THE SECOND 'CHIME' SPEECH SEPARATION AND RECOGNITION CHALLENGE: AN OVERVIEW OF CHALLENGE SYSTEMS AND OUTCOMES

Emmanuel Vincent[⋆], *Jon Barker*[†], *Shinji Watanabe*[‡],
Jonathan Le Roux[‡], *Francesco Nesta*[§], *Marco Matassoni*[‖]

[⋆]Inria, Villers-lès-Nancy, France; [†]University of Sheffield, Sheffield, UK;
[‡]MERL, Cambridge, MA, USA; [§]Conexant Systems, Newport Beach, CA, USA; [‖]FBK-Irst, Trento, Italy

ABSTRACT

Distant-microphone automatic speech recognition (ASR) remains a challenging goal in everyday environments involving multiple background sources and reverberation. This paper reports on the results of the 2nd 'CHiME' Challenge, an initiative designed to analyse and evaluate the performance of ASR systems in a real-world domestic environment. We discuss the rationale for the challenge and provide a summary of the datasets, tasks and baseline systems. The paper overviews the systems that were entered for the two challenge tracks: small-vocabulary with moving talker and medium-vocabulary with stationary talker. We present a summary of the challenge findings including novel results produced by challenge system combination. Possible directions for future challenges are discussed.

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

1. INTRODUCTION

The distant microphone scenario remains one of the major unsolved challenges of automatic speech recognition (ASR) research. There are two main components to this problem: first, target speech signals are subject to the effect of room acoustics so that recorded signals correspond to the original speech signals convolved with the room impulse responses. This effect, widely known as reverberation, cannot be easily predicted because there is no control over the talker-receiver geometry or room characteristics. Second, the target speech is mixed with other sound sources in the environment creating a potentially complex acoustic noise background. Separate research communities have worked on different aspects of these problems and fresh approaches are rapidly emerging [1, 2, 3].

In 2011 the 1st CHiME challenge [4] was held with the aim of progressing distant microphone ASR by bringing together researchers from the signal processing, speech processing and machine learning communities. The challenge involved recognition of utterances that were reverberantly mixed into stereo (binaural) backgrounds recorded in the living room of a family home involving noise sources such as concurrent speakers, TV, game console, footsteps, and distant noise from outside or from the kitchen. The 1st challenge deliberately concentrated on a small-vocabulary recognition task and was designed in such a way as to be easily accessible beyond the traditional ASR community. The challenge attracted 13 systems which employed a wide range of signal enhancement and robust acoustic modelling strategies. The challenge was a success in as much as the best performing system arose from a large multidisciplinary team with the expertise to co-optimise the front-end signal processing and the statistical 'back end' [5]. Extended versions of many of the 1st CHiME challenge systems are presented in a recent Special Issue of Computer Speech and Language [6].

Following the success of the 1st challenge, and with the support of the IEEE AASP, MLSP and SL Technical Committees, a second CHiME challenge was designed that would build on the first by stepping closer toward the demands of a realistic application. Two separate limitations of the 1st challenge were considered: first, the 1st challenge used a small vocabulary recognition task which lowered the bar for participation but which presented the danger of promoting techniques that fail to generalise to less constrained larger-vocabulary tasks. Second, the target talker had been mixed into the backgrounds using a fixed room impulse response, i.e., failing to model the variability caused by talker movement – one of the key design problems for distant microphone ASR. The new challenge was carefully designed to balance the need to address these issues with the desire to grow complexity in small steps and provide some 'backward-compatibility' with the previous challenge edition.

The 2nd CHiME challenge attracted entrants from 13 groups. These groups presented their work at a dedicated workshop[1] that was held in conjunction with ICASSP 2013 and details of the individual systems can be found in the workshop proceedings [7–19]. The purpose of this paper is to provide an overview to this body of work, to compare the performances of the CHiME systems, to provide some novel results on CHiME system combination and to draw conclusions for the future of distant microphone ASR and for the

[1]The workshop was made possible by financial support from our industrial sponsors: Conexant Systems, Adobe, Audience, Google and MERL.

design of future evaluations.

