sources. It differentiated itself from past noise-robust ASR challenges [5–8] by considering more realistic noise conditions and from concurrent source separation challenges [9] by assessing the results in terms of ASR. Thirteen systems were submitted [10–22] which cover a wide range of signal enhancement and robust acoustic modeling techniques. The absolute keyword accuracy achieved by the best system was only 3% below that of a human listener. Further analysis showed that multi-condition training and spatial enhancement are the most effective single strategies but that the resulting performance improvements are not additive and that careful combination of these and other strategies is needed for further improvement.

In order to maximize scientific insight, a number of simplifications were brought to the task so as to keep it tractable and ensure a diversity of submissions. A key decision was to focus on the realism of the noise background while employing an unrealistically simple target speech signal. Surveyed for their opinion, the challenge entrants highlighted three main additional dimensions of difficulty to be considered in future challenges: variability of speaker location, vocabulary size and speech naturalness. Indeed, ASR systems can be surprisingly sensitive to speaker location and it is well known that systems optimized for small vocabulary read speech often fail to scale to larger vocabulary spontaneous speech.

This paper is intended as a reference on the ongoing 2nd 'CHiME' Challenge supported by the IEEE AASP, MLSP and SL Technical Committees. We extend the difficulty of the 1st Challenge in the first two dimensions above, such that the target speech conditions become closer to those in [6, 7] but a realistic multisource noise background is retained as opposed to a single interfering speaker. In order to avoid too large an increase in difficulty, two separate tracks have been proposed: a small vocabulary task with small speaker movements and a medium vocabulary task without speaker movements.

The structure of the rest of the paper is as follows. In Section 2, we detail the creation of the datasets and define the tasks to be addressed. In Section 3, we describe the baseline recognizers provided together with the data and report their performance when trained either from clean, reverberated or noisy data. We conclude in Section 4.

2. DATASETS AND TASKS

The configuration considered by the 2nd 'CHiME' Challenge is that of speech from a single target speaker being binaurally recorded in a domestic environment. Three datasets are provided for each task: a training set, a development set and a test set. Following [4, 9], these data were generated by convolving clean speech signals with binaural room impulse responses (BRIRs) and mixing them with noise backgrounds.

[*]Now at Conexant Systems, Newport Beach, CA, USA.

2.1. Noise and BRIR recordings

The BRIRs and the noise backgrounds were recorded in the same domestic living room using two ear microphones built into a B&K head and torso simulator (HATS) placed at a fixed position [4]. About 14 h of noise backgrounds were collected in chunks of 0.5 to 1.5 h over a period of several days. These include the major sources of noise in a typical family home: concurrent speakers, TV, game console, footsteps, and distant noise from outside or from the kitchen. The BRIRs were measured via the usual sine sweep method for 121 different positions covering a horizontal square grid of 20 cm side centered on the position 2 m directly in front of the HATS, with a grid step of 2 cm. A fixed gain was applied to the estimated BRIRs so that the level after convolution approximately matched that of a human speaker at a natural conversational level.

2.2. Track 1: small vocabulary

As in the 1st Challenge, the small vocabulary task relies on the Grid speech corpus [23]. The target utterances are 6-word sequences read by 34 speakers of the form <command:4> <color:4><prepos.:4><letter:25><digit:10><adverb:4>, where the numbers in brackets indicate the number of choices per word. The task is to recognize the letter and digit tokens. Success is measured by the keyword recognition rate, that is the percentage of correctly recognized tokens.

The temporal placement of the utterances within the noise background was controlled in order to produce mixtures at 6 different ranges of signal-to-noise ratio (SNR): -6, -3, 0, 3, 6 and 9 dB[1]. This was achieved by randomly scanning the background recordings and picking a time interval in the desired SNR range for each utterance. In comparison to conventional robust ASR evaluations [5] which operate by rescaling the speech and noise signals, this mixing procedure is ecologically more valid, although it does not yet account for the Lombard effect as in [6, 7]. The backgrounds at 9 dB are dominated by quasi-stationary ambient sources, while those at -6 dB typically involve nonstationary, sudden sound events.

