INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Parallel Neural Network Features for Improved Tandem Acoustic Modeling

Zoltán Tüske, Wilfried Michel, Ralf Schlüter, Hermann Ney

Human Language Technology and Pattern Recognition, Computer Science Department,
RWTH Aachen University, 52056 Aachen, Germany
{tuske,michel,schlueter,ney}@cs.rwth-aachen.de

Abstract

The combination of acoustic models or features is a standard approach to exploit various knowledge sources. This paper investigates the concatenation of different bottleneck (BN) neural network (NN) outputs for tandem acoustic modeling. Thus, combination of NN features is performed via Gaussian mixture models (GMM). Complementarity between the NN feature representations is attained by using various network topologies: LSTM recurrent, feed-forward, and hierarchical, as well as different non-linearities: hyperbolic tangent, sigmoid, and rectified linear units. Speech recognition experiments are carried out on various tasks: telephone conversations, Skype calls, as well as broadcast news and conversations. Results indicate that LSTM based tandem approach is still competitive, and such tandem model can challenge comparable hybrid systems. The traditional steps of tandem modeling, speaker adaptive and sequence discriminative GMM training, improve the tandem results further. Furthermore, these "old-fashioned" steps remain applicable after the concatenation of multiple neural network feature streams. Exploiting the parallel processing of input feature streams, it is shown that 2-5% relative improvement could be achieved over the single best BN feature set. Finally, we also report results after neural network based language model rescoring and examine the system combination possibilities using such complex tandem models.

Index Terms: Speech recognition, feature combination, tandem, neural network, recurrent, feed-forward, LSTM, MLP, GMM, ASR

1. Introduction and Related Work

In practice, different knowledge sources are available for automatic speech recognition (ASR) which usually are complementary. Thus, combination can lead to improvements in performance. In ASR, ensemble combination can be carried out on multiple levels: on feature level, e.g. by simple feature concatenation [1, 2, 3]; on model level, e.g. via linear or log-linear combination of acoustic and/or language model scores [4, 2, 5, 6]; or on system level, e.g. by recognizer output voting error reduction (ROVER) [7] or confusion network combination (CNC) [8]. The recent advances in neural network (NN) based deep learning techniques led to large variations in models which represent similar learning capacities. Recently, several investigations have demonstrated that new state-of-the-art recognition results can be achieved by simple score fusion of multiple deep models [9, 10, 11]. The way to introduce diversity in the ensemble and how to combine them to improve the classification accuracy has been an active research field for several decades, e.g. [12, 13, 14, 15, 16, 17].

In the hybrid approach, neural networks are used to model tied-triphone HMM state posterior probabilities. The hybrid approach attracted much attention and recently evolved into a de facto standard [18]. However, with the tandem approach an equally powerful method to introduce neural networks into acoustic modeling (AM) still coexists. Though being introduced later than the hybrid approach, the tandem approach was the first to lead to considerable improvements over the former state-of-the-art Gaussian mixture HMM approach using neural networks. In the tandem approach [19, 20], neural networks trained on phonetic targets are used as input to Gaussian mixture HMMs. Exploiting its complementarity, the tandem modeling technique can be used in score combination with the hybrid approach [21]. The tandem approach also allows the use of the well established speaker adaptation frame-work [22, 23]. Furthermore, in recent keyword search evaluations (the task is only implicitly related to minimizing word error rate) the tandem models usually resulted in better performance than the hybrid ones [24]. In a previous study it also has been shown, that Gaussian models can be seen as a generalized softmax output layer, thus allowing for joint end-to-end optimization of NN and GMM.

Simple score fusion of hybrid and/or tandem acoustic models have been investigated in the literature already, e.g. in [21, 9, 10, 11]. In a recent evaluation, we observed that bottleneck (BN) representations of two feed-forward NNs could successfully be combined within the tandem approach [25]. Thus, in this paper we extend our investigation and aim at incorporating various BN feature extractors into the GMM by simple concatenation. Integrating the GMM into the hybrid framework, our approach can also be seen as a late fusion of the hidden representation of neural networks in contrast to early fusion e.g. of [26]. Diversity of the BN representations is achieved by various signal processing techniques, non-linearities, hierarchical structures, as well as feed-forward and recurrent neural network topologies. The study also re-investigates the tandem approach with state-of-the-art, deep, bi-directional long short-term memory (LSTM) neural networks [27, 28, 29]. Feature space speaker adaptation on top of the feature concatenation is also considered which can also be interpreted as speaker dependent fusion of the BN features. Furthermore, comparative investigations on the effect of NN language model (NNLM) based lattice rescoring and system combination are provided for better comprehension of the proposed BN combination method. Based on [30], this work can also be considered as an initial step towards a complex speaker adaptively trained neural network based acoustic model.