The structure of the rest of the paper is as follows. In Section 2, we summarise the design of the datasets and define the tasks that the challenge addresses and in Section 3 we briefly describe the baseline recognisers and report their performance. (For a detailed account of the challenge set-up readers are referred to [20]). Section 4 will provide an overview of the systems that were submitted. Section 5 provides a summary of the system performances and challenge outcomes. Section 6 concludes with a discussion of directions for future challenges.

2. CHALLENGE DESIGN

The challenge considers a single target talker speaking in a noisy domestic environment recorded using binaural microphones. The data are generated by convolving clean target speech signals with binaural room impulse responses (BRIRs) and mixing the result with the noise backgrounds. The BRIRs and noise backgrounds were recorded in the same domestic living room using a B&K head and torso simulator (HATS). In the 1st CHiME challenge [4] the target speech was taken from a small vocabulary and was added into the backgrounds using a fixed and constant BRIR. The 2nd challenge increased the difficulty along two alternative directions: Track 1 investigated the effect of introducing small speaker movements; Track 2 employed a more demanding medium vocabulary target speech corpus. For each task, competitors were provided with separate training, development and test sets and a set of instructions to constrain fair use of the data.

2.1. Second challenge, Track 1: small vocabulary

The small vocabulary track followed the design of the 1st CHiME challenge and employed the Grid speech corpus [21]. This corpus consists of a collection of 34 speakers each reading 1,000 simple 6-word utterances of the form <command:4> <color:4> <prepos.:4> <letter:25> <digit:10> <adverb:4> where the numbers in the brackets indicate the number of word choices. The task is to report the letter and digit tokens and performance is measured as the percentage of tokens recognised correctly.

The clean utterances were convolved with the BRIRs so as to mimic a speaker at a distance of approximately 2 m in front of the HATS. However, in contrast to the 1st CHiME challenge, the precise simulated location was changed from utterance to utterance within a box of dimension 20 cm by 20 cm and a time-varying BRIR was used to model small 5 cm translational head movements occurring during the utterance. The time-varying BRIRs were constructed by measuring the true BRIRs on a 2-D grid with a resolution of 2 cm and then using linear interpolation to estimate BRIRs at a finer spacing. See [20] for further details.

The level of the reverberated utterances matched that of conversational speech spoken live in the room. These utterances were then positioned at selected quieter or louder times within the background recordings so as to produce noisy utterances at 6 different signal-to-noise ratios (SNRs): -6, -3, 0, 3, 6 and 9 dB. Note, this approach allows the target speech to remain at natural levels (i.e., it is not arbitrarily scaled to produce the desired SNRs) but it also means that the SNR settings tend to have different types of noise background.

We generated a development set and a test set using two separate sets of 600 utterances, each of these utterances being used at each of the 6 SNRs. Utterances were added to the continuously-recorded noise backgrounds at positions such that no utterances overlapped. A 17,000 utterance training set was produced using 500 utterances from each of the 34 Grid talkers made available as clean, reverberated and noisy recordings.

2.2. Second challenge, Track 2: medium vocabulary

The medium vocabulary task was constructed using the Wall Street Journal (WSJ0) 5k vocabulary read speech corpus [22]. The recognition task is to transcribe the entire utterance and performance is evaluated in terms of word error rate (WER).

The data were mixed using the same approach as employed in the 1st CHiME challenge, i.e., similar to that described in the previous section except that each utterance is convolved with a fixed BRIR recorded at precisely 2 m directly in front of the HATS. As before, SNR was controlled via the temporal positioning of the utterance. The WSJ0 utterances can be quite long so SNR was defined to be the median value of the segmental SNRs computed over segments of 200 ms. However, because of the large size of the WSJ0 corpus, it was found to be impossible to achieve the -6 to 9 dB range of SNRs using temporal positioning alone and at the same time preventing the noise signals in different mixtures to partially overlap. Therefore, when necessary, a limited amount of rescaling on the speech signal was applied. If the rescaling was still not sufficient for generating all the mixtures, overlapping noise sequences were also included in the search.

