
Proceedings Paper:
Xiong, F., Goetze, S. orcid.org/0000-0003-1044-7343 and Meyer, B.T. (2017) Combination strategy based on relative performance monitoring for multi-stream reverberant speech recognition. In: Proceedings of 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017). International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 05-09 Mar 2017, New Orleans, LA, USA. IEEE , pp. 4870-4874. ISBN 9781509041183

https://doi.org/10.1109/ICASSP.2017.7953082

COMBINATION STRATEGY BASED ON RELATIVE PERFORMANCE MONITORING FOR MULTI-STREAM REVERBERANT SPEECH RECOGNITION

Feifei Xiong[⋆†], *Stefan Goetze*[⋆†], *Bernd T. Meyer*[‡]

[⋆]Fraunhofer Institute for Digital Media Technology IDMT,
Project Group Hearing, Speech and Audio Technology (HSA), Oldenburg, Germany
[†]Cluster of Excellence Hearing4all, Oldenburg, Germany
[‡]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

ABSTRACT

A multi-stream framework with deep neural network (DNN) classifiers is applied to improve automatic speech recognition (ASR) in environments with different reverberation characteristics. We propose a room parameter estimation model to establish a reliable combination strategy which performs on either DNN posterior probabilities or word lattices. The model is implemented by training a multilayer perceptron incorporating auditory-inspired features in order to distinguish between and generalize to various reverberant conditions, and the model output is shown to be highly correlated to ASR performances between multiple streams, i.e., relative performance monitoring, in contrast to conventional mean temporal distance based performance monitoring for a single stream. Compared to traditional multi-condition training, average relative word error rate improvements of 7.7% and 9.4% have been achieved by the proposed combination strategies performing on posteriors and lattices, respectively, when the multi-stream ASR is tested in known and unknown simulated reverberant environments as well as realistically recorded conditions taken from REVERB Challenge evaluation set.

Index Terms— Reverberant speech recognition, multi-stream, posteriors, performance monitoring, weighted system combination

1. INTRODUCTION

Current automatic speech recognition (ASR) systems provide good performance in many scenarios, especially for matched training and test conditions. On the other hand, performance often severely degrades when additive noise or reverberation result in mismatched data, which remains to be a challenging topic for the speech community [1, 2]. As one approach to this challenge, multi-stream frameworks have been proposed [3, 4], which usually involve independent classifiers trained on different data representations (e.g., multi-band frequency processing, data from different recording environments or with different feature extraction schemes) with a subsequent combination of potentially complementary decisions to achieve an optimal result. One crucial issue in multi-stream frameworks is the combination strategy, which consists of *what to combine*, for instance on the basis of feature recombination [5], hidden Markov model (HMM) state levels [3], neural network posterior probabilities [6], or (lattice or word) hypothesis level [7, 8], and *how to combine*, e.g. frame-wise average of all streams [4, 9] or applying stream-specific weights determined by performance monitoring [10, 11]. In this paper we focus on combination strategies operating on the deep neural network (DNN) posterior probabilities as well as on word lattices from multiple streams. Specifically, state-of-the-art DNN-based ASR [12] is used to create our multi-stream framework; a combination can be performed either on the DNN posteriors or on the individual decoded/transcribed result represented as lattice. In general, recognition performance can be increased by assigning higher weights to more reliable streams [13, 14], and therefore it is highly valuable to be able to monitor the classifier performance on unknown test data in order to determine such combination weights.

There are several ways for performance monitoring of ASR systems so that stream-specific weights can be derived. Based on the observation that high noise levels often increase the entropy of DNN posterior distributions, inverse entropy has been introduced as a means of performance monitoring, in which streams with lower entropy are assigned higher weights [6]. A statistical analysis of phoneme posteriors between training and test data has been conducted in [15], where large divergence between these two statistics indicate possible degradation of the classifier performance. A mean temporal distance (MTD) measure of phoneme posteriors was proposed in [11], which is based on the intuition that distant *clean* posterior vectors will be rather different (since they are likely to belong to different phoneme classes), while the difference should be smaller for noisy vectors. This has later been applied for stream selection in multi-stream ASR [16, 17]. Alternatively, autoencoders have been employed in [18] to learn characteristics of the training data, and the reconstruction error obtained with test data was used to monitor the performance of the corresponding classifier.

