THE AUTOMATIC SPEECH RECOGITION IN REVERBERANT ENVIRONMENTS (ASpIRE) CHALLENGE

Mary Harper

IARPA

mary.harper@iarpa.gov, incisiveanalysis@iarpa.gov

ABSTRACT

In this paper, we describe the ASpIRE (Automatic Speech recognition In Reverberant Environments) challenge, which asked participants to construct automatic speech recognition systems that were robust to a variety of acoustic environments and recording scenarios without having access to matched training and development data. We discuss the performance of the systems evaluated in the challenge, summarize how those systems were constructed, and draw conclusions about what contributed to the performance levels of the systems on the highly variable, noisy, reverberant evaluation data set constructed for this challenge.

Index Terms— reverberation, speech recognition, mismatch, robustness

1. INTRODUCTION

Much progress has been made at reducing Word Error Rate (WER) on speech to text (STT) over the past 25 years, as can be observed in Figure 1 [1]. However, when a microphone is far from the speaker (e.g., in meeting rooms, distant talking command and control recordings), STT performance can be severely degraded. The NIST meeting room evaluations, shown in pink in Figure 1, demonstrate that distant microphones represent a significant challenge to the speech research community. The effect of close versus distant microphones can be observed by comparing the curves for close lapel microphones with distant microphones and distant microphone arrays. For meeting room conditions with non-stationary noise and reverberation, the measured WER values for distant microphones and arrays are significantly higher than those for the close lapel microphones.

STT performance also degrades dramatically in the face of mismatch between training and test data conditions [2]. Despite the common wisdom that mismatch can be easily addressed by collecting and transcribing matched condition training and/or adaptation data, when STT is used in the field, the time, effort, and cost of transcribing data for the new conditions become prohibitive. Mismatch can occur, for example, due to differences in background noise, recording conditions (different microphones and rooms), and speaker characteristics. Changes in reverberation across rooms can be a significant source of mismatch. The International Computer Science Institute (ICSI) meeting room [3] study described in [4] used Efron's bootstrap [5] to analyze the sources of error in an STT system; when the conditions were matched (even if far-field), it was observed that model errors dominated, but in mismatched conditions, features were neither invariant nor separable, which caused significant error in addition to model errors. Research using the Augmented Multi-party Interaction (AMI) meeting room corpus [6] and the Multi-channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV) corpus [7,8], for example, has also shown that STT performance degrades for distant microphones when data used for training are mismatched with data used in testing.

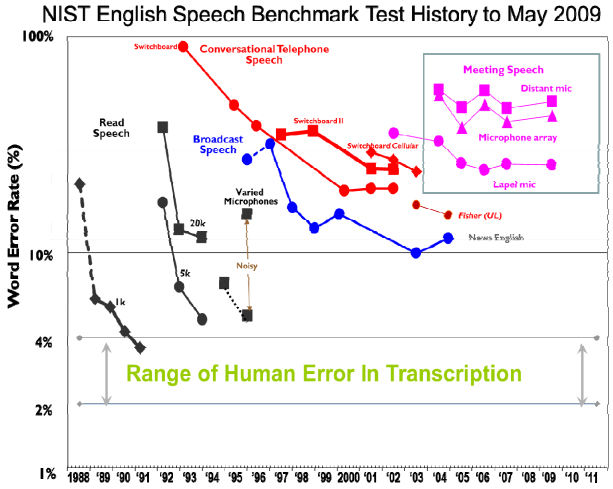

Figure 1. NIST STT Benchmarks (Meeting Room results depicted in pink).

1.1. Related Speech Recognition Challenges

Several prior challenges have involved speech recognition in reverberant environments. The 1st CHiME Speech Separation and Recognition challenge [9] posed the task of recognizing digits and letters in spoken sentences mixed with reverberant backgrounds that were binaurally recorded in a domestic environment with no variability in the location of the speaker. The challenge developers carefully constructed the development and evaluation data sets so they covered six different noise levels over different utterances containing digits and letters from the Grid speech corpus [10]. The utterances were prompted command and control phrases read by the speakers; the resulting speech was not conversational and had limited vocabulary. The challenge was designed to cover important aspects of distant microphone speech recognition while limiting the complexity of the challenge to enable participation by a wide range of participants.

The 2nd CHiME Challenge [11,12] evaluated the performance of STT systems in a domestic environment on a small-vocabulary task similar to that of the 1st CHiME challenge, except that the Grid speech audio was modified to simulate small speaker

movements, and an additional medium-vocabulary STT task was added that used the Wall Street Journal 5K vocabulary read speech corpus (WSJ0) [13] without speaker movements. The difficulty of the challenge was increased by introducing head movements and increasing vocabulary size.

