University of Szeged

Ph.D thesis

Training Methods for Deep Neural Network-Based Acoustic Models in Speech Recognition

Author:
Tamás Grósz

Supervisor:
Dr. László Tóth

PhD School in Computer Science
MTA-SZTE Research Group on Artificial Intelligence
Institute of Informatics
University of Szeged

Szeged, 2018

1 Introduction

Automatic Speech Recognition (ASR) is a key topic of speech technology, where the goal is to transcribe an audio recording (an *utterance*) in an automatic way. For decades the traditional ASR systems used Hidden Markov Models (HMM) with Gaussian Mixture Models (GMM) and, until very recently, these HMM/GMM models represented the state-of-the-art technology in ASR. Nowadays, with the advent of Deep Neural Networks (DNN) the original HMM/GMM models have been replaced by the new HMM/DNN hybrids (shown in Figure 1) [1]. DNNs are a new type of Artificial Neural Networks, which differ in one important aspect from the previous ones, namely that they have many hidden layers. The addition of extra hidden layers creates several problems that make the training of these networks hard. So besides adding new hidden layers, other modifications are also needed, like changing the activation function of the neurons or the learning algorithm itself.

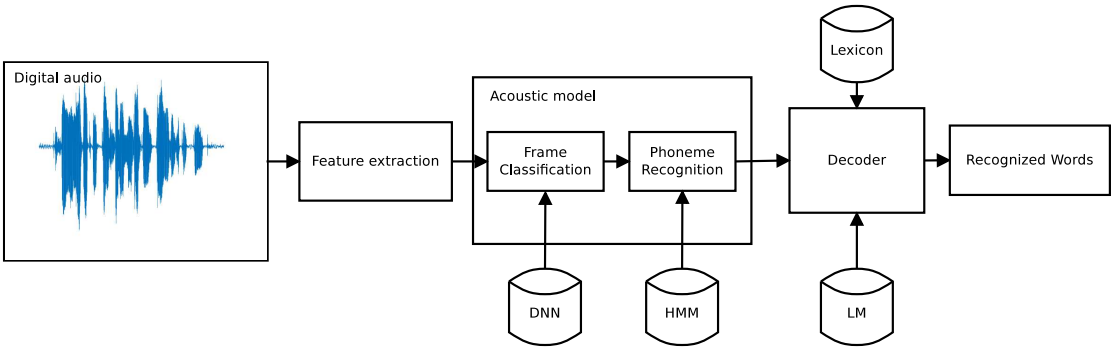

Figure 1: The standard workflow of a HMM/DNN-based ASR system.

The new HMM/DNN hybrids are now routinely used in state-of-the-art ASR systems, but they inherited many of the algorithms from their predecessors (the standard HMM/GMM systems). However, the optimality of these algorithms is not guaranteed with the new models. In this dissertation we describe how we modified some of these earlier methods in speech recognition, so that they better suit the new DNN-based acoustic models. Our main goal is to create new solutions that allow the training of HMM/DNN acoustic models without relying on GMMs during the training process. To achieve the GMM-free training of a HMM/DNN hybrid, we have to solve two key problems, namely the initial alignment of the frame-level state labels and the creation of context-dependent (CD) states.

The methods proposed here will be evaluated using various English and Hungarian corpora, but for the sake of continuity, the Szeged Hungarian Broadcast News Corpus [2] will be used in all chapters as a large vocabulary continuous speech recognition (LVCSR) task.

2 A Comparison of Deep Neural Network Training Methods for Large Vocabulary Continuous Speech Recognition

The second chapter focuses on comparing the performance of four DNN training algorithms. The first one is the original algorithm proposed by Hinton et al. [3] (*DBN*), and the second one is called discriminative pre-training (*DPT*) by Seide et al. [4]. Both of these methods apply a pre-training phase before they finetune the DNN. Deep Rectifier Network [5] (*RECT*), our third approach, differs greatly from the previous two in the sense that it modifies the activation of the hidden neurons instead of the training process. The fourth training algorithm that we examined is a regularisation method called Dropout [6], which simply turns off neurons during training. The Dropout method was applied with standard sigmoid networks (*Sigmoid-DO*) and with rectified ones as well (*RECT-DO*).