In order to make the task more realistic, the clean utterances were convolved with time-varying BRIRs mimicking small head movements within the aforementioned horizontal square grid. The parameterization of the movements was kept simple in order to allow analysis of the results as a function of the movement amplitude and speed: the target speaker was assumed to be static at the beginning of each utterance, then to move once, and finally to be static again. The movements follow a straight left-right line at fixed front-back distance from the HATS and each movement covers a distance of at most 5 cm at a speed of at most 15 cm/s. These movements were implemented by interpolating the set of recorded BRIRs in the same way as the Roomsimove toolbox[2]: for each front-back distance, a finer left-right grid of 2.5 mm step was designed and, for each point of this grid, the corresponding BRIRs were estimated by linear interpolation of the BRIRs recorded on the coarser 2 cm grid; each time sample of the clean speech signal was then convolved with the BRIRs associated with the point of the finer grid that is closest to the source position at that instant. This operation is an approximation of the true time-varying BRIRs. In order to validate this approximation, we conducted a simulation using non-binaural RIRs for simplicity. We generated the RIRs at each point of the finer grid using the source image method [24] assuming similar room geometry and reverberation time and we computed the worst-case modeling error achieved by linearly interpolating the filters over the whole grid, as measured in terms of SNR after convolution with a speech signal. We compared the results to the alternative interpolation procedure that consists of computing the FFT of the RIRs, unrolling phase according to the time delay of arrival, linearly interpolating phase and magnitude, and computing the inverse FFT in a way similar to [25]. Although phase interpolation yields perfect interpolation in the case of pure delay filters, the results in Fig. 1 show that linear interpolation performs better in the case of reverberant RIRs and that it achieves a worst-case SNR of 19 dB for a coarse grid step of 2 cm. Therefore, the modeling noise is at least 10 dB lower than the background noise, which is a reasonable approximation given that it is hardly feasible in practice to record RIRs on a finer grid.

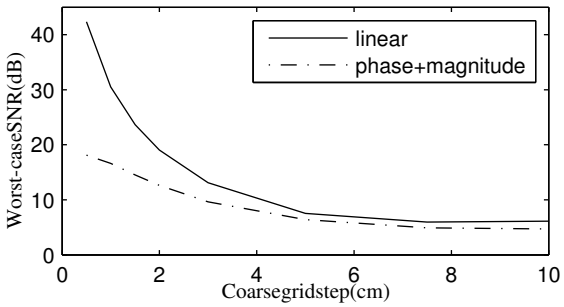

Fig. 1. Comparison of linear vs. phase interpolation for the simulation of time-varying RIRs.

We generated 600 noisy test utterances and 600 noisy development utterances at each of the 6 SNRs, where the same utterances are used for all SNRs and they do not overlap with each other within each dataset. We also provided a training set containing 500 utterances of each of the 34 Grid talkers in clean, reverberated and noisy conditions. All data were sampled at 16 kHz and are freely available under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 license both as isolated utterances and as embedded utterances including 5 s or more background before and after.

[1] In order to better match the perceived SNR, the SNRs were computed from high-pass filtered versions of the signals with a cutoff frequency of 80 Hz. Each SNR must be understood as a distribution of values with a standard deviation on the order of ± 1 dB.

[2] http://www.irisa.fr/metiss/members/evincent/Roomsimove.zip

2.3. Track 2: medium vocabulary

The medium vocabulary task relies on the Wall Street Journal (WSJ0) 5k vocabulary read speech corpus [26] also used in [5–7]. The task is to recognize all words. Success is measured in terms of word error rate (WER), that is the number of word substitutions, insertions and deletions as a fraction of the number of target words.