The development set employs 409 noisy utterances constructed from 10 speakers forming the "no verbal punctuation" part of the WJS0 speaker-independent 5k vocabulary development set. The test set comprises 330 noisy utterances from 8 other speakers from the Nov92 ARPA WSJ evaluation set. The test and development sets are provided at each of the 6 SNRs. The training set includes 7138 noisy utterances constructed from 83 speakers forming the WSJ0 SI-84 training set with each utterance at an SNR randomly selected within the -6 to 9 dB range.

2.3. Instructions

A number of challenge rules were imposed on all entrants. These rules were designed so that the systems would be

broadly comparable, but were kept sufficiently open so as not to artificially favour any one technique or research community. The rules can be summarised as follows. The systems were allowed to exploit knowledge of the temporal placement of the utterances (i.e., no automatic voice activity detection was required), of the surrounding acoustic background, of the speaker identity (for Track 1) and of the speaker movements (for Track 1). They were forbidden from exploiting the SNR labels, the fact that the same utterances are used at each SNR, the fact that the same noise background is used in the development set and the final test set, or the fact that the same utterances are used within the clean, reverberated and noisy training set (note that this rules out so-called "stereo data" approaches which employ clean and noisy versions of the same utterances), the fact that the BRIRs are identical between different test utterances (for Track 2) or the fact that the noise signals in the test utterances may temporally overlap (for Track 2). Systems should use the language models provided. All parameters should be tuned on the provided training and development sets and run on the final test set only once.

3. CHALLENGE BASELINES

For each of the two challenge tracks a baseline recognition system was produced. These systems are summarised below and full details are provided in [20].

The binaural signals were first downmixed to a single channel by averaging. Feature vectors were constructed at a 10 ms frame period from overlapping 25 ms signal windows. Frames were parameterised using a 39 dimensional feature vector composed of 12 Mel-cepstral coefficients plus log-energy together with their deltas and accelerations. Cepstral mean normalisation was employed.

The **Track 1** baseline system is the same as that used in the first CHiME challenge. Each of the 51 words in the Grid vocabulary is modelled with a left-to-right HMM with 2 states per phoneme. Each state is modelled using a 7-component Gaussian mixture model (GMM) with components having diagonal covariance. The language model is fixed according to the simple syntax of the Grid utterances. HTK scripts are provided for building first speaker-independent models and then speaker-dependent models using the challenge-defined 17,000 utterance training set. Recognition is performed using the HVite Viterbi decoder and no pruning.

The **Track 2** baseline follows the recipe in [23]. The system employs 39 phonemes plus silence (sil) and short pause (sp) models. Each phone is modelled using a 3-state HMM with each state modelled as an 8-component diagonal covariance GMM (16-component for the silence model). Triphone states are clustered and tied, reducing the number of independent states down to 1860. The standard WSJ 5K non-verbalised closed bigram language model is employed. Training scripts are provided for re-estimating model parameters from a clean speech acoustic model, but with no change to the model topology. Recognition is performed using HVite with a pruning threshold.

Table 1. Strategies employed by the Track 1 systems.

	Signal	Feat.	Inference
Geiger et al. [7]	X	X	X
Moritz et al. [12]	X	X	X
Meutzner et al. [11]	X	X	X
Ma and Barker [10]	X	X	X
Tran et al. [18]	X	X	X
Nesta et al. [14]	X	X	X
Gemmeke et al. [8]	X		
Mowlaee et al. [13]	X	X	
Sivaraman et al. [15]		X	X
Yilmaz et al. [19]		X	X
Stadtschnitzer et al. [16]	X		X

Table 2. Strategies employed by the Track 2 systems.