2. Speech Recognition Tasks

In order to carry out our research we trained and tested our acoustic models on the following corpora. From the Quaero project we chose the German data set, which we still actively use in the International Workshop on Spoken Language Translation (IWSLT) evaluations [31, 25]. Besides broadcast news and European parliamentary sessions, the training corpus also contains talk shows and interviews which also include spontaneous speech. The 150-hour training corpus was recorded at 16 kHz sampling rate. We report word error rates (WER) on the 3-hour

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-1747

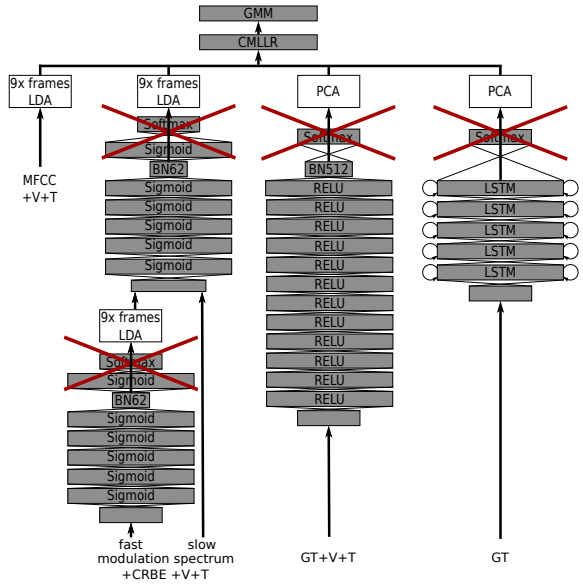

Figure 1: *Parallel bottleneck features in tandem with speaker adapted Gaussian mixture model.*

development and evaluation set of the Quaero evaluation 2012. Moreover, recognition performance is also measured on the 3.5-hour development and evaluation set of the German ASR track of IWSLT 2016. These Microsoft Speech Language Translation (MSLT) test data contain bilingual Skype calls. Our second set of experiments were carried out on the narrow band telephone conversation task of Switchboard which needed some adjustment of the feature extraction pipeline. The acoustic models were trained on the 300-hour SWB training corpus, the lexicon size was limited to 30k words, and the language model was trained only on the transcription of the acoustic training data and the Fisher corpora. The developed models were optimized on Hub5'00, and evaluated on Hub5e'01, RT'03s. For further details we also refer to [32, 25] and [30]. In both tasks, 10% of the training data was selected for cross-validation purpose.

3. Acoustic Modeling and Features

3.1. Cepstral Features

Depending on the recognition task, we extracted 16-dimensional Mel-frequency cepstral coefficients (MFCC) or 15-dimensional Gammatone (GT [33]) features. The features were segmentwise mean-and-variance normalized and also appended with voicedness (V) [2], and tone features (T).

3.2. NN Features

In this study we experimented with three neural networks presented below, and also shown in Fig. 1. The 12-layer feed-forward ReLU (ReLU-FF) and the bidirectional recurrent LSTM (LSTM-RNN) networks were also used as hybrid model for comparison. The NNs were trained on the same Viterbi alignment as the mixture models using cross-entropy criterion. On the German task, the feed-forward networks were initialized multilingually using four Quaero languages (Polish, German, English, French) and then fine-tuned to German.

3.2.1. Hierarchical Feed-Forward MRASTA NN (HMRAS-FF)

Similar to [34, 35], a deep hierarchical BN feature extractor was trained on a concatenation of modulation spectrum of three different critical band energy (CRBE) streams. The input to the networks was also augmented with the central CRBE frame and 9 frames of voicedness (V) and tone (T) features. The three CRBE streams were extracted from GT [33], PLP [36], and MFCC pipelines (20-dimensional for Quaero, 15-dimensional for Switchboard data). The hierarchy was built from two 6-layer networks, and additional sigmoid-BN layers were inserted before the last hidden layer at each level. The BN layers contain 62 nodes and the other hidden layers have 2000 sigmoidal units. This model always estimated 1500 tied-triphone classes, and the hierarchy was also jointly optimized. Unless otherwise stated, in the tandem experiments the LDA transformed BN output of the hierarchy was always concatenated with the cepstral features of Sec. 3.1.