The previously mentioned research is focusing on *absolute* performance monitoring of one specific stream classifier, and the *relative* comparisons between streams are implicit. Consequently, additional rules are required to determine the combination weights for a multi-stream ASR framework. In order to provide an *explicit* weighting knowledge dedicated for the combination strategy in a multi-stream system, in this study we propose a relative performance monitoring which considers all streams at once. Instead of exploring noise robustness, we investigate the applicability of multiple streams in different reverberant situations with minor stationary additive background noise. This was motivated by our previous research in which room characteristics were reliably estimated via a discriminative multi-layer perceptron (MLP) incorporating auditory-inspired spectro-temporal features to predict room parameters (such as reverberation time T_{60}) [19], for which the MLP output was found to be correlated to ASR performance [20]. We refer to this approach as ROom Parameter Estimator (ROPE) model and test its applicability in DNN stream weighting to obtain robustness

This work was partially funded by Google via a Google faculty award to Hynek Hermansky and by the Cluster of Excellence 1077/1‘ “Hearing4all’.

against reverberation. The ROPE algorithm was shown to accurately classify different reverberation effects and to generalize to unseen data [21]. Hence, we assume that the posteriors of the ROPE model output correlate with the relative performances between all DNN streams given test data, which can be utilized straightforwardly for combination weights in a multi-stream DNN/HMM framework trained on several specific reverberation conditions.

The remainder of this paper is organized as follows: Section 2 introduces the proposed multi-stream ASR system that employs the ROPE algorithm for relative performance monitoring of all streams. The combination strategies on DNN posterior probabilities and word lattices are briefly described in Section 3. The experimental procedure is outlined in Section 4 and the results with discussion are presented in Section 5 before Section 6 concludes the paper.

2. MULTI-STREAM ASR FRAMEWORK

As depicted in Fig. 1, each DNN classifier produces one stream, which is trained on a specific reverberant condition represented by convolving with a specific room impulse response (RIR) as well as adding the stationary background noises. Traditional mel-filterbank (FBANK) features are fed to DNNs and the posterior probabilities are computed from each DNN stream by forward-passing a test utterance that are usually subject to the common performance monitoring approaches e.g. MTD method [11]. In sequence, the lattices are generated based on the posteriors during the HMM decoder [22] and the final recognized transcription will be used for evaluation.

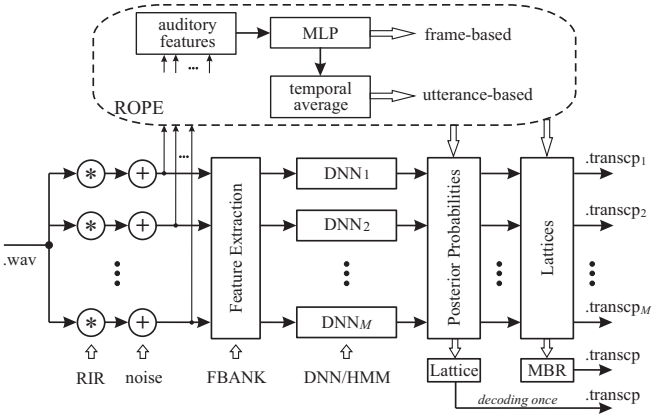

Fig. 1. System structure of the multi-stream ASR framework with M expert DNNs, each trained on a specific reverberant condition.

Assuming that DNNs share the same decision tree for the context-dependent tied states during training, the posterior probabilities given a test data can be combined [4] in order to obtain better acoustic log-likelihoods before decoding. Besides this, it is also typical to perform combination on hypothesis level to obtain complementary knowledge from each stream. Here we investigate the lattice combination using MBR decoding [7] which was shown to improve over the traditional system combination methods such as ROVER [8] or CNC [23]. Lattice combination allows combination of systems with different decision trees but at the cost of multiple decoding operations compared to posterior combination which needs only one decoding operation.

Rather than assigning equal combination weights to all streams which might yield mediocre results, we propose the ROPE model to determine the combination weights that are correlated with the stream performances. An overview of the ROPE processing scheme is depicted in the upper panel of Fig. 1. Diagonal 2D Gabor features [19] are extracted from the reverberant signals and an MLP is trained to map these auditory-inspired inputs to different M classes, where each class represents one of M specific reverberant conditions. With this procedure, the MLP output probabilities of each test item can be interpreted as a measure of similarity of the acoustic test condition and the reverberated room conditions covered by the M expert DNNs. We test if the classification output is a good predictor for room-dependent stream selection by using the classification result directly as stream weights. The MLP generates one estimate per time step, i.e., a frame-based estimate is obtained. The utterance-based estimate is derived from this by simple temporal averaging of the MLP output posterior probabilities.

