The CHiME challenges: Robust speech recognition in everyday environments

Jon Barker, Ricard Marxer, Emmanuel Vincent, Shinji Watanabe

Chapter 1
The CHiME challenges: robust speech recognition in everyday environments.

Jon Barker, Ricard Marxer, Emmanuel Vincent and Shinji Watanabe

Abstract The CHiME challenge series has been aiming to advance the development of robust automatic speech recognition for use in everyday environments by encouraging research at the interface of signal processing and statistical modelling. The series has been running since 2011 and is now entering its 4th iteration. This chapter provides an overview of the CHiME series including a description of the datasets that have been collected and the tasks that have been defined for each edition. In particular the chapter describes novel approaches that have been developed for producing simulated data for system training and evaluation, and conclusions about the validity of using simulated data for robust speech recognition development. We also provide a brief overview of the systems and specific techniques that have proved successful for each task. These systems have demonstrated the remarkable robustness that can be achieved through a combination of training data simulation and multicondition training, well-engineered multichannel enhancement and state-of-the-art discriminative acoustic and language modelling techniques.

Jon Barker
University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK e-mail: j.p.barker@sheffield.ac.uk

Ricard Marxer
University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK e-mail: r.marxer@sheffield.ac.uk

Emmanuel Vincent
Inria, France e-mail: emmanuel.vincent@inria.fr

Shinji Watanabe
MERL, US e-mail: watanabe@merl.com

1.1 Introduction

Speech recognition technology is becoming increasingly pervasive. In particular, it is now being deployed in home and mobile consumer devices where it is expected to work reliably in noisy, everyday listening environments. In many of these applications the microphones are at a significant distance from the user, so the captured speech signal is corrupted by interfering noise sources and reverberation. Delivering reliable recognition performance in these conditions remains a challenging engineering problem.

One of the commonest approaches to distant speech recognition is to use a multichannel microphone array. Beamforming algorithms can then be used to capture the signal from the direction of the target talker while suppressing spatially distinct noise interferers. Although beamforming is a mature technique, the design and evaluation of algorithms is often performed by signal processing researchers optimising speech enhancement objectives. Conversely, builders of speech recognition systems are often disappointed when they try to use beamforming algorithms 'off the shelf' with little idea how to properly optimise them for recognition.

The CHiME challenges were designed with the goal of building a community of researchers that would span signal processing and statistical speech recognition and make progress to robust distant microphone speech recognition through closer collaboration. They were also prompted by a perceived gap in the speech recognition challenge landscape. Most challenges were being designed around lecture hall or meeting room scenarios where, although there might be considerable reverberation, the environment is essentially quiet, e.g., [19, 24, 25]. Other challenges model more extreme noise levels, but these typically use artificially mixed-in noise and pre-segmented test utterances thus providing no opportunity to learn the structure of the noise background or to observe the noise context prior to the utterance, e.g., [14, 21, 26]. In contrast, the CHiME challenges were designed to draw attention toward the noise background by providing speech embedded in continuous recordings and accompanied by considerable quantities of matched noise-only training material.

The 1st CHiME challenge was launched in 2011 and the series is now entering its 4th iteration. Over that time the challenges have developed from small highly-controlled tasks towards more complex scenarios with multiple dimensions of difficulty and greater commercial realism. This chapter provides an account of this development providing a full description of the task design for each iteration and an overview of findings arising from analysis of challenge systems.

1.2 The 1st and 2nd CHiME challenges (CHiME-1 and CHiME-2)

The 1st and 2nd CHiME challenges [5, 33] were conducted between 2011 and 2013 and were both based on a 'home automation' scenario, involving the recognition of command-like utterances using distant microphones in a noisy domestic environment. They both used simulated mixing allowing the choice of speech and background materials to be separately controlled.

1.2.1 Domestic noise background

The noise backgrounds for the 1st and 2nd CHiME challenges were taken from the CHiME Domestic Audio dataset [6]. This data consists of recordings made in a family home using a B & K head and torso simulator type 4128 C. The head has built-in ear simulators that record a left and right ear signal that approximate the signals that would be received by an average adult listener.

The CHiME challenge used recordings taken from a single room – a family living room – over the course of several weeks. The living room recordings were made during 22 separate morning and evening sessions typically lasting around one hour each and totalling over 20 hours. The manikin remained in the same location throughout. Major noise sources are those that are typical of a family home: television, computer games console, children playing, conversations, some street noise from outside and noises from adjoining rooms including washing machine noise and general kitchen noise.

The Domestic Audio dataset is also distributed with binaural room impulse response (BRIR) measurements that were made in the same recording room. The BRIRs were estimated using the sine sweep method [11] from a number of locations relative to the manikin. For each location several BRIR estimates were made. For the particular location 2 m directly in front of the manikin (i.e., at an azimuthal angle of 0°) estimates were made with variable 'room settings': with a set of floor-length bay window curtains opened or closed, and with the door to the adjoining hallway open or closed.

1.2.2 The speech recognition task design

The 1st and 2nd CHiME challenges (CHiME-1 and CHiME-2) both employed artificial mixing of the speech and noise background in order to carefully control the target signal-to-noise ratio. CHiME-1 used a small vocabulary task and a fixed speaker location. CHiME-2 had two tracks extending CHiME-1 in two separate directions: speaker motion and vocabulary size. In all tasks utterances were embedded

within complete unsegmented CHiME Domestic Audio recording sessions. Participants were supplied with the start and end times of each test utterance (i.e., speech activity detection was not part of the task) and they were also allowed to make use of knowledge of the surrounding audio before and after the utterance, (e.g., to help estimate the noise component of the speech and noise mixture.)

