SINGLE AND MULTI-CHANNEL APPROACHES FOR DISTANT SPEECH RECOGNITION UNDER NOISY REVERBERANT CONDITIONS: I^{2}R'S SYSTEM DESCRIPTION FOR THE ASpIRE CHALLENGE

Jonathan Dennis, Tran Huy Dat

Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way, Singapore 138632

ABSTRACT

In this paper, we introduce the system developed at the Institute for Infocomm Research (I^{2}R) for the ASpIRE (Automatic Speech recognition In Reverberant Environments) challenge. The main components of the system are a front-end processing system consisting of a distributed beamforming algorithm, that performs adaptive weighting and channel elimination, a speech dereverberation approach using a maximum-kurtosis criteria, and a robust voice activity detection (VAD) module based on using the sub-harmonic ratio (SHR). The acoustic back-end consists of a multi-conditional Deep Neural Network (DNN) model that uses speaker adapted features combined with a decoding strategy that performs semi-supervised DNN model adaptation using weighted labels generated by the first-pass decoding output. On the single-microphone evaluation, our system achieved a word error rate (WER) of 44.8%. With the incorporation of beamforming on the multi-microphone evaluation, our system achieved an improvement in WER of over 6% to give the best evaluation result of 38.5%.

Index Terms— ASpIRE Challenge, mismatched conditions, reverberation, distant speech recognition, beamforming

1. INTRODUCTION

Recognising speech in far-field microphone recordings is a significant challenge in Automatic Speech Recognition (ASR) due to the high levels of both noise and reverberation present in the signal. The difficulty is partially caused by the mismatch between training and testing environments, where it is often impractical to transcribe any significant amount of data when deploying a system in the field. These problems are the basis of the ASpIRE challenge [1], which further introduces mismatch by containing the training data to be conversational telephone speech. While previous techniques may separately address individual aspects of the task, such as robust voice activity detection [2, 3], distant speech enhancement [4], and DNN acoustic model adaptation [5, 6], the ASpIRE challenge provides an opportunity to bring together these techniques with the goal of improving the WER in real-life mismatched environments.

The task in the ASpIRE challenge is to build an ASR system for English that is trained on conversational telephone speech, but is robust to a variety of unknown acoustic environments and recording scenarios, without having access to matched training and development data. The testing data was collected using multiple simultaneous microphones placed in a wide range of locations in seven different rooms of various different shapes, sizes, surface properties, and noise sources. The challenge therefore offered two evaluation conditions: (1) the Single Microphone condition, where a single microphone is selected randomly, and (2) Multi Microphone condition, where a subset of the microphones is available for use in multi-channel enhancement techniques.

In this paper, we propose a system for recognising speech in the presence of far-field speech and mismatched testing conditions such as those in the ASpIRE challenge. We combine robust front-end processing and speech enhancement methods with state-of-the-art techniques for recognition, such as the discriminative training of the acoustic model [7, 8] and Recurrent Neural Network (RNN) languages model rescoring [9]. To specifically overcome the problem of mismatch between training and testing, we employed three strategies: (1) a multi-conditional training dataset was created by artificially adding noise and reverberation recorded in an office environment for a variety of room sizes, (2) a robust voice-speech detection algorithm was developed, based on the sub-harmonic ratio (SHR) of the acoustic spectrum, that required no prior knowledge of the testing conditions, and (3) a semi-supervised DNN model adaptation was employed to reduce the mismatch between training and testing.

2. SYSTEM OVERVIEW

Figure 1 shows a schematic diagram of the proposed recognition system used by I^{2}R in the ASpIRE challenge. It consists of the following modules:

(Multi-Mic) **Distributed beamforming** using a correlation-weighted delay and sum approach with adaptive channel elimination (see Section 3.5).

Speech dereverberation based on maximising the kurtosis of the Linear Prediction (LP) residual (see Section 3.3).