	Signal	Feat.	Inference
Tachioka et al. [17]	X	X	X
Nesta et al. [14]	X	X	X
Geiger et al. [7]	X	X	X
Hurmalainen et al. [9]	X		X

4. SUBMITTED SYSTEMS

13 teams participated in the challenge; 9 of which evaluated their system in Track 1 only, 2 in Track 2 only and 2 in both tracks. The systems typically combined multiple strategies that can be individually grouped under three headings roughly corresponding to a sequence of processing stages: target signal enhancement, robust feature extraction and robust statistical modelling/inference. Tables 1 and 2 summarise the basic strategies employed by each system.

4.1. Target enhancement strategies

The first processing stage consists of target signal enhancement. This is typically achieved by forming a time-frequency representation of the signal and applying a linear filter in each time-frequency bin. The filter parameters are estimated by either exploiting the spatial and/or spectral diversity of the speech and noise sources.

Spatial diversity is based on the fact that the target speech and interfering noise sources have different spatial locations. This includes beamforming approaches [16] or exploiting interaural phase and level differences [11, 10, 17]. Filter parameters can be learnt from the data, e.g. by constructing inter-channel level difference (ILD) and interchannel time differ-

ence (ITD) histograms for the target and background [17, 11] or steered to fit the known location of the target speech source [10].

Spectral diversity is based on the assumption that the speech and noise have differing spectra. Techniques included building separate GMMs of speech and noise, non-negative matrix factorisation (NMF) (e.g. [7]) and exemplar based enhancement [8]. For example, [7] describes the noisy spectrogram as a sum of a speech and noise spectrogram each estimated from a separate sparse dictionary. Dictionary weights are estimated and the speech signal is estimated by Wiener filtering.

Spatial and spectral cues can be used in conjunction to potentially allow separation in situations where either cue alone would fail. The straightforward approach is to use one followed by the other. For example a delay-and-sum beamformer followed by a codebook-driven spectral enhancement [13] or a spatial dictionary based blind source extraction followed by spectral filtering [14]. Alternatively, spatial and spectral cues can be applied together by using a joint probabilistic framework capable of capturing correlations between the two (e.g. [18, 12]).

4.2. Feature extraction strategies

Feature extraction strategies aim to provide invariance to the background noise that remains after the target enhancement stage. A wide variety of approaches have been employed including normalized modulation cepstral coefficients [15], Gabor filterbank features [12], gammatone frequency cepstral coefficient (GFCC) features [14, 10], nonnegative sparse classification (NSC) features [7], recurrent neural network (BLSTM) features [7], and vocal tract variable trajectories [15]. These features have selective sensitivity to speech-like patterns in either frequency and/or time. A separate strategy is to apply feature transforms to either decorrelate features, e.g. principal component analysis (PCA) (e.g. [15]) or to increase the discriminating power of the recogniser, e.g. linear discriminant analysis (LDA) (e.g. [11]) and feature-space maximum mutual Iinformation (f-MMI) (e.g. [17]). Some systems gain performance by using multiple features modelled either as separate streams [7] or by feature concatenation prior to dimensionality reduction [15].

4.3. Robust modelling/inference

The baseline recognition systems performed decoding (i.e., converting feature streams into word sequences) using a conventional HMM-GMM recogniser. Most CHiME systems adapted this background recogniser in some manner. The most commonly employed strategy was noise adaptive training, i.e., simply retraining the models on noisy speech processed by the target-enhancing front-end. Systems also used discriminative or speaker-adaptive techniques either during training (to improve the models, e.g. [12]) or built into the decoding objective (to compensate for model mismatch, e.g. [17]). A small number of systems employed some form of uncertainty propagation – modelling noisy observations as distributions (e.g. uncertainty decoding [14] and fragment decoding [10]). One system employed a purpose-built decoding algorithm compatible with a dictionary of variable-length exemplars [19]. Finally, four teams improved performance using system combination either at the feature level using either multistream decoding [7] or feature vector concatenation [15], or at the decision level using recogniser output voting error reduction (ROVER) [17, 11].