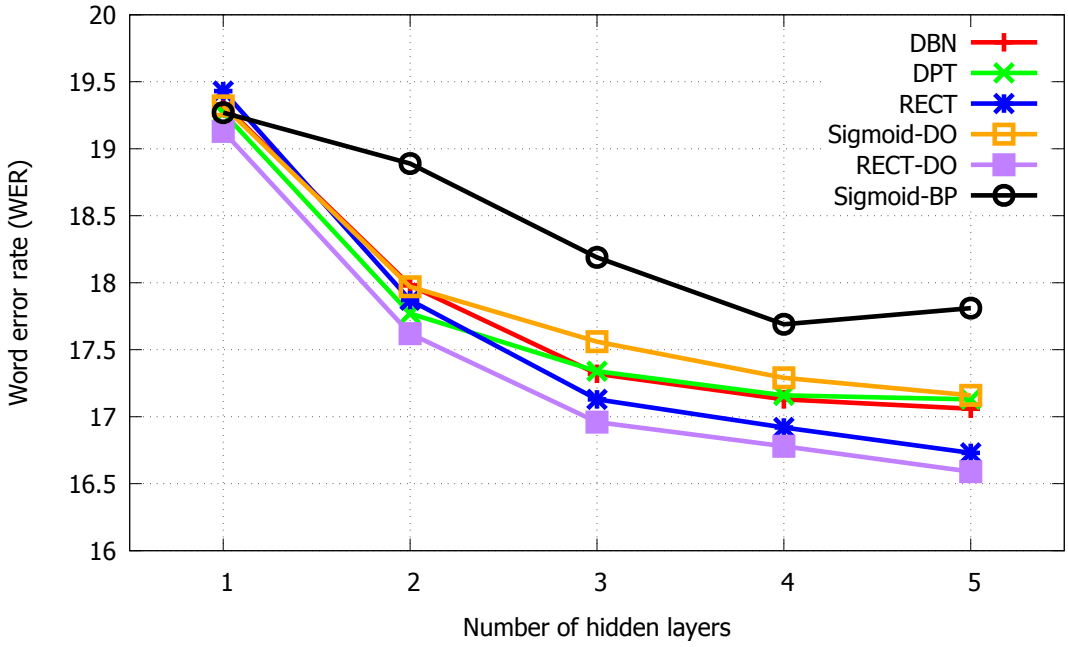

Figure 2: Word error rates for the broadcast news corpus as a function of the number of hidden layers.

In our experiments, we compared the recognition accuracies of these methods on the Szeged Hungarian Broadcast News Corpus. Figure 2 shows the word error rates (WER) got by using different methods. As can be seen, the four algorithms yielded quite similar recognition results, but rectifier networks achieved better accuracy scores and their training was considerably faster. Based on these findings, in my later experiments deep rectifier networks became the preferred choice.