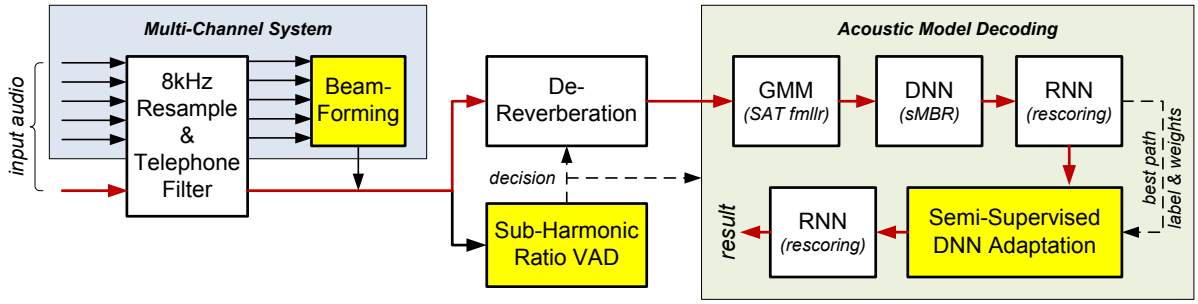

Fig. 1. Overview of the proposed distant speech recognition system for noisy reverberant conditions.

Robust Voice Activity Detection based on the harmonic to sub-harmonic ratio (SHR) feature providing robust detection of voiced speech in noise (see Section 3.4).

Multi-conditional Speech Recogniser using a hybrid DNN architecture, with sequential minimum bayes risk (sMBR) discriminative training (see Section 4).

Semi-supervised adaptation of the DNN to the environment by performing further iterations of cross-entropy training using semi-supervised weighted labels generated by the first pass decoding (see Section 4.4).

The following sections describe each component in detail.

3. FRONT-END SIGNAL PROCESSING

The following subsections describe the front-end processing steps to generate the audio and the speech/non-speech decision for ASR decoding. The following components are detailed: (1) 8kHz resampling and filtering, (2) multi-conditional data generation, (3) dereverberation, (4) voice activity detection, and (5) adaptive beamforming.

3.1. Down-Sampling and Telephone Filter

The training data supplied for this challenge was the Fisher English corpus of conversational telephone speech [10]. The sampling rate for this corpus is 8 kHz, compared to 16 kHz and 48 kHz for the development and evaluation data respectively. In addition to down-sampling, a narrowband telephone filter is applied to the waveform to further reduce the mismatch in recording equipment. The Fisher English corpus, recorded over the telephone channel, has a limited bandwidth of approximately 300-3400 Hz, hence we use a simple filter to normalise all recordings to this narrow bandwidth. Looking at the results for the single-development set in Table 1, in particular the first two rows in each section, it can be seen that there is a consistent improvement in WER of approximately 1% when using the telephone filter as opposed to using a plain 8 kHz down-sampling.

3.2. Multi-Conditional Training Data

To reduce the mismatch between training and testing, we generate multi-conditional training samples by artificiality adding noise and reverberation to the Fisher Corpus. Since the exact testing conditions are not known in advance, we apply a number of impulse responses and noise samples to provide a broad range of data for training. Both the noise and impulse responses are measured in an office environment in small, medium and large room sizes, with T60 times of approximately 0.25, 0.5 and 0.7 seconds respectively. Noise is distributed across a range of signal-to-noise ratios (SNRs), with a different SNR selected uniformly from the range [0-20] dB for each recorded conversation.

The effect of the multi-conditional training data can be seen in Table 1, in particular by comparing the upper and lower box for the GMM and DNN acoustic model, which contain results for with clean and multi training (MCT) conditions respectively. It can be seen that multi-conditional training gives a consistent boost in performance of around 5-8% for both the GMM and DNN acoustic model.

3.3. Speech Dereverberation

In practical distant speech recognition, such as the ASpIRE challenge, we do not have knowledge of either the clean signal, or the room impulse response (RIR), hence this scenario is called Blind Derevereberation. In our system we perform dereverberation using an algorithm that attempts to maximise the kurtosis of the LP residual [4, 11]. The idea is that kurtosis, which is a degree of peakedness of a distribution, can be seen as a measure of the reverberation as a degraded signal will have a lower kurtosis due to the time-spreading impact of the multi-path signal. Hence, the algorithm uses a gradient ascent method to maximise the LP residual kurtosis, and hence increase the peakedness of the signal to make it more like clean speech. In our system we use the VAD to only allow filter updates during regions of speech, since there can be significant regions of silence during one side of a two-party conversation.

