FEATURE COMBINATION AND STACKING OF RECURRENT AND NON-RECURRENT NEURAL NETWORKS FOR LVCSR

Christian Plahl[1], *Michael Kozielski*[1], *Ralf Schlüter*[1] *and Hermann Ney*[1,2]

[1]Human Language Technology and Pattern Recognition,
Computer Science Department, RWTH Aachen University, Aachen, Germany
[2]Spoken Language Processing Group, LIMSI CNRS, Paris, France
{plahl,schlueter,ney}@cs.rwth-aachen.de

ABSTRACT

This paper investigates the combination of different short-term features and the combination of recurrent and non-recurrent neural networks (NNs) on a Spanish speech recognition task. Several methods exist to combine different feature sets such as concatenation or linear discriminant analysis (LDA). Even though all these techniques achieve reasonable improvements, feature combination by multi-layer perceptrons (MLPs) outperforms all known approaches. We develop the concept of MLP based feature combination further using recurrent neural networks (RNNs). The phoneme posterior estimates derived from an RNN lead to a significant improvement over the result of the MLPs and achieve a 5% relative better word error rate (WER) with much less parameters.

Moreover, we improve the system performance further by combining an MLP and an RNN in a hierarchical framework. The MLP benefits from the preprocessing of the RNN. All NNs are trained on phonemes. Nevertheless, the same concepts could be applied using context-dependent states.

In addition to the improvements in recognition performance w.r.t. WER, NN based feature combination methods reduce both, the training and the testing complexity. Overall, the systems are based on a single set of acoustic models, together with the training of different NNs.

***Index Terms*—** feature combination, multi-layer perceptron, recurrent neural networks, long-short-term-memory, speech recognition

1. INTRODUCTION

In recent years a large number of different acoustic features have been developed in the area of speech recognition. In order to benefit from these different acoustic features, lattice or N-best-list system combination methods [1] have been shown to be the most promising approach for years [2]. Other feature combinations like concatenation of the features or linear discriminant analysis (LDA) are suboptimal [3, 4, 5]. The best system combination performance is achieved, when several complementary subsystems are combined, resulting in high computational costs to train all subsystems.

The high computational costs are reduced, when the different acoustic features are combined by a neural network (NN) [5]. The systems trained on multi-layer perceptron (MLP) based posterior estimates outperform all other feature combination methods, and achieve even better recognition results w.r.t. the WER as system combination of the individual subsystems [5]. In this paper, we develop the NN based feature combination approach further, using recurrent neural networks (RNNs) especially the long-short-term-memory (LSTM) concept [6]. LSTMs have not yet been trained on a large amount of data for large vocabulary continuous speech recognition (LVCSR) tasks. We will show that the best performance is achieved when the recurrent and non-recurrent networks are combined in a hierarchical framework.

Probabilistic features derived from a NN have recently become a major component of current state-of-the-art recognition systems for speech as well as for image and handwriting recognition [7, 8, 9, 10]. Whereas in speech recognition the *tandem* approach [11] has been the only method to include NN based features in the Gaussian Hidden Markov Model (GHMM) framework and to improve the GHMM baseline at the same time, the hybrid approach [12] becomes competitive, when the network is trained on context-dependent HMM states [13, 14, 15] in combination with deep neural networks. All experiments in this paper are conducted using the tandem approach.

2. NEURAL NETWORK TOPOLOGIES

2.1. Recurrent Neural Networks

In this paper we investigate recurrent neural networks (RNNs) to combine several sets of short-term features. RNNs are similar to feed-forward networks, e.g. MLPs, but consist of a backward directed loop. The output of a previous time step is looped back and used as additional input. Therefore, contextual information do not have to be encoded explicitly into the feature vector any more. This general structure of an RNN is shown in Figure 1, where the network is unfolded in time. In speech recognition, these RNNs have been used for the first time in [16] for phoneme modeling. Nevertheless, the RNNs as well as other network topologies have been outperformed by the concept of HMMs. Nowadays, RNNs have become interesting again for speech recognition [17, 18]. We use an extension of the RNNs, the long-short-term-memory (LSTM) concept. Previously, the LSTMs have been applied to small conversational speech recognition tasks [17], but not to LVCSR tasks.

The training of an RNN is performed using the back propagation through time (BPTT) algorithm, which is an extension of the conventional back propagation training algorithm.