1.2.2.1 CHiME-1: Small vocabulary

CHiME-1 was based on the small vocabulary Grid corpus task[7]. This is a simple command sentence task that was initially designed for measuring the robustness of *human* speech recognition in noisy conditions. The corpus consists of 34 speakers (18 male and 16 female) each uttering 1000 unique 6-word commands with a simple fixed grammar. Each utterance contains a letter-digit grid reference. These two words are considered as the target keywords and performance is reported in terms of keyword correctness.

The Grid data was split such that 500 utterances per speaker were designated as training data and the remaining utterances were set aside as test data. From the test data, test sets of 600 utterances (about 20 utterances per speaker) were defined. To form noisy test utterances, Grid test set speech was convolved with the CHiME BRIRs and then added to a 14 hour subset of the CHiME background audio. Temporal locations were selected such that the 600 utterances did not overlap and such that the mixtures had a fixed target SNR. By varying the temporal locations it was possible to achieve test sets with SNRs of -6, -3, 0, 3 and 6 dB. Separate test sets were produced for development and final evaluation. The final evaluation test set was released close to the deadline for submitting final results.

For training purposes participants were supplied with a reverberated version of the 17,000 utterance CHiME training set, plus a further 6 hours of background recording. The background audio was from the same room but made up from different recording sessions to those that had been used for the test data. Likewise, a different instance of the 2 m and 0° BRIR was used. No restrictions were placed on how this data could be used for system training.

1.2.2.2 CHiME-2 Track 1: Simulated motion

CHiME-2 Track 1 was designed in response to criticism that the fixed impulse responses used in CHiME-1 made the task too artificial. To test this claim, variability was introduced into the training and test set BRIRs. Specifically, the effect of small speaker movements was simulated. To do this, a new set of BRIRs were recorded on a grid of locations around the 2 m and 0° location used for CHiME-1. The grid had a size of 20 cm by 20 cm grid and a 2 cm resolution requiring a total of 121 (i.e., 11×11) BRIR measurements.

To simulate motion, first interpolation was used to increase the resolution of the BRIR grid in the left-right direction down to a 2.5 mm step size. Then for each utter-

ance a random straight line trajectory was produced such that the speaker moved at a constant speed of at most 15 cm/s over a distance of at most 5 cm within the grid. Then each sample of the clean utterances was convolved with the impulse response from the grid location that was closest to the speaker at that instant.

As with CHiME-1, a 17,000 utterance training set was provided and separate 600 utterance development and final test sets. All utterances had simulated motion. The test sets were produced at the same range of SNRs as CHiME-1.

1.2.2.3 CHiME-2 Track 2: Medium vocabulary

CHiME-2 Track 2 extended CHiME-1 by replacing the small vocabulary Grid task with the medium vocabulary 5,000 word Wall Street Journal (WSJ) task . The data were mixed in the same way as per CHiME-1 with a fixed BRIR at 2 m directly in front of the manikin. As with CHiME-1 different instances of the BRIR were used for training, development and final test sets. SNRs were defined as the median value of segmental SNRs computed over 200 ms windows to be compatible with SNRs used in other WSJ tasks. It was found that because the WSJ utterances are longer than Grid utterances there were fewer periods of CHiME background where low SNRs could be sustained. Hence, some signal rescaling had to be employed to obtain the lowest SNRs. Also it was not possible to follow the rule that temporal locations should be chosen such that test utterances would not share some portions of the background.

The training, data consists of 7138 reverberated utterances from 83 speakers forming the WSJ0 SI-84 training set. Development data is 409 utterances from the 10 speakers forming the "no verbal punctuation" part of the WSJ0 speaker-independent 5k development set. The final test set was 330 utterances from 12 different speakers forming the Nov92 ARPA WSJ evaluation set.

1.2.3 Overview of system performances

The CHiME-1 challenge attracted participation from 13 teams. A broad range of strategies were employed that could be grouped under target enhancement, robust feature extraction and robust decoding. A full review of the systems can be found in [5]. Generally, systems that delivered the best performance successfully combined an enhancement stage (exploiting both spatial and spectral diversity) and a robust decoder, either using some form of uncertainty propagation, an adapted training objective (e.g., MLLR, MAP, bMMI) or simply a multi-condition training strategy using speech plus background mixtures.

For comparison, the challenge was published with a 'vanilla' HMM/GMM baseline system trained on the reverberated speech with MFCC features. This non-robust system scored 82% keyword correctness at 9 dB with performance falling to 30% at -6 dB. Listening tests established Human performance to be 98% at 9 dB and falling

to 90% at -6 dB. Scores for the submitted systems were broadly spread between the non-robust baseline system and the human performance. The overall best performing system [8] made only 57% more errors than the human with correctness varying between 86% and 96% across the SNR range. Analysis of the top-performing systems indicated that the most important strategies were multi-condition training, spatial diversity-based enhancement and robust training.