The performance improvement of the dereverberation is

Acoustic Model	MCT	Tel	AFE	MaxK	Dev WER
GMM-SAT *(8k resample)*					61.5
		X			60.0
		X	X		56.8
		X		X	58.6
	X				53.7
	X	X			53.1
	X	X	X		54.4
	X	X		X	51.4
DNN-SAT-SMBR *(8k resample)*					50.8
		X			49.8
		X	X		48.0
		X		X	48.4
	X				42.3
	X	X			41.1
	X	X	X		43.9
	X	X		X	**38.8**

Table 1. Comparison of front-end processing on the Single-Microphone development set, where the DNN is trained on a 100hr subset of the Fisher corpus *(MCT = multi-conditional training, Tel = 8 kHz telephone filter, AFE = advanced front end denoising, MaxK = maximum kurtosis dereverberation)*.

shown in Table 1, where it is compared to the extended Advanced Front End (AFE) [12]. It can be seen that while AFE provides improvement only for the clean training models, the maximum kurtosis dereverberation gives between 1-2% reduction in WER for clean and multi-conditional models.

3.4. SHR-based Voice Activity Detection

Our system uses the harmonic to sub-harmonic ratio (SHR) [2, 3] as a feature for voiced speech detection. Unlike using a simple summation of the harmonic energy, the SHR is normalised by the sub-harmonic energy, and hence is more robust particularly in mismatched and severe noise conditions.

To compute the SHR within each short-time windowed frame, using a frame length of 32 ms, the amplitude spectrum $E(f)$ is first computed. For voiced segments of speech, $E(f)$ has strong peaks at the harmonics of the fundamental frequency $F0$. From this spectrum, the summation of harmonic amplitude (SHA) and summation of sub-harmonic amplitude (SSA) is computed for each frequency in the range $[F0_{min}, F0_{max}]$ as follows:

$$SHA(f) = \sum_{k=1}^{N_{harm}} \sum_{a=-\Delta}^{\Delta} E(k.f + a) \quad (1)$$

$$SSA(f) = \sum_{k=1}^{N_{harm}} \sum_{a=-\Delta}^{\Delta} E((k - \frac{1}{2}).f + a) \quad (2)$$

where only the first N_{harm} harmonics are taken into account in the summation, and a window of $\Delta = 1$ neighbouring bins are included in the summation to account for inharmonicity.

Finally, the harmonic to sub-harmonic ratio (SHR) is the ratio of the two, as follows:

$$SHR(f) = \frac{SHA(f)}{SSA(f)} \quad (3)$$

where the maximum value $\max_f (SHR(f))$ is taken as the value of the feature for each frame, $SHR[t]$.

An adaptive threshold is defined to provide automatic setting of the VAD for each conversation to achieve a consistent result in mismatched conditions. This takes overlapping 30 second segments, and computes an initial threshold as follows:

$$thres_{init}[t] = \mu(SHR[t]) \quad \text{for } t \pm \tfrac{NW}{2} \quad (4)$$

where μ denotes the mean, and NW is the number of frames within the segment window. An improved estimate is then computed based on frames within the segment, denoted t', that have lower SHR than the initial threshold:

$$thres[t] = \mu(SHR[t']) + 2 * \sigma(SHR[t']) \quad (5)$$

where σ is the standard deviation, such that $2*\sigma$ captures 95% of the distribution. The output decision is further smoothed to join together segments with a gap of less than 1.5 seconds, and to apply a hangover of length 1 second to ensure that unvoiced speech at the start and end of the segments are not missed.

3.5. Correlation-weighted Beamforming

Our multi-microphone system utilises a correlation-weighted multi-microphone beamforming technique, based on time delay of arrival (TDOA) using generalised cross correlation (GCC) with phase transform (PHAT) [13]. Previous studies have demonstrated that the popular PHAT weighted cross-correlation gives excellent performance even in noisy and reverberant environments [14, 15], hence its application in this challenge. The approach also incorporates adaptive channel elimination and rejection strategies to improve the quality of the output signal [16].