The 3rd CHiME challenge [14] added new variability to the distant microphone speech recognition task by using a mobile tablet device to capture live speech with an array of six microphones positioned around the tablet frame in noisy environments rather than by using remixing as in the previous two challenges. Training, development, and evaluation sets were drawn from the WSJ0 corpus sentences presented on a tablet and read by American English speakers in cafes, at street junctions, on public buses, and in pedestrian areas.

The REVERB (REverberant Voice Enhancement and Recognition Benchmark) challenge [15] provided an evaluation framework with common datasets, tasks, and evaluation metrics for both speech enhancement and STT for speech recorded from stationary distant-talking speakers with 1-channel, 2-channel, or 8-channel microphone arrays placed at a fixed distance from the speakers. The training data consisted of clean recordings and simulated recordings (convolved with room impulse responses together with additive background noise) of sentences in the WSJ0 corpus read by a British English speaker (WSJCAM0) [16]. Two tasks were evaluated: (1) Enhancement of reverberant speech with single-channel and multi-channel de-reverberation techniques measured with both objective and subjective metrics; (2) STT performance measured using WER. Evaluation data consisted of simulated recordings of spoken words from WSJCAM0 [16] and from the Multi-Channel WSJ Audio Visual (MC-WSJ-AV) corpus [7,8], which contains utterances recorded in a single noisy and reverberant room with the above-mentioned microphone configurations.

1.2. The ASpIRE Challenge

The ASpIRE challenge encouraged participants to apply and refine state-of-the-art STT techniques to speech recordings of native speakers of American English where mismatch between training and test was high. Participants were expected to build automatic speech recognizers for English that were trained on conversational telephone speech but were robust to a variety of unknown acoustic environments and recording scenarios, without having access to matched training and development data.

For this challenge, we collected and transcribed a new evaluation speech database, the Mixer 8 Pilot corpus, recorded for the Intelligence Advanced Research Projects Activity (IARPA) by the Linguistic Data Consortium (LDC) and transcribed by Appen Butler Hill. The goal of the collection was to capture the types of variability that can be observed when applying STT in the field. The data were collected using a set of different microphones placed in a wide range of locations in seven different rooms (some classrooms and some office space) with various different shapes, sizes, surface properties, and noise sources. Speakers were also recorded from several different positions in each room. With the data simultaneously recorded using multiple distant microphones, the ASpIRE challenge offered two evaluation conditions:

1. **The Single Microphone Condition** tested the ability to mitigate noise and reverberation on sessions recorded across seven different rooms on a single distant microphone (selected randomly).
2. **The Multiple Microphone Condition** tested the ability to mitigate noise and reverberation on the same sessions as the single microphone condition recorded with a set of single distant microphones placed differently across the seven rooms.

The ASpIRE Challenge was launched on November 17, 2014. InnoCentive developed and supported the challenge website (https://www.innocentive.com/ar/challenge/9933624) and MIT Lincoln Laboratory provided test-and-evaluation support. InnoCentive made available a real-time online scoring utility to provide participants with system scores on the development and development-test data. In addition, a leaderboard was kept up to date on the challenge web site to allow participants competing in the single microphone condition to determine how their performance compared to top-scoring solutions on the development-test set.

The evaluation data were first made available to participants for the Single Microphone Condition on February 11, 2015, and evaluation submissions were due by February 18, 2015, to be eligible for award. The evaluation data were made available to participants for the Multiple Microphone Condition on February 19, 2015, and evaluation submissions were due by February 26, 2015, to be eligible for award. Participants were allowed to submit system outputs to either or both conditions. ASpIRE was a "Reduction-to-Practice Challenge"; participants were required to submit a system description of their solution together with the output from their STT system on the evaluation data to be eligible for a prize. Separate awards for the best system in the single microphone and the multiple microphone conditions were made after the winning systems were validated on a small set of new data, once the evaluation was completed.

Compared to prior speech recognition challenges, the ASpIRE challenge addressed far field microphone recordings and introduced the following conditions:

1. Conversational speech was used rather than read speech, increasing difficulty.
2. The vocabulary of the data sets was not controlled or limited; hence, the vocabulary was large and the development and evaluation data contained words that were not seen in training.
3. Evaluation data were explicitly designed to differ substantially from training data, as well as from the development data, to measure system robustness.
4. No information was provided for the audio files that might enable systems to make use of microphone type, room configuration, speaker position, or speaker identity.