Database	Method		Dev. set	Test set
TIMIT [7]	Monostate (39)	CTC + DRN	**26.69%**	**28.60%**
		MMI + DRN	27.70%	30.94%
		Hand-labeled	27.26%	29.35%
		Forced Alignment	27.10%	28.92%
	Monostate (61)	CTC + DRN	**26.07%**	**27.34%**
		MMI + DRN	25.16%	27.89%
		Hand-labeled	26.42%	27.94%
		Forced Alignment	25.92%	27.55%
	Tristate (183)	CTC + DRN	23.20%	24.41%
		MMI + DRN	**20.32%**	**22.76%**
		Hand-labeled	22.75%	24.7%
		Forced Alignment	22.78%	24.48%
Audiobook [8]	Monostate (52)	CTC + DRN	17.85%	16.55%
		MMI + DRN	**16.95%**	**16.12%**
		Forced Alignment	17.76%	16.98%
	Tristate (156)	CTC + DRN	12.58%	11.67%
		MMI + DRN	**10.08%**	**9.67%**
		Forced Alignment	12.53%	11.96%
Broadcasts [2]	Monostate (52)	CTC + DRN	25.96%	**25.58%**
		MMI + DRN	35.66%	65.26%
		Forced Alignment	**25.82%**	25.64%
	Tristate (156)	CTC + DRN	21.62%	21.23%
		MMI + DRN	**20.74%**	**20.42%**
		Forced Alignment	22.13%	21.74%

Table 1: The phoneme error rates got for the different DRN training methods.

3 Sequence Training Methods for Deep Rectifier Neural Networks in Speech Recognition

After determining our preferred choice of DNN, we turned our attention to the task of flat start training, which is the first step of training a speech recognition system. The goal of flat start is to create time-aligned context independent labels for the database. Our aim here was the comparison of two sequence training approaches that could be used to train randomly initialised DNNs without having force-aligned labels. The first one was the Connectionist Temporal Classification (CTC) [9] and the second one was the Maximum Mutual Information (MMI) method [10]. Both of them were

used to train Deep Rectifier Networks (DRNs). We proposed several modifications to the standard MMI method, which were essential to make it suitable for the flat start process. The key modifications that we propose in order to make MMI training suitable for DNN flat start are:

1. The frame-level phonetic targets are determined by a forward-backward search.
2. We employ only phoneme-level transcripts and CI phoneme states.
3. We do not apply state priors or language model.
4. The denominator is estimated by just using the most probable decoded path.
5. We measure the error on a hold-out set; when it increases after a training iteration, we restore the parameters of the network and decrease the learning rate.

In the experimental part, we evaluated the two methods on several phone recognition tasks, Table 1 shows the results we got. For all the databases we tested, we found that the sequence training methods gave better results that those obtained with force-aligned training labels produced by an HMM/GMM system. From the experimental results, it was also clear that the MMI-based approach using tri-state models gave better results than the CTC-based one. Furthermore, DRNs trained with CTC could not produce forced-aligned labels. Based on these findings, we concluded that MMI was the better algorithm for flat start training.

4 A GMM Free Training Method for Deep Neural Networks

Next, we modified the standard state-tying algorithm with the goal of getting rid of its GMM dependency. The context-dependent states used to train DNNs are usually obtained using the standard tying algorithm, even though it is based on likelihoods of Gaussians, hence it is more appropriate for HMM/GMMs. Recently, however, several new refinements have been published which seek to adapt the state tying algorithm to the HMM/DNN hybrid architecture.

Some of the new methods change only the input of the clustering algorithm, by feeding the output or the activations of the neurons in the last hidden layer to the clustering method while the whole state tying algorithm remains intact [11, 12, 13, 14] . Other studies proposed novel decision criteria as well for the clustering method, which better suit the new input provided by a DNN[15, 16].

In an article [15], we proposed a KL-divergence-based approach. We evaluated it along with three other state-tying methods on the same LVCSR tasks, and compared

their performance under the same circumstances. We combined them with our MMI-based flat start method, and showed that the whole training procedure of context-dependent HMM/DNNs can be performed without using GMMs.

Method	Dev.	Test
Iterative CE	28.63%	20.47%
MMI	15.78%	10.07%
MMI+CE	15.43%	9.64%

Table 2: WERs got by using different CI flat start methods on the WSJ.

To test our algorithms the 81-hour long Wall Street Journal (WSJ) English read speech corpus [17] (specifically the `si-284` set) was chosen as it is a well-known and widely used corpus. The experimental results confirmed that the MMI based flat start approach is far better than the procedure of iterative CE DNN training and re-alignment (see Table 2). Furthermore, as can be seen in Table 3, the replacement of the decision criterion used during state clustering is also beneficial for DNN training.