3. COMBINATION STRATEGY

Since DNN posteriors are in the frame domain, both frame-based and utterance-based mode can be applied for posterior combination. Let $P_m(s,t)$ denote the mth DNN posterior matrix $\boldsymbol{P}_m$ at the HMM state s and time frame t, and $w_m(t)$ be the corresponding combination weight. The combined posteriors $\boldsymbol{P}$ can be represented as the sum over all M weighted posteriors,

$$P(s,t) = \sum_{m=1}^{M} w_m(t)\, P_m(s,t)\,. \tag{1}$$

Regarding the weight vector $\boldsymbol{w}(t) = [w_1(t), \ldots, w_M(t)]$, two different approaches for selection can be tested: (a) a weighting of all streams, where $\sum_{m=1}^{M} w_m(t) = 1$, and (b) a winner-takes-all strategy, i.e., $w_m(t) = 1$ if $m = \arg\max w_m(t), \forall m$, else $w_m(t) = 0$, which is effectively a stream-selection scheme (as applied in [17]).

Similarly, the combined lattice $\mathcal{L}$ for utterance-based processing can be written as

$$\mathcal{L} = \sum_{m=1}^{M} \overline{\boldsymbol{w}}_m\, \mathcal{L}_m\,, \tag{2}$$

where $\overline{\boldsymbol{w}}_m$ denotes the utterance-based weight of the mth stream. Note that the total cost of all paths in the individual lattice $\mathcal{L}_m$ will be removed before the union, and a sequent MBR decoding [7] is applied to achieve the weighted system combination.

4. EXPERIMENTAL SETUP

4.1. Speech Data

We use the WSJCAM0 British English corpus [24] as database of clean (anechoic) speech, which contains 7861 utterances for training and another 363 for test at a sampling rate of 16 kHz. In order to create the expert DNN streams which represent various reverberant conditions, 6 realistic recorded RIRs were chosen to generate the training sets (convolved with clean speech), which cover typical room sizes (small, medium and large) and speaker-to-microphone distances (near and far) inspired by the REVERB challenge settings [1]. The clean condition training (cln) is generated by using the clean speech, and the multi-condition training (mc) involves all 6 RIRs convolved with the same amount of clean speech data for fair comparison. In addition, an extended multi-condition training set (mc-ext) is generated by including all speech data from 6 expert training sets, which can be considered as an extreme baseline presumably since an extension of training data could further improves DNN classification scores [25]. In order to test the applicability of the proposed multi-stream system to various reverberant conditions, we use 4 test sets: Set A contains 6 types of test conditions, from the

chosen 6 specific RIRs used for stream training to evaluate matched training-test conditions. Set B includes 6 additional reverberant conditions, with the respective RIRs recorded in the same room as used for training/Set A but at different positions, which can be considered as mild mismatched training-test conditions. Set C includes further 6 reverberant conditions with the respective RIRs recorded in different rooms from Set A or B, in order to evaluate completely mismatched conditions. Real recordings from the REVERB challenge evaluation set (Set Real) are also adopted to demonstrate the effectiveness of the proposed system as well as the generalization of the proposed relative performance monitoring. Note that the corresponding background noises are added with a signal-to-noise-ratio of approx. 20 dB. Fig. 2 displays the acoustic parameters and labels of the RIRs employed by all test conditions in the form categorized by reverberation time T_{60} and direct-to-reverberant ratio DRR.

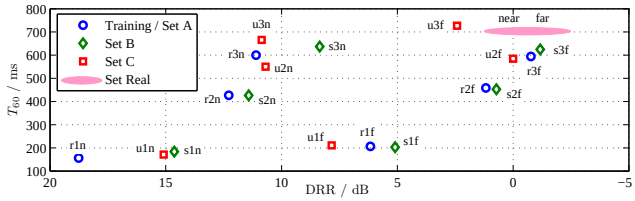

Fig. 2. T_{60} and DRR distribution of the RIRs selected for experiments. Training sets / Test A: r1n, r1f, r2n, r2f, r3n, r3f; Set B: s1n, s1f, s2n, s2f, s3n, s3f; Test C: u1n, u1f, u2n, u2f, u3n, u3f; Test Real: near, far (only estimated T_{60} and DRR).

