The PASCAL CHiME Speech Separation and Recognition Challenge

Jon Barker, Emmanuel Vincent, Ning Ma, Heidi Christensen, Phil Green

The PASCAL CHiME Speech Separation and Recognition Challenge

Jon Barker[a], Emmanuel Vincent[b], Ning Ma[a], Heidi Christensen[a], Phil Green[a]

[a]*Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK*
[b]*INRIA, Centre de Rennes - Bretagne Atlantique, 35042 Rennes Cedex, France*

Abstract

Distant microphone speech recognition systems that operate with human-like robustness remain a distant goal. The key difficulty is that operating in everyday listening conditions entails processing a speech signal that is reverberantly mixed into a noise background composed of multiple competing sound sources. This paper describes a recent speech recognition evaluation that was designed to bring together researchers from multiple communities in order to foster novel approaches to this problem. The task was to identify keywords from sentences reverberantly mixed into audio backgrounds binaurally-recorded in a busy domestic environment. The challenge was designed to model the essential difficulties of multisource environment problem while remaining on a scale that would make it accessible to a wide audience. Compared to previous ASR evaluation a particular novelty of the task is that the utterances to be recognised were provided in a continuous audio background rather than as pre-segmented utterances thus allowing a range of background modelling techniques to be employed. The challenge attracted thirteen submissions. This paper describes the challenge problem, provides an overview of the systems that were entered and provides a comparison alongside both a baseline recognition system and human performance. The paper discusses insights gained from the challenge and lessons learnt for the design of future such evaluations.

Keywords:
Speech recognition; Source separation; Noise robustness

Email addresses: j.barker@dcs.shef.ac.uk (Jon Barker), emmanuel.vincent@inria.fr (Emmanuel Vincent)

1. Motivation

There has been much recent interest in distant speech recognition systems (Wölfel and McDonough, 2009), i.e. systems which unobtrusively capture and recognise speech using microphones positioned perhaps several metres away from the target speaker. Such systems free the user from the constraints of close-talking microphones thus providing potential for unfettered and hence more natural man-machine speech communication. Robust distant microphone speech technology would enable a host of powerful applications including human-robot communication, voice-controlled home automation systems and speech monitoring and surveillance systems.

Unfortunately, the ease with which humans process distant speech belies the fact that the task presents some uniquely difficult challenges: chief among these are the problems of reverberation and additive background noise. The everyday living environments in which we wish to be able to deploy speech recognition technology often possess highly dynamic and complex acoustic backgrounds made up of multiple competing sound sources. The speech we wish to attend to is just one component of an acoustic mixture. Further, in indoor settings, acoustic reflections from walls, floors and ceilings etc, produce reverberation (i.e. a series of closely spaced echoes) that significantly adds to the difficulty of recovering a description of the speech signal from the mixture.

The source separation problems that are inherent in distant speech recognition have been widely addressed by the signal processing community. The topics of blind and then semi-blind source separation have emerged as research problems in their own right with their own associated conferences and evaluations (Makino et al., 2007; Vincent et al., 2012). Techniques developed in this research community should be of direct relevance to distant speech recognition. However, evaluations in this field have typically focussed on *reconstruction* of separated signals which is not the goal of speech recognition. On the other hand the speech recognition community has traditionally paid too little attention to the issue of source separation and could benefit greatly from recent advances made within the source separation community. Bridging the gap between these fields is not trivial: naive attempts lead to suboptimal decoupled systems which treat separation and recognition as independent consecutive processing stages.

One of the primary objectives of the Pascal CHiME speech separation and recognition challenge has been to draw together the source separation and speech recognition communities with the hope of stimulating fresh and more deeply coupled approaches to distant speech recognition. To this end the task has been designed to be widely accessible while capturing the difficulties that make dis-

tant speech recognition a hard problem. Compared to the still widely reported Aurora 2 speech recognition task (Pearce and Hirsch, 2000), the CHiME task is more challenging along a number of dimensions: like Aurora 2 it is built around a small vocabulary speech corpus but it contains many acoustically confusable utterances that rely on finer phonetic distinctions than those required to disambiguate Aurora's digit sequences; the target utterances have been reverberantly mixed into complex multisource noise backgrounds recorded in real everyday living environments; the exploitation of spatial source separation is enabled by the provision of two-channel 'binaurally recorded' signals that mimic the signals that would be received by the ears of a human situated in the recording environment.

The PASCAL CHiME speech separation and recognition challenge also builds on the earlier *PASCAL speech separation challenge* (Cooke et al., 2010). This earlier challenge considered recognition of speech in artificial speech-plus-speech mixtures. The challenge was remarkable in that the best-performing system was able to produce super-human performance (Hershey et al., 2010). Without detracting from the elegance of this winning system, it should be noted that its super-human performance can be explained in part by the narrowness of the task: the very specific training given to the ASR system allowed it to be better adapted to a task for which the human listeners were given no specific training (i.e. they had to rely on their general speech processing abilities). The new PASCAL challenge, by better corresponding to the speech-in-noise task faced by humans in everyday listening, is likely to serve as a fairer comparison of human versus machine speech recognition ability.

The initial submissions to the CHiME challenge were presented at a dedicated workshop that was held as a satellite event of Interspeech 2011. Thirteen groups presented results on the task (Delcroix et al., 2011; Gemmeke et al., 2011; Hurmalainen et al., 2011; Kallasjoki et al., 2011; Kim et al., 2011; Koldovský et al., 2011; Kolossa et al., 2011; Ma et al., 2011; Maas et al., 2011; Nesta and Matassoni, 2011; Ozerov and Vincent, 2011; Vipperla et al., 2011; Weninger et al., 2011). Revised and extended versions of eight of these papers are presented in this special issue.