Flat start strategy	Clustering method	Development	Test
Iterative CE	MFCC + Likelihood	11.02%	8.20%
	DNN + Likelihood	11.48%	7.64%
	DNN (hidden) + Likelihood	11.05%	7.81%
	Kullback-Leibler	10.47%	**7.27%**
	Entropy	**10.24%**	**7.27%**
MMI	MFCC + Likelihood	8.58%	6.13%
	DNN + Likelihood	8.7%	6.47%
	DNN (hidden) + Likelihood	8.85%	6.04%
	Kullback-Leibler	**8.06%**	**5.72%**
	Entropy	**8.03%**	5.92%
MMI + CE	MFCC + Likelihood	8.79%	5.97%
	DNN + Likelihood	9.14%	6.45%
	DNN (hidden) + Likelihood	9.43%	6.77%
	Kullback-Leibler	8.5%	**6.15%**
	Entropy	**8.09%**	6.20%

Table 3: WER values obtained on the development and test sets, got by using the different flat-start and CD state tying methods.

Lastly, we examined our best Hungarian HMM/DNN system to see what type of errors are most common. For this, we collected the word errors and their local context, then we manually categorised and analysed them. Our conclusion was that a new metric is needed to measure the accuracy of Hungarian ASR systems, since the current one (WER) treats some errors more seriously than human readers do.

5 Training Context-Dependent DNN Acoustic Models using Probabilistic Sampling

Next, we turned our attention to the CD training phase of the ASR system. In the current HMM/DNN speech recognition systems, the purpose of the DNN component is to estimate the posterior probabilities of tied triphone states. It is well-known that the distribution of the CD states is uneven, meaning that we have a markedly different number of training samples for the various states. This imbalance in the training data is a source of suboptimality for most machine learning algorithms, and DNNs are no exception to this.

Here, we experimented with the so-called probabilistic sampling method [18] that applies downsampling and upsampling at the same time, to improve the accuracy of CD acoustic models. This re-sampling method defines a new class distribution for the training data, which is a linear combination of the original and the uniform class distributions, and the λ parameter determines the weights of the two distributions. As an extension to previous studies [18, 19], we also proposed a new method to re-estimate the class priors, which is required to remedy the mismatch between the training and the test data distributions introduced by re-sampling.

Figure 3 shows the results we got with probabilistic sampling on the TED-LIUM corpus. Clearly, dividing the DNN outputs by the original priors gives worse results as λ increases, and we found that small λ values (here 0.4) work best. Additionally, with the use of the adjusted priors, the models became more robust. Using the modified probabilistic sampling algorithm we achieved relative word error rate reductions of 5% and 6% on two fair-sized corpora (TED-LIUM [20] and AMI [21]). We also showed that this re-sampling method can improve our GMM-free system outlined in the previous chapter. Our experimental results strongly suggest that the re-estimation of the priors is essential to handle the mismatch between the training and the test data distributions introduced by the re-sampling step. These adjusted priors made the re-sampling method more robust, and the recognition results varied only slightly as the class distribution was shifted with a bigger λ value, towards a uniform distribution.

We also managed to apply DRNs, trained with probabilistic sampling, on several paralinguistics tasks successfully, and these tasks were part of the Computational Paralinguistics ChallengE (ComParE) series. The main goal in paralinguistics is to extract