4.2. ASR and Performance Monitoring

Following the ASR training procedure in our previous study [25], we used FBANK features (40 dimension) with a context window length of 11 frames as DNN input and an auxiliary GMM/HMM system was trained using MFCC features on clean condition data, resulting in 2090 dimensional context-dependent triphone states as DNN output. The text prompts of the utterances are based on WSJ 5K corpus [26], from which a trigram language model was generated.

Auditory-inspired diagonal 2D Gabor features (600 dimension) are used as MLP input in the ROPE algorithm. The optimal number of hidden units was estimated based on the amount of training data and set to 8192 units, and the number of output neurons corresponds to the number of room parameter classes (and consequently DNN streams, i.e. $M = 6$). For comparison, we employed the MTD method [11] for performance monitoring which evaluates the mean cumulative divergence between posterior probabilities $P_m(s,t)$. However, it is only applied to a single stream and the combination weights for all streams are required to be further determined. As suggested in [11, 16], we used

$$w_m = \frac{1/\mathcal{D}_m}{\sum_{m=1}^{M} 1/\mathcal{D}_m} \tag{3}$$

as the combination weights where $\mathcal{D}_m$ denotes the absolute difference between the MTD of all training data of the mth stream and the MTD of the given test utterance. The cumulative range is chosen from 200 ms to 800 ms in steps of 50 ms, with all frames of one utterance being used as each center frame [11, 16, 17], i.e., w_m is the utterance-based weight for Eq. (1)-(2). Further, we also explored the frame-based MTD here by applying a window of 800 ms to the posteriors, so that the temporal context to compute $\mathcal{D}_m(t)$ for frame-based weights $w_m(t)$ is limited to $[t - 800\,\text{ms}, t]$, instead of the whole utterance. As a result, 800 ms delay will be introduced in frame-based mode, which is comparable to the temporal Gabor filter length of 790 ms (also refer to [19, 21]) used in ROPE.

5. RESULTS

5.1. Performance Monitoring

Fig. 3 (a) shows the absolute Pearson's correlation coefficients between the word error rates (WERs) of each test set from 6 expert streams and the relative performance monitoring represented by the utterance-based combination weights $\overline{w}_m$ derived by MTD and ROPE. For Sets A, B, and C, ROPE leads to much higher correlation coefficients than MTD, while both methods provide similarly high coefficients (above 0.8) for Set Real. This is also reflected by the three examples shown in Fig. 3 (b), where for a matched test set such as 'r1n', ROPE is capable of selecting the best stream with a very high probability, but on the other hand, might decrease the correlation since other probabilities (which are close to zero) may be not proportional to WERs of the corresponding streams. Meanwhile, MTD tends to yield nearly uniform distributed stream weights with similar $\mathcal{D}_m$ in Eq. (3), resulting in rather low correlation to WERs (which should result in less reliable performance monitoring). It seems that the MTD-based correlation increases with the increasing reverberation in the test conditions, e.g. from 'r1n' to 'u3n' to 'far', which is outperformed by ROPE.

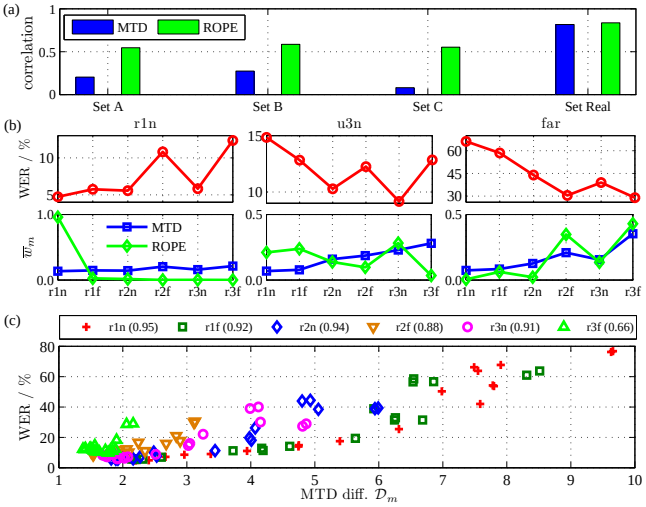

Fig. 3. (a) Absolute Pearson's correlation coefficients between WERs of multiple streams and the utterance-based stream weights $\overline{w}_m$ derived by MTD and ROPE; (b) three correlation examples with 'r1n, u3n, far'; (c) MTD based *absolute* performance monitoring for each of all 6 expert streams with correlation values in parentheses.