The first step uses the VAD select the longest continuous speech portion for global alignment and reference channel selection. The reference channel is selected using a metric based on the cross-correlation of the speech signal between each channel, i, and all the other channels, $j = 1, \ldots, M \;\; j \neq i$. We find the peak in the cross-correlation for each channel pair within a one second window, and sum the coefficients as follows:

$$\overline{xcorr_i} = \sum_{j=1, j\neq i}^{M} xcorr[i, j] \quad (6)$$

where M is the number of channels, and $xcorr[i, j]$ is the standard cross- correlation coefficients (no PHAT weighting)

between channels i and j over the speech region selected by the VAD. The channel i with the highest average cross-correlation is chosen as the reference channel.

Given two signals $x_i(n)$ and $x_{ref}(n)$, the GCC-PHAT is used to estimate the TDOA every 250 ms over an analysis window of 500ms, as follows:

$$\hat{R}^i_{PHAT}(d) = \mathcal{F}^{-1}\left(\frac{X_i(f)X_{ref}(f)^\star}{\mid X_i(f)X_{ref}(f)^\star \mid}\right) \quad (7)$$

where $X_i(f)$ and $X_{ref}(f)$ are the Fourier transforms of the two signals and $\mathcal{F}$ is the inverse Fourier transform. The correlation coefficients between channel i and the reference channel are written as $\hat{R}^i_{PHAT}(d)$, where d is the estimated delay in terms of the number of samples in the signal. The TDOA can then be estimated as follows:

$$TDOA^i_1 = \arg\max_d \left(\hat{R}^i_{PHAT}(d)\right) \quad (8)$$

where the label 1 is applied to indicate that this is the 1-best delay. In practice we compute the $N = 4$ best maxima in the cross-correlation coefficients, and use a simple continuity filter to select the maxima most likely to belong to the same speaker as the previous analysis window.

After TDOA alignment, the final step is an adaptive weighting and channel elimination strategy based on the cross-correlation $xcorr[i, j]$ of the aligned channels. Starting from an even weighting in the first analysis window, $W_m[c = 1] = \frac{1}{M}$, the weights are updated continuously as follows:

$$W_m[c] = (1-\alpha)W_m[c-1] + \alpha\frac{\overline{xcorr}_m[c]}{\sum_{m=1}^{M}\overline{xcorr}_m[c]} \quad (9)$$

where $\alpha = 0.05$ is the adaptation ratio and $\overline{xcorr}_m[c]$ is the average cross-correlation from the TDOA aligned frame c using the approach in (6).

The adaptive channel elimination strategy is applied to remove noise and distortions affecting a small number of microphones. The value of $\overline{xcorr}_m[c]$ is used as a measure of the channel quality, and the following threshold is applied to remove poor quality frames:

$$\overline{xcorr}_m[c] < \frac{1}{M}\sum_{m=1}^{M}(\overline{xcorr}_m[c]) - \beta \quad (10)$$

where segments matching this criteria have weights $W_m[c] = 0$ with the subsequent weights normalised to sum to 1. The parameter β controls the strength of the channel rejection, and was empirically set to 0.04 in our experiments.

A further channel rejection strategy is also used, whereby if any channel repeatedly fails to meet the elimination criteria, it will be removed completely and the beamforming reinitialised without this channel. The criteria was set as:

$$\sum_{c=1}^{C} failure > \frac{C}{\gamma} \quad (11)$$

where C is the total number of frames in the utterance and $\gamma = 4$ was the empirical parameter set to control the trade-off in channel rejection.

4. ACOUSTIC MODELLING AND DECODING

The following subsections describe the acoustic modelling and decoding strategy used in the proposed ASR system, as shown in Figure 1. The following components are detailed: (1) feature extraction and GMM-HMM, (2) DNN acoustic modelling, (3) language modelling and rescoring, and (4) semi-supervised DNN adaptation.

4.1. Feature Extraction and Auxiliary GMM-HMM

The acoustic models (both GMM-HMM and DNN) are trained on 13-dimensional MFCCs, without energy, which are mean normalised over the speech segments extracted from each conversation for the speaker. Later, these features are spliced by ± 3 frames adjacent to the central frame and projected down to 40 dimensions using linear discriminant analysis (LDA).

