Proceedings Paper:
Xiong, F., Goetze, S. orcid.org/0000-0003-1044-7343 and Meyer, B.T. (2017) On DNN posterior probability combination in multi-stream speech recognition for reverberant environments. 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. 5250-5254. ISBN 9781509041183

https://doi.org/10.1109/ICASSP.2017.7953158

ON DNN POSTERIOR PROBABILITY COMBINATION IN MULTI-STREAM SPEECH RECOGNITION FOR REVERBERANT ENVIRONMENTS

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
{feifei.xiong, s.goetze}@idmt.fraunhofer.de, bernd.t.meyer@jhu.edu

ABSTRACT

A multi-stream framework with deep neural network (DNN) classifiers has been applied in this paper to improve automatic speech recognition (ASR) performance in environments with different reverberation characteristics. We propose a room parameter estimation model to determine the stream weights for DNN posterior probability combination with the aim of obtaining reliable log-likelihoods for decoding. The model is implemented by training a multi-layer perceptron to distinguish between various reverberant environments. The method is tested in known and unknown environments against approaches based on inverse entropy and autoencoders, with average relative word error rate improvements of 46% and 29%, respectively, when performing multi-stream ASR in different reverberant situations.

Index Terms— Reverberant speech recognition, multi-stream, neural network, posterior probability, weighted 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 can severely degrade when additive noises 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, a multi-stream framework has been proposed [3, 4], which usually involves independent classifiers trained on different data representations (e.g., multi-band frequency processing, data from various recoding environments or with different feature extraction schemes) with a subsequent combination of potentially complementary decisions to achieve an optimal result. One crucial issue in such multi-stream frameworks is the combination strategy, which can be performed for instance on the basis of feature recombination [5], hidden Markov model (HMM) state levels [3], neural network posterior probabilities [6], or word level combinations/confusions [7].

In this work we focus on combination strategies of deep neural network (DNN) posterior probabilities from multiple streams. Specifically, state-of-the-art DNN-based ASR [8] is used to create a multi-stream framework that is based on DNN output posterior probabilities which form individual streams; after stream combination (for which different methods are explored), the merged stream is used for decoding. The technique for stream merging is a crucial design choice for multi-stream ASR. A relatively simple approach is to calculate the frame-wise average of all streams, as has been proposed in [4, 9]. When knowledge about the stream quality is available, performance can be increased by assigning higher weights to reliable streams [10]. There are several ways to obtain stream-specific weights in *frame*-based mode that are compared in this study: Inverse entropy is a well-established and successful confidence measure in which streams with a low entropy are assigned a high weight [6]. Recently, autoencoders have been proposed for stream weighting, where the reconstruction error of DNN test data activations is used to select or weigh streams [11]. In cases when some streams carry detrimental information, it might be better to pursue a winner-takes-all approach, which has also been explored in autoencoder approaches [12], and is also investigated in this study.

The previously mentioned research on multi-stream frameworks is mainly focusing on noise robustness in ASR. In this study we investigate the applicability of multiple streams in situations with different reverberation conditions. This was motivated by our previous research in which room characteristics were reliably estimated via a discriminative multi-layer perceptron (MLP) [13] to predict room parameters (such as reverberation time T_{60}) [14] or specific environments, where the MLP softmax output was used to combine multiple ASR systems at word level [15]. We refer to this approach as ROom Parameter Estimator (ROPE) model and test its applicability in DNN stream weighting to obtain robustness against reverberation. We assume that DNN performance for each stream correlates with the mismatch level between training and reverberant test data. The ROPE algorithm was shown to accurately classify different reverberation effects and to generalize to unseen data. Hence, the posterior probabilities of the ROPE model output are used as combination weights for each stream in the multi-stream DNN/HMM framework trained on several specific reverberation conditions.

The remainder of this paper is organized as follows: Section 2 briefly introduces the proposed ASR system that employs stream-weighted DNN processing. The combination strategies based on inverse entropy (InvEnt) and autoencoders (AEnc), as well as the proposed ROPE algorithm are described in Section 3. The experimental procedure is outlined in Section 4 and the results with discussion are presented in Section 5. Section 6 concludes the paper.