When considering absolute performance monitoring in a single stream, MTD produces high correlations with WERs (as expected [11]) as shown Fig. 3 (c). Lower $\mathcal{D}_m$ in the MTD method means closer performance to the best result in its specific stream, however, it does not always guarantee better performance if compared to other streams for relative performance monitoring. For instance, stream 'r3f' always yields lower $\mathcal{D}_m$, i.e. higher w_m in Eq. (3), which is specially preferable to test conditions with similarly high reverberation such as Set Real, but it is not a good option for other conditions with low reverberation in Set A, B, C. This explains the high correlations for Set Real and lower correlations for Set A, B, C in Fig. 3 (a).

5.2. ASR Performance

The single-stream ASR systems perform the best between 6 expert streams when test and training data are matched, as listed for Set A

Table 1. ASR WERs on different training and test reverberation conditions for single- and multi-stream systems that take into account 6 single streams and employ different combination strategies including a DNN posterior probability combination and MBR lattice combination. The WER of the clean test signal with clean condition training model is 4.40%.

	Training \ Test	Set A							Set B							Set C							Set Real		
		r1n	r1f	r2n	r2f	r3n	r3f	avg.	s1n	s1f	s2n	s2f	s3n	s3f	avg.	u1n	u1f	u2n	u2f	u3n	u3f	avg.	near	far	avg.
Single-stream	cln	6.73	15.35	20.09	59.11	32.42	78.89	35.43	7.86	11.43	26.26	57.46	50.87	78.91	38.79	7.27	15.30	17.61	60.43	26.29	75.74	33.69	73.01	74.88	73.94
	mc	5.47	6.27	6.51	9.33	7.66	12.26	7.91	5.61	6.27	6.71	9.72	8.91	14.38	8.60	5.54	6.83	8.39	16.77	8.86	21.21	11.26	31.04	32.38	31.71
	mc-ext	4.76	5.20	5.69	7.76	6.27	9.88	6.59	4.81	5.59	5.91	8.11	7.25	11.06	7.12	4.88	6.12	7.06	14.65	8.10	18.70	9.91	31.81	30.01	30.91
	r1n	4.73	8.55	13.96	54.23	25.44	76.22	30.52	5.08	7.13	17.47	53.74	42.01	76.74	33.69	4.93	9.00	11.05	50.33	14.84	67.68	26.30	63.78	66.17	64.97
	r1f	5.74	5.44	11.37	32.90	19.33	60.97	22.62	5.84	5.81	14.11	31.44	31.42	63.68	25.38	5.62	6.96	11.18	38.83	12.82	56.68	22.01	56.50	58.51	57.50
	r2n	5.56	7.56	5.86	19.77	7.69	38.91	14.22	7.32	5.79	7.23	18.11	11.30	39.37	14.85	5.66	7.28	8.50	26.16	10.28	38.49	16.06	44.36	43.96	44.16
	r2f	10.81	8.96	9.11	9.15	10.16	15.87	10.67	10.66	8.47	8.84	9.99	11.05	17.74	11.12	10.27	10.08	11.76	16.60	12.25	20.97	13.65	29.80	30.49	30.14
	r3n	5.83	7.69	6.25	15.96	7.15	27.31	11.69	6.03	7.17	6.83	14.67	8.79	28.93	12.07	6.18	7.79	8.27	22.06	9.16	30.10	13.92	40.02	38.96	39.49
	r3f	12.35	11.23	10.81	9.86	11.47	11.05	11.12	12.69	10.67	10.37	10.08	11.64	12.35	11.30	12.10	11.89	12.54	14.65	12.84	18.24	13.71	28.68	29.00	28.84
Posterior Combination	Sum	5.76	6.62	7.15	13.13	8.72	21.19	10.42	6.03	6.66	8.13	13.45	11.99	23.99	11.70	5.62	7.47	8.66	19.23	9.71	26.17	12.81	33.98	33.46	33.72
