Neuronale Netze in der automatischen Spracherkennung - ein Paradigmenwechsel?

Neural Networks in Automatic Speech Recognition - a Paradigm Change?

R. Schlüter, P. Doetsch, P. Golik, M. Kitza, T. Menne, K. Irie, Z. Tüske, A. Zeyer

Lehrstuhl Informatik 6, RWTH Aachen University, 52074 Aachen, Germany, Email: schlueter@cs.rwth-aachen.de

Abstract

In der automatischen Spracherkennung, wie dem maschinellen Lernen allgemein, werden die Strukturen der zugehörigen stochastischen Modellierung heute mehr und mehr auf unterschiedliche Formen künstlicher neuronaler Netze umgestellt. Dieser Erneuerungsprozess, der schon vor nahezu 30 Jahren begann, führte in den vergangenen 10 Jahren zu erheblichen Verbesserungen in der Erkennungsgenauigkeit. Sowohl in der akustischen Modellierung von Sprache, als auch der a-priori Modellierung von Sprache auf der Textebene markieren künstliche neuronale Netze heute den Stand der Technik in der Spracherkennung für großes Vokabular, und weitere Verbesserungen werden erwartet. Diese Arbeit gibt einen Überblick über den aktuellen Stand der Forschung im Bereich der auf künstlichen neuronalen Netzen basierenden Modellierung von Systemen zur automatischen Spracherkennung. Dies beinhaltet Diskussionen zu den folgenden Themen: neuronale Netzwerktopologien und Zelltypen, Training und Optimierung, Auswahl der Eingabemerkmale, Adaption und Normalisierung, Multifunktionstraining, sowie neuronale Netzmodellierung statistischer Sprachmodelle. Ungeachtet der deutlichen Fortschritte in der Spracherkennung mittels neuronaler Netze, bleiben jedoch weiterhin viele offene Fragen zu klären, bevor eine vollständig konstistente und eigenständige Modellierung durch neuronale Netze in der Spracherkennung erreicht wird. Diese Arbeit schliesst mit einer Diskussion offener Probleme sowie potentieller zukünftiger Forschungsrichtungen, insbesondere bzgl. der Integration neuronaler Netze in den Entscheidungsprozess der automatischen Spracherkennung. Diese Arbeit ist eine Überarbeitung eines auf der SPECOM 2016 präsentierten Übersichtsartikels [Schlüter & Doetsch$^+$ 2016].

In automatic speech recognition (ASR), as in general in the area of machine learning, the structures of the corresponding stochastic modeling more and more are changing to various forms of neural networks. This process of renewal that started about 30 years ago, lead to considerable improvements in recognition performance in the past 10 years. Both in acoustic modeling of speech, as well as the a priori language modeling on textual level, artificial neural network based modeling now marks the state-of-the-art for large vocabulary continuous speech recognition, and further improvements are to be expected. This work gives an overview of current research activities in neural network based modeling for automatic speech recognition systems. This includes discussions of network topologies and cell types, training and optimization, choice of input features, adaptation and normalization, multitask training, as well as neural network based statistical language modeling. Despite the clear progress obtained with neural network modeling in speech recognition, many questions remain to be investigated, yet to obtain a consistent and self-contained neural network based modeling approach that ties in with the former state-of-the-art. We will conclude by a discussion of open problems as well as potential future directions, especially w.r.t. to the integration of neural networks into automatic speech recognition decision process. This work is a revised version of a review article presented at SPECOM 2016 [Schlüter & Doetsch$^+$ 2016].

1. Introduction: Neural Networks in ASR

Even though artificial neural networks (ANN) have been known for long, their application to automatic speech recognition (ASR) remained a limited area of research for quite some time. An efficient learning algorithm for the free parameters of neural networks by error backpropagation was introduced in [Rumelhart & Hinton$^+$ 1986]. A few years later, neural networks with a single (non-linear) hidden layer have been shown to have the universal approximator property, meaning that, similar to *Gaussian* mixture models, they are capable of approximating any continuous function to any level of accuracy [Hornik & Stinchcombe$^+$ 1989]. In [Lippmann 1989], an overview of early approaches to automatic speech recognition using neural network based modeling is given. Notable trends at that time include the time-delay neural network approach [Waibel & Hanazawa$^+$ 1989], the softmax operation for probability normalization on the output layer of neural networks [Bridle 1989], and the introduction of the hybrid concept to hidden *Markov* models (HMM) by interpreting neural network outputs as class posteriors (in the context of the squared error criterion), and using them to model HMM emission probabilities [Bourlard & Wellekens 1989]. Finally, first results which were competitive to standard *Gaussian* mixture HMMs on the Wall Street Journal task were presented using a recurrent neural network [Robinson & Hochberg$^+$ 1994]. Nevertheless, only with the new millennium, approaches including neural network modeling started to outperform the former state-of-the-art. Nowadays, ANN-based ASR systems show considerable improvements of 30% and more relative in word error rate (WER) over *Gaussian* mixture based HMMs, e.g. [Seide & Li$^+$ 2011b]. On specific tasks, ANN-based ASR system even are reported to perform at near-human performance [Saon & Kurata$^+$ 2017, Xiong & Wu$^+$ 2017].

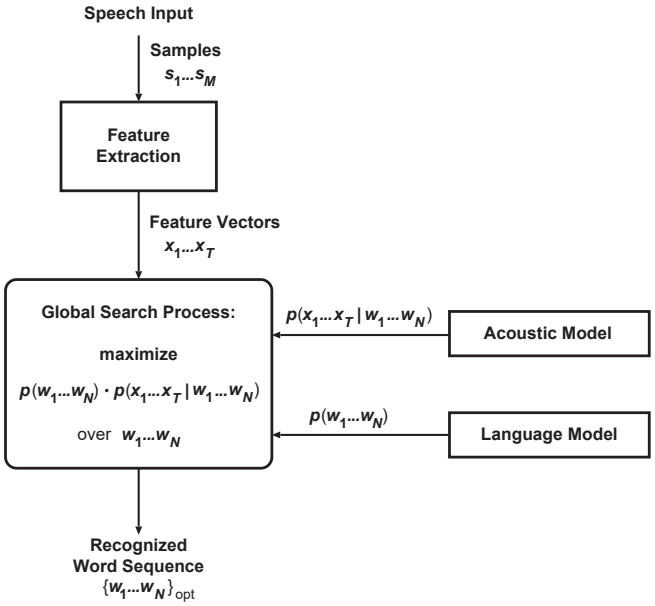

Figure 1: General architecture of a statistical speech recognition system.

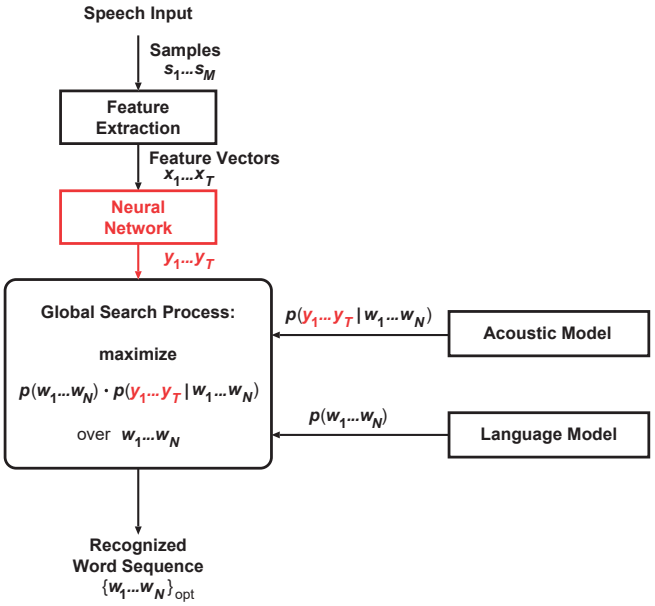

Figure 2: Architecture of a statistical speech recognition system including a neural network based feature transformation[Hermansky & Ellis$^+$ 2000, Grézl & Karafiát$^+$ 2007, Fontaine & Ris$^+$].

In this work, we provide an overview of recent developments and results obtained using ANNs of various flavors in both acoustic and language modeling. In Sec. 2, general approaches to ANN-based acoustic modeling are compared. Sec. 3 presents network topologies and cell types, followed by a discussion of corresponding training criteria in Sec. 4, regularization methods in Sec. 5, and corresponding optimization methods in Sec. 6. Sec. 7 gives an overview of ANN-based language modeling. In Sec. 8, input features for ANN-based ASR and ANN-based learning of the corresponding signal processing are discussed. Sec. 9 shows how multitask learning introduces generalization across multiple languages. Sec. 10, discusses current approaches to adaptation and normalization for ANN-based ASR. Finally, Sec. 11 discusses recent developments especially in decoding using neural network based modeling, followed by general conclusions in Sec. 12.