CHiME-2 outcomes are reviewed in [32]. CHiME-2 Track 1 attracted participation from 11 teams with some overlap with the earlier CHiME-1 challenge. The top performing team achieved a score very similar to the best performance achieved on CHiME-1. Further, teams that made a direct comparison between CHiME-1 and CHiME-2 achieved equal scores on both tasks. It was concluded that the simulated small speaker movements caused little extra difficulty. Track-2 received only 4 entrants with one clear top performer [29] achieving WERs ranging from 14.8% to 44.1% from 9 dB to -6 dB SNR. Achieving this performance required a highly optimised system using spatial enhancement, a host of feature-space transformations, a decoder employing discriminative acoustic and language models and ROVER combination of system variants. Spectral diversity based enhancement, that had performed extremely well in the small vocabulary task, was found to be less useful in Track 2.

1.2.4 Interim conclusions

The CHiME-1 and CHiME-2 challenges clearly demonstrated that distant microphone ASR systems need careful optimisation of both the signal processing and the statistical back-end. However, the challenge design left several questions unanswered.

Are results obtained on artificially mixed speech representative of performance on real tasks? The artificial mixing is useful in allowing SNRs to be carefully controlled but it raises questions about the realism of the data. First, the challenges use studio recorded speech from the Grid and WSJ corpora. Although this speech is convolved with room impulses responses to model the effects of reverberation, speech read in a studio environment will differ in other significant ways from speech spoken and recorded live in noise. Second, it is likely that the range of SNRs used is not representative of SNRs observed in real distant microphone speech applications. Third, the simulation does not capture the channel variability of real acoustic mixing, where many factors will have an impact on the BRIRs.

How can evaluation be designed so as to allow fairer cross-team comparisons? One problem with both CHiME-1 and CHiME-2 was that the lack of a state-of-the-art baseline left every team to develop systems from the ground up. This led to an interesting diversity of approaches but reduced the opportunity for scientifically controlled comparison. Further, although the noise background training data was specified, there were no restrictions on how it could be used. Systems that employed multi-condition training and generated larger noisy training datasets had higher per-

formance. Tighter control of the training conditions could have allowed for more meaningful comparison.

1.3 The 3rd CHiME challenge (CHiME-3)

The 3rd CHiME challenge was designed in response to feedback from the earlier challenges. Several priorities were identified. First, the domestic setting of the earlier challenge had been consider rather narrow and there was desire to broaden the range of noise environments. Second, it was decided to move away from the binaural microphone configuration towards a more traditional microphone array setting that would have greater commercial relevance. The CHiME-3 scenario was therefore chosen to be that of an automatic speech recognition application running on a mobile device that would be used in noisy everyday settings. In order to make the task challenging, a lap-held tablet computer was selected as the target device for which it was estimated microphone distances would be in the range of 30-40 cm (i.e., considerably greater than typical distance for mobile phone usage). Finally, to answer questions that had been raised about the validity of using simulated mixing for training and testing systems, a direct 'simulated versus real' data comparison was built into the task design.

1.3.1 The mobile tablet recordings

The CHiME-3 speech recordings were made using a 6-channel microphone array constructed by embedding Audio-technica ATR3350 omnidirectional lavalier microphones around the edge of a frame designed to hold a Sumsung Galaxy tablet computer. The array was designed to be held in landscape orientation with three microphones positioned along the top and bottom edges as indicated in Figure 1.1. All microphones faced forward except the top-central microphone which faced backward and was mounted flush with the rear of the frame.

The microphone signals were recorded sample-synchronously using a 6-channel TASCAM DR-680 portable digital recorder. A second TASCAM DR-680 was used to record a signal from a Beyerdynamic condenser close-talking microphone (CTM). The recorders were daisy-chained together to allow their transports to be controlled via a common interface. There was a variable delay between the units of up to 20 ms. All recordings were made with 16 bits at 48 kHz and later downsampled to 16 kHz.

Speech was recorded for training, development and test sets. Four native US talkers were recruited for each set (two male and two female). Speakers were instructed to read sentences that were presented on the tablet PC while holding the device in any way that felt natural. Each speaker recorded utterances first in an IAC single-walled acoustically isolated booth and then in each of the following environments:

on a bus (BUS), on a street junction (STR), in a café (CAF) and in a pedestrian area (PED). Speakers were prompted to change their seating/standing position after every ten utterances. Utterances that were misread or read with disfluency were re-read until a satisfactory rendition had been recorded.

1.3.2 The CHiME-3 task design: real and simulated data

The task was based on the WSJ0 5K ASR task, i.e., remaining comparable with CHiME-2 track 2. For the training data, 100 utterances were recorded by each speaker in each environment, totalling 1600 utterances selected at random from the full 7138 WSJ0 SI-84 training set. Speakers assigned to the 409 utterance development set or the 330 utterance final test set each spoke a 1/4 of each set in each environment resulting in 1636 (4×409) and 1320 (4×330) utterances for development and final testing respectively.