5. RESULTS

5.1. Submitted systems

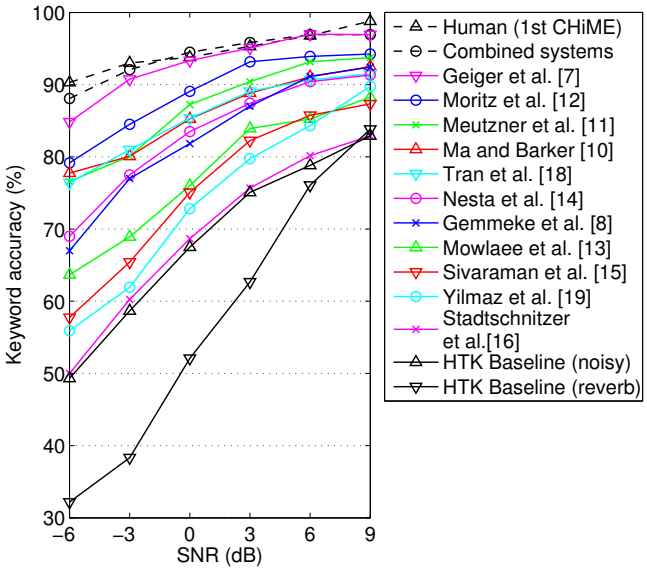

Fig. 1. Keyword accuracy for the Track 1 systems compared to a human listener and to the ASR baselines. A system combination performance is also shown (see Section 5.2).

Results of the 11 systems entered for Track 1 and the 4 systems for Track 2 are presented in Figures 1 and 2 respectively, in which systems are ordered by decreasing performance. Also shown are performances for the baseline systems trained on noise-free reverberated speech or noise-added speech and, for Track 1, the human performance that was measured on the similar 1st CHiME challenge [6]. For Track 1 the system performances are evenly spread within the range from just above the baseline to just below human performance. The performance curves for both Track 1 and 2 are roughly parallel indicating that individual systems are broadly optimised across the SNR range rather than specialising on specific SNRs.

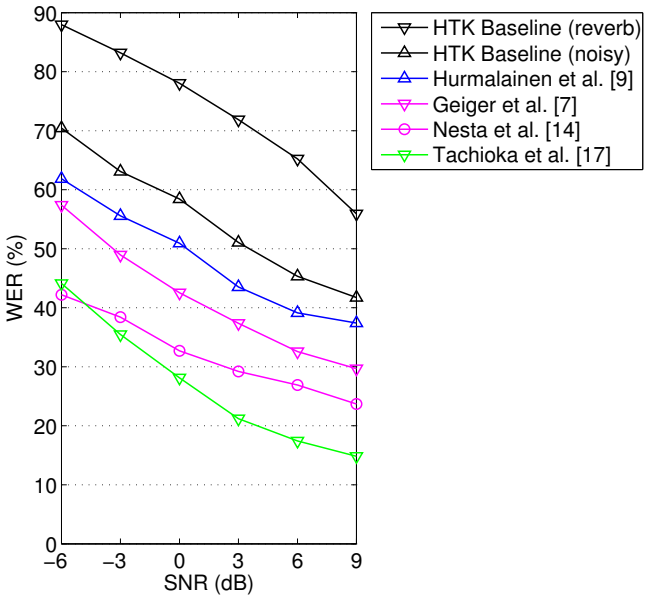

Fig. 2. WERs for the Track 2 systems compared to the ASR baselines.

Careful analysis shows that the strategies which are most effective for both tracks are spatial diversity based enhancement and noise adaptive training. Spectral diversity based enhancement also performed extremely well in the small vocabulary setting of Track 1, but less so in Track 2 due to its relative novelty in this setting. By contrast, careful design of the ASR back-end played a major role in performance in Track 2 but it had a smaller impact in Track 1.