In the next section, the data used in the challenge are described. The evaluation metric and scoring procedures for the ASpIRE challenge are discussed in Section 3. Results are presented in Section 4 and conclusions in Section 5.

2. ASPIRE DATA

To maximize mismatch, ASpIRE participants were expected to train their speech recognizers on conversational telephone speech, as described in Section 2.1. Testing was on far-field microphone recordings. Microphone recordings were provided as development data to participants (see Section 2.2), but these differed

substantially from the recordings developed for the final evaluation (see Section 2.3).

2.1. Training Set

The training set for the ASpIRE challenge was the Fisher conversational telephone training corpus [17]:

1. LDC2004S13 Fisher English Training Speech Part 1: 5,850 complete conversations, each lasting up to 10 minutes;
2. LDC2004T19 Fisher English Training Transcripts Part 1: transcripts for LDC2004S13 conversations with 12% transcribed at LDC and the rest by BBN and WordWave;
3. LDC2005S13 Fisher English Training Speech Part 2: 5,849 complete conversations, each lasting up to 10 minutes; and
4. LDC2005T19 Fisher English Training Transcripts Part 2: transcripts for LDC2005S13 conversations with 12% transcribed at LDC and the rest by BBN and WordWave.

This data set consisted of approximately 2,000 hours of transcribed telephone speech (8 KHz sampling rate). Subjects were primarily native speakers of American English recruited from all parts of the United States to ensure dialectal diversity. A small number of English speakers from outside the U.S., as well as non-native speakers of English, were also recruited for the corpus. Gender was balanced (53% female, 47% male). Each subject was assigned a topic to speak about that was randomly selected from a list that changed daily.

ASpIRE required that participants use only these data and algorithmic transformations of these data for training their systems. Participants were only allowed to use the transcripts provided for the challenge as described above to prepare language models and pronunciation lexicons.

2.2. Development Sets

The development data sets consisted of a 15 hour subset of the 1,425 multi-microphone sessions in the *Mixer 6 corpus* (LDC2013S03) [18] collected by LDC in 2009-2010. Native American English speakers were recorded while making 10 minute long telephone calls on a daily topic announced at the start of the call on 15 microphones installed similarly in two different office rooms at LDC. Care was taken to ensure that microphones were placed similarly in terms of distance, mounting, and orientation in both rooms, and that microphone levels were checked and calibrated. Microphone recordings were captured at 24 kHz, 16-bit. The development data were divided into a 5-hour development set and a 10-hour development-test set (dev-test). The files in each set were transcribed by Appen Butler Hill using the LDC transcription guidelines in [19].

The development sets were chosen to provide a good representation of microphone recordings in real rooms with transcription conventions matching those of the evaluation set. However, the recording environment of the evaluation set differed substantially from the development data in that there were a greater number of rooms, different microphones, and different placements of speakers with respect to the microphones.

Development transcripts were not allowed for either acoustic model or language model training or for supervised adaptation. Challenge participants were also prohibited from listening to or transcribing the development-test speech data.

2.2.1. Single Microphone Condition

ASpIRE_single_dev contained five hours of recorded speech sessions with transcripts. These data were provided for optimization, training selection, and tuning purposes. For each session selected, one microphone from a set of 12 microphones was randomly chosen. Although 14 microphones were available; microphones 12 and 14 were consistently poor in quality and so were not used in the single microphone condition.

ASpIRE_single_dev_test contained ten hours of recorded speech sessions. Transcripts were withheld but participants could evaluate their progress using the online scoring tool and monitor their progress relative to other scores posted on the leaderboard. For each session selected, one microphone from a set of 12 microphones was randomly chosen.

2.2.2. Multiple Microphone Condition

ASpIRE_multi_dev contained audio and transcripts for the same recording sessions as ASpIRE_single_dev. However, these recordings were made from six different microphones selected from among the set of all 14 possible microphones placed around the room, with different sessions using different sets of microphones.

ASpIRE_multi_dev_test contained audio and transcripts for the same recording sessions as ASpIRE_single_dev_test. However, these recordings were made from six different microphones selected from among 14 possible microphones placed around the room, with different sessions using different sets of microphones. Transcripts were withheld but participants could evaluate their progress using the online scoring tool.

2.3. Evaluation Sets

The evaluation set contained 120 5-minute sessions from the *Mixer 8 Pilot Corpus*. Unlike *Mixer 6*, which captured the variability of two different room types, the *Mixer 8 Pilot Corpus* was collected to test whether increasing the number and type of rooms and varying microphone placement would have a greater impact on error than simply changing microphones. By contrast to *Mixer 6*, where the need to instrument rooms up front for the entire period of the collection limited the variability of the resulting corpus, for the *Mixer 8 Pilot Corpus*, LDC used portable collection platforms to enable the collection of speech in a variety of rooms with different room dynamics.