Section 2 will detail the design of the challenge. Section 3 will provide an overview of the systems that were submitted. This overview can also be considered as a concise tutorial of noise-robust ASR in multisource conditions, including a categorisation of the various approaches and links to practical systems for further details. A summary of the machine results and a comparison to human performance on the same task is presented in Section 4. Finally, Section 5 will

conclude with a discussion of directions for future challenges.

2. Challenge design

The challenge task is to recognise keywords from simple target sentences that have been mixed with noise backgrounds. The acoustic mixing has been performed in a manner that simulates the effect of the target sentences having been recorded in a real multisource noise environment. The sections that follow describe the preparation of the noisy speech material, the design of the evaluation task and the baseline system that was provided to challenge entrants.

2.1. Data

The multisource noise backgrounds are taken from the CHiME corpus (Christensen et al., 2010). This corpus contains recordings made in domestic environments using a B&K head and torso simulator (HATS) type 4128 C – a mannequin with built-in left and right ear simulators that record signals that are an approximation of the acoustic signals that would be received by the ears of an average adult listener. The data for the challenge is composed of approximately 14 hours of audio collected during evening and morning recording sessions from a single domestic living room. The major noise sources in the environment are those of a typical family home: two adults and two children, TV, footsteps, electronic gadget sounds (laptops, games console), toys, some traffic noise from outside and noises arriving from a kitchen via a connecting hallway (see Figure 1).

The target speech utterances are the same as those used in the 1st Pascal Speech Separation Challenge, namely, 600 utterances taken from the Grid corpus (Cooke et al., 2006). This corpus consists of 34 speakers (18 male and 16 female) reading sentences which are simple six-word commands obeying the following syntax ,

```
($command $color $preposition $letter $number $adverb)
```

where each word can have the following alternatives,

```
$command =  bin | lay | place | set;
$colour = blue | green | red | white;
$prep = at | by | in | with;
$letter = A | B | C | ... | U | V | X | Y | Z;
$number = zero | one | two ... seven | eight | nine;
$coda = again | now | please | soon;
```

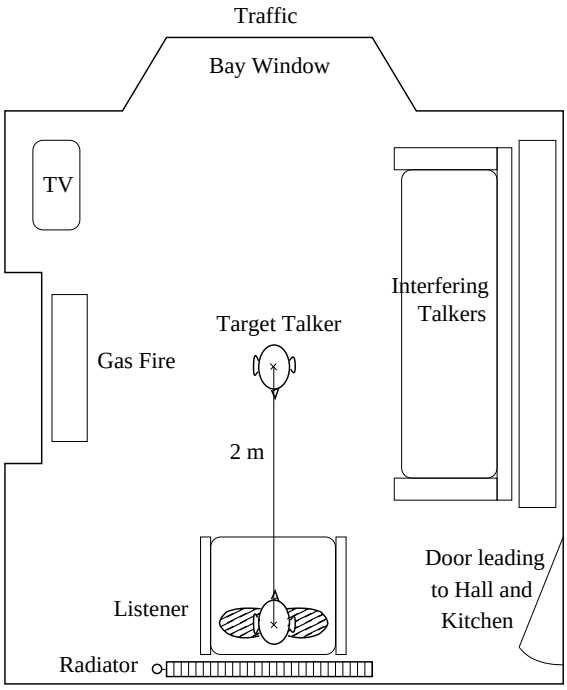

Figure 1: Plan of the CHiME recording setting showing location of the binaural mannequin and the most significant noise source.

The letter W is not used as it is the only letter with a polysyllabic name in English.

The 600-utterance Grid test set then mixed with the CHiME background. The single-channel Grid utterances were first convolved with binaural room impulse responses (BRIR) that are supplied with the CHiME domestic audio corpus. The BRIRs were measured for a position 2 metres directly in front of the HATS. The measurements were made using Farina's sine sweep method (Farina, 2000). An empirically determined gain was applied to the Grid utterances so that the level after convolution approximately matched that of a sequence of Grid utterances that were spoken 'live' at a natural conversational level in the actual room from which the CHiME acoustic backgrounds were recorded. The temporal placement of the reverberated Grid utterances within the 14 hours of CHiME data was controlled in a manner which produced mixtures at 6 different SNRS (-6, -3, 0, 3, 6, 9 dB) resulting in a total of 3,600 test utterances. None of the Grid utterances overlap.

Note, both the speech targets (after application of the BRIRs) and the noise

backgrounds are two channel signals, so the usual definition of SNR requires some generalisation. Here it has been defined as,

$$\mathrm{SNR}_{\mathrm{db}} = 10 \log_{10} \left(\frac{E_{s,l} + E_{s,r}}{E_{n,l} + E_{n,r}} \right) \quad (1)$$

where l and r refer to the left and right channels and s and n to the speech and noise backgrounds. The energy E is computed as the sum of the squared sample amplitudes measured for either the speech of background signals between the start and end points of the utterance. In order that SNRs better reflected the perceived noisiness of the mixtures, the SNR computation employed high-pass filtered versions of the signals in which energy below 80 Hz had been removed.

In contrast to conventional robust ASR evaluations such as Aurora 2, the SNRs have not been controlled by artificially scaling the speech or noise amplitudes, but instead by choosing different noise backgrounds for each SNR point. Mixing in this way better mimics the effect of recording in a natural live environment, but it means that some caution is needed when comparing results at different SNRs, i.e. the backgrounds at the different SNR levels are very different in their nature: whereas at 9 dB the backgrounds are dominated by ambient and quasi-stationary sources, at -6 dB the backgrounds are more likely to contain highly non-stationary acoustic events such as shouts or doors slamming.