2. Neural Networks in Acoustic Modeling

The introduction of neural networks to acoustic modeling can be divided into the *hybrid*, and the *tandem* approach. In the *hybrid* approach [Bourlard & Wellekens 1989, Bourlard & Morgan 1993], HMM emission probabilities are modelled explicitely using (appropriately renormalized) neural networks representing phoneme class posteriors, thus dropping the need for *Gaussian* mixture models (GMM). In [Robinson & Hochberg$^+$ 1994], an early success was obtained on the Wall Street Journal task using recurrent neural networks within the hybrid approach. Nevertheless, GMMs still remained the prevalent acoustic modeling scheme for large vocabulary continuous speech recognition for at least a decade. Meanwhile, the *tandem* approach was proposed [Hermansky & Ellis$^+$ 2000], where the frame-wise phoneme classifier neural network output is post-processed and used as (additional) input feature to a conventional GMM-based recognition system as indicated in Fig. 2. A powerful extension to the tandem approach was suggested in [Grézl & Karafiát$^+$ 2007], where the output of one of the hidden layers rather than the output layer is used. Since the size of hidden layers is not constrained, this leaves more freedom to choose its size, position among the layers, and activation function for feature extraction. Due to the usually reduced layer size, this discriminatively learned representation of input data is termed *bottleneck features*.

Universally considered a breakthrough for hybrid modeling, the authors in [Seide & Li$^+$ 2011b] presented very good results on the widely used conversational telephone speech recognition task *Switchboard*, beating a strong speaker adapted and discriminatively trained GMM baseline. The key to their success was twofold: first, a large number of context-dependent phoneme state targets, instead of monophone targets was modelled. Second, *deep neural networks* (DNN) with a number of large hidden layers were employed. The authors were also among the first groups to transfer training of large models to a *graphics processing unit* (GPU), that offers a great speed-up over CPUs on matrix operations. Both tandem and hybrid approaches have been shown to outperform standard GMM baselines trained on cepstral features [Tüske & Sundermeyer$^+$ 2012]. In principle, both methods can be considered a deep stack of non-linear models trained to optimize different objective functions following different schemes for parameter updates [Tüske & Tahir$^+$ 2015]. From an even more general point of view, both models consist of a classifier and some representation learning

mechanism, although there is no explicit distinction of the two. In [Tüske & Michel$^+$ 2017], both approaches are compared on the Switchboard part of the Hub5'00 test corpus. A hybrid LSTM-RNN model achieved a word error rate (WER) of 10.8%, while a comparable tandem system yielded 10.9% WER. However, current state-of-the-art ASR systems [Saon & Kurata$^+$ 2017, Xiong & Wu$^+$ 2017] mainly concentrate on hybrid modeling, presumably for its more direct acoustic modeling.

3. Neural Network Topologies

The network topology has a major influence on performance. This includes optimization of the number of hidden layers, their size, the choice of activation functions, the cell types, the connectivity between the layers. Recently, a variety of novel activation functions have been suggested in the literature, including Maxout [Goodfellow & Warde-Farley$^+$ 2013], Exponential Linear Unit [Clevert & Unterthiner$^+$ 2016], and many modifications. Of these, the piece-wise linear function *rectified linear unit* (ReLU) [Nair & Hinton 2010] turned out to be very effective. Although unbounded, the ReLU does not violate the universal approximator property of the networks [Sonoda & Murata 2016]. ReLU networks allow to train very deep models even for hard optimization problems, as shown e.g. in [Tüske & Golik$^+$ 2014b] and [Golik & Tüske$^+$ 2015a]. *Bottleneck layers* are not only useful for the tandem approach, but also in a hybrid scenario, allowing to reduce the number of trainable parameters, increasing the processing speed and the generalization ability of the acoustic models [Wiesler & Richard$^+$ 2014]. Hierarchical stacking of multiple neural networks [Valente & Vepa$^+$ 2007] provides another choice in network design that has been shown to provide further gains in performance [Tüske & Schlüter$^+$ 2013]. *Convolutional neural networks* (CNNs) provide another design choice, where a layer with local connectivity shares the weights of its *receptive field* across the positions in the input, to support learning of more robust position-independent features [LeCun & Boser$^+$ 1990]. This concept was applied to speech recognition by defining the convolution layer over a spectrogram, approximated by critical band energies [Abdel-Hamid & Mohamed$^+$ 2012]. Very deep convolutional topologies like the so-called residual network architectures, and the layer-wise context expansion with attention (LACE) both were reported to perform almost as well as recurrent network structures in [Xiong & Wu$^+$ 2017]. In [Golik & Tüske$^+$ 2015a], convolution in time was used even to jointly learn feature extraction and acoustic model on raw waveform input. Highway networks [Srivastava & Greff$^+$ 2015] provide a way to improve information flow across layers and enables training of even deeper networks. Highway networks also were combined with recurrent neural networks in [Zhang & Chen$^+$ 2015]. *Recurrent neural networks* (RNNs) provide another powerful extension of the topology by introducing recurrent connections. Here, hidden layers perform a time-dependent operation taking its own output at the previous time step as input in addition to the output from the previous hidden layer. Optimization of RNNs requires *backpropagation through time* (BPTT), which is known to have the exploding/vanishing gradient effect [Hochreiter & Bengio$^+$ 2001]. Various approaches were suggested to overcome this problem, like gradient clipping [Pascanu & Mikolov$^+$ 2012], or second-order optimization methods [Wiesler & Li$^+$ 2013, Wiesler & Richard$^+$ 2014]. However, the most prominent variant is to modify the recurrent model itself, which led to the long short-term memory (LSTM) [Hochreiter & Schmidhuber 1997] model which allows for better gradient flow via gating units. Further analyses and variants of LSTMs are presented in [Chung & Gülçehre$^+$ 2014, Jozefowicz & Zaremba$^+$ 2015, Greff & Srivastava$^+$ 2015] and [Breuel 2015]. Recently, LSTMs were shown to clearly outperform feed-forward acoustic models [Sak & Senior$^+$ 2014, Geiger & Zhang$^+$ 2014], and can also be stacked to form deep LSTM networks [Graves & Mohamed$^+$ 2013]. Also, bidirectional LSTMs were shown to outperform unidirectional LSTMs [Graves & Schmidhuber 2005]. Bidirectional LSTMs can also be used in online recognition setups [Zeyer & Schlüter$^+$ 2016]. In recent work [Zeyer & Doetsch$^+$ 2017], based on a fast LSTM implementation [Doetsch & Zeyer$^+$ 2017], we observed an improvement from 15.3% WER for an highly optimized DNN to 13.1% WER using LSTM modeling on a 50h subset of the English Quaero task. Similarly, on the larger Switchboard task, in [Tüske & Michel$^+$ 2017] we observed an improvement from 12.3% WER using feed-foward DNN based acoustic models to 10.8% WER using recurrent LSTM acoustic models.

4. Training Criteria

Early approaches to neural network acoustic model training applied the *squared error* criterion [Bourlard & Wellekens 1989]. Nevertheless, the common approach is to minimize frame-level *cross entropy* over a training set. The cross entropy criterion was shown to be more robust to poor initialization, than the squared error criterion [Golik & Doetsch$^+$ 2013]. The training criterion in *connectionist temporal classification* (CTC) overcomes the necessity for allophone alignment by integrating over all alignments [Graves & Fernández$^+$ 2006], similar to *Baum-Welch* training. Frame-level cross entropy does not take the word level into account, treating all frames independently and with equal weight. Also, both the cross entropy and the frame classification error on a held-out data set are only loosely correlated with the target evaluation measure for speech recognition, word error rate. In the tandem approach, GMM/HMM training can be done with discriminative training [He & Deng$^+$ 2008, Heigold & Schlüter$^+$ 2012], but usually does not include further optimization of the underlying neural network, even though joint training would be possible [Tüske & Tahir$^+$ 2015, Tüske & Golik$^+$ 2015]. Sequence training [Kingsbury 2009] is the corresponding approach realizing discriminative training criteria for hybrid modeling. The criteria are the same in both cases, comprising *maximum mutual information* (MMI), *minimum phone error* (MPE) (cf. e.g. [He & Deng$^+$ 2008]), and *state-level minimum Bayes risk* (sMBR) criterion [Kings-

bury & Sainath$^+$ 2012]. A direct comparison of MMI and MPE can be found e.g. in [Wiesler & Golik$^+$ 2015]. For *encoder-decoder* models (cf. Sec 11), a promising approach using upper estimates on the training error are presented in [Bahdanau & Serdyuk$^+$ 2015]. In [Tüske & Golik$^+$ 2015], MPE training results are presented on the Switchboard task. MPE training reduces the WER of a cross entropy trained hybrid model from 13.7% to 12.6%. Further, sequence discriminative training has been found to play a crucial role for performing keyword search [Golik & Tüske$^+$ 2015b].

5. Regularization

Neural networks usually have a huge amount of trainable parameters and are thus prone to overfitting, i.e. they fit very good to the training data but perform badly on unseen data. Regularization aims at better generalization by avoiding overfitting in various ways. One approach would be to balance the amount of trainable parameters with the amount of training data available. Nevertheless, it has been shown that depth boosts neural network performance, both theoretically [Montufar & Pascanu$^+$ 2014] and experimentally [Zeyer & Doetsch$^+$ 2017]. A straight-forward approach to reduce the amount of parameters is to use a reduced matrix representation such as linear bottlenecks [Wiesler & Richard$^+$ 2014], cf. Sec. 3. Also, constraints can be introduced on the parameters or to penalize them in various ways, where minimizing the L_1 or L_2 norm is the simplest solution. This is usually added as an additional term to the optimization criterion. Another class of methods stochastically modifies the network architecture so that the overall model can be seen as an ensemble model of all the stochastic variants. The most prominent method of this kind is dropout [Srivastava & Hinton$^+$ 2014], where hidden nodes are randomly dropped. Another recent promising method is stochastic depth [Huang & Sun$^+$ 2016], where hidden layers are randomly dropped. In [Zeyer & Doetsch$^+$ 2017] we studied the effect of different regularization methods for LSTM based acoustic models and we found a combination of L_2 and dropout to perform best.