The *Mixer 8 Pilot Corpus* was collected in seven rooms of different shapes and sizes using eight different distant microphones, positioned differently within each room, with subject speakers positioned in one of 2-3 locations in each room (3 for larger rooms). The room properties are summarized in Table 1.

Room	Description	Volume (ft^3)	# Pos.
117	Recording Room	1013	2
477	Small Office	1278	2
481	Conference Room	1759	2
126	Recording Room	1776	3
478	Conference Room	3496	3
460	Seminar Room	3547	3
470	Conference Room	13205	3

Table 1: Properties of rooms in the Mixer 8 Pilot Corpus, including room number, brief description, approximate volume, and number of different subject speaker positions tested.

For the evaluation corpus, 39 native speakers of North American English (22 males and 17 females) were recorded. Each subject participated in a single day of recording, with 5-6 different conversations taking place across several different pre-determined positions in one small room (117, 477, or 481) and one larger room (126, 478, 460, or 470). The acoustic environments in this corpus were constructed within seven rooms using eight different microphones, varying microphone height and orientation across rooms, and varying distances between microphones and subject speakers.

An example of a room layout used in the collection is shown in Figure 2. Eight microphones (denoted Microphones 1-8 in Table 2) and one simultaneous telephone-recording system were used to capture one side of a telephone call between each participant and an interviewer. Microphones 1, 2, 6, and 7 were omnidirectional microphones, while all others had a directional response. Microphone 3, 6, and 7 used wireless channels; all others were wired. Microphone positions and orientations were fixed for each room, but across sessions subjects were placed at one of 2-3 different positions in each room to further increase variability. Interviewers elicited freeform conversations from subjects on a set of predetermined topics. Microphone recordings were captured at 48 kHz, 24-bit.

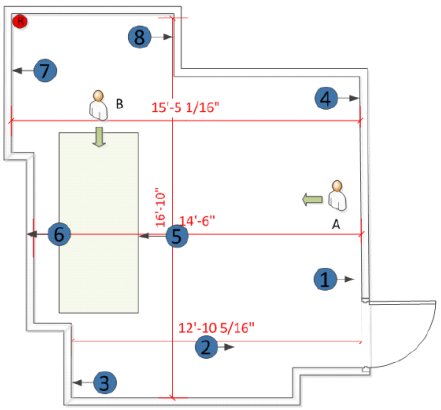

Figure 2. Speaker locations and microphone positions and orientations in room 470.

Microphone	Model	Features
1	Earthworks M23	Flat frequency response, omnidirectional, measurement applications
2	DPA 4090	High sensitivity, flat frequency response, omnidirectional, condenser, high quality studio applications
3	Samson SAC02	High sensitivity, directional, pencil microphone, low-cost home studio applications
4	RODE NT6	Directional, miniature microphone, various applications
5	Shure MX185	Diaphragm condenser microphone, directional, used as a lavalier
6	Sony ECMAW3	Blue tooth microphone, omnidirectional, miniature electret condenser microphone element, home video applications
7	Canon WM-V1	Blue tooth microphone, omnidirectional, prosumer camcorder/outdoor applications
8	Audio Technica AT8035	Shotgun microphone, directional, high off-axis rejection, outdoor recording applications

Table 2: Microphones used in the Mixer 8 Pilot Corpus.

All evaluation files were transcribed by Appen Butler Hill using the LDC transcription guidelines in [19]. As with the development data, two evaluations sets were created, one for each of the microphone conditions:

ASpIRE_single_eval contained ten hours of recorded speech sessions from the *Mixer 8 Pilot Corpus* that were selected to evaluate final submissions to the challenge under the Single Microphone Condition. For each session selected, one microphone from the set of eight different microphones placed around the room was randomly chosen.

ASpIRE_multi_eval contained ten hours of recorded speech sessions from the *Mixer 8 Pilot Corpus* that were selected to evaluate final submissions to the challenge under the Single Microphone Condition. Six of the eight microphones for a session were provided, with different sessions using different sets of microphones.

Challenge participants were prohibited from listening to or transcribing the evaluation data.

