The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments

Heidi Christensen, Jon Barker, Ning Ma, Phil Green

Department of Computer Science, University of Sheffield, United Kingdom

{h.christensen, j.barker, n.ma, p.green}@dcs.shef.ac.uk

Abstract

We present a new corpus designed for noise-robust speech processing research, CHiME. Our goal was to produce material which is both **natural** (derived from reverberant domestic environments with many simultaneous and unpredictable sound sources) and **controlled** (providing an enumerated range of SNRs spanning 20 dB). The corpus includes around 40 hours of background recordings from a head and torso simulator positioned in a domestic setting, and a comprehensive set of binaural impulse responses collected in the same environment. These have been used to add target utterances from the Grid speech recognition corpus into the CHiME domestic setting. Data has been mixed in a manner that produces a controlled and yet natural range of SNRs over which speech separation, enhancement and recognition algorithms can be evaluated. The paper motivates the design of the corpus, and describes the collection and post-processing of the data. We also present a set of baseline recognition results.

Index Terms: Data collection, Binaural, Spatialisation

1. Introduction

Despite much research and investment, automatic speech recognition (ASR) technology is still not an integral part of our everyday lives. One of the greatest barriers to the uptake of ASR is the lack of robustness to interfering noise sources. In most cases this has led to a reliance on close-talking microphones to deliver acceptable performance. However, head mounted microphones are least appropriate in precisely the situations where speech communication could be most useful – human computer interactions in informal, everyday environments (e.g. the home) where keyboard-and-screen interfaces are inconvenient. The CHiME (Computational Hearing in Multisource Environments) project wishes to address these issues by building a general statistical framework for computational hearing that can recognise speech from recordings made by distant microphones in acoustically 'cluttered' environments (i.e. multiple, simultaneous sound sources).

Speech technology has been advanced over the last 30 years by a series of successful evaluation exercises. These usually take the form of tightly specified speech processing challenges that invite open competition and allow direct comparison of algorithms. Some of these tasks such as TIMIT and Aurora have become standards with long shelf lives and are continuing to drive research long after their originators may have foreseen. Given the potential impact of such datasets, extreme care has to be taken that new tasks are well aligned with the demands of real application scenarios, so that optimising algorithms on the task results in real benefit for present and future systems operating in the wild. With this in mind, we have carefully considered the CHiME challenge design in terms of several operating criteria.

Noise background: A wide variety of algorithms exist for handling 'special case' noise backgrounds (stationary noise or slowly adapting noise, speech plus speech, speech babble, noise with a predictable temporal structure [1, 2, 3, 4, 5]), however, these algorithms can be very brittle and often fail badly in more general conditions. We wish to record data with a complexity that is representative of everyday listening conditions. We call our data 'acoustically cluttered', meaning that there may be many noise sources simultaneously active and each source may have a very different characteristic. At the same time, we also wish to constrain the domain: the noise sources are not totally arbitrary but associated with the particular environment in which they were recorded. With sufficient data it should be possible to build a meaningful model of the noise environment.

It is noted that many recent European and American large vocabulary ASR (LVASR) projects have published substantial speech datasets gathered in meeting (AMI[6, 7],ICSI[8]) and lecture (CHIL[9]) scenarios. These settings have been of interest to LVASR because they provide 'good' conditions in which to listen to speech. Compared to less formal settings (e.g. homes, social gatherings etc.) the speech SNRs are typically lower, instances of overlapping speech are rarer and the noise backgrounds are less acoustically varied and less cluttered.

Noise level: We wish the SNRs encountered in the recognition task to be natural and representative of those found in a real application, i.e. if the SNR is low it will be because the speech co-occurs with background noise that naturally has high energy. This contrasts with tasks like Aurora [10] where the same noise segment is added to speech at a range of SNRs, e.g. speech might be added to the noise of a busy cafeteria at 20 dB to produce a mixed signal that would never be heard in any real situation. However, while ensuring the naturalness of the mixtures, we still wish to be able to carefully measure and control the SNR so that we can evaluate ASR algorithms over an appropriately spaced set of difficulty settings.