2.2. Task and evaluation metric

The task is to recognise the letter and digit spoken in each noisy Grid utterance. Systems are scored according to the percentage of the tokens that they recognise correctly at each SNR level.

Participants were provided with a development test set containing 3,600 stereo 16 bit WAV files (600 utterances $\times$ 6 noise levels) available at either 16 kHz or 48 kHz. Each file contains a single end-pointed noisy utterance. The development set was also made available in an unsegmented form, i.e. with the Grid utterances embedded in the continuous CHiME audio. The unsegmented data is accompanied by an annotation file storing the temporal position (start sample and duration) of the utterances to be recognised. Participants were permitted to use the annotation file to segment the utterance prior to recognition (i.e. the task did not consider the challenge of speech detection). Participants were also permitted to make free use of the unsegmented development and test set data, e.g. to learn online background models from the immediate acoustic context of each utterance.

In order to train acoustic speech models a 17,000-utterance training set was provided containing 500 utterances of each of the 34 Grid talkers. The training utterances were provided with reverberation but free of additive noise. The reverberation was performed via convolution with one of the CHiME BRIRs measured at 2 m. Note although the position of the BRIR was matched to that used in construction of the test set, the response was measured at a different time and with a different room configuration, e.g. doors open/closed, curtains drawn/undrawn.

The speaker identity of the utterances in the training and test set was provided and entrants were permitted to use this knowledge in their systems, e.g. by constructing speaker-dependent models. The test utterances were labelled according to SNR in order to facilitate reporting of the results, but participants were not allowed to assume prior knowledge of SNR in their systems.

A further 6 hours of CHiME background audio was released to allow entrants to train background models if they wished. This data was made up of a number of recording sessions made in the same CHiME room but which had not been used during the construction of the test data, i.e. there was no overlap between this data and the audio occurring in the backgrounds of the test set.

Shortly before the challenge deadline a final test set was released to competitors. This test set employed a previously unseen selection of 600 Grid utterances. These utterances were mixed into the CHiME audio using the same procedures as the development set. Again a 2 m distant BRIR was employed but one recorded at a different time from the instances used in either the development or training sets. The same 14 hours of CHiME background was employed, but the random nature of the mixing process meant that the utterances would have been placed at different temporal locations within the recording sessions. Entrants were instructed that they could tune system parameters on the development set but should only run their systems once on the final test set.

2.3. Baseline system

A baseline system was constructed and made available to challenge participants. This system served to demonstrate the performance that would be representative of a conventional non-robust recogniser in which no effort is made to deal with the mismatch between the noise-free training data and noise-contaminated test data. Equally importantly, it was also made available as a default recogniser for participants whose techniques produced 'cleaned' time-domain speech signals. Providing this recogniser as a tool greatly increased the accessibility of the challenge to researchers outside the speech recognition research community.

The recogniser was constructed using HTK 3.4.1 (Young et al., 2006) and using the word-level HMM topologies that were standardised in the 1st PASCAL speech separation challenge (Cooke et al., 2010), i.e., each of the 51 words in the Grid vocabulary is modelled with an HMM with a left-to-right and no skip topology where the number of states is determined using a rule of two states per phoneme. The emission probability for each HMM state is modelled using a Gaussian Mixture Model with 7 components each component having a diagonal covariance matrix.

The models were trained using a conventional 39-dimensional MFCC representation, i.e. 12 Mel-cepstral frame coefficients plus a frame energy term, augmented by temporal differences and accelerations. Features were extracted at a 100 Hz frame-rate. Prior to feature extraction the binaural training data was reduced to a single channel by averaging the left and right ear signals. Training proceeded in two stages. First, a single set of speaker-independent models was trained from a flat start using the full 17,000 utterances of reverberant but noise-free training data. Second, speaker-dependent models for each of the 34 speakers were constructed by applying further iterations of Baum-Welch parameter estimation using the 500 utterances belonging to the specific speaker.

This baseline system performed well on the noise-free data achieving a recognition accuracy of 96%.

3. Submitted systems

Thirteen systems were submitted by research teams in Europe and Asia, which combine several processing strategies at one or more levels:

- *target enhancement*,
- *robust feature extraction*,
- *robust decoding*.

Certain systems exploit the available speech-free background to train noise models, while others rely on the mixed utterances only. Table 1 summarises the strategies adopted by each team. In the following, we provide more details about the strategies employed by each system at each of these three levels.

3.1. Target enhancement

The first level of processing consists of enhancing the target signal. Due to the time-varying nature of the target and the background, this is typically

	Enhanced target	Robust features	Robust decoder	Trained noise model
(Delcroix et al., 2011)	X		X	X
(Maas et al., 2011)	X		X	
(Weninger et al., 2011)	X	X	X	X
(Nesta and Matassoni, 2011)	X	X	X	
(Kolossa et al., 2011)	X	X	X	
(Hurmalainen et al., 2011)		X	X	X
(Ozerov and Vincent, 2011)	X		X	X
(Ma et al., 2011)	X	X	X	
(Koldovský et al., 2011)	X			
(Kim et al., 2011)	X	X	X	
(Gemmeke et al., 2011)	X		X	X
(Vipperla et al., 2011)	X		X	X
(Kallasjoki et al., 2011)	X	X	X	

Table 1: Overview of the processing strategies employed by the submitted systems.

achieved by representing the input noisy signal in the time-frequency domain and applying a linear filter in each time-frequency bin. The range of employed filters looks very diverse at first. All systems combine a spatial filter resulting from a fixed or an adaptive beamformer (Koldovský et al., 2011) with a spectral filter such as a highpass or lowpass filter, a Wiener filter (Ozerov and Vincent, 2011), or a binary or soft time-frequency mask (Delcroix et al., 2011). These filters can be applied in the short time Fourier transform domain, in the short-time mel spectrum domain (Gemmeke et al., 2011) or via a gammatone filterbank (Ma et al., 2011). Finally, their implementation can be tuned in many ways including oversubtraction (Koldovský et al., 2011), spectral floor/offset (Maas et al., 2011), temporal smoothing (Nesta and Matassoni, 2011) and use of magnitude ratios as opposed to power ratios (Weninger et al., 2011).