3. STT SCORING

Challenge systems were required to produce a verbatim, case-insensitive transcript of all uttered lexical items within the audio input according to Standard Normal Orthographic Representation (SNOR) rules. Transcripts were required to be whitespace-separated case-insensitive lexical tokens. Non-alphabetic characters were not to be transcribed, except for apostrophes for contractions and possessives and hyphens for hyphenated words and fragments. Spelled letters were to be represented as a letter followed by a period (e.g., "a. b. c."). The format for the system output was a Conversation Time Marked (CTM) file of the lexical tokens and their begin- and end-time for all recordings. For scoring purposes, three types of tokens were processed: (1) Scored tokens that must be recognized; (2) Optionally deletable tokens that may be omitted by the STT system without penalty; and (3) Non-scored tokens. Scored tokens included all words transcribed as specified in the LDC transcription guidelines [19]. An example of an optionally deletable token would be a word fragment, which appears in reference transcripts with a leading or trailing hyphen (e.g., /-tter/ or /th-/). System tokens matching the beginning or ending of the fragment's text were scored as correct (e.g., /latter/ would be scored as a match to /-tter/ and /theory/ to /th-/). Non-scored tokens were expected to be removed from the CTM output file before submission for scoring. Examples of non-scored tokens would be unintelligible speech or sounds such as coughing, laughing or sneezing.

A global map file (GLM) was made available to participants on the challenge website in order to transform both the reference and system output token strings via a set of rules to normalize spelling variants prior to scoring. For example, the GLM rules expanded contractions in the system output to all possible expanded forms, thereby generating several alternative token strings in the system output. These rules may also split a token string into two or more strings (e.g., compound words, hyphenated words).

Once the GLM was applied, the scorable lexical token sequences from the system CTM and reference STM were aligned using a global minimization of a Levenshtein distance function which weighted the cost of correct words, insertions, deletions and substitutions as 0, 3, 3 and 4 respectively. Given this alignment, word error rate (WER) of the STT output of a system was calculated as follows:

$$WER = \frac{(N_{Del} + N_{Ins} + N_{Subst.})}{N_{Ref}}$$

where

N_{Del} = the number of unmapped reference tokens (i.e., deletion errors),

N_{Ins} = the number of unmapped STT output tokens (i.e., insertion errors),

N_{Subst} = the number of mapped STT output tokens with non-matching reference spelling (i.e., substitution errors), and

N_{Ref} = the maximum number of reference tokens.

The NIST Scoring Toolkit (SCTK) was used to evaluate the performance of system-generated STT output in CTM format. The following command was used to score a CTM-formatted file with hubscore (hubscr) using segment time marked (STM) formatted reference file:

% hubscr.pl -V -g glm_filename.glm –l english –h hub4 -r ref_filename.stm hyp_filename1.ctm

4. RESULTS AND DISCUSSION

Five prize-eligible primary systems were submitted to the single microphone condition challenge. In addition to submitting their primary system (the system that was, in their judgment, their best), teams were allowed to submit other systems for independent scoring. These other systems are identified as contrast systems in Table 3, which reports the WER scores on dev-test and evaluation data for all the systems that were submitted with evaluation outputs[1], as required by the challenge rules. Other teams submitted system outputs for development and development-test scoring, but did not complete the challenge by submitting evaluation outputs, so their scores are not reported in Table 3. Only one prize-eligible submission was received for the multiple microphone condition.

Prize winning systems [20] appear in green in Table 3. In the single microphone condition challenge, three prizes were awarded, one to each of the top three scoring systems, which all scored within 1% of each other. The prize in the multiple microphone condition was awarded to the only system submitted to that challenge. It is significant to note that this multiple microphone system was by far the most accurate system submitted to the ASpIRE Challenge in terms of WER. This team's developers were able to derive benefit from six microphones compared to a single microphone system.

WER scores on the dev-test data set were consistently better than scores from the same system on the evaluation data set. Two factors may explain this observation. First, during the development period, the challenge participants had feedback on the performance of their systems on the dev-test data. The dev-test data had ground truth against which teams could measure their performance continually. Moreover, WER scores on the dev-test set were accessible via the InnoCentive scoring server. By contrast, the systems were only scored once on evaluation data at the end of the challenge. Second, there was a greater diversity of recording conditions in the evaluation set than in the dev-test set. This increased diversity also explains why the error rates were higher on the evaluation data set compared to the dev-test data set. In general, systems with the best scores on the dev-test set also tended to perform the best on the evaluation set; however, this was not always the case (e.g., system 15 and 16 performed as well as or better than systems 4, 6, 8, and 11 on dev-test but not on evaluation data). Systems that were consistently high scoring across test sets were clearly more robust to the new recording conditions in the *Mixer 8 Pilot corpus*.