2.1.1. Bi-directional Recurrent Neural Networks

Whereas all RNNs have access to the full past history, the access to future frames is limited. Future context can only be included in the network by delaying the output or encoding the future frames in the feature vector, resulting in better recognition performance [18]. Instead, we train a forward and a backward directed RNN to provide all past and all future frames to the RNN. The forward direct network scans the input sequence in normal order, whereas the backward directed RNN processes the input sequence in opposite direction.

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

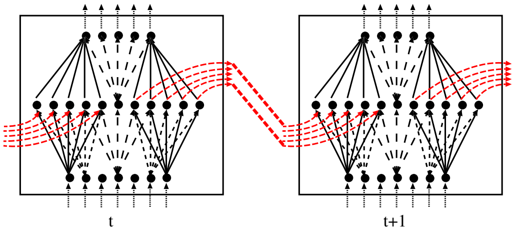

Fig. 1. Structure of a recurrent neural network unfolded in time. The recurrent connections are the dashed connections, marked in red, going from time step t to $t + 1$.

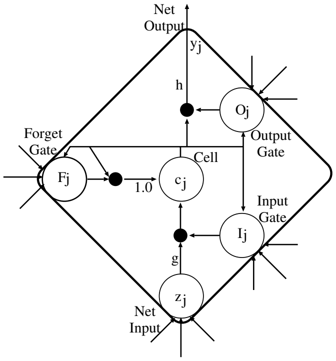

Fig. 2. LSTM unit. The inner cell c_j is controlled by different gating units: forget gate (F_j), input gate (I_j) and output gate (O_j). The input of the LSTM cell contains feed-forward as well as recurrent connections.

The final output layer combines the forward and backward directed RNNs and therefore makes use of the whole input sequence. Due to the limited capacity to model long-term dependencies in classical RNNs, bi-directional RNNs (BRNNs) [19] do not perform much better than RNNs with a delayed output.

2.1.2. *Long-Short-Term-Memory*

The main disadvantage of the concept of RNNs is the vanishing gradient problem, which has been analyzed in detail in [6]. When the error of the network is back propagated trough the time, the error blows up or decays exponentially over time. In order to avoid this effect, the unit has been re-designed resulting in the LSTM concept [6]. As shown in Figure 2 the core of a LSTM unit is controlled by several gating units. While the input and the forget gate influences the input and the output respectively, the forget gate controls the cell state.

Compared to classical RNNs, the LSTM-RNNs are able to learn temporal sequences of 1000 time steps or more [6]. This ability to model large long-temporal dependencies is sufficient to cope with the temporal dependencies in speech. The concepts of LSTM-RNNs have been successfully applied to text and handwriting recognition [9] as well as to acoustic modeling [17, 20] and language modeling [21]. Nevertheless, LSTM-RNNs have not yet been used for acoustic modeling in LVCSR systems when a large amount of data is available.

2.2. Hierarchical Neural Networks

In a hierarchical framework, several NNs are stacked together. Each NN is trained on the output of a previous NN [22, 23]. In addition to the NN based features from the previous network, other features can be provided as well. The temporal context of each network in the hierarchy can be selected independently. Each network in the hierarchical framework is a feature detector, providing features representing localized detectors at the start and global feature detectors at the end of the hierarchy. Presenting less significant features in a later stage of the hierarchy can improve the overall system performance [23].

The main motivation for creating a hierarchy of recurrent and non-recurrent networks is the fact that RNNs provide good features, but the training of RNNs is very time consuming, especially the training on context-dependent states. In our experiments the training time of the LSTM-RNNs is 4 times larger than the MLP training. Using the RNN as a preprocessing step to provide features, the information encoded in the RNN can be efficiently used, e.g. by MLPs.

3. NEURAL NETWORK FEATURE COMBINATION

In current speech recognition systems, NN based probabilistic features are important to obtain the best performance. Therefore, optimizing the NN based probabilistic features has been one of the main research areas in the last years. The type of the input features has been under investigation as well as the best topology or structure of a NN. As an alternative to the short-term features [24], [25] introduce features based on long temporal context. These features contain a temporal context of up to one second and provide complementary information [8, 25]. As shown in [26], the hierarchical bottle-neck structure seems to be a very good NN topology for the tandem approach. The hierarchical bottle-neck features combine the advantages of the bottle-neck approach [27] and the hierarchical framework [22].