The above strategies are insufficient, however, and achieving good performance requires combination of multiple features or systems, each of which involves modifications at every stage of ASR. Indeed, the top systems for the two tracks are highly complex systems. For Track 1, the best system is using exemplar-based enhancement followed by multiple feature streams that combine the advantages of MFCC, BLSTM and sparse-coding features and a decoder employing noise-adaptive training and MAP speaker adaptation [7]. The performance of this system is only marginally poorer than that of the human listener. The top Track 2 system is using spatial enhancement, a host of feature-space transformations (LDA, MLLT, MLLR) and then a decoder employing discriminative acoustic and language models plus a ROVER combination of system variants [17]. This system has a WER that is just 49% that of the multicondition-trained baseline system, and at 9 dB WER is reduced from 41.7% to just 14.8%.

5.2. System combination

CHiME challenge entrants were also asked to submit recognition transcripts for both the development and final test sets. Access to these transcripts allowed us to perform system combination experiments for the Track 1 systems.

System outputs were combined using a standard weighted-voting technique. Letter and digit tokens were considered independently. For a given input utterance the output token for the combined system was determined by taking a weighted vote accross the individual system outputs. Each classifier's vote was weighted by its logodds of beings correct. The logodds were determined by measuring the classifier's letter and digit recognition performances on the development set and averaging across all SNRs (i.e., the combination was SNR-independent and thus did not break the rule that systems should not exploit knowledge of the SNR).

For each N, we evaluated all possible combinations of N systems among the top $N+1$ systems. Using the development set, the best performance was attained by combining all but the third system among the top 5 systems, i.e., the 4 systems in [7], [12], [10] and [18]. This is consistent with the phoneme confusion metrics in [24] which show that the third system in [11] is the least different from the other top 4 systems and therefore least likely to be useful in a combination. Final Track 1 test set results for the combination of these 4 systems are shown in Figure 1. Averaged over SNRs, the keyword accuracy achieved by the combined system is 94.11% compared to 92.99% for the best single system and 94.67% for the human. The combined system outperforms the human by approximately 0.5% (absolute) at the intermediate 0 and 3 dB noise levels.

6. CONCLUSIONS AND FUTURE DIRECTIONS

The purpose of the 2nd CHiME challenge was to separately investigate the demands on systems of introducing speaker motion and of increasing the complexity of the speech recognition task. It has been found that *small* speaker movements do not significantly increase the task difficulty. The best performing system achieved a score that was similar to that achieved by best system for the 1st challenge and had an error rate that was only 30% (relative) greater that that of the human listener. Teams that directly compared their system's performance on the CHiME 1 and CHiME 2 datasets achieved equal performance on each [10]. On the other hand, increasing vocabulary size did significantly increase the challenge difficulty. Performance gains relative to the baseline systems were substantially poorer and the best WERs at -6 dB remained at over 40%.

The most effective single strategies to address the challenge turned out to be spatial diversity based enhancement and noise adaptive training. Spectral diversity based enhancement was also beneficial with a small vocabulary, while improvements to the ASR back-end were essential with a larger vocabulary. Nevertheless, it is remarkable that, in either case, the best results were obtained from the combination of highly complicated and tuned systems resulting from collaborative efforts.

It is clear that the challenges reported here are still highly artificial. For example, if the speaker location was not approximately fixed in time and space then we would need to solve other high-level problems such as speaker tracking, speaker identification and speech activity detection. Further, we would need solutions to the problems caused by variability in speaker-receiver geometry. Future editions of the challenge will attempt to move closer to realistic conditions but we need to make advances while remaining aware of the need to retain involvement from a broad community of researchers. This will be best achieved by extensive consultation and cross-community discussion.

7. REFERENCES

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

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

[3] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*, Wiley, 2012.

[4] J. Barker and E. Vincent, Eds., *Computer Speech and Language*, vol. 27, 2013.

[5] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S.-J. Hahm, and A. Nakamura, "Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation," in *Proc. CHiME-2011*, Florence, Italy, Sept. 2011, pp. 12–17.

[6] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.