The data were mixed in the same way as the 1st 'CHiME' Challenge, i.e., clean speech utterances were convolved with the BRIRs corresponding to the fixed position directly in front of the HATS. Different recording conditions were used for the training, development and test datasets with the door being open/closed and the curtains being drawn/undrawn. The temporal placement of the utterances was controlled in a similar way as above in order to produce mixtures at the same range of SNRs. Since WSJ0 features longer utterances, the SNR for a given utterance was defined as the median value of the segmental SNR computed over segments of 200 ms. Also, due to the increased amount of data, nonoverlapping temporal placement of the utterances in the development and test sets was no longer feasible hence limited signal rescaling and overlap were allowed when necessary.

The development set includes 409 noisy utterances from 10 speakers, forming the "no verbal punctuation" (NVP) part of the WSJ0 speaker-independent 5k vocabulary development set. The test set comprises 330 noisy utterances from 12 other speakers, forming the Nov92 ARPA WSJ evaluation set. The training set includes 7138 reverberated utterances from 83 speakers forming the WSJ0 SI-84 training set. Both the development and the test utterances are released at each of the 6 SNRs while a noisy training set is provided by mixing each utterance at one random SNR, uniformly distributed in the defined range. All the noisy utterances are provided both in isolated and in embedded form. All data were sampled at 16 kHz and are available under agreement with the Linguistic Data Consortium (LDC).

2.4. Instructions

A set of instructions was provided in order to keep the task as close to an application scenario as possible, avoid unvoluntary overfitting and allow systems to be broadly comparable. The systems are allowed to exploit knowledge of the temporal placement of the utterances, of the surrounding background, of the speaker identity (for task 1) or of the speaker movements (also for task 1). However, they cannot exploit the SNR labels in the test set, the fact that the same utterances are used at each SNR, the fact that the same noise backgrounds are used in the development and test sets, the fact that the same utterances are used within the clean, reverberated and noisy training sets[3], the fact that the BRIRs are identical between different test utterances (for task 2) or the fact that the noise signals in the test utterances may temporally overlap (also for task 2). All parameters should be tuned on the provided training and development sets using the provided language models and the system should be run only once on the test set. Besides these rules, entrants are left entirely free in the development of their system, so as not to artificially disadvantage one research community over another.

[3]Note that this forbids so-called "stereo data" approaches, which assume the availability of synchronised clean and noisy data.

3. BASELINES

For each of the two tracks, a baseline ASR system based on HTK [27] was made available so as to lower the entry bar for researchers outside the SL community and demonstrate the performance achievable with neither signal enhancement nor advanced robust acoustic modeling techniques. This system includes both training and decoding scripts. Trained acoustic models were provided for clean, reverberated and noisy data. An alternative baseline ASR system based on Kaldi was made available for Track 2 and is separately described in [28].

3.1. Acoustic features

The speech waveforms are parameterized into a sequence of standard 39-dimensional Mel-frequency cepstral coefficient (MFCC) vectors: 12 Mel-cepstral coefficients processed by cepstral mean normalization (CMN), plus logarithmic frame energy and delta and acceleration coefficients (MFCC_E_D_A_Z). The MFCCs are extracted from 25 ms time frames with a step size of 10 ms. Prior to feature extraction, the input binaural signals are downmixed to mono by averaging the two channels together. Note that, although this downmixing operation leads to a small degradation of WER (1.5% on average) for Track 2 compared to the front end in [29], we decided to use it in order to allow comparison of the results across tracks and with the 1st 'CHiME' Challenge.

3.2. Small vocabulary recognizer

The baseline system for Track 1 is identical to that of the 1st 'CHiME' Challenge. Each of the 51 words in the Grid vocabulary is modeled with a left-to-right HMM with 2 states per phoneme. The emission probability for each state is represented as a Gaussian mixture model (GMM) with 7 components with diagonal covariance. The language model is fixed according to the Grid syntax.

Training proceeds in two stages: a speaker-independent model is first trained from a flat start using the full 17,000-utterance training set with HCompV, HERest and HHEd; a speaker-dependent model is then derived for each of the 34 speakers by applying further Baum-Welch iterations on the 500 utterances belonging to that speaker using HERest. Exact Viterbi decoding is performed using HVite.