2. MULTI-STREAM ASR FRAMEWORK

Unlike previous studies on multi-stream frameworks utilizing multiple acoustic features [16, 17, 11], we employ the multi-stream framework involving several *expert* DNNs (FBANK features as input), which are trained on different reverberant data sets to generate posteriors suitable for dealing with various reverberant situations. As

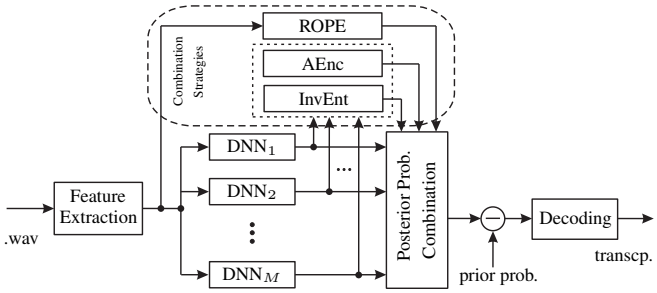

Fig. 1. System structure of the multi-stream ASR framework with M DNNs, each trained on a specific reverberant condition. InvEnt and AEnc, as well as the proposed ROPE-based approaches are employed for stream combination.

depicted in Fig. 1, each DNN classifier trained on a specific reverberant condition produces one stream. Posterior probabilities are computed from each DNN stream by forward-passing a test utterance, which is then subject to the combination strategies InvEnt and AEnc. In contrast to this, the proposed ROPE-based combination strategy directly uses the acoustic features calculated from the test utterance to produce the combination weight. The combined posterior probabilities are obtained by applying the combination weight to each DNN posterior probabilities using the sum rule [4] (which produced better results than the product rule [4] in our pilot experiments) for stream merging. Prior triphone probabilities learned from the training set are subtracted in logarithm domain from the posterior probabilities [18] before they are passed to the HMM decoder.

3. STREAM COMBINATION AND WEIGHTING

Since the combination strategies operate on DNN posterior probabilities, the frame-based mode can be applied for real-time applications: Let $P_m(s,t)$ denote the mth DNN posterior probability matrix $\boldsymbol{P}_m$ at the HMM state s and time frame t, and $w_m(t)$ be the corresponding combination weight at time frame t. The combined posterior probability $\boldsymbol{P}$ obtained from the sum rule [4] can be represented as the sum over all M weighted posterior probabilities, i.e.,

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

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. Further, the utterance-based mode can be implemented by simple temporal averaging across the whole utterance for all frames t.

3.1. Inverse Entropy

The entropy of the DNN softmax output is used as the confidence measure as proposed in [6], where the weight $w(t)$ is reciprocal to the value of entropy $\boldsymbol{e}_{\mathrm{InvEnt}}(t) = -\sum_{s=1}^{S} P(s,t)\log_2(P(s,t))$ and S denotes the HMM state dimension. The corresponding weight vector $\boldsymbol{w}(t)$ is computed as

$$w_m(t) = \frac{1/e_{\mathrm{InvEnt},m}(t)}{\sum_{m=1}^{M} 1/e_{\mathrm{InvEnt},m}(t)}\,. \tag{2}$$

3.2. Autoencoders

The combination rule using autoencoders was proposed in [11] based on the observation that matched test data yields low reconstruction errors compared to mismatched test data when an autoencoder was trained on DNN pre-softmax output by minimizing the reconstruction error to capture the distribution of the DNN activations from each stream. In other words, a high reconstruction error indicates a large mismatch between the given test data and the respective DNN stream, so that a low weight should be assigned to this stream. As suggested in [11], the elements of the weight vector $\boldsymbol{w}(t)$ are given by

$$w_m(t) = \frac{1/\left\|e_{\mathrm{AEnc},m}(s,t)\right\|^2}{\sum_{m=1}^{M} 1/\left\|e_{\mathrm{AEnc},m}(s,t)\right\|^2}\,, \tag{3}$$

where the reconstruction error $e_{\mathrm{AEnc},m}(s,t)$ is calculated as the difference between the autoencoder input and output for each stream and $\|\cdot\|$ denotes ℓ_2-norm operation across all the HMM state s.

3.3. Room Parameter Estimator

An overview of the ROPE processing scheme is depicted in Fig. 2: First, reverberant signals are constructed from clean (anechoic) speech convolved with measured room impulse responses (RIRs) [19]. An MLP is trained to map the input (acoustic FBANK features) to different reverberant conditions, where the labels represent one of M specific reverberant conditions. Since the number of MLP output neurons equals the number of DNN expert streams, the MLP outputs can be used directly as DNN stream weights. With this procedure, we investigate if the classes that represent specific reverberation conditions are suitable to predict which stream will perform best for a given test item. The MLP generates one estimate per time step, i.e., a frame-based estimate is directly obtained. The utterance-based estimate is obtained by simple temporal averaging of the MLP output.