6. Optimization

Neural network weights usually are trained by using error backpropagation and stochastic gradient descent (SGD) on the corresponding training criterion [Rumelhart & Hinton$^+$ 1986]. Improving the initial starting point of the optimization was addressed by so called *pretraining* techniques [Hinton & Osindero$^+$ 2006]. Here, the weights are initialized layer-by-layer using unsupervised restricted Boltzmann machines or supervised methods like *discriminative pretraining* [Seide & Li$^+$ 2011a]. As an alternative to SGD, also batch methods are possible, usually applying second-order information, e.g. LBFGS, Rprop, or the Hessian-Free approach [Wiesler & Richard$^+$ 2014]. Second-order optimization can also be transferred to SGD by normalizing mean and variance within a batch [Wiesler & Richard$^+$ 2014, Gülçehre & Bengio 2014]. The step size for the adjustment is usually treated as hyperparameter and various methods have been proposed to improve the estimate of the optimal step size, like AdaGrad [Duchi & Hazan$^+$ 2010], Adadelta [Zeiler 2012], or Adam [Kingma & Ba 2014], which was reported to give very stable optimization. Also, gradient clipping and noise addition are commonly used methods to improve convergence. Table 1 shows the performance of these methods on an ASR task. In order to benefit from modern computing architectures like GPUs, algorithms have to be designed to perform simple operations on large amounts of data in parallel. The workload of each single operation can be distributed on several machines. More commonly, training times are reduced and a direct extension to gradient descent is obtained by *data parallelism*, where samples are partitioned into batches and considered as single step in the optimization procedure, c.f. e.g. [Dean & Corrado$^+$ 2012, Doetsch & Zeyer$^+$ 2017].

Table 1: Word error rate on 50 hours of English spoken sentences from the Quaero corpus using different optimization methods. The results are taken from [Zeyer & Doetsch$^+$ 2017].

Systems	WER [%]
Gradient Descent	15.0
AdaGrad	15.6
Adadelta	15.1
Adam	14.8
+ gradient noise	14.6

7. Neural Networks in Language Modeling

The earliest ANN-based approach to language modeling known to us was proposed in [Nakamura & Shikano 1989] and termed *NETgram*. However, the first application of ANN-based language models in ASR only appeared in [Bengio & Ducharme$^+$ 2000]. While competitive results against the conventional count-based models were already reported with feed-forward ANN-based n-gram models, neural networks have become truly popular for language modeling only after introducing recurrent neural networks into language modeling [Mikolov & Karafiát$^+$ 2010]. As opposed to the count model and feed-forward ANN, the recurrent neural network effectively can handle unlimited context by compressing it into a fixed size context vector. This property is an elegant solution to the context length problem which is fundamental in language modeling. Finally, the long short-term memory recurrent network was first applied for language modeling in [Sundermeyer & Schlüter$^+$ 2012], which is considered the state-of-the-art architecture for language modeling today [Melis & Dyer$^+$ 2017]. In practice, the combination of count-based and neural network-based approaches gives best results in ASR within the hybrid approach. Linear interpolation is the most popular and effective combination method (a small improvement by an alternative back-off level combination has been reported in [Chen & Liu$^+$ 2015]). Due to the high computational complexity of neural language models and the context-induced search complexity, ANN-based lan-

guage models mostly are applied in a rescoring step using N-best lists, or lattices [Sundermeyer & Tüske+ 2014] generated using count models. Overall, the relative improvements from a 4-gram count model in perplexity of about 30% and in word error rate of 16% were reported in [Sundermeyer & Ney+ 2015] by using a neural network with two LSTM layers. Further results can be found in Tables 2 and 3. More recently, on the Switchboard

Table 2: Standalone perplexities on the Quaero English 2012. Training data of 50M running words. The results are taken from [Tüske & Irie+ 2016, Irie & Tüske+ 2016].

model type	# of layers	PPL	CPU time
4-gram count model	–	163.7	30min
10-gram MLP	1	136.5	1 week
	2	130.9	–
Recurrent NN	1	125.2	–
LSTM-RNN	1	107.8	3 weeks
	2	100.5	–

Table 3: Interpolated perplexities and word error rate on Quaero English 2013 [Tüske & Irie+ 2016, Irie & Tüske+ 2016]. 250h acoustic training data, 3B/50M running words for training count/neural models. CN decoding.

train data	model type	# of layers	PPL	WER [%]
3.1B	4-gram count model	–	131.2	12.4
50M	+ 10-gram MLP	1	112.5	11.5
		2	110.2	11.3
	+ Recurrent NN	1	108.1	11.1
	+ LSTM-RNN	1	96.7	10.8
		2	92.0	10.4

corpus, we observed a similar reduction by 15% relative in WER, i.e. from 9.8% for a 4-gram count-based language model down to 8.3% for an LSTM-based language model [Irie & Lei+ 2018]. In another recent study, we investigated the use of recurrent model topologies to model long, but limited context. We observed that the effective context length needed to obtain optimum performance levelled around 20-40 words [Tüske & Schlüter+ 2018].

8. Acoustic Features and Feature Learning

Despite similarities in standard feature extraction pipelines like MFCC, PLP, or Gammatone [Davis & Mermelstein 1980, Hermansky 1990, Schlüter & Bezrukov+ 2007], they can be expected to complement each other when combined. Using these features as concatenated input to neural networks, about 5-10% relative WER reduction was observed on a broadcast news and conversation task in [Plahl & Kozielski+ 2013, Tüske & Golik+ 2014b]. Further investigation also revealed that the optimal cepstral/critical band energy features for MLP requires higher resolution (up to 50-60 dimensional) than for GMM (15-20 dimensional) [Tüske & Golik+ 2015].

Recent studies demonstrated that feature extraction be

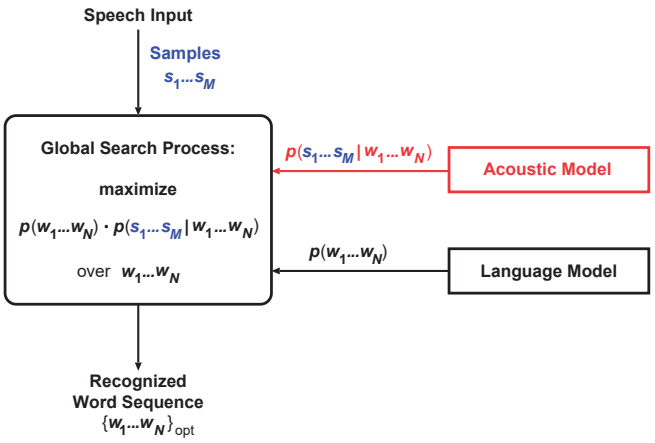

Figure 3: Integration of signal processing for acoustic feature generation and acoustic modeling.

integrated into the acoustic model and learned completely from data [Tüske & Golik+ 2014b, Sainath & Weiss+ 2015a], as depicted in Fig. 3. Interestingly, the model training leads to auditory-like filterbanks. Fig. 4 shows the convolutional processing in the first hidden layer, and Fig. 5 shows the corresponding learned filter weights in the time domain and transformed into the frequency domain, respectively [Golik & Tüske+ 2015a]. Fig 6 shows the effective convolution in time and frequency resulting from the second hidden convolutional layer applied in [Golik & Tüske+ 2015a]. Fig. 7 shows examples of corresponding filters learned directly from the speech samples as part of the overall acoustic model training. For details, cf. [Golik & Tüske+ 2015a]. Using more training data, in [Tüske & Michel+ 2017] it is shown that even standard feed-forward neural network structures can be used to learn the acoustic feature extraction and still obtain comparable WER to using manually designed features. For sufficient amounts of training data, models trained on the raw time signal can even outperform standard preprocessing, even for multichannel scenarios [Sainath & Weiss+ 2015b].

However, this usually requires a large amount of transcribed speech. In low-resource scenarios, well-

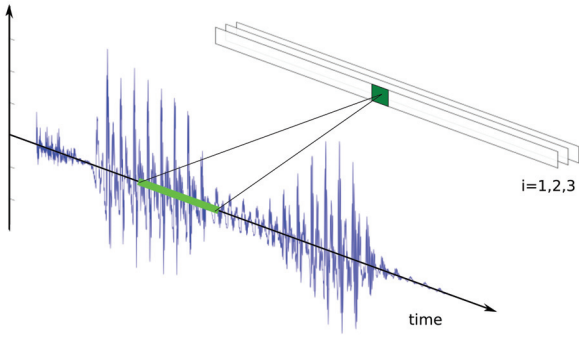

Figure 4: Convolution in time.