The live-recorded training data was supplemented with 7138 simulated noisy utterances constructed by artificially adding the WSJ training set to a separately recorded 8 hours of noise background (2 hours from each of the environments). Techniques for simulation were included as part of the baseline described in the next section. Participants were encouraged to try and improve on the baseline simulation technique under the assumption that reducing mismatch between simulated training data and real test data would lead to better ASR performance. In order to extend the scientific outcomes of the challenge, a simulated development and test set was also produced. Given that previous CHiME challenges had used only simu-

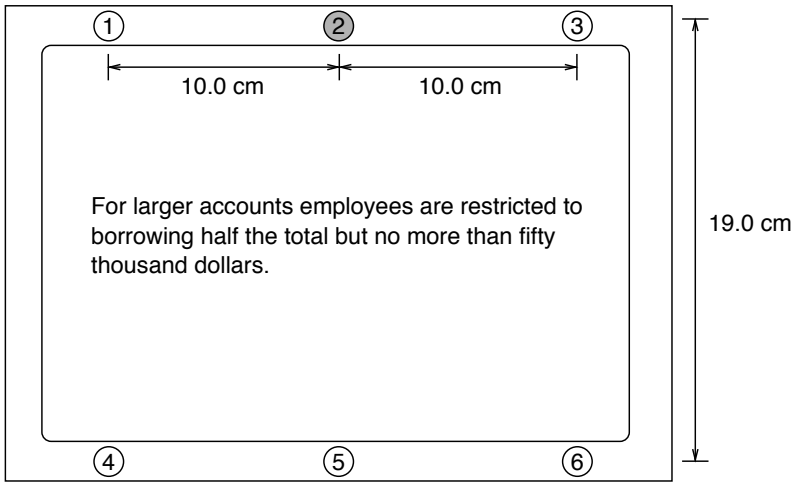

Fig. 1.1 The geometry of the 6-channel CHiME-3 microphone array. All microphones are forward facing except for channel 2 (shaded gray) which faces backwards and is flush with the rear of the 1 cm thick frame.

lated data, it was important to know whether the performance of a system evaluated using simulated data is a good predictor of performance on real data.

Additional rules were imposed in order to keep systems as comparable as possible. Chiefly, participants were asked to tune system parameters using only the development data and to report results on the final test data. Any language model was allowed as long as it was trained from official WSJ language model training data. New simulation techniques for training data were not allowed to expand the amount of training data and had to keep the same pairing between utterances and segments of noise background. A constraint of 5 seconds was placed on the amount of audio context that could be used preceding an utterance.

1.3.3 The CHiME-3 baseline systems

The CHiME-3 challenge was distributed alongside baseline systems for training data simulation, multichannel speech enhancement and automatic speech recognition. These systems are outlined below and are described in greater detail in [4].

1.3.3.1 Simulation

The simulation baseline software was designed for adding clean WSJ speech to microphone array noise recordings in such a way as to model the effects of speaker and tablet motion. The procedure for mixing is performed in two stages.

First, a set of STFT-domain time-varying impulse responses (IR) between the close-talking microphone (considered to be clean speech) and each of the other microphones are estimated in the least-squares sense. Estimates are made in each frequency bin and in blocks of frames partitioned such that each partition contains a similar amount of speech. The SNR at each tablet microphone can then be estimated.

In a second stage, the spatial position of the speaker is tracked in each of the CHiME training data recordings. To do this signals are first represented in the complex-valued STFT domain using 1024 sample, half-overlapped sine windows. The position of the speaker is encoded by a nonlinear SRP-PHAT pseudo-spectrum. The peaks of the pseudo-spectrum are tracked using the Viterbi algorithm. A time-varying filter modelling direct sound between the speaker and the microphones is then constructed.

Original WSJ training utterances are then convolved with filters estimated from CHiME training utterances. An additional equalization filter is applied that is estimated as the ratio of the average power spectrum of CHiME booth recordings and the average power spectrum of WSJ training data. Finally, the equalised recordings are rescaled to match the estimated real training data SNRs and are then mixed with noise backgrounds taken from the separate 8 hour set of noise-only recordings.

1.3.3.2 Enhancement

The baseline enhancement system is designed to take the 6 channel array recordings and produce a single channel output with reduced background noise, suitable for input into the ASR system.

The baseline system is based on a minimum variance distortionless response (MVDR) beamforming approach. The target talker is tracked using the peaks in the nonlinear SRP-PHAT pseudo-spectrum (as used in the simulation component). The multichannel noise covariance matrix is estimated from 400 ms to 800 ms of context prior to the utterance. MVDR with diagonal loading is then employed to estimate the target speech spectrum.

Some of the CHiME test recordings are subject to microphone failures. These can be caused by microphone occlusion during handling, or vibrations leading to intermittent connection failures (particularly in the BUS environment). The baseline system applied a simple energy-based criteria to detect microphone failure and ignore failed channels.

1.3.3.3 ASR

Two Kaldi-based ASR baseline systems were provided: a lightweight GMM/HMM system for rapid experimentation and a state-of-the-art DNN baseline for final benchmarking.

The GMM/HMM system employed 13th order MFCCs to represent individual frames. Feature vectors were then formed by concatenating three frames of left and right context and compressing to 40 dimensions using linear discriminative analysis with classes being one of 2500 tied tri-phone HMM states. A total of 15,000 Gaussians were used to model the tied states. The system also employed maximum likelihood linear transformation and feature-space maximum likelihood linear regression with speaker adaptive training.

The DNN baseline employed a network with 7 layers and 2048 units per hidden layer. Input was based on the 40-dimensional filterbank frames with 5 frames of left and right context (i.e., a total of $11 \times 40 = 440$ input units). The DNN is trained using standard procedures described in [31]: pretraining using restricted Boltzmann machines; cross entropy training; sequence discriminative training using the state-level minimum Bayes risk criterion.