[7] J. T. Geiger, F. Weninger, A. Hurmalainen, J. F. Gemmeke, M. Wöllmer, B. Schuller, G. Rigoll, and T. Virtanen, "The TUM+TUT+KUL approach to the 2nd CHiME challenge: Multi-stream ASR exploiting BLSTM networks and sparse NMF," in *Proc. CHiME-2013*, Vancouver, Canada, June 2013, pp. 25–30.

[8] J. F. Gemmeke, A. Hurmalainen, and T. Virtanen, "HMM-regularization for NMF-based noise robust ASR," in *Proc. CHiME-2013*, Vancouver, Canada, June 2013, pp. 47–52.

[9] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Compact long context spectral factorisation models for noise robust recognition of medium vocabulary speech," in *Proc. CHiME-2013*, Vancouver, Canada, June 2013, pp. 13–18.

[10] N. Ma and J. Barker, "A fragment-decoding plus missing-data imputation ASR system evaluated on the 2nd CHiME challenge," in *Proc. CHiME-2013*, Vancouver, Canada, June 2013, pp. 53–58.

[11] H. Meutzner, A. Schlesinger, S. Zeiler, and D. Kolossa, "Binaural signal processing for enhanced speech recognition robustness in complex listening environments," in *Proc. CHiME-2013*, Vancouver, Canada, June 2013, pp. 7–12.

[12] N. Moritz, M. R. Schädler, K. Adiloglu, B. T. Meyer, T. Jürgens, T. Gerkmann, B. Kollmeier, S. Doclo, and S. Goetze, "Noise robust distant automatic speech recognition utilizing NMF based source separation and auditory feature extraction," in *Proc. CHiME-2013*, Vancouver, Canada, June 2013, pp. 1–6.

[13] P. Mowlaee, J. A. Morales-Cordovilla, F. Pernkopf, H. Pessentheiner, M. Hagmüller, and G. Kubin, "The 2nd 'CHIME' speech separation and recognition challenge: Approaches on single-channel source separation and model-driven speech enhancement," in *Proc. CHiME-2013*, Vancouver, Canada, June 2013, pp. 59–64.

[14] F. Nesta, M. Matassoni, and R. F. Astudillo, "A flexible spatial blind source extraction framework for robust speech recognition in noisy environments," in *Proc. CHiME-2013*, Vancouver, Canada, June 2013, pp. 33–38.

[15] G. Sivaraman, V. Mitra, and C. Y. Espy-Wilson, "Fusion of acoustic, perceptual and production features for robust speech recognition in highly non-stationary noise," in *Proc. CHiME-2013*, Vancouver, Canada, June 2013, pp. 65–70.

[16] M. Stadtschnitzer, D. Stein, and R. Bardeli, "Employing stochastic constrained LMS algorithm for ASR frontend processing," in *Proc. CHiME-2013*, Vancouver, Canada, June 2013, pp. 71–72.

[17] Y. Tachioka, S. Watanabe, J. L. Roux, and J. R. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," in *Proc. CHiME-2013*, Vancouver, Canada, June 2013, pp. 19–24.

[18] D. T. Tran, E. Vincent, D. Jouvet, and K. Adiloğlu, "Using full-rank spatial covariance models for noise-robust ASR," in *Proc. CHiME-2013*, Vancouver, Canada, June 2013, pp. 31–32.

[19] E. Yılmaz, J. F. Gemmeke, and H. Van hamme, "Noise-robust automatic speech recognition with exemplar-based sparse representations using multiple length adaptive dictionaries," in *Proc. CHiME-2013*, Vancouver, Canada, June 2013, pp. 39–43.

[20] E. Vincent, J. Barker, S. Watanabe, J. L. Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. ICASSP 2013*, Vancouver, Canada, May 2013, IEEE.

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

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

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

[24] S. R. M. Prasanna, B. Yegnanarayana, J. P. Pinto, and H. Hermansky, "Analysis of confusion matrix to combine evidence for phoneme recognition," Tech. Rep. RR 07-27, IDIAP, 2007.