Training	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
Clean	10.58	11.17	13.33	17.75	21.17	24.42
Reverb	32.17	38.33	52.08	62.67	76.08	83.83
Noisy	49.33	58.67	67.50	75.08	78.83	82.92

Table 1. Baseline test set keyword accuracy (%) for Track 1.

Training	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
Clean	93.56	93.05	91.67	89.05	84.79	79.21
Reverb	87.97	83.19	78.05	71.87	65.23	55.91
Noisy	70.43	63.09	58.42	51.06	45.32	41.73

Table 2. Baseline test set WER (%) for Track 2.

3.3. Medium vocabulary recognizer

The baseline system for Track 2 follows the recipe in [29]. The number of phonemes is 41: 39 phones plus 1 silence (sil) and 1 short pause (sp) model. The output distributions of sp and sil have their parameters tied. The number of clustered triphone HMM states is 1860 and is relatively smaller than the conventional setup (more than 2000 states). Each HMM has three output states with a left-to-right topology with self-loops and no skip. Each HMM state is represented by a GMM with 8 components for phoneme-based HMMs and 16 for silence-based HMMs. The standard WSJ 5K non-verbalized closed bigram language model is considered.

The provided training scripts only re-estimate the HMM-GMM parameters from a clean speech acoustic model, and do not change the model topology for simplicity. Decoding is performed using HVite with a pruning threshold.

3.4. Parameter tuning

We did not fine tune the features (0th MFCC vs. log-energy, other features than MFCCs), the acoustic model topology (triphone HMM clustering and number of GMM components) and the search parameters (language model weight, insertion penalty), as the optimal tuning is highly dependent on the enhancement technique used. Other LVCSR decoders could also be considered. Readers interested in providing suggestions or advice in this regard are welcome to contact us.

3.5. Baseline results

Tables 1 and 2 report the performance of the baseline systems trained on clean, reverberated or noisy data as a function of the SNR. The best results are achieved by training on noisy data for all SNRs except the 9 dB SNR condition in Track 1.

These figures must be compared to a keyword accuracy of 97.25% and 95.58% and to a WER of 7.49% and 18.40% when decoding the clean or reverberated utterances underlying the test set using clean or reverberated acoustic models respectively. While reverberation alone increases the error rate by a factor of 1.6 to 2.5, background noise further increases it by a factor of 2.3 to 11 depending on the SNR and it is therefore the main issue to be solved by the challenge entrants.

Training	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
Clean	13.25	12.33	17.08	20.75	27.50	34.08
Reverb	30.33	35.33	49.42	62.75	75.00	82.50

Table 3. Baseline test set keyword accuracy (%) for the 1st 'CHiME' Challenge.

Comparison with the baseline results of the 1st 'CHiME' Challenge in Table 3 shows the impact of speaker movements and vocabulary size. Speaker movements decrease the keyword accuracy by 4% on average with clean training, but they increase it by 2% with reverberated training due to the averaging effect they induce on the spectral differences between training and test. Larger vocabulary size increases the error rate by a factor of 1.3 to 3.2 with reverberated training depending on the SNR. Of course, the impact of speaker movements and vocabulary size may be different on more advanced systems and this is what the Challenge will seek to determine.

4. CONCLUSION

The series of 'CHiME' Challenges pursues the endeavor of evaluating robust ASR systems in real-world environments involving multisource background noise. The 2nd edition has increased the difficulty along two axes: small speaker movements and vocabulary size. Precise instructions have been provided to allow comparison of systems and maximize scientific insight. The submitted systems and the results will be unveiled at the 2nd 'CHiME' Workshop.

5. ACKNOWLEDGMENT

We would like to thank the IEEE AASP, MLSP and SL Technical Committees for supporting this challenge, and Conexant Systems, Audience, Adobe Systems, Google and MERL for sponsoring the 2nd 'CHiME' Workshop.

6. REFERENCES

[1] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind speech separation*, Springer, 2007.

[2] M. Wölfel and J. McDonough, *Distant Speech Recognition*, Wiley, 2009.

[3] D. Kolossa and R. Häb-Umbach, Eds., *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*, Springer, 2011.