3.2.2. Feed-Forward NN with Rectified Linear Units (ReLU-FF)

To train this model, high-resolutional (50-dimensional for Quaero, 40-dimensional for Switchboard) GT cepstral features were also extracted. Here, we used a square DCT transformation matrix similar to [30]. The neural network was trained on 17-frame context of GT voicedness and tone features. Its 12 hidden layers contain 2000 nodes each and use rectified linear unit (ReLU) non-linearities [37]. The last layer was low-rank factorized by a 512-dimensional linear bottleneck [38]. This linear BN output was used in the tandem experiments after PCA transformation. Using the same target as the GMM, during the training $l_2 = 0.05$ regularization and classical momentum was applied.

3.2.3. LSTM Recurrent MLP (LSTM-RNN)

LSTM recurrent neural networks have resulted in significant gains and achieve state-of-the-art results in many tasks [29, 39, 40, 41]. Our bi-directional recurrent network consists of five LSTM layers. Each layer has 500 and 600 nodes in the English and German tasks, respectively, for each direction. Training was performed by Adam optimization with incorporated Nesterov momentum, using the RETURNN toolkit [42]. We used an l_2 normalization parameter of 0.01, and employed a dropout rate of $p = 0.05$ on the outputs of the hidden nodes. The rest of the training setup is the same as for ReLU-FF. The tandem model was trained on the PCA transformed output of the (1000 or 1200-dimensional) final hidden layer.

Table 1: *Speaker independent word error rate comparison of various acoustic models using frame-wise training criteria.*

Modeling	NN type	Training criterion	German				English	
			Quaero		IWSLT		Hub5'00	
			Dev	Eval	Dev	Eval	CH	SWB
Hybrid	ReLU-FF	CE	14.6	16.9	27.4	25.0	25.1	12.3
	LSTM-RNN		**13.9**	**15.3**	**24.7**	**22.7**	21.0	**10.8**
Tandem	HMRAS-FF	ML	14.4	16.2	27.6	24.8	26.2	13.3
	ReLU-FF		14.9	17.2	28.9	26.1	25.3	12.6
	LSTM-RNN		14.2	15.4	25.3	22.9	**20.7**	10.9

3.3. Gaussian Mixture HMM

The acoustic models were based on the standard 3-state left-to-right Hidden Markov Model (HMM) topology. The German and Switchboard systems used 4500 and 9000 generalized tied-triphone states, respectively. The acoustic models were trained on the Viterbi alignment generated by the previous systems: [43, 30]. Speaker adaptive training was carried out using the constrained version of the MLLR transformation (CMLLR) [44]. The CMLLR matrices for the test data were estimated unsupervised on the speaker independent recognition output. The final speaker adaptive models were also enhanced by minimum phone error (MPE) discriminative training [45].

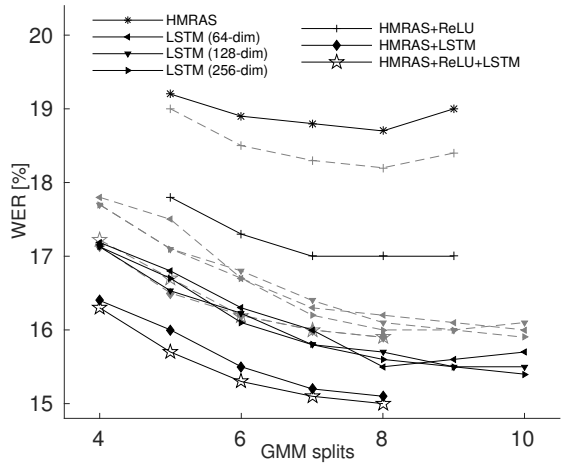

Figure 2: *Effect of GMM splitting and feature concatenation. Word error rate (WER [%]) measured on Hub'00 (SWB+CH). Gray dashed and black solid lines indicate speaker independent and speaker adaptive modeling, respectively.*