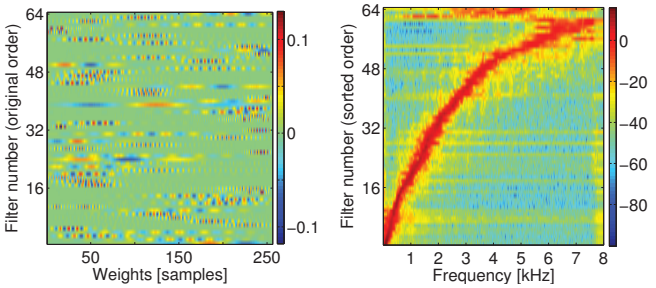

Figure 5: First hidden convolutional filters: original weights in time domain (left) and transformed into frequency domain (right).

established static standard feature extraction procedures still show a significant performance margin over data-driven feature extraction. Using 50 hours of broadcast news and conversations data showed about 10% relative performance loss [Golik & Tüske+ 2015a, Tüske & Golik+ 2015] for data driven feature extraction, even if the modeling was informed by standard feature extraction steps. In acoustically challenging recognition tasks, with only a few hours of speech available, well-established preprocessing steps like RASTA filtering [Hermansky & Morgan 1994] of critical band energies and feature combination can still be beneficial [Tüske & Golik+ 2014a].

Also for multichannel scenarios, beamforming can be supported by neural network based model. In [Heymann & Drude+b], BLSTM-based ask estimation is introduced for generalized eigenvalue based beamforming successfully. In [Heymann & Drude+a], the mask-estimation even is included in the overall acoustic model training. The mask estimation even can be done speaker adaptive, as shown in [Menne & Ney 2018].

9. Multilingual Modeling

Cepstral features typically capture formant related information and are a good starting point to develop acoustic models for any language. As neural networks have become a major component of recent HMM based ASR techniques, it was observed that neural network based posterior features possess language independent properties to a certain degree. This can be exploited in the tandem approach [Stolcke & Grézl+ 2006, Tóth & Frankel+ 2008, Plahl & Schlüter+ 2011]. Taking advantage of multiple language acoustic data poses the question of how to

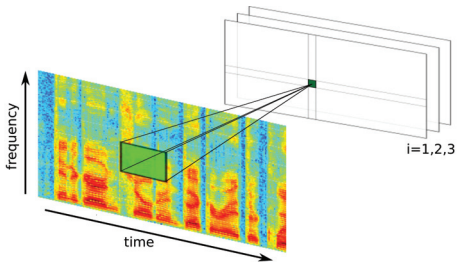

Figure 6: Convolution in time and frequency.

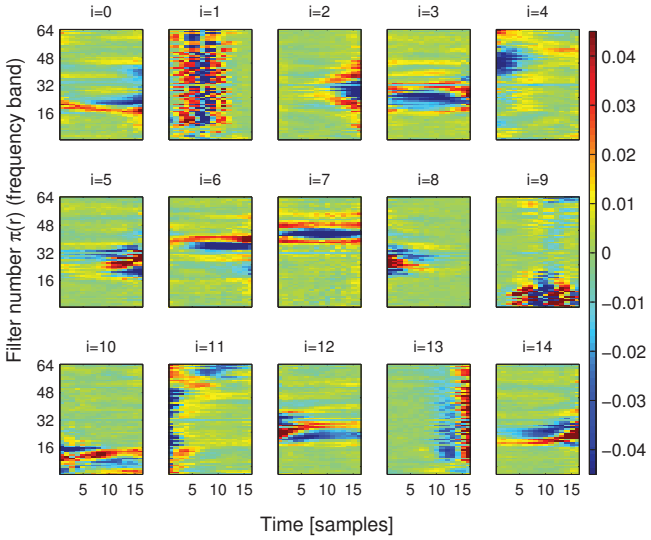

Figure 7: Examples of learned time-frequency filters.

handle differing phoneme sets. Language independent mappings can be done using international phonetic alphabets like IPA or SAMPA [Schultz & Waibel 1997], or by various data driven approaches [Byrne & Beyerlein+ 2000, Thomas & Ganapathy+ 2010]. Nevertheless, this often introduces ambiguities. In another approach, a joint phoneme set was generated by having language dependent phonemes [Grézl & Karafiát+ 2011], although this might introduce unnecessary discrimination between similar phonemes. The inherent layer-wise structure of an MLP also allows to train the model on multiple languages by sharing only hidden layers across languages [Scanzio & Laface+ 2008], which forces the network to learn a *language independent* representation on a deeper level. This multiple output training is closely related to multi-task training proposed by [Caruana 1993] and to subspace methods [Burget & Schwarz+ 2010]. One of the biggest advantages of multilingual modeling is that the training can be done without knowing the target language. Thus, system development on a new, unseen language becomes more efficient and even leads to significant performance gains. In [Tüske & Nolden+ 2014], multilingual features gave relative gains of 9-11% in WER, and 30-40% in key word search on low resource tasks for a number of quite different languages within the Babel project [Babel]. The usefulness of multilingual models later was confirmed for very low amounts of only 3h target language training data [Golik & Tüske+ 2015b]. Even with using 28 languages for training, still additional benefits were observed [Golik & Tüske+ 2017]. Even for tasks with large training sets, significant improvements were observed. Between 3-7% relative WER reduction was observed on broadcast news and conversation LVCSR tasks for four languages using between 110h and 320h for training per language, and 700h overall for training the multilingual net [Tüske & Schlüter+ 2013]. In summary, multilingual features and corresponding initialization schemes provide an efficient acoustic modeling framework for unseen qalanguages, and could reduce system development

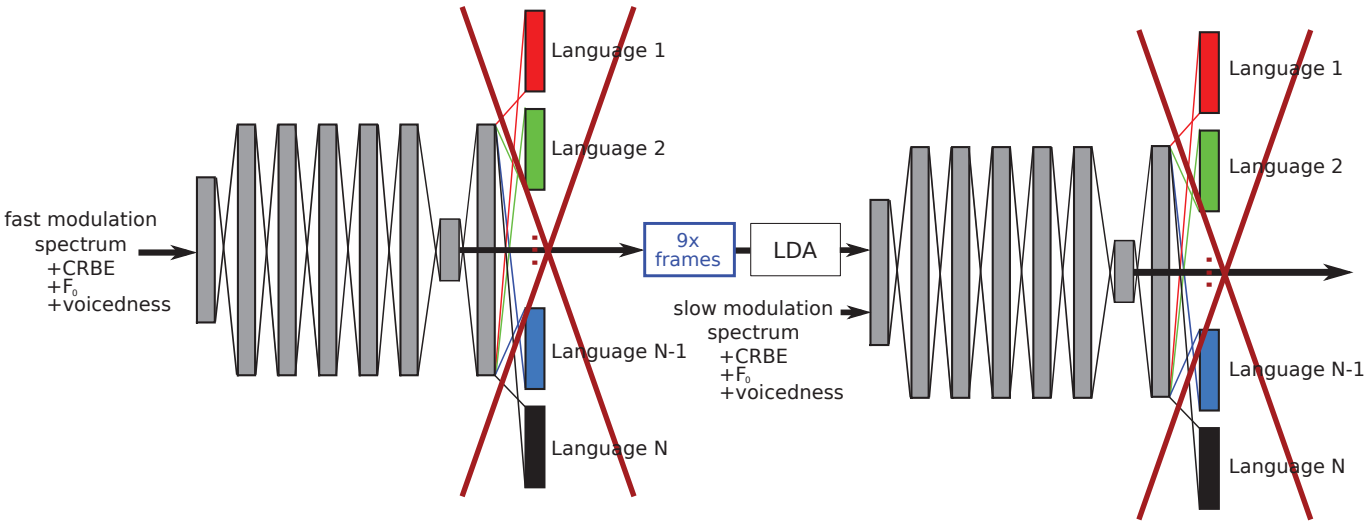

Figure 8: Architecture of a hierarchical multilingual bottleneck multi-layer perceptron.

time and costs significantly.

10. Adaptation and Normalization

Adaptation and normalization w.r.t. to speaker, environment, or recording channel usually provides significant improvements in ASR. Using GMMs, vocal tract length normalization (VTLN), maximum likelihood linear regression (MLLR), as well as feature space/constrained MLLR (CMLLR), together with speaker adaptive training work well, and have already been transferred to neural network based ASR successfully [Grézl & Karafiát$^+$ 2007, Schaaf & Metze 2010, Seide & Li$^+$ 2011a]. Although many approaches still necessitate GMMs as target models in the background, optimization within the neural network structure also is possible [Seide & Li$^+$ 2011a]. E.g., for the real test set on the noisy, multichannel CHiME3 task we observed an improvement from 8.6% to 6.9% WER using CMLLR over an LSTM-based hybrid system and using beamforming kindly provided by the authors of [Heymann & Drude$^+$b]. I-vectors are a well known low-dimensional speaker representation used in the domain of speaker recognition/verification, which can be used to inform neural network acoustic models with speaker information. Using i-vectors, in [Saon & Soltau$^+$ 2013] a 5-6% relative improvement in WER over a DNN baseline trained on already speaker adapted features was obtained on the Switchboard task. Speaker cluster adaptive training based on i-vectors is proposed in [Prasad & Sim 2016]. In [Samarakoon & Sim$^+$ 2017], speaker representations also are used to interpolate between speaker clusters. Neural networks can also be used to generate a variety of adaptation codes, by using speakers or environmental conditions as classification targets for a network [Miao & Metze 2015, Qian & Tan$^+$ 2016]. Besides CMLLR, different affine transformations can be used throughout the network for speaker adaptation [Li & Sim 2010, Xue & Li$^+$ 2014], also using recurrent LSTM models [Liu & Wang$^+$ 2016], and in [Huang & Lu$^+$ 2018] also for bidirectional LSTM models for the case of speech synthesis. In [Wang & Wang 2017], unsupervised adaptation of batch-normalized acoustic models applying scaling and shifting is proposed. In [Swietojanski & Renais 2016], learning hidden unit contributions is combined with speaker adaptive training without requiring intermediate speaker dependent features or modeling. [Kitza & Schlüter$^+$ 2018] shows a systematic study of using affine speaker adaptation transforms throughout the layers of a deep BLSTM, and a comparison to learning hidden unit contributions for speaker adaptation.