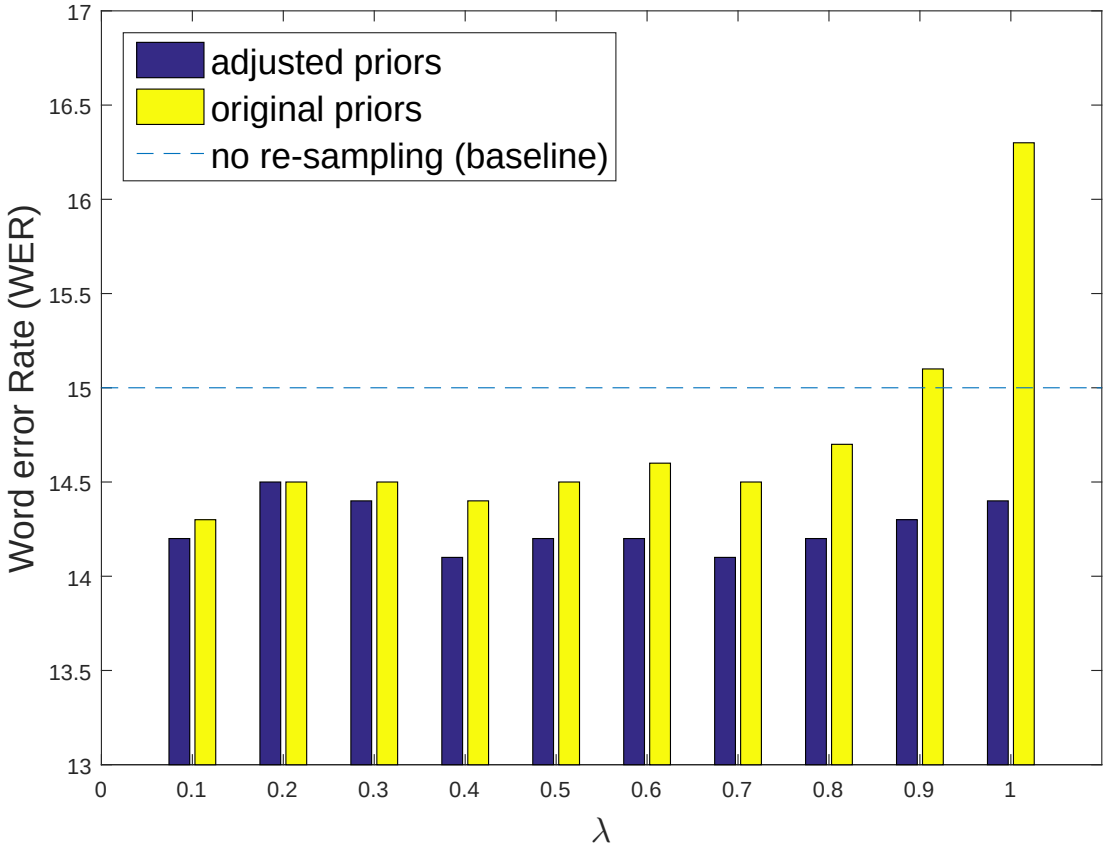

Figure 3: Word error rates got for the test set of the TED-LIUM corpus using probabilistic sampling.

and identify phenomena present in the audio signal other than the words uttered. In 2014, we created a system to detect the intensity of cognitive and physical load of the speaker [22]. Later, we examined the possibility of detecting deceit from speech [23]. Last year, we won the Cold Challenge, where our system had to separate healthy speakers from those who had a cold [24].

6 Conclusions and future directions

In this thesis, we successfully adapted the standard methods of the old HMM/GMM acoustic models to better suit the new HMM/DNN hybrid. We revised both the initial training phase (flat start) and the CD state-tying phase, and introduced new strictly DNN-based solutions to these problems. By combining these methods, we created a new training pipeline that does not depend on GMMs at all. We also demonstrated that the final training phase could be improved by employing a simple re-sampling method. On the Szeged Hungarian Broadcast News corpus, a traditional HMM/GMM gave a WER of 20.07%, the best DNN that still relies on GMMs produced a WER of only 16.59%; while our best GMM-free system managed to achieve a WER of 15.79%.

Naturally, many experiments have been left for the future, mainly due to lack of

time or because they lay outside the scope of the present study. The following list presents some of the possible future research directions.