Single Microphone				
System ID	**Team**	**Primary?**	**Dev-Test WER**	**Evaluation WER**
13	A	Primary	27.1	44.3
14	B	Primary	27.5	44.3
4	C	Primary	29.9	44.8
11	D	Primary	39.8	52.7
15	E	Primary	27.6	53.4
8	A	Contrast	29.0	43.9
9	A	Contrast	27.4	44.0
6	D	Contrast	40.0	52.8
16	E	Contrast	27.9	54.1
7	D	Contrast	40.0	54.4
12	D	Contrast	39.9	54.7
17	D	Contrast	39.4	50.7
Multiple Microphone				
System ID	**Team**	**Primary?**	**Dev-Test WER**	**Evaluation WER**
18	C	Primary	28.2	38.5

Table 3. Results from the ASpIRE Challenge rank ordered by the WER score on the evaluation data set.

4.1. Single Microphone System Designs

Team A [21]: The single microphone systems submitted by Team A used a Rover combination of systems [22]. System 13, the primary system, was a 26-way combination; contrast system 8 was a 2-way system combination; and contrast system 9 was a 7-way combination of Gaussian Mixture Models (GMMs) and Deep Neural Networks (DNNs). To minimize the mismatch between the training and test conditions, these systems were developed using a combination of neural network-based speech enhancement together with multi-condition training. The developers added additional copies of training data convolved with noise and room impulse responses in an attempt to better model the evaluation recording conditions. The approach used the parallel clean and noisy, reverberant data to learn a transformation that mapped the more challenging speech to clean speech, while also using the reverberant and noisy portion to train acoustic models to address challenging audio conditions. Although the 26-way combination had the lowest WER on the dev-test, the 2-way system, while simpler, performed slightly better on the evaluation set.

Team B [23]: System 14, the single microphone system submitted by Team B, was a multi-splice DNN system [24]. It used 6 hidden layers with an effective input context of t-16 to t+12, that was designed to tackle longer-term interactions between the direct speech signal and reverberation by processing longer, asymmetric durations of speech. The system was trained on all of the Fisher data and on multiple additional versions of the data created by distorting the audio with various room impulse responses and noise. The system used 40-d Mel Frequency Cepstral Coefficients (MFCCs) together with i-vectors [25]. The i-vectors characterized both the speaker and environment of the recording as inputs to the multi-splice DNN. In addition, the system benefitted from learned pronunciation probabilities and word position-dependent silence probabilities. The resulting system was able to deliver among the lowest WERs on both the dev-test and evaluation sets.

Team C [26]: System 4 was among the top performing systems on both the dev-test and evaluation sets. This system was a DNN

[1] Some teams did not publish on their submission and are anonymous in this paper due to the ASpIRE challenge agreement.

trained on 500 hours of clean training data augmented with data created by adding noise and reverberation. The system used restricted Boltzmann machine (RBM*)* pre-training, cross-entropy training, and sequential minimum Bayes risk discriminative training. A tri-phone GMM was used to generate alignments and speaker-adapted features. Speech activity detection made use of the harmonic to sub-harmonic ratio [27] as a feature for voiced speech detection; this approach was expected to be robust in mismatched and severe noise conditions. The speech files were down sampled to 8 kHz, a telephone filter was applied, followed by a dereverberation algorithm that used kurtosis as a measure of the reverberation [28]. Evaluation data was used for semi-supervised retraining [29].

Team D [30]: System 11 team D's primary system, was a 5-way Rover combination [22] of three convolutional deep neural nets (CDNNs) [31, 32] and 2 DNNs. While this system was trained on reverberant transformations of the Fisher data and with a wide range of speech features (e.g., Damped Oscillator Coefficients, Gammatone Filterbank coefficients, and i-vectors) and speech enhancement techniques (e.g., Non-negative Matrix Factorization [33]), its speech activity detection system was trained on data (from the DARPA RATS program) that was highly mismatched with the recording conditions of both Mixer 6 and Mixer 8. The primary system used a Recurrent Neural Network (RNN) language model (LM) for rescoring in addition to a more standard 4-gram LM. The contrast systems were similar to system 11 except that system 6 did not use the RNN LM; system 7 replaced one DNN with a GMM and eliminated the RNN LM; system 12 replaced one DNN with a GMM and used slightly different neural network configurations; and system 17 replaced Gammatone Filterbank coefficients with Normalized Modulation Coefficients and eliminated the RNN LM.