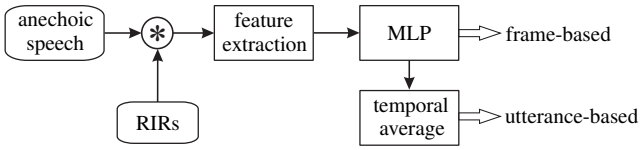

Fig. 2. Overview of the ROPE model setup to determine the combination weight vector $\boldsymbol{w}(t)$.

4. EXPERIMENTAL SETUP

4.1. Speech Data and Impulse Responses

We used the WSJCAM0 British English corpus [20] as database of clean (anechoic) speech. It contains 7861 utterances for training and another 1088 for test at a sampling rate of 16 kHz, in conformance with the REVERB Challenge [1]. We used the collected realistic recorded RIRs [21] to generate various reverberant conditions which were categorized by reverberation time T_{60} and direct-to-reverberate ratio DRR. To create the expert DNN streams, 8 types of training sets with the same amount of speech data were selected: The clean condition training is generated by using the clean speech, and the multi-condition training involves 44 generalized RIRs (convolved with clean speech) with T_{60} ranging from 100 ms to 900 ms and DRR ranging from -4 dB to 18 dB, which covers a wide range of RIRs that occur in real life scenarios. Six RIRs were chosen to create six additional training sets, which cover typical room sizes (small, medium and large) and speaker-to-microphone distances (near and

far). In order to test various reverberant conditions, we used 2 test sets: Set A contains seven types of test conditions, including clean and the chosen six specific RIRs used for stream training. In other words, Set A can be used to evaluate matched training-test conditions. Set B includes ten additional types of reverberant conditions, including six types that are similar to the above mentioned six training RIRs (mild), two types with moderate mismatch and two types represent severely mismatched conditions, as plotted in Fig. 3. As a result, eight versions of training data (clean, six specific reverberant conditions and multi-condition), are considered in the multi-stream framework ($M = 8$ in Fig. 1), and 17 test conditions (seven from Set A and ten from Set B) are used to evaluate the effectiveness of the combination strategies in Section 3 w.r.t. ASR performance under various reverberant environments.

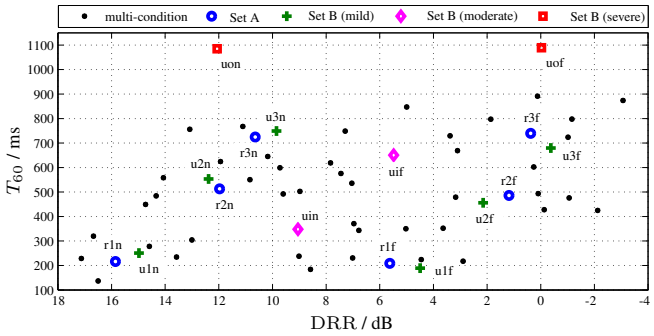

Fig. 3. T_{60} and DRR distribution of the selected RIRs for experiments. Together with clean (cln) and multi-condition (mc), six specific conditions (r1n, r1f, r2n, r2f, r3n, r3f) were chosen to construct the multi-stream training set. The test set contains Set A including seven test conditions (cln and six specific RIRs from training set) and Set B including six mild (u1n, u1f, u2n, u2f, u3n, u3f), two moderate 'inner' (uin, uif) and two severe 'outer' conditions (uon, uof).

4.2. ASR, Autoencoders and ROPE

For DNN/HMM pre-training, we use a 7-layer deep belief network with 2048 neurons for the hidden layers trained as the stack of the restricted Boltzmann machine [22]. DNNs are then fine-tuned using a cross-entropy cost function to classify feature vectors into triphone states provided by an auxiliary GMM triphone system trained with the maximum likelihood criterion. Similar to previous studies [23], 40-dimensional FBANK features with a context window length of 11 frames were used as DNN input and the GMM/HMM system was trained using MFCC features with LDA-MLLT (40-dimensional) on clean condition data, resulting in 2090-dimensional context-dependent triphone states. 5% of the training data served as validation set, while the remaining data was used for training. The text prompts of the utterances are based on the WSJ 5K corpus [24], from which a trigram language model was generated.