1.4 The CHiME-3 evaluations

A total of 26 systems were submitted to the CHiME-3 challenge all of which achieved a lower test set WER than the 33.4% scored by the baseline DNN system. This section presents the performance of the top systems and provides an overview of the strategies that were most effective for reducing WERs.

1.4.1 An overview of CHiME-3 system performances

Table 1.1 presents the results of the top 10 overall best systems. Most of the best systems achieved WERs in the range 13% down to 10%. The overall best system achieved a WER of just 5.8%, significantly better than the 2nd placed system. The table also shows WERs broken down by noise environment. For most systems the highest WERs were observed in the BUS environment and the lowest in STR, with CAF and PED lying somewhere in between. However, there are notable exceptions, for example the best system has a WER of just 4.5% in the CAF environment.

Table 1.1 Overview of the top 10 systems submitted to the CHiME-3 Challenge. The left side of the table summarizes the key features of each system indicating where systems have differed from the baseline with respect to training (Tr); multichannel enhancement (ME); single channel enhancement (ME); feature extraction (FE); feature transformation (FT); acoustic modelling (AM); language modelling (LM); system combination (SC). The right hand side reports WERs for the final test set overall (Ave.) and for each environment individually. Results are shown for the real data test set only. For performance on the simulated data see [4].

System	Tr	ME	SE	FE	FT	AM	LM	SC	BUS	CAF	PED	STR	Ave.
Yoshioka et al. [36]	X	X		X		X	X		7.4	4.5	6.2	5.2	5.8
Hori et al. [15]		X	X	X	X		X	X	13.5	7.7	7.1	8.1	9.1
Du et al. [9]		X		X	X	X		X	13.8	11.4	9.3	7.8	10.6
Sivasankaran et al. [27]	X	X		X	X		X		16.2	9.6	12.3	7.2	11.3
Moritz et al. [18]	X			X	X		X		13.5	13.5	10.6	9.2	11.7
Fujita et al. [12]		X	X	X	X			X	16.6	11.8	10.0	8.8	11.8
Zhao et al. [37]	X	X		X	X				14.5	11.7	11.5	10.0	11.9
Vu et al. [34]		X	X		X		X		17.6	12.1	8.5	9.6	11.9
Tran et al. [30]	X	X		X	X		X		18.6	10.7	9.7	9.6	12.1
Heymann et al. [13]	X	X					X		17.5	10.5	11.0	10.0	12.3
DNN Baseline v2		X			X		X		19.1	11.4	10.3	10.3	12.8
DNN Baseline									51.8	34.7	27.2	20.1	33.4

1.4.2 An overview of successful strategies

Analysis of results shows that no single technique is sufficient for success. Systems near the top of the table modified multiple components whereas systems that improved one or two components performed consistently poorly. Best performance required a combination of improved multichannel processing, good feature normalisation and improvement to the baseline language model. The most commonly employed strategies are reviewed briefly below.

1.4.2.1 Strategies for improved signal enhancement

Good target enhancement is crucial for success and nearly all teams attempted to improve this component of the baseline. Many teams replaced the baseline's super-directive MVDR beamformer with a conventional delay and sum beamformer, e.g., [15, 27, 23]. Others kept the MVDR framework but tried to improve the estimates of the steering vector [36], or the speech and noise covariances [13]. Another popular strategy was to add a post-filter stage, for example spatial coherence filtering [20, 3] or dereverberation [36, 10]. A smaller number of teams used additional single channel enhancement stages after the array processing, e.g., NMF based source separation [34, 2], but these approaches were found to have a more marginal benefit.

1.4.2.2 Strategies for improved statistcal modelling

Most teams adopted the same feature design as the baseline design, i.e., MFCC features for the initial alignment stages followed by filterbank features for the DNN pass. However, good speaker/environment normalisation was found to be important. Whereas the baseline only applied explicit speaker normalising transforms for the HMM/GMM training, it was found that it was also advantageous to improve normalisation for DNN training. Strategies included performing utterance-based feature mean and variance normalisation [9, 12, 37, 35] and augmenting DNN inputs with pitch-based features [9, 35, 16]. The most successful strategies were found to be fMLLR [15, 27, 18, 34, 30] and augmentation of DNN inputs with either i-vectors [18, 38] or bottleneck features extracted from a speaker classification DNN [28]. Using both fMLLR and feature vector augmentation provided additive benefits [38, 28, 20].

For acoustic modeling most teams adopted the DNN architecture provided by the baseline system. Notable alternatives included convolutional neural networks, e.g. [35, 16, 2] and Long Short Time Memory (LSTM) networks, e.g. [20, 2, 17]. A comparison of submission performances did not demonstrate any clear advantage to any particular architecture, and indeed, some of the best systems employed the baseline architecture. Where alternative architectures were employed they were often used in combination, e.g. [36, 9, 38].

Most teams implemented a language model rescoring stage using a more sophisticated model than the 3-gram model used by the baseline decoder. All teams doing so were able to achieve significant performance enhancements. Language models used for rescoring included DNN-LMs [34], LSTM-LMs [10] or, most commonly, recurrent neural network language models (RNN-LMs) [36, 27, 28, 22].