4. Language Modeling

For the German recognition tasks, 5-gram Kneser-Ney smoothed language models were trained. On the broadcast news domain, the vocabulary consisted of a mixture of 300k full words and word fragments [31], and the language model contained about 180M n-grams. The LM developed for the IWSLT task modeled the distribution of 377k different full words. For the IWSLT evaluation, we also trained an LSTM language model on a subset of the LM data [25]. The embedding layer mapped the input into a 300-dimensional space, the two LSTM layers contained 300 nodes. Due to the large vocabulary size, 200 word-classes were used to speed up the training. For Switchboard, a single-layer LSTM network was trained. The embedding and the hidden layer size was set to 1000, and word class approximation was not used. In addition, a 20-gram ReLU feed-forward network was also trained similar to [46]. The embedding had a size of 128 and the network contained four 1024-dimensional hidden layers which were low-rank factorized by 256-dimensional linear bottlenecks. During recognition, the lattices were rescored using the `rwthlm` [47] toolkit.

5. Experimental Results

5.1. Speaker Independent Baseline Results

In the first set of experiments we trained speaker independent tandem models on each of the aforementioned MLP features, and compared their performance to the hybrid approach in Table 1. Obviously, the LSTM network performed best. On the German task, the multilingual initialization of the ReLU-FF network accounts for about 5% rel. improvement of Quaero and IWSLT results. As can be seen, the tandem approach lags slightly behind the hybrid one, except for the CallHome test. The 0-2% relative performance gap between them might be bridged by frame-wise discriminative training of the Gaussian models, as demonstrated in [30]. The two feed-forward BN features showed mixed results. Whereas ReLU-FF performed better on English telephone speech, the HMRAS-FF clearly outperformed it on the German Quaero and IWSLT tasks. MPE training of the best hybrid models resulted in 14.9% and 22.7% WER on the eval sets of Quaero and IWSLT tasks. On Hub'00 the sequence discriminatively trained LSTM model achieved 10.5% and 20.5% WER on the SWB and CHM subsets.

Table 2: *WER comparison of single and combined tandem models before and after ML-SAT training.*

NN type			SAT	German				English	
HMRAS FF	ReLU FF	LSTM RNN		Quaero		IWSLT		Hub5'00	
				Dev	Eval	Dev	Eval	CH	SWB
×				14.4	16.2	27.6	24.8	26.2	13.3
			×	13.8	15.8	25.6	23.0	24.7	12.7
	×			14.9	17.2	28.9	26.2	25.3	12.6
			×	14.4	16.5	26.8	24.4	23.6	12.0
		×		14.2	15.4	25.3	22.9	20.7	10.9
			×	13.9	15.3	24.3	22.4	19.9	10.9
×	×			14.2	16.0	26.6	24.3	24.5	12.8
			×	13.6	15.5	25.1	22.7	22.9	12.1
×	×	×		13.5	14.8	24.1	21.8	20.8	10.7
			×	**12.9**	**14.5**	**23.5**	**21.1**	**19.7**	**10.3**

5.2. Speaker Adaptive Training and Feature Combination

In the second set of experiments we concatenated the MLP features and also switched to speaker adaptive Gaussian model training. The results are summarized in Table 2. We observed that in general the ML-SAT tandem models are on par with the speaker independent hybrid models. Speaker independent results show that concatenation of the two feed-forward BNs resulted in 2% relative improvement over the best single one. On Hub5'00, we overall observed improvement which concentrated only on the CallHome part (3% relative improvement). We measured a similar range of improvements after speaker adaptive training. Providing also the LSTM features for the concatenation (HMRAS+ReLU+LSTM), further gains were observed. 5-7% relative improvement was observed over the single best (LSTM) features after speaker adaptive training on the German tasks. On the Hub5'00 we measured on average 2% relative improvement. Concatenation of only the HMRAS-FF and LSTM-RNN resulted in similar but slightly less improvements, see Fig. 2. We also carried out an investigation for the optimal size of the various feature streams, results are presented in Table 4. In these experiments the transformation of the cepstral features (GT+V+T) and BN part of HMRAS-FF was optimized separately. Experiments revealed that larger feature space might result in better speaker independent results, but speaker adaptation is more sensitive to the input feature size. Optimizing the mixture splitting showed that usually the larger models are the better independent of the feature space dimension, cf. Fig. 2.