	Product	5.71	6.34	6.71	14.94	8.89	27.95	11.75	5.71	6.44	7.98	14.96	11.76	31.04	12.98	5.61	6.84	8.61	21.51	9.50	30.61	13.78	38.29	39.33	38.81
	MTD Fr All	5.88	6.86	6.83	12.33	8.35	16.81	9.51	5.93	6.64	8.03	12.18	10.69	18.58	10.34	6.06	7.45	8.96	18.92	9.67	23.04	12.35	32.55	32.04	32.29
	MTD Fr Max	7.49	7.47	8.17	13.16	10.25	15.96	10.41	7.00	7.03	8.83	13.03	12.35	17.33	10.92	7.44	8.05	9.98	19.13	11.71	23.46	13.29	32.67	31.97	32.32
	MTD Utt All	6.30	6.76	7.25	10.94	8.45	15.81	9.25	6.15	6.89	8.00	11.62	10.59	17.23	10.08	5.95	7.45	8.98	16.96	9.71	22.07	11.85	30.73	30.22	30.47
	MTD Utt Max	9.13	8.54	8.78	10.64	10.18	11.76	9.83	9.44	8.22	9.20	10.50	11.28	12.93	10.26	9.25	9.64	10.76	**14.87**	12.35	**17.92**	12.46	28.68	29.10	28.89
	ROPE Fr All	**4.51**	5.44	5.83	9.22	7.06	11.11	**7.19**	**4.95**	6.15	7.06	10.52	9.42	13.59	8.61	5.20	6.91	**8.45**	16.33	9.42	19.31	**10.93**	29.35	29.41	29.38
	ROPE Fr Max	4.61	5.39	5.91	**9.03**	7.10	11.13	**7.19**	4.98	6.13	7.06	10.57	9.35	13.77	8.64	5.17	6.84	8.79	17.01	9.66	19.58	11.17	29.51	30.35	29.93
	ROPE Utt All	4.71	**5.35**	**5.76**	9.37	**6.95**	11.18	7.22	5.00	6.17	**7.05**	10.50	9.05	13.70	8.57	5.03	**6.83**	8.74	16.81	**9.33**	20.18	11.15	**27.12**	**28.53**	**27.82**
	ROPE Utt Max	4.71	5.39	5.88	9.13	7.10	**10.99**	7.20	5.01	**5.81**	7.17	**10.06**	**8.79**	**12.35**	**8.19**	**4.88**	7.18	10.59	16.38	11.11	19.36	11.58	28.90	29.17	29.03
MBR	Equal weights	5.29	6.05	6.23	11.71	7.81	18.50	9.26	5.42	6.22	7.10	11.60	9.79	20.67	10.13	5.30	**6.45**	**8.42**	18.13	9.10	23.85	11.87	33.44	33.29	33.36
	MTD Utt All	6.27	6.54	6.71	9.71	7.64	12.43	8.21	6.27	6.45	7.22	10.10	9.50	13.74	8.88	6.08	7.15	8.59	14.77	9.55	18.92	10.84	28.04	29.41	28.72
	MTD Utt Max	9.01	8.05	8.71	10.32	9.81	11.60	9.58	9.10	8.00	9.20	10.25	10.89	12.57	10.00	9.16	9.55	10.33	**14.70**	12.04	17.75	12.25	27.85	29.03	28.44
	ROPE Utt All	**4.52**	5.35	**5.81**	**8.79**	**7.08**	**10.84**	**7.06**	**4.84**	**5.76**	**6.69**	**9.71**	8.91	12.26	**8.02**	**4.68**	6.61	8.66	15.08	**9.03**	**18.89**	**10.49**	**27.08**	**28.49**	**27.78**
	ROPE Utt Max	4.59	**5.34**	5.88	8.81	7.12	10.86	7.10	4.90	5.79	7.10	9.83	**8.78**	**11.88**	8.04	4.78	7.13	10.74	16.20	11.10	19.19	11.52	28.11	29.03	28.57

in Table 1. In general, scenarios never seen during training can still benefit from expert streams covering similar reverberant conditions, e.g. the WERs from diagonal positions of Set B and C are smaller than the performance of the multi-condition system (mc). The extended multi-condition system (mc-ext) provides even lower WERs of approx. 1.5% on average, since it generalizes better to unseen data than 'mc'. However, the improvement for Set Real is rather small, indicating that the effectiveness might be limited by solely extending training data size for improving the ASR robustness against specific reverberant situations, particularly for severely mismatched cases.