Prior to DNN training, an auxiliary GMM-HMM is first trained to provide speaker adaptive transforms (SAT) and the initial alignments for training the subsequent DNN system by forced alignment, which inherits the same tied-state structure. To train the GMM-HMM, a bootstrap training approach is used, whereby the system is first trained on the original clean data, and the subsequent alignments are used to initialise the multi-conditional training. Hence, basic monophone, triphone and LDA GMM-HMM systems were first trained on 10, 30 and 100 hours respectively of randomly selected utterances. For multi-conditional training, the clean LDA alignments are first used to train a LDA GMM-HMM on 100 hours of utterances, followed by SAT training on the full dataset to give a final SAT GMM-HMM system with 7557 tied triphone states and 300k Gaussians.

4.2. DNN Acoustic Modelling

The multi-conditional DNN acoustic model is trained on top of SAT features that are spliced ± 5 frames and rescaled to have zero mean and unit variance. The DNN has 5 hidden layers, where each hidden layer has 2k sigmoid neurons, and a 7557 dimensional softmax output layer. The hidden layer weights are initialised using layer-wise restricted Boltzmann machine (RBM) pretraining, using 100 hours of randomly selected utterances from the Fisher corpus. After pretraining, fine-tuning is performed to minimize the per-frame cross-entropy between the labels and network output. The first stage of fine-tuning was performed using the same 100 hour subset as for pretraining with a learning rate of 0.002 and halving beginning when the network improvement slows. This then generated alignments for a larger 500 hour subset, to perform

Audio Set	RNN_1	DNN-Adapt	RNN_2	WER
Single-Development				35.2
	X			34.2
	X	X		33.2
	X	X	X	32.4
Single-Development-Test				32.0
	X			31.4
	X	X		29.9
	X	X	X	29.3
Single-Evaluation				47.6
	X			46.8
	X	X		45.2
	X	X	X	**44.8**
Multi-Development				25.0
	X			24.1
	X	X		22.8
	X	X	X	22.3
Multi-Development-Test				31.5
	X			31.1
	X	X		29.7
	X	X	X	29.1
Multi-Evaluation				41.7
	X			41.1
	X	X		39.0
	X	X	X	**38.5**

Table 2. Step-by-step results on the ASpIRE challenge for each step in the decoding system, across each of the different microphone test conditions. The best-performing multi-conditional DNN-SAT-SMBR acoustic model is compared across all experiments.

a second stage of fine-tuning. Finally, the 500 hour DNN is re-trained by sequence-discriminative training to optimise the state minimum Bayes risk (sMBR) objective. Two iterations are performed with a fixed learning rate of 1e-5. The Kaldi toolkit is used for all experiments [17]

4.3. Language Model and Rescoring

Two language models are trained, both using the full amount of data in the Fisher English corpus. The first is a 4-gram model, trained using the "Kaldi LM" package [17], which has a perplexity of approximately 69. The second is a RNN model trained using "RNNLM-0.3e" [9], with 20k words, 300 hidden units, 300 classes, and 2000m direct connections. The RNN language model has a perplexity of approximately 60, and is used to rescore the output decoding lattice, with interpolation weight 0.3 against the 4-gram LM. Note that the CMU pronouncing dictionary [18] was used, limited to the words that appear in the Fisher English corpus.

4.4. Semi-supervised DNN adaptation

While the multi-conditional training helps to reduce mismatch between training and set, a semi-supervised DNN adaptation technique is utilised to further reduce the mismatch between training and testing conditions [5, 6]. Additional iterations of fine-training of the DNN requires a frame-level label, and potentially also a confidence measure, and these are generated based on the initial output of the system after RNN rescoring, as shown in Figure 1.

The frame-level confidence c_{frame_i} is extracted from the lattice posteriors $\gamma(i, s)$, which express the probability of being in state s at time i. The decoding output gives us the best path state sequence, $s_{i,1best}$, and the confidence values are the posteriors under this sequence, as follows [6]:

$$c_{frame_i} = \gamma\left(i, s_{i,1best}\right) \tag{12}$$

The best path state sequence and confidence measures are then used as the target labels and weightings respectively for additional iterations of DNN fine-tuning. In our experiments, all weights in the network are updated, as our experiments suggested this performed better than adapting only the first layer of the DNN. The learning rate is 0.0008, with halving performed each iteration until no improvement is observed.