5.3. Effect of NNLM Lattice Rescoring

Next, we carried out lattice rescoring experiments using the MPE trained acoustic models. The rescored results also include confusion network (CN) decoding [48]. After the MPE training of the German models, HMRAS+ReLU concatenation achieved 23.4%, and 20.9% WER on the IWSLT dev and eval sets, whereas HMRAS+ReLU+LSTM resulted in 22.6% and 20.5% WER (2-3% relative better). After NNLM rescoring and CN decoding, the performance difference between the two combinations were still 1-3% relative. Overall, the MRAS+ReLU+LSTM resulted on the dev and eval set of the IWSLT task in 21.4% and 19.1% WER. On the Quaero dev and eval sets, the best combined features (MRAS+ReLU+LSTM) achieved 12.5% and 14.0% after SAT+MPE. Compared to the acoustic model developed in [31], it is 2.5% absolute better using the same test conditions. Experimenting with the concatenated BN features, we also observed that the improved feature space decreased the effect of sequence-discriminative training on the tandem model.

The rescoring experiments on Switchboard are presented in Table 3. It can be observed that the effect of NNLM rescoring and

Table 3: *NNLM rescoring and CN decoding of single and parallel BN tandem systems. WER measured after SAT+MPE on standard English telephone conversation tasks.*

AM				NNLM			Hub5'00		Hub5e'01			RT'03s	
HMRAS FF	ReLU FF	LSTM RNN	multi-ling.	ReLU FF	LSTM RNN	multi-pass	CH	SWB	SWB	SWB2 p3	SWB Cell	SWB	FSH
		×					19.1	10.5	11.4	14.4	18.8	19.6	12.7
				×			17.7	9.2	10.1	12.8	17.1	17.9	11.6
					×		17.2	9.2	10.2	12.6	16.9	17.5	11.1
				×	×		17.2	9.0	9.9	12.6	16.9	17.6	11.1
×		×					18.8	10.0	11.1	13.6	18.0	18.4	12.1
					×		17.2	8.9	10.0	12.2	16.4	17.0	11.1
×	×	×					18.7	10.2	11.1	13.6	18.1	18.7	12.1
				×			17.5	8.8	9.7	12.3	16.4	17.1	11.1
					×		17.3	8.7	9.8	12.3	16.3	16.9	10.8
				×	×		17.1	**8.5**	9.6	11.9	16.1	16.8	10.8
				×	×	×	16.9	**8.5**	**9.4**	11.9	16.1	16.6	10.6
×	×	×	×	×	×	×	**16.0**	8.7	9.5	**11.7**	**15.7**	**15.8**	**10.1**

Table 4: *Effect of the LDA/PCA size. Results measured on the English Hub5'00, after SAT using ML tandem models.*

LDA size		PCA size		Hub5'00	
GT+V+T	HMRAS FF	ReLU FF	LSTM RNN	CH	SWB
-	-	-	64	20.2	**10.8**
			128	20.1	10.9
			256	**19.9**	10.9
30	45	-	100	**19.7**	**10.5**
40	55		80	20.0	10.4
	55		100	**19.7**	10.7
45	64		64	20.4	10.4
			128	19.6	10.8
20	30	40	110	19.7	**10.3**
		30	100	**19.6**	10.4
			80	19.7	**10.3**
	40	40	80	19.9	10.4
30	40	40	80	19.8	10.5
40	50	55	55	20.5	10.6
45	64	64	64	20.4	10.7

feature combination is not additive. E.g. rescoring improved the WER on RT'03s more than 10% relative using the LSTM tandem acoustic model. However, rescoring resulted in only 8% relative improvement using the combined HMRAS+ReLU+LSTM features. Overall, on the evaluation sets (Hub5e'01 and RT03s) we measured 3-5% relative WER reduction even after rescoring with the combined feed-forward ReLU and LSTM-RNN LMs. We also note that the performance difference between feed-forward and recurrent NNLM also reduces with stronger AM. In order to improve the CMLLR estimation multi-pass rescoring was also investigated with HMRAS+ReLU+LSTM system. Applying rescoring on the speaker independent recognition output we indeed could improve the adaptation and obtained 0.5% lower WER on Hub5'00 after the second decoding pass. However, as can be seen in Table 3, in the end the second rescoring step showed only small improvement over the single pass rescoring approach. An additional experiment was carried out where the feed-forward networks (HMRAS-FF and ReLU-FF) were multilingually initialized on about 1800 hours of speech of 28 languages and fine-tuned on the 300-hour SWB corpus. The detail of the language resources can be found in [35, 24, 49]. On average additional 2-5% relative improvement was measured on all three test sets (Hub5'00, Hub5e'01, RT'03s) but not on all subsets. For comparison with results of other groups we kindly refer the reader to e.g. [10, 50, 39, 51, 11].