A more fundamental view is to categorise the filters according to the set of cues that are used to estimate their parameters. This results in three enhancement strategies exploited by five, four, and three systems respectively:

- *Spatial diversity-based enhancement*, based on the assumption that the target and the background have different spatial positions. This includes beamforming (Kolossa et al., 2011) or Independent Component Analysis (ICA) (Nesta and Matassoni, 2011) followed by Wiener post-filtering, and clustering of Interaural Time and Level Differences (ITD/ILD) (Kim et al., 2011). The ITD and ILD of the target, or equivalently its steering vector in

beamforming terminology, may be either fixed to the center of the sound scene or estimated from the noisy signal under geometrical constraints.

- *Spectral diversity-based enhancement*, based on the assumption that the target and the background have different spectra. This includes multiple pitch tracking (Ma et al., 2011), Gaussian Mixture Models (GMM), Nonnegative Matrix Factorisation (NMF), and exemplar-based enhancement via e.g. Nonnegative Matrix Deconvolution (NMD) (Vipperla et al., 2011). GMM, NMF and NMD represent the target and the background by specific spectra which are learned from reverberated speaker-dependent clean speech and speech-free background.

- *Combined spatial and spectral diversity-based enhancement*, coupling the above two strategies. The coupling can be achieved either by chaining *e.g.* ITD clustering and exemplar-based enhancement (Kallasjoki et al., 2011) or by designing of joint probabilistic framework for ITD and GMM (Delcroix et al., 2011) or ITD, ILD and NMF (Ozerov and Vincent, 2011). This results in increased robustness and applicability to all mixtures in theory, regardless of whether the target and the background have the same direction or the same spectra.

3.2. Robust feature extraction

The second level of processing consists of extracting features that are robust to the background noise or to what remains of it after the target enhancement front-end. Two complementary strategies can be distinguished, which are used by five and two systems respectively:

- *Robust features*, such as Gammatone Frequency Cepstral Coefficients (GFCC) (Nesta and Matassoni, 2011), Mel spectra (Hurmalainen et al., 2011), or additional framewise word estimates generated by a Recurrent Neural Network (RNN) (Weninger et al., 2011). The purpose of these features is respectively to improve robustness to spectrum underestimation thanks to wider filters, concentrate noise in fewer coefficients, and model the long-range context.

- *Robustifying feature transformations*, such as Maximum Likelihood Linear Transformation (MLLT) (Kallasjoki et al., 2011) and Linear Discriminant Analysis (LDA) (Kolossa et al., 2011). These transformations decorrelate the features or reduce their dimension so as to increase the likelihood or the discriminating power of the recogniser.

3.3. Robust decoding

The final level of processing is to transform the sequence of features into a sequence of words. The difficulty is that the features exhibit different values than those in clean speech due to the background noise or to what remains of it. The decoding relies most often on a conventional HMM-GMM recogniser. Four complementary strategies can then be used to enhance the performance of the recogniser, which are employed by eight, six, five and four systems respectively:

- *Multi-condition training*, that is training the decoder on unprocessed noisy speech (Kim et al., 2011) or noisy speech processed by the target enhancement front-end (Gemmeke et al., 2011) at all SNR levels. Alternatively, when the amount of noisy data is insufficient, the decoder can be trained on clean data and adapted to the noisy data (Ozerov and Vincent, 2011).

- *Robust training*, that is adapting the training objective to the amount of training data and the task at hand. This strategy, which is not specific to noisy data, includes noise-dependent setting of the dimension of the GMM acoustic model (Maas et al., 2011), Maximum Likelihood Linear Regression (MLLR), Maximum A Posteriori (MAP), and/or mean-only speaker adaptation (Maas et al., 2011), and discriminative training using the differenced maximum mutual information (dMMI) (Delcroix et al., 2011).

- *Noise-aware decoding*, that is exploiting confidence measures about the feature values estimated as part of the feature extraction process, so as to focus on the most reliable ones. Existing measures represent the confidence in each feature either by a probability between 0 and 1, or by a distribution over its value. The former representation leads to channel-attentive decoding (Kim et al., 2011), while the latter leads to a range of modified decoding objectives known as Uncertainty Decoding (UD) (Kallasjoki et al., 2011), Modified Imputation (MI) (Kolossa et al., 2011), "missing data" decoding methods such as fragment decoding (Ma et al., 2011) or Dynamic Variance Adaptation (DVA) (Delcroix et al., 2011).

- *System combination*, that is running multiple decoders exhibiting different error patterns and fusing them so as to keep the most popular or reliable output at each instant. The fusion can be conducted either at the HMM level, an approach known as multistream decoding (Weninger et al., 2011), or by applying a voting scheme to the outputs such as the Recogniser Output Voting Error Reduction (ROVER) (Vipperla et al., 2011).

A completely different strategy termed *model combination* consists of jointly decoding the target and the background without any target enhancement front-end. The system of Hurmalainen et al. (2011) relies on this strategy, whereby both the target and the background are represented by exemplar-based models and the estimated exemplar activations are transformed into state likelihoods via a trained mapping, which are then decoded via a standard HMM.