The concept of NN based feature combination used in this paper is simple. The different short-term feature streams are first combined and the super feature vector is used as input for the NN training. During the training, the NN selects the most relevant information out of the features to discriminate the phoneme classes. Even though the best results are obtained using the bottle-neck concept [27], we keep the network as simple as possible. We have trained networks with just one hidden layer based on phoneme classes. Without any loss of generality, the same concept could be used to train on context-dependent states or bottle-neck features, where similar results are expected.

Due to the non-linear output activation of the NNs, the feature transformation includes non-linear parts. As we have shown in [5] this non-linearity is important to overcome the limitation of the LDA approach.

3.1. Input Features

The different recurrent and non-recurrent networks are trained on MFCCs, PLPs, or Gammatone (GT) features [28]. The features are augmented with first order temporal derivatives and the second order temporal derivative of the first dimension, resulting in a 33 dimensional feature vector for MFCCs and PLPs and 31 components for GT features. The final feature streams are globally normalized by mean and variance. In order to simplify the feature extraction, additional transformations are skipped. In the hierarchical framework, the non-recurrent NN is trained on the posterior estimates of the RNN augmented by the RNN input features.

While we have extended the feature vector for the training of non-recurrent networks by a temporal context of ±4 frames, past and future frames are given by the recurrent bi-directional structure of the LSTMs. Depending on the number of features combined, the LSTM-RNN is trained on a 33, 66 or 97 dimensional feature vector. The input dimension for the classical NNs varies from $33 \times 9 = 297$ (single feature set) up to 1170 (all feature sets).

3.2. Training

The trainings of all networks are performed using NNs with just one hidden layer. The BRNNs are based on the LSTM structure with a hidden layer of size 200. The non-recurrent NN is an MLP consisting of 4000 units in the hidden layer. Both, the LSTM-RNN and the MLPs are trained on the 33 phoneme classes of the Spanish data set. Depending on the number of feature streams combined, the number of parameters learned during the NN training varies from 400k to 500k and from 300k to 6M for LSTM-RNN and MLPs respectively.

During the training of the NNs, the learning rate η is adjusted corresponding to the frame classification performance on a cross-validation set. A momentum term is included in the weight update rule to avoid large changes.

The final phoneme posterior estimates derived from the NNs are transformed by logarithm. Within a sliding window of size 9, the 33 dimensional log posterior estimates are transformed by LDA and reduced to 45 components. In the acoustic front-end these reduced log posterior features are augmented with the LDA reduced short-term MFCC features to a 90 dimensional input.

4. ACOUSTIC MODELING

As in [5] the systems differ only in the NN features used in the acoustic front-end. The LDA reduced NN posterior estimates are augmented by LDA reduced MFCCs, which are transformed by VTLN. We have performed the training of the NNs as well as the acoustic models on the same 160h of Spanish audio data.

The acoustic models for all systems are based on triphones with a cross-word context, modeled by a 6-state left-to-right HMM. A decision tree based state tying is applied resulting in a total of 4500 generalized triphone states. The acoustic models consist of Gaussian mixture distributions with a globally pooled diagonal covariance matrix. In the end, the acoustic model contains of 1.1M mixture densities.

In order to compensate for speaker variations we use constrained maximum likelihood linear regression speaker adaptive training (SAT/CMLLR) In addition, during recognition, maximum likelihood linear regression (MLLR) is applied to the means of the acoustic models. For computational reasons we have not included discriminative training. Experiments show that we gain additional 5-10% by discriminative training even with NN based features.

5. EXPERIMENTAL SETUP

Approximately 160 hours of Spanish Broadcast news and speech data collected from the web are used both for training the NN phoneme posterior estimates and for training the GHMMs. The evaluation of the systems is performed on the development corpus of 2010 (dev10) and the evaluation corpora of 2010 (eval10) and 2009 (eval09). Each of these corpora contains around 3h of speech. During recognition the parameters have been tuned on the dev10 corpus. All the data for training as well as for recognition are provided within the Quaero project.

We use a 4-gram language model (LM) during recognition consisting of 60k words. The LM is trained on the final text editions and verbatim transcriptions of the European Parliament Plenary Sessions, and data from the Spanish Parliament and Spanish Congress, provided within the TC-STAR project. All LM data provided within the Quaero project are included as well as the acoustic transcriptions.