5.4. System combination experiment

In the final experiment carried out on the German IWSLT task, we tested the improved tandem system in combination with other systems developed for the IWSLT'16 evaluation campaign. In [25], four systems were developed using two different acoustic models (hybrid, tandem) and two language models (full-word, hybrid sub-word). As can be seen in Table 5, the feature combined tandem alone is already better than the CN combined submission. Adding this new tandem system to the CN combination, we still observed 1% absolute gain on the evaluation set which indicates strong complementarity between the systems.

Table 5: *Using improved tandem system in confusion network system combination. WER measured on the German IWSLT sets.*

	#Systems	dev	eval
HMRAS+ReLU+LSTM	1	21.4	19.1
IWSLT'16 submission [25]	4	21.6	19.6
Combination	5	**20.3**	**18.1**

6. Conclusions

Similar to other combination techniques, it was demonstrated that efficient combination of complementary knowledge sources is possible also within the tandem framework. Combination through concatenation of several neural networks lead to consistent improvement over the best single system. Our approach also demonstrated that the speaker adaptive transformation and fusion of the multiple NN features is also possible and further improves the results. The combined system outperformed the best single system even after strong neural network language model rescoring, by 2-5% relative. Having possibly heterogeneous networks available, the proposed method is especially attractive due to the quite efficient training of the subsequent GMM, e.g. when developing an ASR system for a new language using existing NN features from other tasks.

Besides score fusion with hybrid acoustic models, in the future we plan to investigate data augmentation, I-vectors, and more diverse neural networks which might improve our tandem and hybrid results further. We will also carry out joint training of the hierarchy similar to [30].

7. Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 694537). The work reflects only the authors' views and the European Research Council Executive Agency is not responsible for any use that may be made of the information it contains. The authors would like to thank Albert Zeyer and Kazuki Irie for training the LSTM acoustic and language models, respectively, on Switchboard data.

8. References

[1] R. Haeb-Umbach and M. Loog, "An investigation of cepstral parameterisations for large vocabulary speech recognition," in *Eurospeech*, 1999, pp. 1323–1326.

[2] A. Zolnay *et al.*, "Using multiple acoustic feature sets for speech recognition," *Speech Communication*, vol. 49, no. 6, pp. 514–525, Jun. 2007.

[3] C. Plahl *et al.*, "Feature combination and stacking of recurrent and non-recurrent neural networks for LVCSR," in *ICASSP*, 2013, pp. 6714–6718.

[4] P. Beyerlein, "Discriminative model combination," in *ASRU*, 1997, pp. 238–245.

[5] B. Hoffmeister *et al.*, "Log-linear model combination with word-dependent scaling factors," in *Interspeech*, 2009, pp. 248–251.

[6] J. Yang *et al.*, "System combination with log-linear models," in *ICASSP*, 2016, pp. 5675–5679.

[7] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *ASRU*, 1997, pp. 347–354.

[8] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *NIST Speech Transcription Workshop*, 2000.

[9] T. Alumäe *et al.*, "The 2016 BBN Georgian telephone speech keyword spotting system," in *ICASSP*, 2017, pp. 5755–5759.

[10] W. Xiong *et al.*, "The Microsoft 2016 conversational speech recognition system," in *ICASSP*, 2017, pp. 5255–5259.

[11] G. Saon *et al.*, "The IBM 2016 English conversational telephone speech recognition system," in *Interspeech*, 2016, pp. 7–11.

[12] D. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[13] L. Xu *et al.*, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 22, no. 3, pp. 418–435, 1992.

[14] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *ICSLP*, vol. 1, 1996, pp. 426–429.