1.4.2.3 Strategies for improved system training

The CHiME-3 challenge was designed so as to let teams experiment with training data simulation. It was stressed that the simulation technique used to make the simulated training data was to be considered as a baseline and the MATLAB source code was made available to all participants. Rules allowed the WSJ and noise background to be remixed as long as each training utterance remained paired with the same segment of noise background.

Despite encouragement, few teams attempted to experiment with alterations to the training data. The only exception were [13] and [35] who achieved significant performance improvements by remixing the training data at a range of SNRs. Although within the rules, this increases the training data quantity rather than just the quality. One other team [27] generated simulated training data in the feature domain using a condition restricted Boltzmann machine but failed to achieve better results. A number of teams generated an expanded training set by simply applying feature extraction directly to the individual channels (i.e., rather than first combining them into a single enhanced signal) [36, 18, 37, 38]. Surprisingly, this produced consistent improvements in performance despite the mismatch between the individual channels and the enhanced signals used for testing.

Techniques for improved training data simulation have remained largely unexplored. Given the relative simplicity of the baseline simulation there is potential for significant advancements in this area.

1.4.3 Key findings

Analysis of CHiME-3 systems indicates that to achieve the highest scores requires complex systems applying multiple recognition passes and the possible combination of multiple feature extractors and classifiers. However, the largest consistently observed gains over the baseline came from three commonly applied techniques. First, replacing the MVDR beamformer with a delay and sum beamformer. (Teams taking this step were using the BeamformIt toolkit beamformer implementation [1] and therefore improvement in WER may be partly due to the manner in which BeamformIt implicitly weights microphones according to their correlation hence making it robust to microphone failures, in addition to the difference between MVDR and delay and sum.) Second, providing better speaker and environment normalisation by employing fMLLR transformed features for training the DNN. Third, adding a language model rescoring stage using a more complex language model, e.g., either a 5-gram model or an RNN language model.

After the challenge, a new baseline system was built that incorporated these three changes. This reduced the baseline WER from 33.4% to 12.8% making it competitive with the top 10 systems (see the row labeled ‘DNN Baseline v2’ in Table 1.2). This system has now replaced the original baseline as the official CHiME-3 baseline distributed with Kaldi.

A secondary goal of the challenge was to investigate the utility of simulated multichannel data either for training systems or for evaluation. Regarding acoustic modeling of noisy data, where comparisons were made it was found that using the simulated data always improved results compared to using real data alone, despite possible mismatch. However, some care is needed with the microphone array processing of the simulated data. The simple nature of the mixing means that array processing that has been optimised for the simulated data can produce overly optimistic enhancements, i.e., enhancements in which the SNRs are not representative of the SNRs that will be achieved when enhancing the real data. This mismatch can lead to poorer system performance and may explain why remixing the simulated training with a broader range of SNRs was beneficial. The problem could be fixed in a more principled fashion by improving the simulation itself, however few teams attempted this so there is more work to be done before conclusions can be drawn.

Finally, considering simulated test data: [4] presents the correlation between system performance on the real and simulated test set across all 26 systems submitted to the challenge. Although the correlation is strong there were observed to be many outlier systems, in particular, systems which achieved very low WERs on the simulated data but proportionally poorer WERs on the real data. This result suggests that extreme caution is needed when interpreting the results of fully simulated challenges.

1.5 Future directions: CHiME-4 and beyond

Although significant progress has been made, distant microphone speech recognition still remains a significant challenge. For modern everyday applications, that are expected to work in a wide variety of noise environments, the root of the problem is potential mismatch between training and test data: it is not possible to anticipate the acoustic environment in which the device will be used when training the system. The CHiME challenges reviewed in this chapter have highlighted two key distant microphone ASR strategies that can address this mismatch problem. First, microphone array processing, which reduces the potential for mismatch by the degree to which it successfully removes noise from the signal. Second, multicondition training which reduces mismatch to the extent that the noise environment can be successfully anticipated.

The solutions to mismatch seen in CHiME systems have proved remarkably effective, particularly in CHiME-3. However, although efforts were made to increase the realism of the evaluation, the challenge design significantly under-represents the degree of mismatch that real systems will have to handle. First, the training and test speech both come from the same narrow and well-represented domain for which it is possible to build well-matched language and acoustic models. Second, the training data has been recorded on the exact same device that is used for testing. This means that not only is the microphone array geometry matched, but so too are the individual microphone channels. Third, the noise environments, although more var-

ied than those employed in many challenges, still only represent four rather narrow situations. Again, the same noise environments were employed in both the training and the test data.

Table 1.2 A summary of the CHiME challenge tasks.

Edition		Channels	Noise	Task	Mixing	SNR (dB)	Distance
CHiME-1		Binaural	Domestic	Grid	Simu static	-6 to 9	2 m
CHiME-2	Track 1	Binaural	Domestic	Grid	Simu moving	-6 to 9	2 m
	Track 2	Binaural	Domestic	WSJ 5K	Simu static	-6 to 9	2 m
CHiME-3		6	Urban	WSJ 5K	Real/Simu	-5 to 0	30-40 cm
CHiME-4	1,CH	1	Urban	WSJ 5K	Real/Simu	-5 to 0	30-40 cm
	2-CH	2	Urban	WSJ 5K	Real/Simu	-5 to 0	30-40 cm
	6-CH	6	Urban	WSJ 5K	Real/Simu	-5 to 0	30-40 cm

One of the aims of the Fred Jelinek Workshop presented in this book was to develop novel solutions to the mismatch problem. In order to emphasise mismatch novel evaluation protocols were developed, in particular cross-corpora evaluation in which the training data from one corpus (e.g., AMI) would be used for building systems to be tested with data from another (e.g., CHiME-3). Inspired by this work, a new iteration of the CHiME challenge (CHiME-4) is now in progress. This iteration will use the same datasets that were constructed for CHiME-3 but has taken two steps towards increasing the mismatch challenge. First, 1 and 2 channel tracks are being introduced that will reduce the opportunity for noise removal in the enhancement stage. Second, the 1 and 2 channel tasks will employ different channel subsets for training and testing.