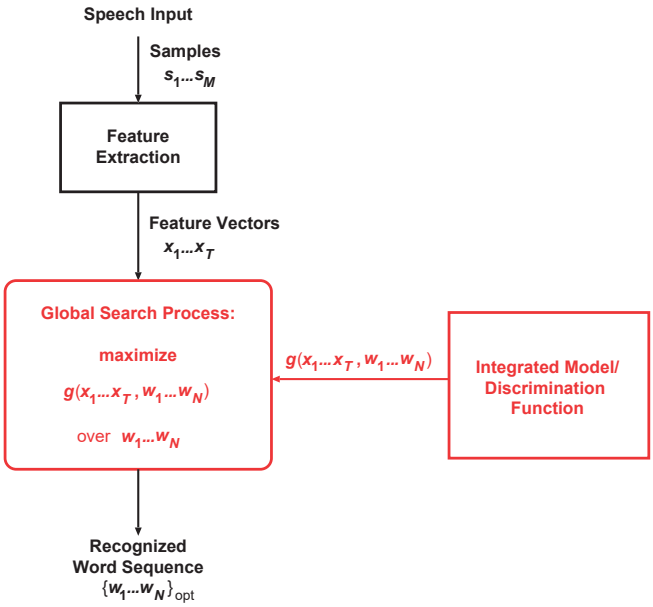

Figure 9: Architecture of a statistical speech recognition system with integrated sequence modeling, like in the encoder/attention/decoder approach [Bahdanau & Chorowski$^+$ 2016, Chan & Jaitly$^+$ 2015], segmental modeling [Lu & Kong$^+$ 2016], or the related inverted HMM approach [Doetsch & Hegselmann$^+$ 2016].

11. Sequence Modeling

While neural networks proved a powerful tool in local and sequential classification tasks, segmentation and length modeling is still entirely done within an HMM framework. Although CTC [Graves & Fernández$^+$ 2006] provides a simpler topology, it could still be seen as realizing a specific HMM topology, where search can be done using standard HMM-based implementations. Recently, a new approach was introduced to integrate segmentation and length modeling into a recurrent topology [Bahdanau & Cho$^+$ 2015]. These so called *end-to-end* systems separate the input and output handling into two different models: an *encoder*, which reads the input and is trained to compute discriminative features from the observations, and a *decoder*, which produces the desired output target sequence label-by-label by utilizing the encoded features. The decoder includes modeling of label (word) context, thus even integrating language modeling to the extent of utilizing labeled acoustic training data, while the encoder is designed to generate significant representations that are neither constrained by input nor output length. The topology of these models is furthermore closely related to generative RNNs [Graves 2013] and has been applied to ASR tasks, already [Bahdanau & Chorowski$^+$ 2016, Chan & Jaitly$^+$ 2015]. In the decoder, length modeling is done by including and hypothesizing an *end-of-sequence* symbol. In its simplest form, the encoder only produces a single final activation that is subsequently used to initialize the decoder. However, the performance of these models quickly degrades even for moderately long sequences (around 10 symbols, depending on the recurrent cell used) [Bahdanau & Cho$^+$ 2015]. It is furthermore commonly seen as unfavorable that an output sequence of arbitrary size is encoded into a fixed size representation and the effective capacity of those representations is not entirely understood. To account for arbitrary input length, several so-called *attention-mechanisms* were developed [Bahdanau & Cho$^+$ 2015, Chan & Jaitly$^+$ 2015]. Here, at each decoder step, an expected input is computed as a normalized and statistically localized linear combination of all features provided by the encoder. End-to-end approaches usually employ beam search with static beams of limited sizes; their recognition results do not yet outperform hybrid HMM approaches [Chan & Jaitly$^+$ 2015] for tasks with limited amount of training data. However, recently encoder-attention-devoder approaches where shown to outperform standard hybrid HMM approaches for tasks using 1000h and more training data in [Chiu & Sainath$^+$ 2017, Zeyer & Irie$^+$ 2018]. Another approach to find a consistent integration of discriminative neural network based acoustic models into a speech recognition framework is the segmental [Lu & Kong$^+$ 2016] or inverted HMM approach [Doetsch & Hegselmann$^+$ 2016]. Here, the standard HMM alignment of HMM states to observation frames is inverted to aligning observation frame segments to output symbols. Where in the standard HMM approach the hidden alignment variables are the HMM states, in the inverted approach the hidden alignment variables are the observation frame segment boundaries. In [Beck & Hannemann$^+$ 2018] competetive results are presented for large vocabulary automatic speech recognition and handwriting recognition for small to medium sized training corpora.

12. Conclusions

In this work, an overview of recent developments in automatic speech recognition using neural network based modeling approaches was presented. Approaches for both acoustic and language modeling already show considerable improvements, especially using recurrent topologies, and combined with discriminative sequence training criteria. Currently, improvements of approx. 50% relative in word error rate are observed when replacing GMM/HMM acoustic models and count-based language models by recurrent long-short term memory (LSTM) based acoustic and language models. More independence from e.g. GMM-based model initialization, or separate modeling aspects like CART would be desirable to obtain more consistent neural network based modeling, which is covered in the area of end-to-end approaches, which e.g. attention approaches relate to. Also, further exploitation of the high potential of neural network based modeling, especially w.r.t. the integration of recurrent models into the decision rule, as well as consistent modeling, training, and decoding w.r.t. the target evaluation measure word error rate seem worthwhile. Further improvements are to be expected from also integrating the search problem together with acoustic and language modeling into a single, joint model, like in encoder-attention-decoder based approaches. However, a better understanding of the learned networks is needed to support training tasks of varying size and quality, and to take advantage of separate training data for acoustic (transcribed speech) and language models (text only).

References

[Abdel-Hamid & Mohamed$^+$ 2012] O. Abdel-Hamid, A. Mohamed, H. Jiang, G. Penn: "Applying Convolutional Neural Networks Concepts to Hybrid NN-HMM Model for Speech Recognition", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4277–4280, Kyoto, Japan, March 2012.

[Babel] Babel: *http://www.iarpa.gov/Programs/ia/Babel/babel.html*, US IARPA Project, 2012–2016.

[Bahdanau & Cho$^+$ 2015] D. Bahdanau, K. Cho, Y. Bengio: "Neural Machine Translation by Jointly Learning to Align and Translate", *Proc. Intern. Conf. on Learning Representations (ICLR)*, San Diego, CA, May 2015.

[Bahdanau & Chorowski$^+$ 2016] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, Y. Bengio: "End-to-End Attention-Based Large Vocabulary Speech Recognition", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4945–4949, Shanghai, China, March 2016.

[Bahdanau & Serdyuk+ 2015] D. Bahdanau, D. Serdyuk, P. Brakel, N.R. Ke, J. Chorowski, A.C. Courville, Y. Bengio: "Task Loss Estimation for Sequence Prediction", *CoRR*, Vol. abs/1511.06456, 2015.

[Beck & Hannemann+ 2018] E. Beck, M. Hannemann, P. Dötsch, , R. Schlüter, H. Ney: "Segmental Encoder-Decoder Models for Large Vocabulary Automatic Speech Recognition", Sept. 2018.

[Bengio & Ducharme+ 2000] Y. Bengio, R. Ducharme, P. Vincent: "A Neural Probabilistic Language Model", *Proc. Advances in Neural Information Processing Systems (NIPS)*, Vol. 13, pp. 932–938, Denver, CO, Nov. 2000.

[Bourlard & Morgan 1993] H.A. Bourlard, N. Morgan: *Connectionist Speech Recognition: a Hybrid Approach*, Kluwer Academic Publishers, Norwell, MA, 1993.

[Bourlard & Wellekens 1989] H. Bourlard, C.J. Wellekens: "Links between Markov Models and Multilayer Perceptrons", in D. Touretzky (Edt.), *Advances in Neural Information Processing Systems I*, pp. 502–510, Morgan Kaufmann, San Mateo, CA, 1989.

[Breuel 2015] T.M. Breuel: "Benchmarking of LSTM Networks", *arXiv preprint arXiv:1508.02774*, Vol., 2015.