4. Results

The thirteen systems described in Section 3 were evaluated on the data of Section 2.1 according the keyword accuracy metric in Section 2.2. Two benchmark systems, namely the ASR baseline in Section 2.3 and a human listener, were also evaluated.

4.1. Human results

The results for a human listener were obtained as follows. The subject was one of the authors who is very familiar with the specific CHiME domestic audio environment. In order to offer fairer comparison than in the first PASCAL Challenge, the noisy utterances were presented in 34 blocks with each block containing just one Grid talker (i.e. to better match the speaker-dependent assumptions exploited by the computational systems). Prior to presenting each block, the listener was allowed to hear six reverberant but noise-free utterances spoken by the target speaker. These presentations allowed the listener to anticipate the target accent and speaking style. Within a block the SNRs were randomised. The mixtures were played starting from two seconds prior to the start of each utterance. It was believed that this lead-in would provide the listener with helpful background context. All listening was performed in an IAC single-alled acoustically-isolated booth using binaural headphone presentation. 200 different utterances were presented in each SNR condition taken from the challenge development and test sets. The listening tests were conducted over 5 sessions each covering 5 or 6 Grid talkers and lasting approximately 30 minutes each.

The resulting keyword accuracies are displayed in Figure 2. Digit recognition is highly reliable and remains 99% correct down to -3 dB. Letter recognition performance falls steadily with increasing noise level at about 1% per dB, from 97% at 9 dB down to 83% at -6 dB. Remarkably, the overall accuracy remains higher than 90% at -6 dB. Note, in order to avoid the listener remembering repeats of utterances, the human results are based on a different 200 utterances at each SNR, whereas the computational systems used the same 600 utterances for each SNR. The difference in the human and machine test sets mean that extra

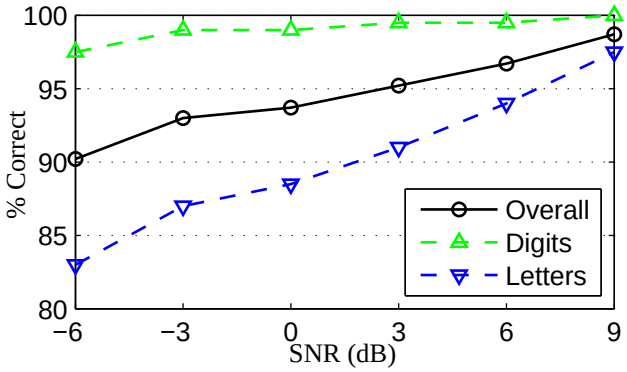

Figure 2: Keyword accuracy achieved by the human listener.

care should be taken when comparing performances and in particular it should be noted that the 95% binomial proportion confidence interval for the human listener results is approximately ±2%.

Detailed inspection of listener responses shows that most of the errors concern highly confusable letters, such as *m* and *n*, *v* and *b*, *s* and *f* or *u* and *e*. Confusion between phonetically distinct letters such as *g* and *q* happen rarely and only when the target is completely masked by the background at the time when the letter is pronounced. Figure 3 shows the letter confusion data for results pooled across the three noisiest settings, i.e. -6, -3 and 0 dB.

4.2. Machine results

Figure 4 shows the keyword accuracy achieved by the submitted systems, compared to the human listener and the ASR baseline. The baseline drastically degrades with increasing noise, from 82% accuracy at a modest 9 dB SNR to 30% at -6 dB SNR. At the lowest SNR the baseline system still performs above chance level (7%) but is unable even to recognise the digit keyword reliably.

The performance of the submitted systems spans the range between the baseline and the human. Caution must be taken when comparing the results of different systems due to the fact that they assumed different scenarios of use, translating into different assumptions regarding e.g. which data could be used for training and validation. However, some broad observations can be made. It can be observed that the systems form two loose groupings. In one group (8 systems) it is observed that the decrease of SNR from 9 dB to 3 dB has very little affect on the performance and that keyword accuracy at -6 dB remains at or above 70 %. In the second group (5 systems), although performance at 9 dB may be competitive there is a marked and steady deterioration in performance so that performance

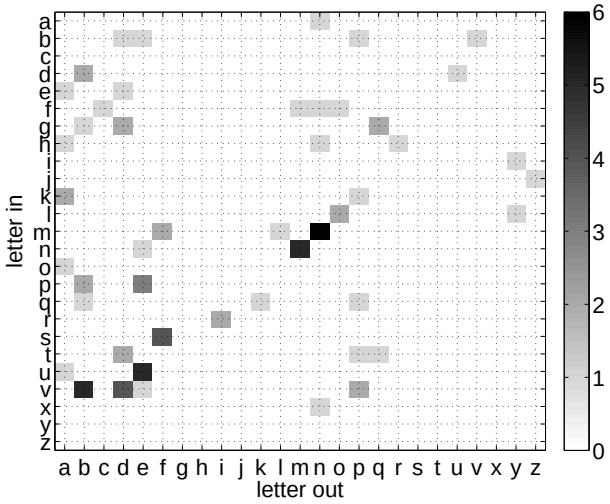

Figure 3: Confusion matrix for human recognition of the Grid utterance letter tokens summed across the -6, -3 and 0 dB SNRs, i.e. 600 test utterances with approximately 24 presentations of each letter. Only recognition errors are shown.

at 3 dB is significantly depressed and the performance at -9 dB falls below 60 %. The overall best-performing system, authored by Delcroix et al. (2011), has the best performance in every SNR condition. The accuracy it achieves of 96% at 9 dB and 86% at -6 dB corresponds to only 57% more keyword errors than the human on average. Given the margin of error for the listening tests, the performance of this system at the intermediate noise levels of 0 and 3 dB is not statistically different from that of the human listener.