To conclude, a summary of all the datasets and tasks comprising the complete CHiME challenge series is presented in Table 1.2. The datasets for all CHiME editions are publicly available and state-of-the-art baseline systems are distributed with the Kaldi speech recognition toolkit.[1]

References

1. Anguera, X., Wooters, C., Hernando, J.: Acoustic beamforming for speaker diarization of meetings. IEEE Transactions on Audio, Speech and Language Processing **15**(7), 2011–2023 (2007)
2. Baby, D., Virtanen, T., Van Hamme, H.: Coupled dictionary-based speech enhancement for CHiME-3 challenge. Tech. Rep. KUL/ESAT/PSI/1503, KU Leuven, ESAT, Leuven, Belgium (2015)
3. Barfuss, H., Huemmer, C., Schwarz, A., Kellermann, W.: Robust coherence-based spectral enhancement for distant speech recognition (2015). ArXiv:1509.06882

[1] Instructions for obtaining CHiME datasets can be found at http://spandh.dcs.shef.ac.uk/chime.

4. Barker, J., Marxer, R., Vincent, E., Watanabe, S.: The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015, Scottsdale, AZ, USA, December 13-17, 2015, pp. 504–511 (2015). DOI 10.1109/ASRU.2015.7404837
5. Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P.: The PASCAL CHiME speech separation and recognition challenge. Computer Speech and Language **27**(3), 621–633 (2013)
6. Christensen, H., Barker, J., Ma, N., Green, P.: The CHiME corpus: a resource and a challenge for Computational Hearing in Multisoure Environments. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010). Makuhari, Japan (2010)
7. Cooke, M., Barker, J., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. Journal of the Acoustical Society of America **120**(5), 2421–2424 (2006). DOI 10.1121/1.2229005
8. Delcroix, M., Kinoshita, K., TNakatani, o., Araki, S., Ogawa, A., Hori, T., Watanabe, S., Fujimoto, M., Yoshioka, T., Oba, T., Kubo, Y., Souden, M., Hahm, S.J., Nakamura, A.: 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. 1st CHiME workshop on Machine Listening in Multisource Environments, pp. 12–17. Florence, Italy (2011)
9. Du, J., Wang, Q., Tu, Y.H., Bao, X., Dai, L.R., Lee, C.H.: An information fusion approach to recognizing microphone array speech in the CHiME-3 challenge based on a deep learning framework. In: IEEE ASRU, Scottsdale, AZ, USA, December 13-17, 2015, pp. 430–435 (2015)
10. El-Desoky Mousa, A., Marchi, E., Schuller, B.: The ICSTM+TUM+UP approach to the 3rd CHiME challenge: Single-channel LSTM speech enhancement with multi-channel correlation shaping dereverberation and LSTM language models (2015). ArXiv:1510.00268
11. Farina, A.: Simultaneous measurement of impulse response and distortion with a swept sine technique. In: Proc. 108th AES convention. Paris, France (2000)
12. Fujita, Y., Takashima, R., Homma, T., Ikeshita, R., Kawaguchi, Y., Sumiyoshi, T., Endo, T., Togami, M.: Unified ASR system using LGM-based source separation, noise-robust feature extraction, and word hypothesis selection. In: IEEE ASRU, Scottsdale, AZ, USA, December 13-17, 2015, pp. 416–422 (2015)
13. Heymann, J., Drude, L., Chinaev, A., Haeb-Umbach, R.: BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge. In: IEEE ASRU, Scottsdale, AZ, USA, December 13-17, 2015, pp. 444–451 (2015)
14. Hirsch, H.G., Pearce, D.: The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP), vol. 4, pp. 29–32 (2000)
15. Hori, T., Chen, Z., Erdogan, H., Hershey, J.R., Le Roux, J., Mitra, V., Watanabe, S.: The MERL/SRI system for the 3rd CHiME challenge using beamforming, robust feature extraction, and advanced speech recognition. In: IEEE ASRU, Scottsdale, AZ, USA, December 13-17, 2015, pp. 475–481 (2015)
16. Ma, N., Marxer, R., Barker, J., Brown, G.J.: Exploiting synchrony spectra and deep neural networks for noise-robust automatic speech recognition. In: IEEE ASRU, Scottsdale, AZ, USA, December 13-17, 2015, pp. 490–495 (2015)
17. Misbullah, A., Chien, J.T.: Deep feedforward and recurrent neural networks for speech recognition (unpublished). Unpublished technical report
18. Moritz, N., Gerlach, S., Adiloglu, K., Anemüller, J., Kollmeier, B., Goetze, S.: A CHiME-3 challenge system: Long-term acoustic features for noise robust automatic speech recognition. In: IEEE ASRU, Scottsdale, AZ, USA, December 13-17, 2015, pp. 468–474 (2015)
19. Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S.M., Tyagi, A., Casas, J.R., Turmo, J., Cristoforetti, L., Tobia, F., , Pnevmatikakis, A., Mylonakis, V., Talantzis, F., Burger, S., Stiefelhagen, R., Bernardin, K., Rochet, C.: The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. Language Resources and Evaluation **41**(3-4), 389–407 (2007)