The autoencoder uses an input layer followed by two 2048-dimensional hidden layers, a 1024-dimensional bottleneck, two 2048-dimensional layers and the output layer. The cost function used to optimize the network parameters is the mean squared error function with sigmoid nonlinearity. The pre-softmax output of DNNs is used as input to the autoencoder. In order to obtain comparable reconstruction performance from autoencoders (which is required for reasonable comparison with a given test set, but different training data), we fine-tune the training procedure for each autoencoder by adjusting the number of epochs to achieve similar training reconstruction error distributions.

In a similar setup, the spliced FBANK features are also used as MLP input to estimate room parameters with the ROPE algorithm (see Section 3.3). The optimal number of hidden units was estimated based on the amount of training data and set to 8192 units. One hidden layer is used and the number of output neurons corresponds to the number of room parameter classes (and consequently DNN streams, i.e. $M = 8$). The cost function used to optimize MLP parameters is the cross-entropy function with sigmoid nonlinearity. All the aforementioned systems are implemented with the Kaldi speech recognition toolkit [25].

5. RESULTS AND DISCUSSION

5.1. Single-stream Results

The single-stream ASR systems achieve the best results when test and training data are matched (cf. Set A in upper left part of Table 1). Not surprisingly, the 'mc' system provides lower word error rates (WERs) on *average* since it generalizes better to unseen data (44 RIRs involved). However, some scenarios never seen during training (Set B) still benefit from specific streams. For instance, 'u1n', 'u2n' and 'uon' result in lower WERs when using specific streams rather than 'mc', indicating that this specific training data has a higher similarity to the test data compared to the generalized 'mc' training set. Results for Set B (mild) for similar pairs of T_{60} and DRR are good, but 'mc' training produces very good results particularly for far-field with low DRR values. For Set B with moderate and severe situations, specific streams can still provide comparable results to the 'mc' stream.

5.2. Evaluation of Combination Strategies

As described in Section 3, we use InvEnt, AEnc and ROPE methods for weight estimation, and apply two decision rules (utterance- and frame-based) and two stream merging schemes (winner-takes-all and stream weighting) in each case, denoted by 'Utt-Max', 'Utt-SW', 'Frame-Max' and 'Frame-SW', respectively. Fig. 4 illustrates the respective average WERs from all 17 test conditions. Lowest WERs are obtained with the proposed ROPE method in all four modes, with an average absolute WERs reduction of 7.5%, 7.2% and 4.4% compared to equal weights, InvEnt and AEnc, respectively. For InvEnt and AEnc, 'Frame' works better than 'Utt' in general, while ROPE performs nearly consistent in these four different modes. A possible explanation is the independent frame processing in InvEnt and AEnc, resulting in isolated noisy frames which might severely affect 'Utt-Max' since inaccurate decision based on maximal averaged probability will result in completely unreliable stream selection for the whole utterance; the ROPE algorithm uses spliced features as input, which provides some temporal smoothing, albeit the splicing window size is moderate with 11 frames in total.

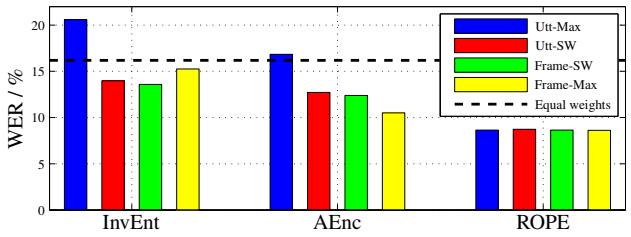

Fig. 4. ASR Performance comparison for the multi-stream systems using three different combination strategies (InvEnt, AEnc, ROPE) for 'Utt' or 'Frame' combinations with 'SW' or 'Max', as well as a baseline with equal weights.

Table 1. ASR WERs on different training and test reverberation conditions for single-stream systems (upper half) and for the multi-stream systems that take into account eight single streams and employ different stream combination strategies (lower half).