Separate analysis of the impact of each strategy for target enhancement, feature extraction or decoding is difficult, since they were not always separately evaluated by their authors. Also, conclusions must be treated with caution as the relative contribution of system components is likely to be highly data- and task-dependent.

Nevertheless, by listing the strategies employed by the 8 top-performing systems, it appears that the most effective single strategies are the most established ones, namely

- multi-condition training,
- spatial diversity-based enhancement,
- robust training,

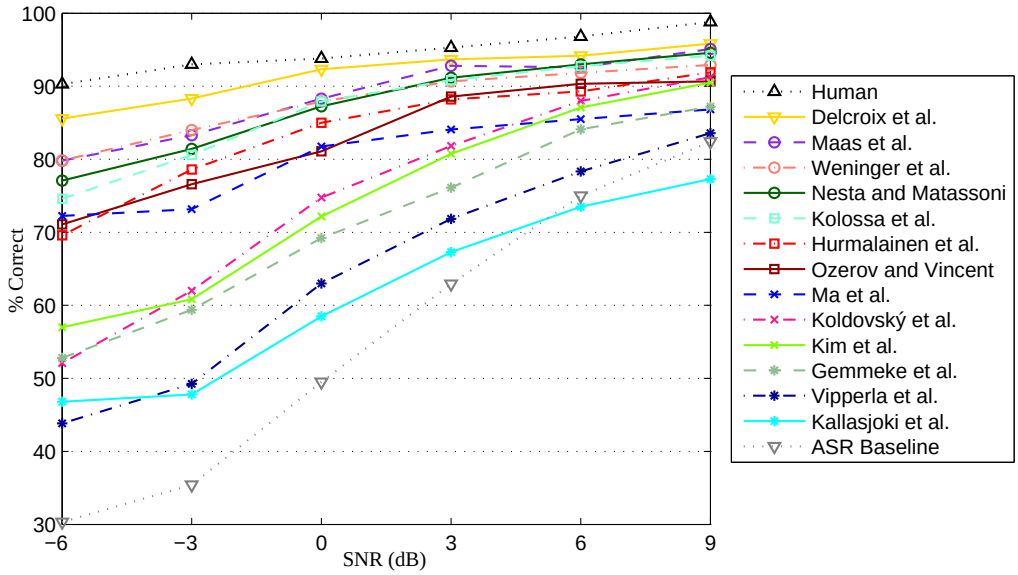

Figure 4: Keyword accuracy achieved by the thirteen submitted systems, compared to the human listener and the ASR baseline.

which are employed by 6, 6 and 5 systems out of 8 respectively. For instance, compared to a baseline of 55.9%, Kolossa et al. (2011), Nesta and Matassoni (2011) and Delcroix et al. (2011) reported average keyword accuracies of 80.6%, 76.8% and 69.0% respectively using multi-condition training, semi-blind source extraction based on ICA and Wiener post-filtering, or dMMI discriminative acoustic model training with automatic model size selection as implemented witin their recognizer SOLON.

The top-performing system of Delcroix et al. (2011) has not succeeded due to the stand-out performance of any one particular component but instead through careful combination of well-engineered signal processing and statistical modelling. The authors present a detailed system analysis that provides an instructive demonstration of the interactions between gains due to each processing stage. Remarkably, their target enhancement stage DOLPHIN (standing for dominance based locational and power-spectral characteristics integration), which exploits combined spatial and spectral diversity, is by itself able to improve average keyword accuracy to 85.1% with SOLON. Likewise, multi-condition training applied with no target enhancement increased the score to 84.7% with SOLON using a 42 times larger multi-condition set than the clean training set. Combining multi-condition training, DOLPHIN-based target enhancement and SOLON-based robust training increased performance to 90.2% – which alone would have

been sufficient to have been the top scoring system.

More recent strategies bring smaller additional improvement. In the case of Delcroix et al. (2011), performance was further increased to 91.1% via MLLR-style model adaptation combined with a simple form of uncertainty decoding known as DVA. The team also experimented with an exemplar-based denoising stage and found it could be gainfully employed in conjunction with systems trained on clean data, but was less compatible with multi-condition training regimes. A final 0.6% improvement to 91.7% was squeezed out by fusing the outputs of three systems that used different selections of these processing stages.

4.3. Discussion: Challenge limitations and future evaluations

4.3.1. Challenge Complexity

The design of speech technology evaluations involves compromise. On the one hand there is a desire that the evaluation should closely model the specific speech application that is motivating the challenge. If the task is oversimplified it encourages artificial toy solutions that lead to research dead-ends when it is found that the systems fail to scale. On the other hand there is the need to design a task that is sufficiently tractable that it engages and carries forward the target research communities. If the task is too realistic then at best competitors will fail to make progress and systems will fail in uninteresting ways – at worst, researchers will be reluctant to engage with the task and the evaluation will be stillborn.

Faced with this compromise the current challenge has aimed to succeed by starting from existing robust ASR evaluations and taking steps which are small but which force novel solutions. A key decision in this respect was to focus on the complexity and realism of the noise background while employing an unrealistically simple target speech signal. The simplicity of the underlying ASR task has allowed the challenge to attract a number of researchers who would not have had the resource to engage in a large vocabulary task. Nevertheless, even this simple ASR task has highlighted the need to co-design and carefully integrate the signal processing front-ends and statistical back-ends of speech recognition systems.