[Bridle 1989] J.S. Bridle: "Probabilistic Interpretation of Feedforward Classification Network Outputs with Relationships to Statistical Pattern Recognition", in F. Fogelmann Soulié, J. Hérault (Edts.), *Nato ASI Series F: Computer and Systems Sciences*, Vol. 68, chapter Neurocomputing, pp. 227–236, Springer, Berlin, Heidelberg, 1989.

[Burget & Schwarz+ 2010] L. Burget, P. Schwarz, M. Agarwal, P. Akayazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey, A. Rastrow, R.C. Rose, S. Thomas: "Multilingual Acoustic Modeling for Speech Recognition Based on Subspace Gaussian Mixture Models", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4334–4337, 2010.

[Byrne & Beyerlein+ 2000] W. Byrne, P. Beyerlein, J.M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, W. Wang: "Towards Language Independent Acoustic Modeling", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 2, pp. 1029–1032, 2000.

[Caruana 1993] R. Caruana: "Multitask Learning: A Knowledge-Based Source of Inductive Bias", *Proc. Intern. Conf. on Machine Learning (ICML)*, pp. 41–48, 1993.

[Chan & Jaitly+ 2015] W. Chan, N. Jaitly, Q.V. Le, O. Vinyals: "Listen, Attend and Spell", *CoRR*, Vol. abs/1508.01211, 2015.

[Chen & Liu+ 2015] X. Chen, X. Liu, M. Gales, P. Woodland: "Investigation of Back-Off Based Interpolation between Recurrent Neural Network and N-Gram Language Models", *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 181–186, Scottsdale, AZ, Dec. 2015.

[Chiu & Sainath+ 2017] C.C. Chiu, T.N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R.J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, M. Bacchiani: "State-of-the-art Speech Recognition With Sequence-to-Sequence Models", *arXiv preprint arXiv:1712.01769*, Vol., Dec. 2017.

[Chung & Gülçehre+ 2014] J. Chung, Ç. Gülçehre, K. Cho, Y. Bengio: "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling", *CoRR*, Vol. abs/1412.3555, 2014.

[Clevert & Unterthiner+ 2016] D. Clevert, T. Unterthiner, S. Hochreiter: "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)", *Proc. Intern. Conf. on Learning Representations (ICLR)*, San Juan, Puerto Rico, May 2016.

[Davis & Mermelstein 1980] S. Davis, P. Mermelstein: "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 28, No. 4, pp. 357–366, Aug. 1980.

[Dean & Corrado+ 2012] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M.A. Ranzato, A. Senior, P. Tucker, K. Yang, Q.V. Le, A.Y. Ng: "Large Scale Distributed Deep Networks", in F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (Edts.), *Advances in Neural Information Processing Systems (NIPS)*, pp. 1223–1231, Nips Foundation (http://books.nips.cc), 2012.

[Doetsch & Hegselmann+ 2016] P. Doetsch, S. Hegselmann, R. Schlüter, H. Ney: "Inverted HMM - a Proof of Concept", *Proc. Advances in Neural Information Processing Systems (NIPS)*, Barcelona, Spain, Dec. 2016.

[Doetsch & Zeyer+ 2017] P. Doetsch, A. Zeyer, P. Voigtlaender, I. Kulikov, R. Schlüter, H. Ney: "RETURNN: The RWTH Extensible Training framework for Universal Recurrent Neural Networks", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5345–5349, New Orleans, LA, March 2017.

[Duchi & Hazan+ 2010] J. Duchi, E. Hazan, Y. Singer: "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization", Technical Report UCB/EECS-2010-24, EECS Department, University of California, Berkeley, Mar 2010.

[Fontaine & Ris+] V. Fontaine, C. Ris, J.M. Boite:

[Geiger & Zhang+ 2014] J.T. Geiger, Z. Zhang, F. Weninger, B. Schuller, G. Rigoll: "Robust Speech Recognition using Long Short-Term Memory Recurrent Neural Networks for Hybrid Acoustic Modelling", *Proc. Interspeech*, pp. 631–635, 2014.

[Golik & Doetsch+ 2013] P. Golik, P. Doetsch, H. Ney: "Cross-Entropy vs. Squared Error Training: a Theoretical and Experimental Comparison", *Proc. Interspeech*, pp. 1756–1760, Lyon, France, Aug. 2013.

[Golik & Tüske+ 2015a] P. Golik, Z. Tüske, R. Schlüter, H. Ney: "Convolutional Neural Networks for Acoustic Modeling of Raw Time Signal in LVCSR", *Proc. Interspeech*, pp. 26–30, Dresden, Germany, Sept. 2015.

[Golik & Tüske+ 2015b] P. Golik, Z. Tüske, R. Schlüter, H. Ney: "Multilingual Features Based Keyword Search for Very Low-Resource Languages", *Proc. Interspeech*, pp. 1260–1264, Dresden, Germany, Sept. 2015.

[Golik & Tüske+ 2017] P. Golik, Z. Tüske, K. Irie, E. Beck, R. Schlüter, H. Ney: "The 2016 RWTH Keyword Search System for Low-Resource Languages", in I.M. Alexey Karpov, Rodmonga Potapova (Edt.), *International Conference Speech and Computer*, Vol. 10458 of *Lecture Notes in Computer Science, Subseries Lecture Notes in Artificial Intelligence*, pp. 719–730, Hatfield, UK, Sept. 2017, Springer Cham, Switzerland.

[Goodfellow & Warde-Farley+ 2013] I.J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, Y. Bengio: "Maxout Networks", *Proc. Intern. Conf. on Machine Learning (ICML)*, Atlanta, GA, June 2013.

[Graves 2013] A. Graves: "Generating Sequences with Recurrent Neural Networks", *CoRR*, Vol. abs/1308.0850, 2013.

[Graves & Fernández+ 2006] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber: "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks", *Proc. Intern. Conf. on Machine Learning (ICML)*, pp. 369–376, New York, NY, 2006, ACM.

[Graves & Mohamed+ 2013] A. Graves, A.r. Mohamed, G. Hinton: "Speech Recognition with Deep Recurrent Neural Networks", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6645–6649, IEEE, 2013.

[Graves & Schmidhuber 2005] A. Graves, J. Schmidhuber: "Framewise Phoneme Classification with Bidirectional Lstm and other Neural Network Architectures", *Neural Networks*, Vol. 18, No. 5, pp. 602–610, 2005.

[Greff & Srivastava+ 2015] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, J. Schmidhuber: "LSTM: A Search Space Odyssey", *arXiv preprint arXiv:1503.04069*, Vol., 2015.

[Grézl & Karafiát+ 2007] F. Grézl, M. Karafiát, S. Kontár, J. Černocký: "Probabilistic and Bottleneck Features for LVCSR of Meetings", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 757–760, Honolulu, HI, April 2007.

[Grézl & Karafiát+ 2011] F. Grézl, M. Karafiát, M. Janda: "Study of Probabilistic and Bottle-Neck Features in Multilingual Environment", *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 359–364, 2011.

[Gülçehre & Bengio 2014] Ç. Gülçehre, Y. Bengio: "ADASECANT: Robust Adaptive Secant Method for Stochastic Gradient", *CoRR*, Vol. abs/1412.7419, 2014.

[He & Deng+ 2008] X. He, L. Deng, W. Chou: "Discriminative Learning in Sequential Pattern Recognition – A Unifying Review for Optimization-Oriented Speech Recognition", *IEEE Signal Processing Magazine*, Vol. , 2008.

[Heigold & Schlüter+ 2012] G. Heigold, R. Schlüter, H. Ney, S. Wiesler: "Discriminative Training for Automatic Speech Recognition: Modeling, Criteria, Optimization, Implementation, and Performance", *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 58–69, Nov. 2012.

[Hermansky 1990] H. Hermansky: "Perceptual Linear Predictive (PLP) Analysis of Speech", *Journal of the Acoustical Society of America*, Vol. 87, No. 4, pp. 1738–1752, 1990.

[Hermansky & Ellis+ 2000] H. Hermansky, D. Ellis, S. Sharma: "Tandem connectionist Feature Extraction for Conventional HMM Systems", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 3, pp. 1635–1638, Istanbul, Turkey, June 2000.

[Hermansky & Morgan 1994] H. Hermansky, N. Morgan: "RASTA Processing of Speech", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 4, pp. 578–589, Oct 1994.

[Heymann & Drude+a] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, R. Häb-Umbach: "Beamnet: End-to-End Training of a Beamformer-Supported Multi-Channel ASR System", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 5325.

[Heymann & Drude+b] J. Heymann, L. Drude, A. Chinaev, R. Häb-Umbach: "BLSTM Supported GEV Beamformer Front-End for the 3rd CHiME Challenge", *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 444.

[Hinton & Osindero+ 2006] G.E. Hinton, S. Osindero, Y.W. Teh: "A Fast Learning Algorithm for Deep Belief Nets", *Neural Computation*, Vol. 18, No. 7, pp. 1527–1554, July 2006.

[Hochreiter & Bengio+ 2001] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber: "Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term De-

pendencies", in J. Kolen, S. Kremer (Edts.), *A Field Guide to Dynamical Recurrent Networks*, IEEE Press, New York, 2001.