[4] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.

[5] D. Pearce and H.-G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP*, 2000, vol. 4, pp. 29–32.

[6] J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, et al., "To separate speech!: A system for recognizing simultaneous speech," in *Proc. MLMI*, 2007, pp. 283–294.

[7] I. Himawan, I. McCowan, and M. Lincoln, "Microphone array beamforming approach to blind speech separation," in *Proc. MLMI*, 2007, pp. 295–305.

[8] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural Speech Separation and Recognition Challenge," *Computer Speech and Language*, vol. 24, pp. 94–111, 2010.

[9] E. Vincent, S. Araki, F. J. Theis, G. Nolte, et al., "The Signal Separation Evaluation Campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[10] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, et al., "Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation," in *Proc. CHiME*, 2011, pp. 12–17.

[11] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based speech enhancement and its application to noise-robust automatic speech recognition," in *Proc. CHiME*, 2011, pp. 53–57.

[12] A. Hurmalainen, K. Mahkonen, J. F. Gemmeke, and T. Virtanen, "Exemplar-based recognition of speech in highly variable noise," in *Proc. CHiME*, 2011, pp. 1–5.

[13] H. Kallasjoki, S. Keronen, G. J. Brown, J. F. Gemmeke, et al., "Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments," in *Proc. CHiME*, 2011, pp. 58–63.

[14] Y.-I. Kim, H.-Y. Cho, and S.-H. Kim, "Zero-crossing-based channel attentive weighting of cepstral features for robust speech recognition: The ETRI 2011 CHiME challenge system," in *Proc. Interspeech*, 2011, pp. 1649–1652.

[15] Z. Koldovský, J. Málek, J. Nouza, and M. Balík, "CHiME data separation based on target signal cancellation and noise masking," in *Proc. CHiME*, 2011, pp. 47–50.

[16] D. Kolossa, R. F. Astudillo, A. Abad, S. Zeiler, et al., "CHiME Challenge: Approaches to robustness using beamforming and uncertainty-of-observation techniques," in *Proc. CHiME*, 2011, pp. 6–11.

[17] N. Ma, J. Barker, H. Christensen, and P. Green, "Recent advances in fragment-based speech recognition in reverberant multisource environments," in *Proc. CHiME*, 2011, pp. 68–73.

[18] R. Maas, A. Schwarz, Y. Zheng, K. Reindl, et al., "A two-channel acoustic front-end for robust automatic speech recognition in noisy and reverberant environments," in *Proc. CHiME*, 2011, pp. 41–46.

[19] F. Nesta and M. Matassoni, "Robust automatic speech recognition through on-line semi blind source extraction," in *Proc. CHiME*, 2011, pp. 18–23.

[20] A. Ozerov and E. Vincent, "Using the FASST source separation toolbox for noise robust speech recognition," in *Proc. CHiME*, 2011, pp. 86–87.

[21] R. Vipperla, S. Bozonnet, D. Wang, and N. Evans, "Robust speech recognition in multi-source noise environments using convolutive non-negative matrix factorization," in *Proc. CHiME*, 2011, pp. 74–79.

[22] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME challenge contribution: NMF-BLSTM speech enhancement and recognition for reverberated multisource environments," in *Proc. CHiME*, 2011, pp. 24–29.

[23] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, pp. 2421–2424, 2006.

[24] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[25] H. Hacıhabiboğlu, B. Günel, and A. M. Kondoz, "Head-related transfer function filter interpolation by root displacement," in *Proc. WASPAA*, 2005, pp. 134–137.

[26] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete," Linguistic Data Consortium, Philadelphia, 2007.

[27] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, et al., *The HTK Book, version 3.4*, University of Cambridge, 2006.

[28] Y. Tachioka, S. Watanabe, and J. R. Hershey, "Effectiveness of discriminative training and feature transformation for reverberated and noisy speech," in *Proc. ICASSP*, 2013.

[29] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments," Tech. Rep., Cavendish Laboratory, University of Cambridge, 2006.