For the posterior combination strategy in multi-stream systems, sum and product rules with equal weights [4] provide mediocre results compared to the individual stream performance, and in general, the sum rule performs better than the product rule (which can be considered as sum rule in the logarithm domain), particularly in high reverberant conditions such as 's3f, u3f' and Set Real. When the combination weights derived from relative performance monitoring are available with two modes (utterance-based 'Utt' and frame-based 'Fr') and two selection rules (all-stream-combine 'All' and winner-takes-all 'Max'), a posterior combination becomes more effective and absolute average WERs are reduced by 1.3% and 3.2% when applying MTD and ROPE, respectively. For the MTD method, 'Utt' generally surpasses 'Fr' due to the insufficient temporal context for 'Fr' (800 ms vs. whole utterance) to accurately calculate $\mathcal{D}_m$ for each frame. 'Max' usually leads to better performance for high reverberant test sets while 'All' tends to achieve similar results as sum rule for other low reverberant conditions due to nearly uniform distributed weights as exemplified in Fig. 3 (b).

In contrast, ROPE based combination weights are capable of improving the ASR performance in all four modes, and performances of almost all the test conditions are close or even superior to the corresponding best result from single streams. Specifically, 'Fr' performs similarly as 'Utt', indicating that the proposed ROPE method is applicable in real-time multi-stream ASR. 'Max' seems to give slightly better results than 'All' when matched or mildly mismatched stream exists such as for Sets A and B. In other words, combinations with potentially detrimental streams might even degrade the final performance. On the other hand, if completely unseen scenarios are tested, 'All' generally surpasses 'Max' since high weights will be assigned to the reliable streams (with complementary knowledge) due to the generalization of ROPE. Their combination would yield a result close to the best one from single streams (e.g. 'u1f, u2f') or even better, e.g., Set Real with absolute WER reduction of 1% on average with 'Utt All', indicating the potential of posterior combination being capable of extracting complementary knowledge from different streams. Meanwhile, 'Max' might increase the risk of selecting a similar stream which however does not produce the lowest error, particularly in utterance-based processing, e.g. in 'u2n, u3n'.

The same trend can be observed for MBR lattice combination, in which mediocre results are obtained when equal weights are applied. Also, weighted combination is effective to improve the multi-stream performance, where in general, 'All' behaves better than 'Max' and ROPE outperforms MTD by absolute 1-2% on average. It is also interesting to observe that lattice combination performs only slightly better than posterior combination strategy, indicating that it might be more preferable to explore combination strategies on DNN posteriors due to its low complexity with only one decoding operation while multiple decoding operations are required to perform lattice combination. Furthermore, the multi-stream system with 'Utt All' incorporating ROPE achieves an absolute WER reduction of 1.05% and 0.15% on average compared to 'mc' and the *oracle* best single stream, respectively. Particularly for Set Real, the proposed system even outperforms 'mc-ext' by absolute 3.13%, albeit WERs are 0.65% higher on average for simulated test Set A, B, C.

6. CONCLUSIONS

In this paper we investigated the effectiveness of a multi-stream DNN/HMM framework for ASR systems in various reverberant environments. In order to determine reliable weights for combination based on either DNN posteriors or word lattices, a dedicated relative performance monitoring was proposed based on a room parameter estimator (ROPE), which exhibits higher correlations of WERs for multiple streams than the conventional mean temporal distance based performance monitoring with an additional rule for weight determination. This resulted in consistent improvements in known and unknown reverberant scenarios, outperforming the baseline systems with equal weights and the mean temporal distance method for stream weighting or selection. We also showed that the proposed system provided better performance than a multi-condition baseline and a very competitive extended multi-condition baseline particularly for realistic evaluation data from REVERB Challenge. Further, although lattice combination generally performed slightly better, posterior combination showed its potential for real-time applications (frame-wise) while providing a low complexity during decoding.

7. REFERENCES

[1] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[2] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, USA, Dec. 2015, pp. 504–511.