Participants were well aware of the limitation of the challenge and many of the papers caveat their conclusions with the need for validation on future and more realistic tasks. Surveyed for their opinion challenge entrants have highlighted three main dimensions of difficulty that need to be explored from the current starting point:

1. **Variability of speaker location** – In the current evaluation the target speaker

remains at a fixed position and orientation with respect to the listener. Further, although room responses were mismatched across training, development and test sets, the same room response was used for every utterance within each set. Although it may be acceptable for a distant speech application to be tuned for a particular 'sweet spot', a practical system would still need to be able to tolerate a good degree of speaker head movement. Previous evaluations using speech that has been recorded live in situ have shown that ASR systems can be surprisingly sensitive to speaker location (Himawan et al., 2008). Further, in a real system there would be considerable additional channel variability caused by other external effects such as changes in humidity, temperature, furniture placement, room occupancy etc.

2. **Vocabulary size** – Employing a small vocabulary size is a convenient way of reducing the complexity of an ASR task: small-vocabulary recognisers are generally easier to build and train; they bypass the need for complex language modelling; they allow recognition experiments to be run with little computational cost. However, there is a very large risk that techniques designed for small vocabularies will fail to scale to larger ones. For example, consider the task of digit recognition. In English digits can be distinguished from each other by observing their vowels alone. A digit recogniser based solely on vowel identification might look fantastically robust but would fail to work when faced with vocabularies containing words distinguished on the basis of, say, their unvoiced consonants. Further, in small vocabulary task lexical constraints may be highly informative, but as the vocabulary size increases the lexical constraint decreases. This shift can undermine conclusions that are drawn from a small vocabulary task.
3. **Speech naturalness** – The current task employed simple command utterances that were recorded in a booth from talkers reading a prompt. The repetitive structure of the utterances and the prompted-recording set-up encouraged a style of speech that lacks a lot of speech's natural variability: speakers fall into a consistent rhythm and intonation pattern and tend to speak at a consistent level with little emotion (other than boredom!). The unnaturally decreased speech variance certainly makes the recognition task easier, but, unfortunately, it might make it easier in ways that favour approaches that do not work well on more natural speech. For example, the surprisingly good performance of 'exemplar-based' approaches on the current task could have been a result of the unnatural degree of similarity between exemplars of the same token. A further limitation of the 'record-

then-mix' approach is that it does not model active-talking effects, like the Lombard effect, in which talkers subtly (both consciously and unconsciously) adapt the timing and quality of their speech to allow themselves to be better heard against the noise background (Cooke and Lu, 2010).

Future challenges could be constructed with an increase of difficulty along any of the dimensions described above. However, as discussed earlier it is important to advance in careful steps and according to feedback from the research communities involved.

4.3.2. Challenge Focus

When designing an ASR challenge it is necessary to carefully consider the 'rules' so that attention can be focussed on key scientific/engineering questions. The difficulty here is that focus is often gained by taking a reductionist view that underestimates the importance of the complex interactions between components. For example, it could be argued that results in the current challenge would have been more 'comparable' if a focus had been placed on enhancement by constraining all competitors to use a pre-defined standard acoustic model structure and/or training regime (i.e. ala the Aurora 2 challenge). However, given it is increasingly clear that it is important to co-optimise the signal processing and statistical modelling, it is also clear that it is impossible to select a 'back-end' that does not unfairly advantaged one system over another. Indeed, a fundamental aim of the challenge was to build bridges between the signal processing and statistical modelling communities and encourage them to work together to develop deeply coupled systems.

Nevertheless, now that some experience has been gained, future challenges based on the CHiME scenario would benefit from tighter constraint. Constraints that could be introduced include

1. **Multi-condition training data**: Many teams employed multi-condition training but each developed their own set of noisy utterances from the noise-free speech and noise backgrounds provided. However, huge differences in the amount of data used are likely to have made a big difference to the effectiveness. Providing a fixed set of data (e.g. based on the regime employed by the winning system) would reduce system variability.
2. **Acoustic context**: The utterances were embedded in continuous audio and no restrictions were placed on the duration of the context that could be employed. Participants were also left free to employ the post-utterance context. At very least it would seem rational to prohibit use of post-utterance

audio that would not be available to a responsive real-time system, for example.

3. **Computational complexity**: In order to make challenges assessable to a wide audience it is necessary to keep the scale reasonably small. However, small tasks can encourage well-resourced participants to employ algorithms that would clearly scale badly and become computationally intractable when applied to data that is closer to real applications. One safeguard against this is to set limits on amount of computation allowed – typically via a 'real-time' factor. At very least, a first step would be to ask participants to report computational cost and an analysis of complexity along with their results.

In order to maximise scientific impact without unduly stifling creativity, future challenges could allow for two sets of results: a compulsory 'closed system' result which adheres to tight competition constraints and allows meaningful cross-system comparisons, and an optional 'open system' result in which rules are relaxed in order to explore unconstrained performance potential.

Some of the above lessons learnt have already been applied to the Second CHiME Speech Separation and Recognition Challenge, which is currently running (Vincent et al., to appear). This challenge extends the difficulty in two separate directions, namely variability of speaker location and vocabulary size, and provides tighter instructions.