6. MULTIPLE FEATURE COMBINATION

6.1. Recurrent Neural Network Feature Combination

In the first experiments the different short-term acoustic features are combined using the concept of bi-directional LSTM-RNNs (BLSTM-RNNs). In preliminary results, not presented here, the BLSTM-RNNs have outperformed the other bi-directional and uni-directional RNNs. The results presented in Table 1 are obtained after feature based speaker adaptation using CMLLR.

Table 1. MLP and BLSTM-RNN feature combination results using a speaker adapted model (SAT/CMLLR). The NNs combine up to three different short-term features using different NN topologies. The log-posterior estimates and the augmented MFCCs are transformed independently of each other by LDA to 45 components each. The baseline system is trained on MFCCs only.

System	NN input		# of NN	Testing corpora (WER [%])		
Input	Type	Size	Params	dev10	eval10	eval09
MFCC	—	—	—	21.6	18.2	16.7
+ MLP	MFCC	297	1.32M	20.4	16.9	15.5
	+ PLP	594	2.51M	20.1	16.6	15.3
	+ GT	873	3.63M	19.8	16.3	15.0
+ BLSTM-RNN						
	GT	31	0.37M	19.9	16.6	15.2
	PLP	33	0.37M	20.0	16.2	15.2
	MFCC	33	0.37M	19.4	15.9	14.9
	+ GT	64	0.43M	19.0	15.4	14.3
	+ PLP	66	0.42M	18.9	15.7	14.5
	+ GT	97	0.48M	19.0	15.4	14.3

As observed by the MLP combinations in [8], adding a second short-term feature stream improves the overall performance. Depending on the feature sets used, the second feature stream decreases the WER of the BLSTM-RNN based posterior estimates by more than 0.4% absolute. This is similar to the gain obtained by the feature combination using MLPs. When we combine all three features sets, no additional improvements are observed for the BLSTM-RNN. Since the three short-term features are produced in a similar way, the combined features cover a lot of redundant information of the speech signal. Nevertheless, without testing each feature combination, the best performance is produced by combining all features by the NN. The additional training effort for the third feature set is negligible, since the size of the input layer is increased only.

The BLSTM-RNN features clearly outperform the MLP results. Moreover, the best MLP result has been beaten by the BLSTM-RNN with just one feature stream. Overall, the BLSTM-RNNs achieve a 1% absolute better WER, which is about 5% relative. Furthermore, note that the BLSTM-RNNs achieve the large improvements with less parameters trained.

6.2. Hierarchical Processing

In the hierarchical processing we have tested both combinations, training an MLP on top of the BLSTM-RNN posterior estimates and training an BLSTM-RNN on the output of an MLP. While the latter combination has not been very successful, the MLP benefits from the BLSTM-RNN based features. The result of the hierarchical processing using just MFCC features are summarized in Table 2. The MLPs trained on the BLSTM-RNN features achieve the same performance as the BLSTM-RNN based features alone, but improves the WER of the previous MLP results by 1% absolute or 4% relative. Since we have not gained anything by this hierarchical combination, the same

Table 3. Speaker adapted tandem recognition results of hierarchical BLSTM-RNN→MLP posterior estimates on Spanish. The MLPs are trained on the posteriors of the BLSTM-RNN combined by the short-term feature streams used in the BLSTM-RNN network as well. The final tandem systems are trained on a 90 dimensional feature vector containing the MFCC features augmented by the NN posteriors.

System	NN input		Total # of parameters		Testing corpora (WER [%])		
Input	Type	Size	NN	GHMM	dev10	eval10	eval09
MFCC	—	—	—	50M	21.6	18.2	16.7
+ MLP	MFCC	297	1.3M	99M	20.4	16.9	15.5
+ BLSTM-RNN	MFCC	33	0.4M	99M	19.4	15.9	14.9
→ MLP	BLSTM-RNN						
	+ MFCC	594	2.8M	99M	19.2	15.8	14.9
	+ GT	873	4.0M	99M	18.9	15.4	14.3
	+ PLP	891	4.1M	99M	19.0	15.7	14.6
	+ GT	1170	5.3M	99M	18.8	15.4	14.2

Table 2. Speaker adapted recognition results of different NN posterior features trained on MFCCs. The posteriors are derived by an MLP, by an BLSTM-RNN or by a hierarchical processing of BLSTM-RNNs and MLPs, marked by →. The NN based features are augmented by the VTLN transformed MFCCs, resulting in a 90 dimensional input, to train the tandem system.