	Training \ Test	Set A							Set B: (mild)						(moderate)		(severe)		Avg.
		cln	r1n	r1f	r2n	r2f	r3n	r3f	u1n	u1f	u2n	u2f	u3n	u3f	uin	uif	uon	uof	
Single-stream	cln	**4.78**	7.26	11.94	25.54	58.24	32.66	79.60	8.66	8.93	17.47	36.96	41.77	79.96	16.81	42.44	28.19	86.28	34.55
	r1n	5.44	**5.19**	8.16	15.23	42.24	22.75	69.24	**5.66**	7.17	11.50	27.26	27.52	68.09	11.67	32.73	19.37	81.86	27.12
	r1f	7.19	6.24	**6.04**	12.45	30.18	20.24	60.07	6.25	6.93	11.03	20.68	21.87	55.57	9.33	26.28	20.13	76.99	23.38
	r2n	7.16	6.17	7.26	**6.79**	19.10	12.00	43.06	6.42	7.51	8.00	14.33	9.54	34.43	7.85	16.42	11.19	57.14	16.13
	r2f	12.13	9.31	7.84	8.82	**8.80**	11.38	23.86	9.38	9.95	9.47	10.73	10.34	17.12	8.55	10.89	13.62	39.94	13.06
	r3n	7.48	6.63	8.21	10.05	19.14	**7.06**	23.82	7.34	8.52	**6.92**	14.02	12.54	30.15	9.16	10.61	**7.86**	38.16	13.39
	r3f	13.35	11.11	10.63	12.34	12.87	9.49	**10.99**	11.58	10.63	12.34	12.87	9.49	16.16	11.51	10.34	10.52	20.81	12.17
	mc	5.85	5.92	6.55	7.38	10.20	8.56	14.37	6.06	**6.64**	7.01	**9.30**	**8.50**	**14.09**	**7.04**	**8.80**	8.16	**19.09**	**9.03**
Multi-stream	Equal weights	6.01	6.45	7.83	10.89	18.65	11.52	33.96	6.24	7.51	8.93	15.63	15.01	35.25	8.99	15.87	12.71	53.66	16.18
	InvEnt Utt-Max	6.14	6.27	7.70	13.75	30.14	16.73	46.03	6.64	7.93	10.36	22.34	20.18	47.75	11.74	23.48	15.44	57.42	20.59
	InvEnt Utt-SW	5.77	5.58	6.24	8.51	17.05	10.11	30.30	5.70	6.57	7.54	13.43	11.74	30.35	7.72	13.28	10.16	47.55	13.97
	InvEnt Frame-SW	5.77	5.42	6.27	8.27	16.31	9.93	28.71	5.68	6.49	7.47	13.28	11.61	29.35	7.49	12.64	9.87	46.44	13.58
	InvEnt Frame-Max	5.82	5.66	6.75	9.46	18.58	10.96	32.34	6.08	6.72	8.19	14.76	13.63	34.00	8.55	14.90	11.00	52.02	15.26
	AEnc Utt-Max	6.87	6.25	6.04	12.03	24.85	16.01	31.09	6.37	7.07	10.32	16.74	18.67	37.35	9.13	21.16	16.79	39.50	16.83
	AEnc Utt-SW	5.87	5.49	6.14	8.18	15.13	9.81	25.30	5.79	6.47	7.49	12.47	11.26	25.84	7.57	12.20	9.81	41.35	12.71
	AEnc Frame-SW	5.98	5.55	6.27	7.88	14.59	9.45	24.70	5.77	6.51	7.39	12.09	10.60	24.72	7.40	11.48	9.46	40.67	12.38
	AEnc Frame-Max	6.61	6.11	6.75	7.62	11.79	9.02	17.03	6.18	6.78	7.46	10.95	9.55	17.83	7.54	10.41	9.18	27.74	10.50
	ROPE Utt-Max	**4.78**	5.18	6.04	6.77	**8.63**	**6.99**	**10.61**	5.76	6.64	6.79	**8.97**	9.33	16.71	7.03	**8.57**	8.05	**20.15**	8.64
	ROPE Utt-SW	4.79	**5.11**	**5.90**	6.70	8.78	7.09	11.29	**5.47**	**6.22**	6.76	9.99	8.58	15.50	7.03	8.88	8.15	22.19	8.73
	ROPE Frame-SW	4.81	5.17	5.92	**6.53**	8.65	7.02	11.08	5.51	6.40	**6.68**	9.60	**8.38**	15.23	**6.93**	9.05	**7.84**	22.22	8.64
	ROPE Frame-Max	4.86	5.15	5.97	**6.53**	8.76	7.01	11.00	5.73	6.60	6.72	9.43	8.66	**15.18**	7.11	9.02	7.95	20.88	**8.62**

As quantified in the lower half of Table 1, simply using equal weights for combination which produces mediocre results and is hence a suboptimal processing strategy. In general, AEnc performs better than InvEnt and ROPE leads to the best results, showing that ROPE surpasses InvEnt and AEnc to determine reliable combination weights. Further, compared to the best single-stream system in most test sets, multi-stream systems can provide comparable results, indicating that the proposed multi-stream framework is robust against different types of reverberation by a weighted combination of complementary DNN posterior probabilities. On the other hand, with comparatively poor results for some high-reverberation scenarios (e.g., 'r3f' 'u3f' and 'uof'), the average performance obtained by InvEnt and AEnc is below the single-stream 'mc'. Although ROPE outperforms 'mc' on average for all four combination techniques, it can be observed that for unseen high-reverberant conditions 'u3f' and 'uof', 'mc' is still advantageous and 'Max' seems to be more preferable than 'SW', indicating that weights of detrimental streams are at least occasionally too high.