5. CHALLENGE EVALUATION

The word error rate (WER) results of the proposed recognition system on the ASpIRE challenge are shown in Table 2. There are three separate testing sets: (1) Development, (2) Development-Test, and (3) Evaluation. Each of these are split into two different experiments: (1) Single, and (2) Multi, depending on how many microphones were available for each conversation during decoding.

5.1. Results and Discussion

The results in Table 2 show that the best results on the Single-Development and Single-Development-Test sets, both subsets of the Mixer 6 corpus (LDC2013S03), are reasonable consistent at 32.4% and 29.3% respectively. Comparing this against the WER result of 39.3% on our development tuning set, which was held out from the simulated multi-conditional data,suggests that the simulated conditions for training, including 0dB noise and a long reverberation time, were harder than development testing conditions. This is confirmed by an average segmental SNR estimate of 12.4dB and 12.1dB for the two sets respectively, using the approach in [19].

While the Development sets aimed to provide a good representation of microphone recordings in real rooms, the Evaluation set differed substantially in that there were a greater number of rooms, different microphones, and different placements of speakers with respect to the microphones.

Training/Processing/Decoding Step	WER Improvement
Clean → Multi-Conditional	5-8%
DNN Training: 100 → 500 hours	3%
Telephone Filtering	1%
Beamforming	6-10%
Dereverberation	1-2%
RNN Language Model Rescoring	0.5-1%
DNN Semi-supervised Adaptation	1-2%

Table 3. Comparison of the approximate WER improvements given by each component of the system.

Our Single-Evaluation result of 44.8% reflects the challenging nature of these conditions, and this is supported by the segmental SNR estimate of 4.8dB for this set.

In the Multi-Development and Multi-Evaluation sets we see a strong improvement of approximately 10% and 6% absolute WER respectively compared to the respective Single sets. This demonstrates the benefits of using multiple microphones when available, and the robustness of the correlation-weighted beamforming approach, which performs well even in very low SNR conditions. However, almost no improvement is observed on the Multi-Development-Test set. The most likely cause of this anomalous result is the audio-bleed present from the other side of the conversation that is not annotated, and was noted by the challenge organisers as a problem. This may be enhanced by the beamforming and be more likely to be picked up by the VAD leading to a higher number of insertions. The Development set does not display this problem, as the ground-truth VAD was provided and used for decoding, hence eliminating the problem on this set.

5.2. Analysis of Word Error Rate Improvements

A summary of the contribution of each processing step to the final WER result is shown in Table 3. It can be seen that the biggest contributions come from the multi-conditional training, as shown in the first segment of the table, where using a DNN with 500 hours of multi-conditional data gives around a 10% improvement.

The multi-microphone beamforming algorithm also gives a significant improvement in performance of between 6-10% on the comparable sets, which highlights the robustness of the approach with distributed microphones under challenging conditions.

Finally, both the signal processing and decoding strategies each give around 1-2% improvement, with the biggest contribution coming from the maximum-kurtosis dereverberation and semi-supervised DNN adaptation, which both help to reduce the mismatch between training and testing conditions.

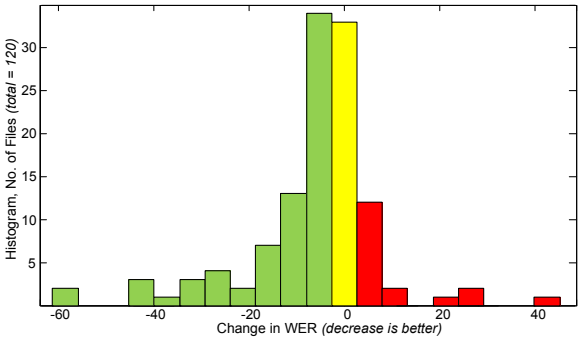

Fig. 2. Histogram of the change in WER for files in the Multi-Evaluation compared to those from the Single-Evaluation.