- Firstly, we should consider applying a new DNN type, namely the Convolutional Neural Network (CNN), since it has provided impressive results both in image processing and speech recognition.
- To further extend the results of this research work, it would be worth examining other sequence learning methods such as minimum phone error (MPE) or state-level minimum Bayes risk (sMBR), and adapt them so they are suitable for flat start training.
- It is worth investigating what would happen if we had more CD clusters in our GMM-free systems. The hypothesis here is that with more states we should get better results, of course, at the cost of increased training and evaluation times.
- It would be interesting to learn how the CD DNNs trained with probabilistic sampling perform after a final sequence discriminative training phase, which is nowadays a common practice.

7 Key points of the Thesis

	[2]	[25]	[26]	[15]	[27]	[28]	[22]	[23]	[24]
I	•								
II/1		•							
II/2		•	•						
III/1				•					
III/2			•	•	•				
IV						•	•	•	•

Table 4: Correspondence between the thesis points and the publications.

In the following we list the key results of the dissertation. Above, Table 4. summarizes the relation between the theses and the corresponding publications.

I. The author compared the performance of four deep learning methods empirically; two of these methods were pre-training algorithms, the third one applied the rectifier activation function and the fourth was a regularisation technique called Dropout. The experiments were also carried out using a Hungarian speech corpus, and this study was among the first to apply a HMM/DNN system to

Hungarian speech recognition. The results indicated that the new HMM/DNN systems can outperform the traditional HMM/GMM system significantly. The conclusion of the experiments was that, although the four algorithms yielded quite similar recognition performances, rectifier networks consistently produced the best results.

II/1. The CTC algorithm was originally proposed for the training of recurrent neural networks, but here the author showed that it can also be used to train conventional feed-forward networks. Using several corpora, deep rectifier networks were trained with the CTC method, in order to determine whether this approach was suitable for the flat start training phase. The results told us that CTC can be used to train randomly initialised networks without time-aligned labels.

II/2. As a competitor, the MMI-based training algorithm was also examined. The author proposed several modifications to the standard MMI, to make it suitable for the task (flat start training). The experimental results indicated that the modified MMI is a far superior alternative to CTC, for training randomly initialised networks without time-aligned labels.

III/1. The author created a new DNN-based state-tying method by changing the decision criterion used by the standard algorithm during the clustering step. Since this new state tying method uses posterior probability vectors produced by DNNs as input, KL-divergence seemed a logical choice for decision criterion. The experimental results also supported this view, as the new method markedly outperformed the original one.

III/2. By combining the MMI-based flat start training algorithm with the KL-divergence-based clustering method, the author built an ASR system that did not rely on GMMs. He compared this GMM-free solution with other recently proposed alternatives, and found that it was competitive with the other approaches used. Furthermore, the results demonstrated empirically that the GMM-free systems were capable of producing better results than those that relied on GMMs.

IV. The author examined the probabilistic sampling method for the training of CD DNNs. He hypothesised that when the training data is re-sampled, the prior probability values need to be re-estimated. He justified this experimentally, and showed that re-sampling with adjusted priors greatly improves the performance of CD DNNs. This re-sampling algorithm was also applied with great success in several paralinguistic tasks.

References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] T. Grósz and L. Tóth, "A comparison of Deep Neural Network training methods for Large Vocabulary Speech Recognition," in *Proceedings of TSD*, pp. 36–43, 2013.

[3] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[4] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proceedings of ASRU*, pp. 24–29, 2011.

[5] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," in *Proceedings of AISTATS*, pp. 315–323, 2011.

[6] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," in *CoRR*, vol. 1207.0580, 2012.

[7] S. S. Lamel L., Kassel R., "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *DARPA Speech Recognition Workshop*, pp. 121–124, 1986.