Recording style: As explained earlier, we have a focus on distant microphone speech recognition. We also acknowledge that sound source separation is an important component of robust speech recognition in everyday listening settings. It is clear that spatial cues are important for facilitating source separation, and to provide these, we need to make recordings with more than one microphone. We have chosen to use a binaural microphone setup (i.e. two microphones in a configuration modelling human ears) rather than employ a larger microphone array. This economy is justified by the self-evident observation that two channels are sufficient for robust speech processing in humans.

The work in this paper was funded by the EPSRC grant EP/G039046/1.

Speech material: Given that the SNR range is dictated by the environment, once the environment has been chosen, the 'difficulty' of the recognition task can best be controlled through the choice of target speech material. We wished to avoid a large vocabulary task because we believe the barriers to noise robust ASR can be studied more efficiently using small vocabulary tasks. However, we wish to select a small vocabulary task that adequately reflects the difficulties of general small vocabulary ASR applications (i.e. one that cannot be solved using principles that do not generalise well). For example, the digit string recognition task that underlies the popular Aurora 2 evaluation framework [1] – still occasionally used as a test of robust ASR systems – can be largely solved on the basis of vowel recognition alone and is therefore too narrow to meet our requirements.

With the above considerations in mind, Section 2 outlines the design of the CHiME corpus. Technical details of the data collection and subsequent post-processing are described in Sections 3 and 4 respectively. Section 5 concludes with the presentation of representative recognition results using baseline ASR systems and discusses plans for distribution of the data.

2. Design

The CHiME background noise is recorded separately from the target speech. The target speech is subsequently artificially added but in a manner that closely simulates the effect of the speech being present in the room. This allows us to readily control the target speech SNR, target talker location, talker characteristics etc.

For the background noise, a domestic environment has been chosen, such as would be encountered in a home automation application. It is a convenient setting for recording data over an extensive period of time. It also provides a surprisingly rich mix of sound sources, some of which may be easy to model (e.g. a washing machine that remains in a fixed position and runs a predictable program) and some which are not (e.g. children running around while talking, screaming and laughing). All recordings are made in one family home. Within the house it was chosen to collect a large number of hours from a relatively small number of environments – presently two rooms: lounge and kitchen. This will allow the data to be useful when studying how a fixed recognition system may learn to adapt to its environment over time.

The target speech for the recognition task has been taken from the Grid[11] corpus. Although small vocabulary, Grid utterances have previously been demonstrated to be a good test of the state-of-the-art for noise robust ASR. As the original Grid recordings are high quality and have very little reverberation they proved to sound very natural when added to the room recordings using the techniques described in Section 4.

The CHiME corpus will be made freely available to the public and when the CHiME challenge is announced (expected to happen in the autumn of 2010) it will be available for download through the challenge web page[12] – the site currently contains some *taster* segments of mixed and unmixed data.

3. Collection

The data has been collected using a B&K head and torso simulator (HATS) and the corpus contains three main parts (summarised in Table 1): i) around 10 hours of background recordings (for development and training) split approximately evenly between the lounge and the kitchen, ii) a set of binaural room impulse responses (BRIRs) from different spatial locations relative to the HATS in the same rooms and iii) around 30 hours of background recordings (*different* to the ones in part i)) into which sentences from the Grid corpus have been mixed after they have been spatialised with the appropriate BRIRs.

Data type	Kitchen	Lounge
Background recordings (dev+train)	5:50:08	5:38:56
Background rec. mixed with Grid	16:05:41	14:41:29
Impulse responses (# of locations)	14	18

Table 1: The CHiME corpus; number of hours recorded and number of binaural room impulse responses (BRIRs).

The house is a typical English Victorian semi-detached house with a relatively high ceiling height - around 365 cm. The lounge (385 cm $\times$ 385 cm) has carpeted floors and plastered ceiling and walls; one wall is dominated by a large, bay window. The kitchen is smaller at approximately 365 cm $\times$ 300 cm with linoleum floors, plastered walls and ceiling and fitted units along most walls. The major noise sources in the environment are those of a typical family home: two adults and two children, TV, kitchen and laundry appliance sounds, foot steps, electronic gadget sounds (laptops, games console), toys, gerbils and noise from the outside such as traffic, voices and birds. Both rooms have a number of fixed noise sources such as washing machine, TV, and kettle which, depending on whether they are on or off, will contribute to the overall noise background. Figure 1 shows the locations of the main, *stationary* noise sources.