[Hochreiter & Schmidhuber 1997] S. Hochreiter, J. Schmidhuber: "Long Short-Term Memory", *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[Hornik & Stinchcombe+ 1989] K. Hornik, M.B. Stinchcombe, H. White: "Multilayer Feedforward Networks Are Universal Approximators", *Neural Networks*, Vol. 2, No. 5, pp. 359–366, July 1989.

[Huang & Lu+ 2018] Z. Huang, H. Lu, M. Lei, Z. Yan: "Linear networks based speaker adaptation for speech synthesis", Vol., March 2018.

[Huang & Sun+ 2016] G. Huang, Y. Sun, Z. Liu, D. Sedra, K. Weinberger: "Deep Networks with Stochastic Depth", *arXiv preprint arXiv:1603.09382*, Vol., 2016.

[Irie & Lei+ 2018] K. Irie, Z. Lei, L. Deng, R. Schlüter, H. Ney: "Investigation on Estimation of Sentence Probability By Combining Forward, Backward and Bi-directional LSTM-RNNs", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, AB, April 2018.

[Irie & Tüske+ 2016] K. Irie, Z. Tüske, T. Alkhouli, R. Schlüter, H. Ney: "LSTM, GRU, Highway and a Bit of Attention: An Empirical Overview for Language Modeling in Speech Recognition", *Proc. Interspeech*, pp. 3519–3523, San Francisco, CA, Sept. 2016.

[Jozefowicz & Zaremba+ 2015] R. Jozefowicz, W. Zaremba, I. Sutskever: "An Empirical Exploration of Recurrent Network Architectures", *Proc. Intern. Conf. on Machine Learning (ICML)*, pp. 2342–2350, 2015.

[Kingma & Ba 2014] D.P. Kingma, J. Ba: "Adam: A Method for Stochastic Optimization", *CoRR*, Vol. abs/1412.6980, 2014.

[Kingsbury 2009] B. Kingsbury: "Lattice-based Optimization of Sequence Classification Criteria for Neural-Network Acoustic Modeling", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3761–3764, Taipei, Taiwan, April 2009.

[Kingsbury & Sainath+ 2012] B. Kingsbury, T.N. Sainath, H. Soltau: "Scalable Minimum Bayes Risk Training of Deep Neural Network Acoustic Models Using Distributed Hessian-free Optimization", *Proc. Interspeech*, Portland, OR, Sept. 2012.

[Kitza & Schlüter+ 2018] M. Kitza, R. Schlüter, H. Ney: "Comparison of BLSTM-Layer-Specific Affine Transformations for Speaker Adaptation", Sept. 2018.

[LeCun & Boser+ 1990] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel: "Handwritten Digit Recognition with a Back-Propagation Network", *Proc. Advances in Neural Information Processing Systems (NIPS)*, Vol. 2, Denver, CO, Nov. 1990.

[Li & Sim 2010] B. Li, K.C. Sim: "Comparison of Discriminative Input and Output Transformations for Speaker Adaptation in the Hybrid NN/HMM Systems", *Proc. Interspeech*, pp. 526–529, Makuhari, Japan, Sept. 2010.

[Lippmann 1989] R.P. Lippmann: "Review of Neural Networks for Speech Recognition", *Neural Computation*, Vol. 1, No. 1, pp. 1–38, March 1989.

[Liu & Wang+ 2016] C. Liu, Y. Wang, K. Kumar, Y. Gong: "Investigations on speaker adaptation of LSTM RNN models for speech recognition", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5020–5024, Shanghai, China, March 2016.

[Lu & Kong+ 2016] L. Lu, L. Kong, C. Dyer, N.A. Smith, S. Renals: "Segmental Recurrent Neural Networks for End-to-End Speech Recognition", *Proc. Interspeech*, pp. 385–389, San Francisco, CA, Sept. 2016.

[Melis & Dyer+ 2017] G. Melis, C. Dyer, P. Blunsom: "On the State of the Art of Evaluation in Neural Language Models", *arXiv preprint arXiv:1707.05589*, Vol., July 2017.

[Menne & Ney 2018] S.R. Menne, Tobias, H. Ney: "Speaker Adapted Beamforming for Multi-Channel Automatic Speech Recognition", Sept. 2018.

[Miao & Metze 2015] Y. Miao, F. Metze: "Distance-Aware DNNs for Robust Speech Recognition", *Proc. Interspeech*, pp. 761–765, Dresden, Germany, Sept. 2015.

[Mikolov & Karafiát+ 2010] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, S. Khudanpur: "Recurrent Neural Network Based Language Model.", *Proc. Interspeech*, pp. 1045–1048, Makuhari, Japan, Sept. 2010.

[Montufar & Pascanu+ 2014] G.F. Montufar, R. Pascanu, K. Cho, Y. Bengio: "On the Number of Linear Regions of Deep Neural Networks", *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2924–2932, 2014.

[Nair & Hinton 2010] V. Nair, G.E. Hinton: "Rectified Linear Units Improve Restricted Boltzmann Machines", *Proc. Intern. Conf. on Machine Learning (ICML)*, pp. 807–814, Haifa, Israel, June 2010.

[Nakamura & Shikano 1989] M. Nakamura, K. Shikano: "A Study of English Word Category Prediction Based on Neural Networks", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 731–734, Glasglow, UK, May 1989.

[Pascanu & Mikolov+ 2012] R. Pascanu, T. Mikolov, Y. Bengio: "On the Difficulty of Training Recurrent Neural Networks", *arXiv preprint arXiv:1211.5063*, Vol., 2012.

[Plahl & Kozielski+ 2013] C. Plahl, M. Kozielski, R. Schlüter, H. Ney: "Feature Combination and Stacking of Recurrent and Non-recurrent Neural Networks for LVCSR", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6714–6718, Vancouver, Canada, May 2013.

[Plahl & Schlüter+ 2011] C. Plahl, R. Schlüter, H. Ney: "Cross-Lingual Portability of Chinese and English Neural Network Features for French and German LVCSR", *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 371–376, 2011.

[Prasad & Sim 2016] A. Prasad, K.C. Sim: "Microphone Distance Adaptation Using Cluster Adaptive Training for Robust Far Field Speech Recognition", *Proc. Interspeech*, pp. 3823–3827, San Francisco, CA, September 2016.

[Qian & Tan+ 2016] Y. Qian, T. Tan, D. Yu, Y. Zhang: "Integrated Adaptation with Multi-Factor Joint-Learning for Far-Field Speech Recognition", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1–5, Shanghai, China, 2016.

[Robinson & Hochberg+ 1994] T. Robinson, M. Hochberg, S. Renals: "IPA: Improved Phone Modelling with Recurrent Neural Networks", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. I, pp. 37–40, April 1994.

[Rumelhart & Hinton+ 1986] D.E. Rumelhart, G.E. Hinton, R.J. Williams: "Learning Representations By Back-Propagating Errors", *Nature*, Vol. 323, pp. 533–536, Oct. 1986.

[Sainath & Weiss+ 2015a] T.N. Sainath, R.J. Weiss, A. Senior, K.W. Wilson, O. Vinyals: "Learning the Speech Front-End with Raw Waveform CLDNNs", *Proc. Interspeech*, pp. 1–5, 2015.

[Sainath & Weiss+ 2015b] T.N. Sainath, R.J. Weiss, K.W. Wilson, A. Narayanan, M. Bacchiani, A. Senior: "Speaker Location And Microphone Spacing Invariant Acoustic Modeling from Raw Multichannel Waveforms", *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 30–36, 2015.

[Sak & Senior+ 2014] H. Sak, A. Senior, F. Beaufays: "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling", *Proc. Interspeech*, pp. 338–342, Singapore, Sept. 2014.

[Samarakoon & Sim+ 2017] L. Samarakoon, K.C. Sim, B. Mak: "An investigation into learning effective speaker subspaces for robust unsupervised DNN adaptation", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5035–5039, New Orleans, LA, March 2017.

[Saon & Kurata+ 2017] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.L. Lim, B. Roomi, P. Hall: "English Conversational Telephone Speech Recognition by Humans and Machines", *arXiv preprint arXiv:1703.02136*, Vol., 2017.

[Saon & Soltau+ 2013] G. Saon, H. Soltau, D. Nahamoo, M. Picheny: "Speaker Adaptation of Neural Network Acoustic Models Using i-Vectors", *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 55–59, Olomouc, Czech Republic, Dec. 2013.

[Scanzio & Laface+ 2008] S. Scanzio, P. Laface, L. Fissore, R. Gemello, F. Mana: "On the Use of a Multilingual Neural Network Front-End", *Proc. Interspeech*, pp. 2711–2714, 2008.

[Schaaf & Metze 2010] T. Schaaf, F. Metze: "Analysis of Gender Normalization Using MLP and VTLN Features", *Proc. Interspeech*, pp. 306–309, 2010.

[Schlüter & Bezrukov+ 2007] R. Schlüter, I. Bezrukov, H. Wagner, H. Ney: "Gammatone Features and Feature Combination for Large Vocabulary Speech Recognition", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 649–652, 2007.