System	NN Input	Testing corpora (WER [%])		
Input	Size	dev10	eval10	eval09
MFCC	—	21.6	18.2	16.7
+ MLP	297	20.4	16.9	15.5
+ BLSTM-RNN	33	19.4	15.9	14.9
→ MLP	33	19.4	16.0	14.9
+ MFCC	66	19.2	15.8	14.9

short-term features are added as input for the MLP training. Now, the hierarchical MLP benefits from the BLSTM-RNN posteriors as well as from the additional MFCC features. Overall, the recognition performance is improved slightly by 0.2% absolute on dev10 and 0.1% absolute on eval10.

6.3. Hierarchical Feature Combination

In the previous section we have observed that the hierarchical framework improves the recognition performance, when the same features are provided in every stage of the hierarchy. In this experiments we have applied the same concept to perform hierarchical NN based feature combinations. We first train a BLSTM-RNN on the combined features and afterwards an MLP on the BLSTM-RNN posterior augmented by the same input features.

We could verify the small improvements on a subset of the training data for all feature sets. When all training data is used, the improvements vanished. As shown in Table 3 the performance of the hierarchical approach is improved slightly, when all feature streams are combined. Even though the number of parameters of the MLP is larger than of the BLSTM-RNN, the performance does not degenerate. The MLP benefits from the preprocessing of the features by the BLSTM-RNN. Overall, the best BLSTM-RNN result is improved by 0.1% absolute, corresponding to 40 words which are recognized correctly.

7. SUMMARY AND CONCLUSION

The aim of this paper was to improve the NN based feature combination approach using recurrent and non-recurrent neural networks. Therefore, we have proposed different NN topologies and combinations of these networks. We showed, that the BRNNs using the LSTM structure clearly outperform the MLP based feature combination approach. Moreover, the BLSTM-RNNs achieved a better performance w.r.t the final WER using much less parameters. When the same input features were used, the BLSTM-RNN reduced the WER by 1% absolute on all corpora. Moreover, the BLSTM-RNN concept has been applied the first time on a large scale LVCSR task.

In the hierarchical framework, the MLP had benefit from the BLSTM-RNNs and improved the performance slightly. Nevertheless, to achieve the best performance, the same short-term features had to be provided in every stage of the hierarchy. On the other hand, the BLSTM-RNNs trained on the MLP based posteriors had shown no improvements. Even more, this hierarchical combination show a degradation in performance.

As a next step, we will investigate the effect of context-dependent states for NN based feature combination. Since the training of an RNN takes 5 times longer as the MLP training, the influence of the RNN posteriors for context-dependent MLP have to be analyzed as well as the bottle-neck concept. Furthermore, we will investigate the best combination of short-term and long-term features using NNs.

8. CONTRIBUTIONS TO PRIOR WORK

In this work, we continued our work on NN based feature combination started in [5]. There we had shown that the MLP based feature combination approach outperform other combination methods, e.g. feature concatenation [4], combination by a LDA transformation [3] or system combination [2]. We combined several short-term features using BLSTM-RNNs [6], which had been applied only to image [9] or small speech recognition tasks [17]. Moreover, we gave a comparison of MLPs and BLSTM-RNNs trained on the same corpus and input features. The BLSTM-RNNs achieved much better WERs with less parameters.

The concept of hierarchical processing of several MLPs was introduced in [22]. We transferred the concept of NN stacking to combine recurrent and non-recurrent NNs. Here, the RNN was used to provide a clever preprocessing of the combined features to improve the MLP results.

9. ACKNOWLEDGMENTS

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation. H. Ney was partially supported by a senior DIGITEO Chair grant from Ile-de-France.

10. REFERENCES

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

[2] A. Zolnay, *Acoustic Feature Combination for Speech Recognition*, Ph.D. thesis, RWTH Aachen University, Aachen, Germany, Aug. 2006.

[3] R. Schlüter, A. Zolnay, and H. Ney, "Feature combination using linear discriminant analysis and its pitfalls," in *Interspeech*, Pittsburgh, PA, USA, Sept. 2006, pp. 345–348.