[3] H. Bourlard, S. Dupont, H. Hermansky, and N. Morgan, "Towards Subband-based Speech Recognition," in *Proc. of European Signal Processing Conference (EUSIPCO)*, Trieste, Italy, Sep. 1996, pp. 1579–1582.

[4] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On Combining Classifiers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[5] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-band Speech Recognition in Noisy Environments," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Seattle, WA, USA, May 1998, pp. 641–644.

[6] H. Misra, H. Bourlard, and V. Tyagi, "New Entropy based Combination Rules in HMM/ANN Multi-Stream ASR," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2003, pp. 741–744.

[7] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes Risk Decoding and System Combination based on A Recursion for Edit Distance," *Journal on Computer Speech and Language*, vol. 25, no. 4, pp. 802–828, Oct. 2011.

[8] J. G. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Santa Barbara, CA, USA, Dec. 1997, pp. 347–354.

[9] K. Kirchhoff and J. A. Bilmes, "Combination and Joint Training of Acoustic Classifiers for Speech Recognition," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[10] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-Stream Adaptive Evidence Combination for Noise Robust ASR," *Speech Communication*, vol. 34, pp. 25–40, 2001.

[11] H. Hermansky, E. Variani, and V. Peddinti, "Mean Temporal Distance: Predicting ASR Error from Temporal Properties of Speech Signal," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 7423–7426.

[12] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[13] B. Hoffmeister, T. Klein, R. Schlüter, and H. Ney, "Frame based System Combination and a Comparison with Weighted ROVER and CNC," in *Proc. Interspeech*, Pittsburgh, USA, Sep. 2006, pp. 537–540.

[14] F. Valente and H. Hermansky, "Combination of Acoustic Classifiers based on Dempster-Shafer Theory of Evidence," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, HI, USA, Apr. 2007, pp. 1129–1132.

[15] N. Mesgarani, S. Thomas, and H. Hermansky, "A Multi-stream Multiresolution Framework for Phoneme Recognition," in *Proc. Interspeech*, Makuhari, Japan, Sep. 2010, pp. 318–321.

[16] E. Variani, F. Li, and H. Hermansky, "Multi-Stream Recognition of Noisy Speech with Performance Monitoring," in *Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 2978–2981.

[17] S. H. Mallidi, T. Ogawa, and H. Hermansky, "Uncertainty Estimation of DNN Classifiers," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, USA, Dec. 2015, pp. 283–288.

[18] S. H. Mallidi, T. Ogawa, K. Veselý, P. S. Nidadavolu, and H. Hermansky, "Autoencoder based Multi-Stream Combination for Noise Robust Speech Recognition," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015, pp. 3551–3555.

[19] F. Xiong, S. Goetze, and B. T. Meyer, "Blind Estimation of Reverberation Time based on Spectro-Temporal Modulation Filtering," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 443–447.

[20] F. Xiong, S. Goetze, and B. T. Meyer, "Estimating Room Acoustic Parameters for Speech Recognizer Adaptation and Combination in Reverberant Environments," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 5559–5563.

[21] F. Xiong, S. Goetze, and B. T. Meyer, "Joint Estimation of Reverberation Time and Direct-to-Reverberation Ratio from Speech using Auditory-Inspired Features," in *ACE Challenge Workshop, satellite event of IEEE-WASPAA*, New Paltz, NY, USA, Oct. 2015.

[22] N. Morgan and H. Bourlard, "An Introduction to the Hybrid HMM/Connectionist Approach," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 24–42, 1995.

[23] G. Evermann and P. Woodland, "Posterior Probability Decoding, Confidence Estimation and System Combination," in *NIST Speech Transcription Workshop*, College Park, MD, USA, 2000.

[24] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAM0: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Detroit, MI, USA, May 1995, pp. 81–84.

[25] F. Xiong, B. T. Meyer, N. Moritz, R. Rehr, J. Anemüller, T. Gerkmann, S. Doclo, and S. Goetze, "Front-End Technologies for Robust ASR in Reverberant Environments - Spectral Enhancement-based Dereverberation and Auditory Modulation Filterbank Features," *EURASIP Journal on Advances in Signal Processing*, vol. 2015:70, pp. 1–18, 2015.

[26] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete," in *Linguistic Data Consortium (LDC)*, Philadelphia, PA, USA, 2007.