20. Pang, Z., Zhu, F.: Noise-robust ASR for the third 'CHiME' challenge exploiting time-frequency masking based multi-channel speech enhancement and recurrent neural network (2015). ArXiv:1509.07211
21. Parihar, N., Picone, J., Pearce, D., Hirsch, H.G.: Performance analysis of the Aurora large vocabulary baseline system. In: Proceedings of the 2004 European Signal Processing Conference (EUSIPCO), pp. 553—556. Vienna, Austria (2004)
22. Pfeifenberger, L., Schrank, T., Zöhrer, M., Hagmüller, M., Pernkopf, F.: Multi-channel speech processing architectures for noise robust speech recognition: 3rd CHiME challenge results. In: IEEE ASRU, Scottsdale, AZ, USA, December 13-17, 2015, pp. 452–459 (2015)
23. Prudnikov, A., Korenevsky, M., Aleinik, S.: Adaptive beamforming and adaptive training of DNN acoustic models for enhanced multichannel noisy speech recognition. In: IEEE ASRU, Scottsdale, AZ, USA, December 13-17, 2015, pp. 401–408 (2015)
24. Renals, S., Hain, T., Bourlard, H.: Interpretation of multiparty meetings: The AMI and AMIDA projects. In: Proceedings of the 2nd Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA), pp. 115–118 (2008)
25. RWCP meeting speech corpus (RWCP-SP01) (2001). URL http://research.nii.ac.jp/src/en/RWCP-SP01.html
26. Segura, J.C., Ehrette, T., Potamianos, A., Fohr, D., Illina, I., Breton, P.A., Clot, V., Gemello, R., Matassoni, M., Maragos, P.: The HIWIRE database, a noisy and non-native English speech corpus for cockpit communication. Online. http://www. hiwire. org (2007)
27. Sivasankaran, S., Nugraha, A.A., Vincent, E., Morales-Cordovilla, J.A., Dalmia, S., Illina, I.: Robust ASR using neural network based speech enhancement and feature simulation. In: IEEE ASRU, Scottsdale, AZ, USA, December 13-17, 2015, pp. 482–489 (2015)
28. Tachioka, Y., Kanagawa, H., Ishii, J.: The overview of the MELCO ASR system for the third CHiME challenge. Tech. Rep. SVAN154551, Mitsubishi Electric (2015)
29. Tachioka, Y., Watanabe, S., Le Roux, J., Hershey, J.R.: Discriminative methods for noise robust speech recognition: a chime challenge benchmark. In: Proc. 2nd CHiME workshop on Machine Listening in Multisource Environments. Vancouver, Canada (2013)
30. Tran, H.D., Dennis, J., Yiren, L.: A comparative study of multi-channel processing methods for noisy automatic speech recognition on the third CHiME challenge (unpublished)
31. Veselý, K., Ghoshal, A., Burget, L., Povey, D.: Sequence-discriminative training of deep neural networks. In: Proc. INTERSPEECH, pp. 2345–2349 (2013)
32. Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M.: The second 'CHiME' speech separation and recognition challenge: an overview of challenge systems and outcomes. In: Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 162–167 (2013)
33. Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M.: The second 'chime speech separation and recognition challenge: Datasets, tasks and baselines. In: Proc. ICASSP (2013)
34. Vu, T.T., Bigot, B., Chng, E.S.: Speech enhancement using beamforming and non negative matrix factorization for robust speech recognition in the CHiME-3 challenge. In: IEEE ASRU, Scottsdale, AZ, USA, December 13-17, 2015, pp. 423–429 (2015)
35. Wang, X., Wu, C., Zhang, P., Wang, Z., Liu, Y., Li, X., Fu, Q., Yan, Y.: Noise robust IOA/CAS speech separation and recognition system for the third 'CHIME' challenge (2015). ArXiv:1509.06103
36. Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Fujimoto, M., Yu, C., Fabian, W.J., Espi, M., Higuchi, T., Araki, S., Nakatani, T.: The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices. In: IEEE ASRU, Scottsdale, AZ, USA, December 13-17, 2015, pp. 436–443 (2015)
37. Zhao, S., Xiao, X., Zhang, Z., Nguyen, T.N.T., Zhong, X., Ren, B., Wang, L., Jones, D.L., Chng, E.S., Li, H.: Robust speech recognition using beamforming with adaptive microphone gains and multichannel noise reduction. In: IEEE ASRU, Scottsdale, AZ, USA, December 13-17, 2015, pp. 460–467 (2015)

38. Zhuang, Y., You, Y., Tan, T., Bi, M., Bu, S., Deng, W., Qian, Y., Yin, M., Yu, K.: System combination for multi-channel noise robust ASR. Tech. Rep. SJTU SpeechLab Technical Report, SP2015-07, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China (2015)