Team E: System 15, Team E's primary submission, was a Rover combination [22] of three subsystems made up of DNNs and/or CNNs with various types of input features along with a single GMM. The first subsystem was a combination of two DNN systems trained on Feature space Maximum Likelihood Linear Regression (FMLLR) transformed Perceptual Linear Prediction and Frequency Domain Linear Prediction (FDLP) features appended with i-vectors estimated on power-normalized cepstral coefficients (PNCC) and FDLP features. The second subsystem was a discriminatively trained HMM-GMM system that used bottleneck features from a DNN trained on FMLLR and i-vector features. The third subsystem combined two hierarchical DNNs, one using bottleneck and i-vector features and the other using bottleneck features from a DNN trained on FMLLR transformed features, i-vectors, and features using an inverse filtering approach to suppress reverberation. It also incorporated a maxout DNN using annealed dropout on log-Mel spectrograms with projections of the input layer constrained to local receptive fields (LRFs) of the input features. The subsystems of system 15 were trained on different 600 hour subsets of a clean version of the training set. A two-step de-noising algorithm was applied to the test sets after down-sampling to 8 kHz, and cepstral mean subtraction was applied during feature extraction. The contrast system 16 was similar to the primary submission except that it added a fourth subsystem consisting of three DNNs to the Rover combination. Both of Team E's systems scored among the best WERs on the dev-test, but did not perform equally well on the evaluation set. Their submissions, like those of group D, used speech activity detection trained on DARPA RATS data.

All of the top performing single microphone systems used multi-condition training by augmenting the clean training data with additional degraded data (noise and room impulses) and performed some form of enhancement and adaptation. In addition, the best-performing systems used speech activity detection appropriate for far field microphone data. In fact, in post-challenge analysis by Lincoln Laboratory, the systems with the best WERs were also more accurate at speech activity detection (see [34]).

4.2. Multiple Microphone System Design

Team C [26] was the only participant to submit to this condition. Fortunately, this group also submitted a top performing single microphone system that helped in the interpretation of the improved performance using multiple microphones. System 18 was comparable to system 4, except that the signal processing pipeline used a beam-forming algorithm [35], and semi-supervised training was based on six times the evaluation data used in the single microphone condition. This system benefited from access to a larger sampling of test conditions in the multiple microphone condition.

5. CONCLUSIONS

The ASpIRE challenge participants exploited a wide range of approaches to make impressive progress toward fieldable solutions to the very hard problem of building robust speech recognition systems when there is no matched training and development data. Neural networks, speech activity detection, multi-condition training, speech enhancement, and unsupervised adaptation all contributed to lower WER under both evaluation conditions. It is clear that lack of matched supervised adaptation data is no longer an insurmountable challenge for making progress on robust speech recognition. The individual lessons learned in the ASpIRE challenge are likely to lead to better systems in the future, and mixtures of methods from each of the teams may contribute to further progress.

The ASpIRE challenge also demonstrated that working continually on the same test data and making progress on that data may not guarantee robustness to data collected in new, but related, recording conditions. Reverberation was clearly important in both the development and evaluation sets; however, microphone variability was greater in Mixer 6 and room variability in Mixer 8. This suggests that new challenges that aim to measure system robustness need to creatively collect new test data with mismatch and then limit testing on these data until after systems are developed.

ACKNOWLEDGMENTS

The author would like to thank Chris Cieri, Fred Goodman, Stephanie Strassel, and Kevin Walker for their help with designing and collecting the Mixer 8 data. Thanks go to Jon Fiscus for his advice on this challenge. Thanks also go to Michael Brandstein, Nicolas Malyska, Jennifer Melot, Jessica Ray, and Wade Shen for advice and help in overseeing the transcription of the data, assessing factors that were likely to impact system performance, building baseline systems, selecting development and evaluation sessions, and scoring the evaluation and validation submissions.

6. REFERENCES

[1] NIST, "Automatic Speech Recognition Evaluations at NIST," 2009. Available: http://itl.nist.gov/iad/mig/publications/ASRhistory/index.html [Accessed: September 19, 2014].

[2] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making Machines Understand Us in Reverberant Rooms: Robustness against Reverberation for Automatic Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[3] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus," in *Proceedings of ICASSP*, 2003.

[4] S. H. K. Parthasarathi, S. Y. Chang, J. Cohen, N. Morgan, and S. Wegmann, "The Blame Game in Meeting Room ASR: An Analysis of Feature Versus Model Errors in Noisy and Mismatched Conditions," in *Proceedings of ICASSP*, 2013.

[5] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[6] T. Hain, V. Wan, L. Burget, M. Karafiat, J. Dines, J. Vepa, G. Garau, and M. Lincoln, "The AMI System for the Transcription of Speech in Meetings," in *Proceedings of ICASSP,* 2007.