5.3. Analysis of Single vs. Multi

Focusing on the most challenging Evaluation testing condition, we now analyse the effect of beamforming by comparing the improvement in scores between each individual file in the Single and Multi Evaluation set.

Figure 2 shows the change in WER between Single and Multi, where a negative change means an improvement in WER. It can be seen that 89 of the 120 conversations showed an improvement in WER of up to 61.1%, with a further 14 files with a WER increase less than 3%. Of the remaining 17 files which had a significantly higher WER for the Multi as compared to the Single, the worst was an increase of 45.3%. An analysis of the contributing factors to this erroneous result, reveals an increase in insertions from 7.1% to 58.3%, which suggests that the audio-bleed phenomenon may not be completely eliminated in the Evaluation condition.

In the case of the other files, it is expected that having multiple audio streams available may not always give an improvement, particularly in the case where the randomly selected Single file is already has the best SNR and low reverberation. This should be the case for approximately for 15 files, i.e. 1/8th of the 120 files.

6. CONCLUSION

In this paper we have presented a complete description of the system developed at I^2R for the ASpIRE challenge. The most important components of the system are a front-end processing system consisting of speech dereverberation and distributed beamforming, which contribute up to 10% improvement in WER, combined with a robust VAD system based on the harmonic-to-subharmonic ratio. Multi-conditional training of the acoustic back-end was another significant factor in improving the WER results, as this considerably reduces the mismatch between training and testing conditions. Finally, our proposed decoding strategy performed both semi-supervised DNN model adaptation and RNN language model rescoring, contributes a further 2-3% WER reduction.

7. REFERENCES

[1] M Harper, "The Automatic Speech recognition In Reverberant Environments (ASpIRE) Challenge," in *Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 1–6.

[2] Xuejing Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 1, pp. I–333.

[3] Thomas Drugman and Abeer Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics.," in *Interspeech*, 2011, pp. 1973–1976.

[4] B.W. Gillespie, H.S. Malvar, and D.A.F. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 2001, vol. 6, pp. 3701–3704, IEEE.

[5] Hank Liao, "Speaker adaptation of context dependent deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7947–7951.

[6] Karel Vesely, Mirko Hannemann, and Lukas Burget, "Semi-supervised training of deep neural networks," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 267–272.

[7] Erik McDermott, Shinji Watanabe, and Atsushi Nakamura, "Discriminative training based on an integrated view of mpe and mmi in margin and error space," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4894–4897.

[8] Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks.," in *INTERSPEECH*, 2013, pp. 2345–2349.

[9] Tomáš Mikolov, "Statistical language models based on neural networks," 2012.

[10] Christopher Cieri, David Miller, and Kevin Walker, "The fisher corpus: a resource for the next generations of speech-to-text.," in *LREC*, 2004, vol. 4, pp. 69–71.

[11] Vikrant Tomar, "Blind dereverberation using maximum kurtosis of the speech residual," 2010.

[12] A Sorin and T Ramabadran, "Extended advanced front end algorithm description, Version 1.1," *ETSI STQ Aurora DSR Working Group, Tech. Rep. ES*, vol. 202, pp. 212, 2003.

[13] Charles H Knapp and G Clifford Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, 1976.

[14] Yong Rui and Dinei Florencio, "Time delay estimation in the presence of correlated noise and reverberation," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004, vol. 2, pp. ii–133.

[15] Cha Zhang, Dinei Florêncio, and Zhengyou Zhang, "Why does phat work well in lownoise, reverberative environments?," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 2565–2568.

[16] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Acoustic beamforming for speaker diarization of meetings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2011–2022, 2007.

[17] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý, "The kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding (ASRU)*. 2011, IEEE.

[18] Carnegie Mellon University, "The carnegie mellon university pronouncing dictionary v07a," in *[Online] http://www.speech.cs.cmu.edu/cgi-bin/cmudict*, 2015.

[19] Tran Huy Dat, Kazuya Takeda, and Fumitada Itakura, "Robust snr estimation of noisy speech based on gaussian mixture modeling on log-power domain," in *ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, 2004.