[Schlüter & Doetsch+ 2016] R. Schlüter, P. Doetsch, P. Golik, M. Kitza, T. Menne, K. Irie, Z. Tüske, A. Zeyer: "Automatic Speech Recognition Based on Neural Networks", *Proc. International Conference Speech and Computer (SPECOM), Lecture Notes in Computer Science, Subseries Lecture Notes in Artificial Intelligence*, Vol. 9811, pp. 3–17, Budapest, Hungary, Aug. 2016.

[Schultz & Waibel 1997] T. Schultz, A. Waibel: "Fast Bootstrapping Of LVCSR Systems With Multilingual Phoneme Sets", *Proc. European Conf. on Speech Communication and Technology (Eurospeech)*, 1997.

[Seide & Li+ 2011a] F. Seide, G. Li, X. Chen, D. Yu: "Feature Engineering in Context-Dependent Deep Neural Networks for Conversational Speech Transcription", *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 24–29, Waikoloa, HI, Dec. 2011.

[Seide & Li+ 2011b] F. Seide, G. Li, D. Yu: "Conversational Speech Transcription using Context-Dependent Deep Neural Networks", *Proc. Interspeech*, pp. 437–440, Florence, Italy, Aug. 2011.

[Sonoda & Murata 2016] S. Sonoda, N. Murata: "Neural Network with Unbounded Activation Functions is Universal Approximator", *Applied and Computational Harmonic Analysis*, Vol., 2016, In Press, Corrected Proof, Available online 17 December 2015.

[Srivastava & Greff+ 2015] R.K. Srivastava, K. Greff, J. Schmidhuber: "Training Very Deep Networks", *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2368–2376, 2015.

[Srivastava & Hinton+ 2014] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov: "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", *The Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1929–1958, 2014.

[Stolcke & Grézl+ 2006] A. Stolcke, F. Grézl, M.Y. Hwang, X. Lei, N. Morgan, D. Vergyri: "Cross-Domain and Cross-Language Portability of Acoustic Features Estimated by Multilayer Perceptrons", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 321–324, 2006.

[Sundermeyer & Ney+ 2015] M. Sundermeyer, H. Ney, R. Schlüter: "From Feedforward to Recurrent LSTM Neural Networks for Language Modeling", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 3, pp. 517–529, March 2015.

[Sundermeyer & Schlüter+ 2012] M. Sundermeyer, R. Schlüter, H. Ney: "LSTM Neural Networks for Language Modeling.", *Proc. Interspeech*, pp. 194–197, Portland, OR, Sept. 2012.

[Sundermeyer & Tüske+ 2014] M. Sundermeyer, Z. Tüske, R. Schlüter, H. Ney: "Lattice Decoding and Rescoring with Long-Span Neural Network Language Models", *Proc. Interspeech*, pp. 661–665, Singapore, Sept. 2014.

[Swietojanski & Renais 2016] P. Swietojanski, S. Renais: "SAT-LHUC: Speaker adaptive training for learning hidden unit contributions", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5010–5014, Shanghai, China, March 2016, IEEE.

[Thomas & Ganapathy+ 2010] S. Thomas, S. Ganapathy, H. Hermansky: "Cross-Lingual and Multistream Posterior Features for low Resource LVCSR Systems", *Proc. Interspeech*, pp. 877–880, 2010.

[Tóth & Frankel+ 2008] L. Tóth, J. Frankel, G. Gosztolya, S. King: "Cross-Lingual Portability of MLP-Based Tandem Features–A Case Study for English and Hungarian", *Proc. Interspeech*, pp. 2695–2698, 2008.

[Tüske & Golik+ 2014a] Z. Tüske, P. Golik, D. Nolden, R. Schlüter, H. Ney: "Data Augmentation, Feature Combination, and Multilingual Neural Networks to Improve ASR and KWS Performance for Low-Resource Languages", *Proc. Interspeech*, pp. 1420–1424, Singapore, Sept. 2014.

[Tüske & Golik+ 2014b] Z. Tüske, P. Golik, R. Schlüter, H. Ney: "Acoustic Modeling with Deep Neural Networks Using Raw Time Signal for LVCSR", *Proc. Interspeech*, pp. 890–894, Singapore, Sept. 2014.

[Tüske & Golik+ 2015] Z. Tüske, P. Golik, R. Schlüter, H. Ney: "Speaker Adaptive Joint Training of Gaussian Mixture Models and Bottleneck Features", *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 596–603, Scottsdale, AZ, Dec. 2015.

[Tüske & Irie+ 2016] Z. Tüske, K. Irie, R. Schlüter, H. Ney: "Investigation on Log-Linear Interpolation of Multi-Domain Neural Network Language Model", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6005–6009, Shanghai, China, March 2016.

[Tüske & Michel+ 2017] Z. Tüske, W. Michel, R. Schlüter, H. Ney: "Parallel Neural Network Features for Improved Tandem Acoustic Modeling", *Proc. Interspeech*, pp. 1651–1655, Stockholm, Sweden, Aug. 2017.

[Tüske & Nolden+ 2014] Z. Tüske, D. Nolden, R. Schlüter, H. Ney: "Multilingual MRASTA Features for Low-Resource Keyword Search and Speech Recognition Systems", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[Tüske & Schlüter+ 2013] Z. Tüske, R. Schlüter, H. Ney: "Multilingual Hierarchical MRASTA Features for ASR", *Proc. Interspeech*, pp. 2222–2226, Lyon, France, Aug. 2013.

[Tüske & Schlüter+ 2018] Z. Tüske, R. Schlüter, H. Ney: "Investigation on LSTM Recurrent N-gram Language Models for Speech Recognition", April 2018.

[Tüske & Sundermeyer+ 2012] Z. Tüske, M. Sundermeyer, R. Schlüter, H. Ney: "Context-Dependent MLPs for LVCSR: TANDEM, Hybrid or Both?", *Proc. Interspeech*, pp. 18–21, Portland, OR, Sept. 2012.

[Tüske & Tahir+ 2015] Z. Tüske, M.A. Tahir, R. Schlüter, H. Ney: "Integrating Gaussian Mixtures into Deep Neural Networks: Softmax Layer with Hidden Variables", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4285–4289, Brisbane, Australia, April 2015.

[Valente & Vepa+ 2007] F. Valente, J. Vepa, C. Plahl, C. Gollan, H. Hermansky, R. Schlüter: "Hierarchical Neural Networks Feature Extraction for LVCSR System", *Proc. Interspeech*, pp. 42–45, Antwerp, Belgium, Aug. 2007.

[Waibel & Hanazawa+ 1989] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang: "Phoneme Recognition: Neural Networks vs. Hidden Markov Models", *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 1, pp. 107–110, April 1989.

[Wang & Wang 2017] Z.Q. Wang, D. Wang: "Unsupervised speaker adaptation of batch normalized acoustic models for robust ASR", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4890–4894, New Orleans, LA, March 2017.

[Wiesler & Golik+ 2015] S. Wiesler, P. Golik, R. Schlüter, H. Ney: "Investigations on Sequence Training of Neural Networks", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*

(ICASSP), pp. 4565–4569, Brisbane, Australia, April 2015.

[Wiesler & Li+ 2013] S. Wiesler, J. Li, J. Xue: "Investigations on Hessian-Free Optimization for Cross-Entropy Training of Deep Neural Networks", *Proc. Interspeech*, pp. 3317–3321, Lyon, France, Aug. 2013.

[Wiesler & Richard+ 2014] S. Wiesler, A. Richard, R. Schlüter, H. Ney: "Mean-normalized Stochastic Gradient for Large-Scale Deep Learning", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 180–184, Florence, Italy, May 2014.

[Xiong & Wu+ 2017] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, A. Stolcke: "The Microsoft 2017 Conversational Speech Recognition System", *arXiv preprint arXiv:1708.06073*, Vol., 2017.

[Xue & Li+ 2014] J. Xue, J. Li, D. Yu, M. Seltzer, Y. Gong: "Singular Value Decomposition based Low-Footprint Speaker Adaptation and Personalization for Deep Neural Network", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6359–6363, Florence, Italy, May 2014.

[Zeiler 2012] M.D. Zeiler: "ADADELTA: An Adaptive Learning Rate Method", *CoRR*, Vol. abs/1212.5701, 2012.

[Zeyer & Doetsch+ 2017] A. Zeyer, P. Doetsch, P. Voigtlaender, R. Schlüter, H. Ney: "A Comprehensive Study of Deep Bidirectional LSTM RNNs for Acoustic Modeling in Speech Recognition", *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2462–2466, New Orleans, LA, March 2017.

[Zeyer & Irie+ 2018] A. Zeyer, K. Irie, R. Schlüter, H. Ney: "Improved Training of End-to-End Attention Models for Speech Recognition", Sept. 2018.

[Zeyer & Schlüter+ 2016] A. Zeyer, R. Schlüter, H. Ney: "Towards Online-Recognition with Deep Bidirectional LSTM Acoustic Models", *Proc. Interspeech*, San Francisco, CA, Sept. 2016.

[Zhang & Chen+ 2015] Y. Zhang, G. Chen, D. Yu, K. Yao, S. Khudanpur, J. Glass: "Highway Long Short-Term Memory RNNs for Distant Speech Recognition", *arXiv preprint arXiv:1510.08983*, Vol., 2015.