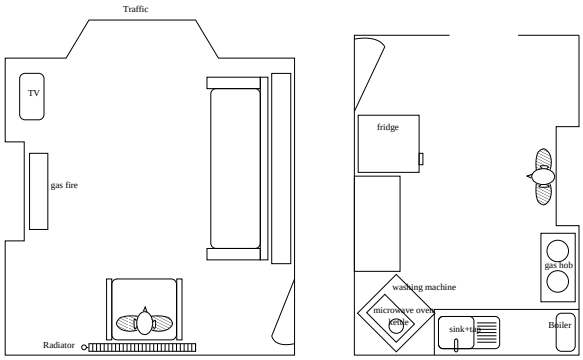

Figure 1: Location of the HATS in relation to the main noise sources in the lounge *(a)* and the kitchen (*(b)*.

3.1. Background recordings

A total of 57 background sessions were recorded with the binaural B&K HATS placed in either of two locations in the house: the lounge (a total of 15 hours collected from 22 individual sessions) and the kitchen (a total of 21 hours from 21 individual sessions). The recording times range from around 7:30 in the morning to 20:00 in the evening, and the data are recorded at the maximum available sampling frequency of 96 kHz and with a precision of 32 bit. The reverberation time was measured using the method of Schroeder integration [13] and both the lounge and kitchen have a T_{60} time of 300 ms.

The equipment was set up permanently so it was easy to turn on when in the house. This also served to minimise any

recording level differences between sessions. Calibration measurements were carried out with a B&K 4231 calibrator after each 'run' of recording days. The calibrator fits snugly around each microphone (after the pinna has been removed) and plays a constant tone; the background sessions have subsequently been normalised according to these measurements.

Figure 2 illustrates the recording set up. The microphones in the B&K HATS are connected to the B&K 4128E amplifier. The amplifier has internal filters that were set up to remove frequency components below 20 Hz and above 20 kHz. The amplifier is connected to a MOTU 8pre box which digitises the signals. The MOTU 8pre box communicates with the MacBook Pro via a firewire cable. The background sessions were recorded directly onto the MacBook harddrive using the 'Audacity' multitrack recording and editing software.

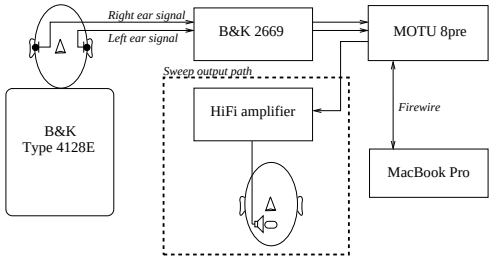

Figure 2: Diagram of HATS based recording equipment. When estimating the BRIRs, the sweep signal is played through the artificial mouth as indicated.

A couple of pilot recordings were carried out to estimate the best recording level for capturing the very wide dynamic range of audio experienced in the two rooms. Gain was adjusted to be as high as possible while avoiding substantial clipping. With the home automation application in mind it was considered reasonable to expect our target system to experience occasional clipping (e.g. during impulsive sound events occurring close to the microphones). As the acoustics of the kitchen was 'louder' than the lounge, the amplification on the MOTU box was turned down slightly for those sessions. The calibration recordings have been used to normalise all recordings to the same relative level in the data that will be distributed.

Figure 3 shows the ratemap of two background recording snippets. They are very representative in that there is speech under which can be heard the slowly varying and stationary sounds of the washing machine with an occasional unpredictable noise such as foot steps or a child hitting sticks together.

3.2. Impulse responses

In each room a number of binaural room impulse responses (BRIRs) have been estimated for various source locations relative to the HATS. The BRIRs are positioned on polar grid, i.e. either moving the sound source on an arc at equi-distance from the HATS (e.g. at 100 cm and with 10° azimuth intervals) or at an equi-angle (at 0° azimuths and with 50 cm distances). All positions are measured from the center point of the HATS head (halfway between the ears).