[4] A. Zolnay, R. Schlüter, and H. Ney, "Acoustic feature combination for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA, Mar. 2005, vol. 1, pp. 457–460.

[5] C. Plahl, R. Schlüter, and H. Ney, "Improved acoustic feature combination for LVCSR by neural networks," in *Interspeech*, Florence, Italy, Aug. 2011, pp. 1237–1240.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *IEEE Transactions on Neural Networks*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[7] M. Sundermeyer, M. Nußbaum-Thom, S. Wiesler, C. Plahl, A. El-Desoky Mousa, S. Hahn, D. Nolden, R. Schlüter, and H. Ney, "The RWTH 2010 Quaero ASR evaluation system for English, French, and German," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 2212–2215.

[8] C. Plahl, B. Hoffmeister, G. Heigold, J. Lööf, R. Schlüter, and H. Ney, "Development of the GALE 2008 Mandarin LVCSR system," in *Interspeech*, Brighton, U.K., Sept. 2009, pp. 2107–2110.

[9] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, May 2009.

[10] S. E. Boquera, M. J. C. Bleda, J. G. Moya, and F. Z. Martinez, "Improving offline handwritten text recognition with hybrid HMM/ANN models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 767–779, Apr. 2011.

[11] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature stream extraction for conventional HMM systems," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2000, pp. 1635–1638.

[12] H. Bourland and N. Morgan, "Connectionist speech recognition: A hybrid approach," *Series in engineering and computer science. Kluwer Academic Publishers*, vol. 247, 1994.

[13] F. Seide, L. Gang, and Y. Dong, "Conversational Speech Transcription using context-dependent Deep Neural Network," in *Interspeech*, Florence, Italy, Aug. 2011, pp. 437–440.

[14] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, and P. Novak, "Making Deep Belief Networks effective for Large Vocabulary Continuous Speech Recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Hawaii, USA, Dec. 2011, pp. 30–35.

[15] Z. Tüske, M. Sundermeyer, R. Schlüter, and H. Ney, "Context-dependent MLPs for LVCSR: TANDEM, hybrid or both?," in *Interspeech*, Portland, OR, USA, Sept. 2012.

[16] T. Robinson, M. Hochberg, and S. Renals, "IPA: Improved phone modelling with recurrent neural networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Apr. 1994, vol. 1, pp. 37–40.

[17] M. Wöllmer, F. Eyben, B. Schuller, and G. Rigoll, "Recognition of spontaneous conversational speech using long short-term memory phoneme predictions," in *Interspeech*, Makuhari, Japan, Sept. 2010, pp. 1946–1949.

[18] O. Vinyals, S. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust ASR," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, Mar. 2012, pp. 4085–4088.

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

[20] M. Wöllmer, B. Schuller, and G. Rigoll, "A novel bottleneck-BLSTM front-end for feature-level context modeling in conversational speech recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Hawaii, USA, Dec. 2011, pp. 36–41.

[21] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Interspeech*, Portland, OR, USA, Sept. 2012.

[22] F. Valente, J. Vepa, C. Plahl, C. Gollan, H. Hermansky, and R. Schlüter, "Hierarchical neural networks feature extraction for LVCSR system," in *Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 42–45.

[23] F. Valente, M. Magimai-Doss, C. Plahl, and S. Ravuri, "Hierarchical processing of the modulation spectrum for GALE Mandarin LVCSR system," in *Interspeech*, Brighton, U.K., Sept. 2009, pp. 2963–2966.

[24] H. Hermansky and S. Sharma, "TRAPs - classifiers of temporal patterns," in *Proc. Int. Conf. on Spoken Language Processing*, Sydney, Australia, Dec. 1998.

[25] H. Hermansky and P. Fousek, "Multi-resolution RASTA filtering for TANDEM-based ASR," in *Interspeech*, Lisbon, Portugal, Sept. 2005, pp. 361–364.

[26] C. Plahl, R. Schlüter, and H. Ney, "Hierarchical bottle neck features for LVCSR," in *Interspeech*, Makuhari, Japan, Sept. 2010, pp. 1197–1200.

[27] F. Grézl, M. Karafiat, S. Kontar, and J. Cernock, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, Apr. 2007, vol. 4, pp. 757–760.

[28] R. Schlüter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, Apr. 2007, vol. 4, pp. 649–652.