[15] J. Kittler *et al.*, "On combining classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[16] H. Misra *et al.*, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *ICASSP*, vol. 2, 2003, pp. 741–744.

[17] L. Deng and J. C. Platt, "Ensemble deep learning for speech recognition," in *Interspeech*, 2014, pp. 1915–1919.

[18] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Norwell, MA, USA: Kluwer Academic Publishers, 1993.

[19] H. Hermansky *et al.*, "Tandem connectionist feature extraction for conventional hmm systems," in *ICASSP*, 2000, pp. 1635–1638.

[20] F. Grézl *et al.*, "Probabilistic and bottle-neck features for LVCSR of meetings," in *ICASSP*, 2007, pp. 757–760.

[21] H. Wang *et al.*, "Joint decoding of tandem and hybrid systems for improved keyword spotting on low resource languages," in *Interspeech*, 2015, pp. 3660–3664.

[22] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[23] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[24] P. Golik *et al.*, "Multilingual features based keyword search for very low-resource languages," in *Interspeech*, 2015, pp. 1260–1264.

[25] W. Michel *et al.*, "The RWTH Aachen LVCSR system for IWSLT-2016 German Skype conversation recognition task," in *IWSLT*, 2016.

[26] H. Soltau *et al.*, "Joint training of convolutional and non-convolutional neural networks," in *ICASSP*, 2014, pp. 5572–5576.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[29] A. Graves *et al.*, "Hybrid speech recognition with deep bidirectional LSTM," in *ASRU*, December 2013, pp. 273–278.

[30] Z. Tüske *et al.*, "Speaker adaptive joint training of Gaussian mixture models and bottleneck features," in *ASRU*, 2015, pp. 596–603.

[31] M. A. B. Shaik *et al.*, "The RWTH Aachen German and English LVCSR systems for IWSLT-2013," in *IWSLT*, 2013, pp. 120–127.

[32] M. Nußbaum-Thom *et al.*, "The RWTH 2009 Quaero ASR evaluation system for English and German," in *Interspeech*, 2010, pp. 1517–1520.

[33] R. Schlüter *et al.*, "Gammatone features and feature combination for large vocabulary speech recognition," in *ICASSP*, 2007, pp. 649–652.

[34] F. Valente and H. Hermansky, "Hierarchical and parallel processing of modulation spectrum for ASR applications," in *ICASSP*, 2008, pp. 4165–4168.

[35] Z. Tüske *et al.*, "Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages," in *Interspeech*, Sep. 2014, pp. 1420–1424.

[36] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[37] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *the 27th Int. Conf. on Machine Learning*, 2010, pp. 807–814.

[38] T. N. Sainath *et al.*, "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," in *ICASSP*, 2013, pp. 6655–6659.

[39] D. Povey *et al.*, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech*, 2016, pp. 2751–2755.

[40] A. Zeyer *et al.*, "A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition," in *ICASSP*, 2017, pp. 2462–2466.

[41] H. Sak *et al.*, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Interspeech*, 2014, pp. 338–342.

[42] P. Doetsch *et al.*, "RETURNN: The RWTH extensible training framework for universal recurrent neural networks," in *ICASSP*, 2017, pp. 5345–5349.

[43] Z. Tüske *et al.*, "Multilingual hierarchical MRASTA features for ASR," in *Interspeech*, 2013, pp. 2222–2226.

[44] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[45] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *ICASSP*, 2002, pp. I–105–I–108.

[46] Z. Tüske *et al.*, "Investigation on log-linear interpolation of multi-domain neural network language model," in *ICASSP*, 2016, pp. 6005–6009.

[47] M. Sundermeyer *et al.*, "`rwthlm` – The RWTH Aachen University neural network language modeling toolkit," in *Interspeech*, 2014, pp. 2093–2097.

[48] L. Mangu *et al.*, "Finding consensus among words: Lattice-based word error minimization," in *Eurospeech*, 1999, pp. 495–498.

[49] P. Golik *et al.*, "The 2016 RWTH keyword search system for low-resource languages," in *Speech and Computer*, accepted for publication, 2017.

[50] I. Medennikov *et al.*, "Improving english conversational telephone speech recognition," in *Interspeech*, 2016, pp. 2–6.

[51] K. Veselý *et al.*, "Sequence-discriminative training of deep neural networks," in *Interspeech*, 2013, pp. 2345–2349.