[7] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The Multi-Channel Wall Street Journal Audio Visual Corpus (MC-WSJ-AV): Specification and Initial Experiments," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005.

[8] M. Lincoln, E. Zwyssig, and I. McCowan, "Multi-Channel WSJ Audio," LDC2014S03, Philadelphia: Linguistic Data Consortium, 2014.

[9] 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.

[10] 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.

[11] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The Second 'CHiME' Speech Separation and Recognition Challenge: Datasets, Tasks and Baselines," in *Proceedings of ICASSP* , 2013.

[12] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The Second 'CHiME' Speech Separation and Recognition Challenge: An Overview of Challenge Systems and Outcomes," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2013.

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

[14] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The Third 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2015.

[15] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The Reverb Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[16] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition," in *Proceedings of ICASSP,* 1995.

[17] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus, A Resource for the Next Generations of Speech-to-Text," in *Proceedings 4th International Conference on Language Resources and Evaluation*, 2004.

[18] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "Mixer 6," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 2010.

[19] Linguistic Data Consortium, "Linguistic Data Consortium Transcription Guidelines (NQTR)," 2006 [online]. Available: https://catalog.ldc.upenn.edu/docs/LDC2010S01/trans_guide_nqrt_span.doc, [Accessed: September 19, 2014].

[20] IARPA, "IARPA Announces Winners of its ASpIRE Challenge," 2015 [online]. Available: http://www.dni.gov/index.php/newsroom/press-releases/210-press-releases-2015/1252-iarpa-announces-winners-of-its-aspire-challenge [Accessed: September 10, 2015].

[21] R. Hsiao, J. Ma, W. Hartmann, M. Karafiat, F. Grezl, L. Burget, I. Szoke, J. H. Cernocky, S. Watanabe, Z. Chen, S. H. Mallidi, H. Hermansky, S. Tsakalidis, and R. Schwartz, "Robust Speech Recognition in Unknown Reverberant and Noisy Conditions," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2015.

[22] J. G. Fiscus, "A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 1997.

[23] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "JHU ASpIRE system: Robust LVCSR with TDNNs, i-vector Adaptation, and RNN-LMs," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2015.

[24] V. Peddinti, D. Povey, and S. Khudanpur, "A Time Delay Neural Network Architecture for Efficient Modeling of Long Temporal Contexts," in *Proceedings of Interspeech*, 2015.

[25] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector Based Speaker Adaptation of Deep Neural Networks for French Broadcast Audio Transcription," in *Proceedings of ICASSP*, 2014.

[26] J. Dennis and H. D. Tran, "Single and Multi-channel Approaches for Distant Speech Recognition under Noisy Reverberant Conditions: I^{2}R's System Description for the ASpIRE Challenge," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2015.

[27] X. Sun, "Pitch Determination and Voice Quality Analysis Using Subharmonic-to-Harmonic Ratio," in *Proceedings of ICASSP*, 2002.

[28] B. W. Gillespie, H. S. Malvar, and D. Florêncio, “Speech Dereverberation via Maximum-kurtosis Subband Adaptive Filtering,” in *Proceedings of ICASSP*, 2001.

[29] H. Liao, “Speaker Adaptation of Context Dependent Deep Neural Networks,” in *Proceedings of ICASSP*, 2013.

[30] V. Mitra, J. Van Hout, W. Wang, M. Graciarena, M. McLaren, H. Franco, and D. Vergyri, “Improving Robustness Against Reverberation For Automatic Speech Recognition,” in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2015.

[31] V. Mitra, W. Wang, and H/ Franco, “Deep Convolutional Nets and Robust Features for Reverberation-robust Speech Recognition,” in *Proceedings of SLT*, 2014.

[32] V. Mitra, W. Wang, H. Franco, Y. Lei, C. Bartels, and M. Graciarena, “Evaluating Robust Features on Deep Neural Networks for Speech Recognition in Noisy and Channel Mismatched Conditions,” in *Proceedings of Interspeech*, 2014.

[33] K. Kumar, R. Singh, B. Raj, and R. Stern, "Gammatone Sub-band Magnitude-domain Dereverberation for ASR," in *Proceedings of ICASSP*, 2011.

[34] J. Melot, N. Malyska, J. Ray, and W. Shen, “Analysis of Factors Affecting System Performance in the ASpIRE Challenge,” in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2015.

[35] X. Anguera, C. Wooters, and J. Hernando, “Acoustic Beamforming for Speaker Diarization of Meetings,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 2011-2023, 2007.