The IRs are determined using the sine sweep method[14, 15] as implemented in Farina's AURORA v.4.3 plugin for ADOBE AUDITION[16]. The process is as follows: i) generate a sine sweep with a frequency progressing from 20 Hz to 20 kHz on a log scale, ii) simultaneously play the sweep signal while

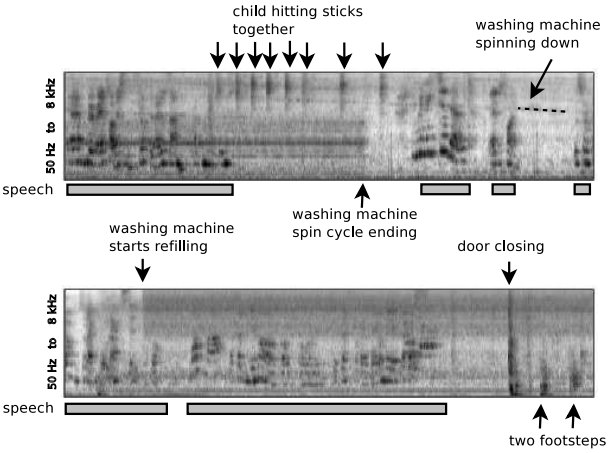

Figure 3: Ratemap snippets of a background noise recordings containing speech, slowly and fast varying noise sources.

recording the response, iii) convolve the recorded response with the inverse of the sweep signal to determine the IR. The sweep response recordings were made using the setup shown in Figure 2. As the BRIRs will be used to add speech to the corpus, we have played the sweep through a B&K 4227 artificial mouth which simulates speech's directionality. For each IR location 3-4 responses were recorded and the least noisy was subsequently determined by visually inspecting the sonogram of the IR. The sweep responses are recorded at 96 kHz and saved as 32 bit floats.

When recording the sweep response using the setup shown in Figure 2, the response will be coloured according to the transfer function of the amplifier and artificial mouth. The amplifier is a standard mid-range hifi amplifier which should only have a small colouration in the typical speech domain. However, because the artificial mouth is designed to mimic the characteristics of a real mouth, the frequency response of that loudspeaker has been designed with a 'speech shaped' response in mind. When estimating the IR from a sweep played by this loudspeaker, and subsequently using the BRIRs to spatialise the Grid utterances, these utterances will therefore be affected twice by a mouth shaped filtering. Further, as the Grid utterances were recorded in an acoustic booth, the effect of the booth will also have an effect on the recordings.

To compensate for these two effects (room and amplifier+artificial mouth audio path) a compensation filter was designed. For this, the setup of Figure 2 was placed in the acoustic booth and the sweep response was measured using the same microphone used for the original Grid recordings. The artificial mouth and microphone were arranged to match the typical microphone/talker geometry in the Grid recordings. The associated IR was found using the sine-sweep method described earlier and from this response a linear phase compensation filter was estimated.

4. Post-processing

All utterances from the training and test sets of the Grid corpus were filtered with the acoustic booth compensation filter (see Section 3.2) and then convolved with a BRIR pair selected from the CHiME corpus. The 2 meter at 0° azimuth response has been chosen for the initial experiments. A gain factor was

then experimentally determined to approximately match the level of the reverberated Grid utterances to the level recorded by the HATS of a talker producing the same Grid utterances in the CHiME room at the position from which the impulse was recorded. The live talker had been instructed to read the utterances as if speaking across the room to the HATS.

After scaling, the reverberated Grid utterances (left and right channels) were then artificially added to the binaural CHiME recordings, i.e. by summing the signal amplitudes. The *lounge* dataset has been used for the experiments reported here. The utterances were mixed at a range of SNRs. Rather than artificially scaling either the Grid or CHiME data, we select an utterance-length segment of CHiME test data that produces the desired SNR when added at its natural level, i.e. the mixtures reflect the range of SNRs that would be encountered in the real environment. We were able to generate SNRs in the range of -6 dB to 18 dB in this manner. The selection algorithm employed ensures that each Grid utterance is mixed into a different and non-overlapping segment of the CHiME test data.

Note, both the signal and the noise have two channels, so definition of SNR needs some generalisation. Here, we have defined it to be,

$$SNR_{dB} = 20 \log_{10} \left(\frac{E_{s,l} + E_{s,r}}{E_{n,l} + E_{n,r}} \right), \qquad (1)$$

where l and r refer to the left and right channels and s and n to the signal and noise. The energy, E, is computed as the sum of the squared sample amplitudes.