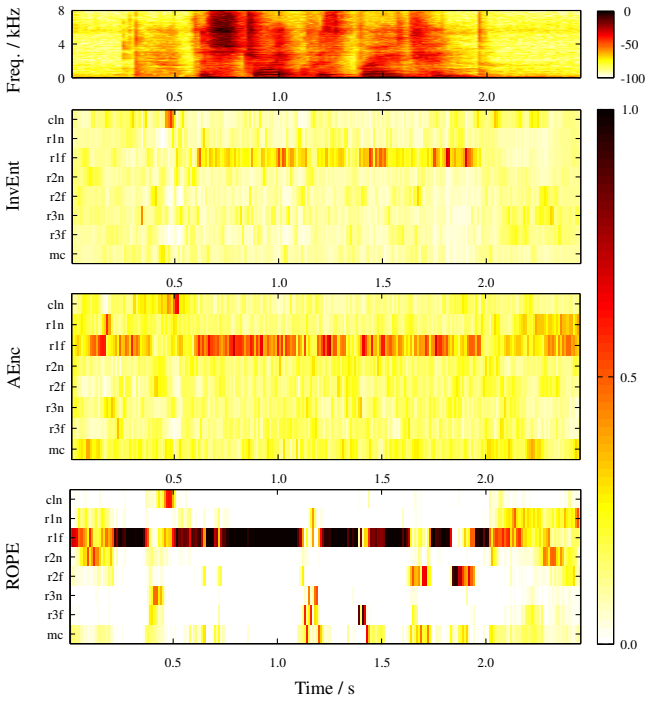

Fig. 5. Example for a speech utterance from reverberation scenario 'r1f'. Top: Spectrogram; Lower three panels: Frame-wise combination weights obtained from the methods InvEnt, AEnc and ROPE.

To gain insight about the differences of the three weight estimation approaches, the output of the models can be visually inspected. Fig. 5 shows one example of the combination weights for an utterance that was reverberated (condition 'r1f'). The best average strategy is to select the corresponding stream 'r1f' (as shown in Table 1), which is represented by the third row in the stream selection output. While all three models exhibit the highest activations for the best class, ROPE provides consistently higher and far less noisy estimates (also on short time scales) than the other two approaches, which reflects the low WER obtained with this method.

6. CONCLUSIONS

This paper presented a new method to determine stream weights for combination of DNN posterior probabilities in a multi-stream DNN/HMM framework to improve ASR robustness in various reverberant environments. This approach resulted in consistent improvements in known and unknown scenarios, outperforming inverse entropy and autoencoders for stream weighting or selection. Stable results were obtained independently of the specific combination strategy (weighting or winner-takes-all), and the temporal context (frame-wise vs. utterance-level), indicating that the method is applicable in real-time ASR. Further comparisons to word-level combination schemes such as ROVER [7] as well as to other methods exploring temporal information to determine stream weights e.g. based on mean temporal distance [26] will be conducted in future work.

7. ACKNOWLEDGEMENTS

This work was funded by the DFG (Cluster of Excellence 1077/1 "Hearing4All" and the SFB/TRR 31 "The Active Auditory System") as well as the BMBF and the EC (project KNOTS, grant no. AAL-2013-6-144). Additional funding was provided via a Google faculty award to Hynek Hermansky. The authors thank Hynek Hermansky and Sri Harish Mallidi for valuable discussions.

8. 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] 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.

[8] 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.

[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] 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.

[12] 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.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation," *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1: Foundations. MIT Press, 1986.

[14] 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.

[15] 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.

[16] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "On using MLP Features in LVCSR," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, Jeju Island, Korea, Oct. 2004.

[17] 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.

[18] 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.

[19] H. Kuttruff, *Room Acoustics*, Spon Press, London, 4th edition, 2000.

[20] 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.

[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] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling using Deep Belief Networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[23] 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.

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

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi Speech Recognition Toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Big Island, HI, USA, Jul. 2011.

[26] 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.