[8] L. Tóth, B. Tarján, G. Sárosi, and P. Mihajlik, "Speech recognition experiments with audiobooks," *Acta Cybernetica*, pp. 695–713, 2010.

[9] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, vol. 385 of *Studies in Computational Intelligence*. Springer, 2012.

[10] X. He and L. Deng, *Discriminative Learning for Speech Recognition*. San Rafael, CA, USA: Morgan & Claypool, 2008.

[11] A. Senior, G. Heigold, M. Bacchiani, and H. Liao, "GMM-free DNN training," in *Proceedings of ICASSP*, 2014.

[12] C. Zhang and P. Woodland, "Standalone training of context-dependent Deep Neural Network acoustic models," in *Proceedings of ICASSP*, pp. 5597–5601, 2014.

[13] A. Senior, G. Heigold, M. Bacchiani, and H. Liao, "GMM-free DNN acoustic model training," in *Proceedings of ICASSP*, pp. 5639–5643, 2014.

[14] M. Bacchiani and D. Rybach, "Context dependent state tying for speech recognition using deep neural network acoustic models," in *Proceedings of ICASSP*, pp. 230–234, 2014.

[15] G. Gosztolya, T. Grósz, L. Tóth, and D. Imseng, "Building context-dependent DNN acousitc models using Kullback-Leibler divergence-based state tying," in *Proceedings of ICASSP*, pp. 4570–4574, 2015.

[16] L. Zhu, K. Kilgour, S. Stüker, and A. Waibel, "Gaussian free cluster tree construction using Deep Neural Network," in *Proceedings of Interspeech*, pp. 3254–3258, Sep 2015.

[17] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of HLT*, pp. 357–362, 1992.

[18] L. Tóth and A. Kocsor, "Training HMM/ANN hybrid speech recognizers by probabilistic sampling," in *Proceedings of ICANN*, pp. 597–603, 2005.

[19] M. Song, Q. Zhang, J. Pan, and Y. Yan, "Improving HMM/DNN in asr of under-resourced languages using probabilistic sampling," in *Proceedings of ChinaSIP*, pp. 20–24, 2015.

[20] A. Rousseau, P. Deléglise, and Y. Estéve, "TED-LIUM: an automatic speech recognition dedicated corpus," in *Proceedings of LREC*, pp. 125–129, 2012.

[21] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[22] G. Gosztolya, T. Grósz, R. Busa-Fekete, and L. Tóth, "Detecting the intensity of cognitive and physical load using AdaBoost and Deep Rectifier Neural Networks," in *Proceedings of Interspeech*, pp. 452–456, 2014.

[23] G. Gosztolya, T. Grósz, R. Busa-Fekete, and L. Tóth, "Determining native language and deception using phonetic features and classifier combination," in *Proceedings of Interspeech*, (San Francisco, CA, USA), pp. 2418–2422, Sep 2016.

[24] G. Gosztolya, R. Busa-Fekete, T. Grósz, and L. Tóth, "DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification," in *Proceedings of Interspeech*, pp. 3522–3526, 2017.

[25] T. Grósz, G. Gosztolya, and L. Tóth, "A sequence training method for Deep Rectifier Neural Networks in speech recognition.," in *Proceedings of SPECOM*, pp. 81–88, Sep 2014.

[26] G. Gosztolya, T. Grósz, and L. Tóth, "GMM-free flat start sequence-discriminative DNN training," in *Proceedings of Interspeech*, (San Francisco, CA, USA), pp. 3409–3413, Sep 2016.

[27] T. Grósz, G. Gosztolya, and L. Tóth, "A comparative evaluation of GMM-free state tying methods for ASR," in *Proceedings of Interspeech*, pp. 1626–1630, 2017.

[28] T. Grósz, G. Gosztolya, and L. Tóth, "Training context-dependent DNN acoustic models using probabilistic sampling," in *Proceedings of Interspeech*, pp. 1621–1625, 2017.