The mixing technique produces natural-sounding mixtures in which it is not obvious that the Grid talker was not present in the room[1]. Note, a shortcoming is that the technique does not model the effect of the noise background on speech production [17]. For example, during noisy background periods, talkers may raise their voices, or may even delay speaking until the background noise is lower.

5. Baseline results and conclusions

Baseline recognition results are included here to give readers a feel for the difficulty of the task (Figure 4). All systems employ the reverberated Grid training data and use word-level HMMs with the standard Grid corpus model topology [18]. The MFCC-based system using 13 cepstral coefficients plus deltas and accelerations is identical to the baseline system used in the recent Speech Separation Challenge [18]. The MFCC-CMN system further applies cepstral mean normalisation during training and testing. Also shown is the performance of a system employing multi-condition training, i.e. the Grid training set was mixed with CHiME using the same techniques as employed for the test set and the same range of SNRs. Models were then trained using this noisy training data. Following [18], the results represent the percentage of Grid keywords (i.e. the letter and digit) that have been recognised correctly.

The CHiME challenge (expected to be announced in autumn 2010 and run through spring 2011) will be a binaural, speech separation recognition task based on the CHiME corpus.

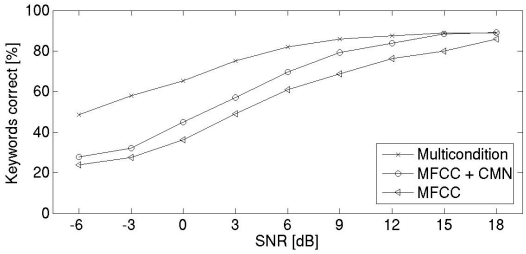

Figure 4: Results for CHiME baseline systems (see text).

6. References

[1] D. Pearce and H.-G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of ISCA ITRW ASR*, 2000.

[2] S. S. Tan and A. M. Ahmad, "Adaptive parallel model combination for reduced environmental mismatch in noisy speech recognition," in *Proc. of Int. Conf. on Electronic Design*, 2008.

[3] Y. Zhao, S. Wang, and K.-C. Yen, "Recursive estimation of time-varying environments for robust speech recognition," in *Proc. of ICASSP'01*, 2001.

[4] N. Krishnamurthy and J. H. L. Hansen, "Babble noise: modeling, analysis, and applications," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, pp. 1394–1407, 2009.

[5] M. L. Seltzer, J. Droppo, and A. Acero, "A harmonic model based front end for robust speech recogntition," in *Proc. of Eurospeech'03*, 2003.

[6] "The AMI project homepage," http://www.amiproject.org, 2007.

[7] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. A. McCowan, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus: a pre-announcement," in *Proc. of MLMI'2005*, 2005.

[8] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in *Proceedings of HLT 2001*, San Diego, 2001.

[9] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnevmatikakis, V., F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet, "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms," *Language Resources and Evaluation*, vol. 41, no. 3–4, pp. 389–407, Dec 2007.

[10] "The AURORA challenge homepage." [Online]. Available: http://aurora.hsnr.de

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

[12] ""the CHiME challenge webpage"," Internet. [Online]. Available: \url{http://www.dcs.shef.ac.uk/spandh/chime/research_challenge.html}

[13] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 36, pp. 409–413, 1964.

[14] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept sine technique," in *In Proc. 108th AES Convention*, Paris, France, 2000.

[15] ——, "Advancements in impulse response measurements by sine sweeps," in *Proc. of 122ed AES Convention*, Vienna, Austria, May 2007.

[16] ——, "AURORA." [Online]. Available: http://www.aurora-plugins.com/Aurora_XP/index.htm

[17] Y. Lu and M. Cooke, "Speech production modifications produced in the presence of low-pass and high-pass filtered noise," *J. Acoust. Soc. Am.*, vol. 126, pp. 1495–1499, 2009.

[18] M. P. Cooke, J. Hershey, and S. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech and Language*, vol. 24, pp. 1–15, 2010.

[1] Examples can be found on the CHiME project web page [12]