5. Conclusion

Distant microphone speech recognition in everyday listening conditions is a challenging goal that will only be achieved with a coordinated and multidisciplinary research effort. This paper has presented a speech recognition challenge that has been motivated by this goal. The task was based on the recognition of simple command sentences reverberantly mixed into binaural recordings of a busy domestic environment containing multiple competing sound sources. The challenge attracted thirteen submissions. The successful systems have employed multiple strategies to increase robustness at each stage of the recognition process and complementarily combined techniques for target enhancement (ITD/ILD clustering, GMM/NMF/NMD...), robust feature extraction (GFCC, RNN...) and robust decoding (multi-condition training, MLLR/MAP adaptation, uncertainty decoding...). The best overall system (Delcroix et al., 2011) was able to recognise the test utterances with an error rate that was only 57% higher than that

of a highly motivated human listener, and with a performance that was not significantly less than human performance at the 0 and 3 dB SNR levels. Although without further controlled experimentation it is hard to draw strong conclusions about which strategies work best, it is clear that multi-condition training and spatial enhancement (e.g., via ITD/ILD clustering) are the most effective single strategies, which can improve the keyword recognition accuracy by more than 20% absolute compared to the baseline. By combining these and other strategies in an appropriate fashion, it is possible to engineer systems that are remarkably robust to substantial degrees of non-stationary noise. However the resulting performance improvements do not add up and an improvement on the order of 10% absolute only can be achieved compared to multi-condition training alone. The paper concluded by discussing the limitation of the current challenge and the key dimensions of difficulty that might be explored in future more realistic evaluations, some of which have been taken into account for the Second CHiME Speech Separation and Recognition Challenge (Vincent et al., to appear).

Acknowledgment

The authors thank the EU Network of Excellence PASCAL (Pattern Analysis, Statistical modeling and Computational Learning) (www.pascal-network.org) for funding to support the Challenge; ISCA (International Speech Communication Association) (www.isca-speech.org) for supporting the workshop at which Challenge results were originally disseminated; the UK EPSRC for funding the CHiME research project which provided the domestic audio recordings.

References

Christensen, H., Barker, J., Ma, N., Green, P., 2010. The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments. In: Proc. Interspeech 2010. Tokyo, Japan, pp. 1918–1921.

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

Cooke, M., Hershey, J., Rennie, S., 2010. Monaural speech separation and recognition challenge. Computer Speech and Language 24, 94–111.

Cooke, M., Lu, Y., 2010. Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. J. Acoust. Soc. Am. 4 (128), 2059–2069.

Delcroix, M., Kinoshita, K., Nakatani, T., Araki, S., Ogawa, A., Hori, T., Watanabe, S., Fujimoto, M., Yoshioka, T., Oba, T., Kubo, Y., Souden, M., Hahm, S.-J., Nakamura, A., 2011. 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 Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 12–17.

Farina, A., 2000. Simultaneous measurement of impulse response and distortion with a swept sine technique. In: Proc. 108th AES Convention. Paris, France, p. 5093.

Gemmeke, J. F., Virtanen, T., Hurmalainen, A., 2011. Exemplar-based speech enhancement and its application to noise-robust automatic speech recognition. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 53–57.

Hershey, J. R., Rennie, S., Olsen, P. A., Kristjánsson, T., 2010. Super-human multi-talker speech recognition: A graphical modeling approach. Computer Speech and Language, 24, 45–66.

Himawan, I., Mccowan, I., Lincoln, M., 2008. Microphone array beamforming approach to blind speech separation. In: Popescu-Belis, A., Renals, S., Bourlard, H. (Eds.), Machine Learning for Multimodal Interaction. Spring Berlin, pp. 295–308.

Hurmalainen, A., Mahkonen, K., Gemmeke, J. F., Virtanen, T., 2011. Exemplar-based recognition of speech in highly variable noise. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 1–5.

Kallasjoki, H., Keronen, S., Brown, G. J., Gemmeke, J. F., Remes, U., Palomäki, K. J., 2011. Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 58–63.

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

Koldovský, Z., Málek, J., Nouza, J., Balík, M., 2011. CHiME data separation based on target signal cancellation and noise masking. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 47–50.

Kolossa, D., Fernandez Astudillo, R., Abad, A., Zeiler, S., Saeidi, R., Mowlaee, P., da Silva Neto, J. P., Martin, R., 2011. CHiME challenge: Approaches to robustness using beamforming and uncertainty-of-observation techniques. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 6–11.

Ma, N., Barker, J., Christensen, H., Green, P., 2011. Recent advances in fragment-based speech recognition in reverberant multisource environments. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 68–73.

Maas, R., Schwarz, A., Zheng, Y., Reindl, K., Meier, S., Sehr, A., Kellermann, W., 2011. A two-channel acoustic front-end for robust automatic speech recognition in noisy and reverberant environments. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 41–46.

Makino, S., Lee, T.-W., Sawada, H. (Eds.), 2007. Blind speech separation. Signals and Communication Technology. Springer.

Nesta, F., Matassoni, M., 2011. Robust automatic speech recognition through on-line semi blind source extraction. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 18–23.

Ozerov, A., Vincent, E., 2011. Using the FASST source separation toolbox for noise robust speech recognition. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 86–87.

Pearce, D., Hirsch, H.-G., Oct. 2000. The aurora experimental framework for the performa nce evaluation of speech recognition systems under noisy conditions. In: Proc. ICSLP '00. Vol. 4. Beijing, China, pp. 29–32.

Vincent, E., Araki, S., Theis, F., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker,

B., Lutter, D., Duong, N., 2012. The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges. Signal Processing 92, 1928–1936.

Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M., to appear. The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines. In: Proc. 2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP).

Vipperla, R., Bozonnet, S., Wang, D., Evans, N., 2011. Robust speech recognition in multi-source noise environments using convolutive non-negative matrix factorization. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 74–79.

Weninger, F., Geiger, J., Wöllmer, M., Schuller, B., Rigoll, G., 2011. The Munich 2011 CHiME challenge contribution: NMF-BLSTM speech enhancement and recognition for reverberated multisource environments. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME). pp. 24–29.

Wölfel, M., McDonough, J., 2009. Distant Speech Recognition. John Wiley and Sons.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2006. The HTK Book, version 3.4. University of Cambridge.