The RWTH Aachen LVCSR system for IWSLT-2016 German Skype conversation recognition task

Wilfried Michel, Zoltán Tüske, M. Ali Basha Shaik, Ralf Schlüter, Hermann Ney

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

Abstract

In this paper the RWTH large vocabulary continuous speech recognition (LVCSR) systems developed for the IWSLT-2016 evaluation campaign are described. This evaluation campaign focuses on transcribing spontaneous speech from Skype recordings. State-of-the-art bidirectional long short-term memory (LSTM) and deep, multilingually boosted feed-forward neural network (FFNN) acoustic models are trained an narrow and broadband features. An open vocabulary approach using subword units is also considered. LSTM and count-based full word and hybrid backoff language modeling methods are used to model the morphological richness of the German language. All these approaches are combined using confusion network combination (CNC) to yield a competitive WER.

1. Introduction

The International Workshop on Spoken Language Translation (IWSLT) is an annual workshop including an evaluation campaign in the tasks of automatic speech recognition (ASR), machine translation (MT), and spoken language translation (SLT), which is the union of the aforementioned. Participants have the opportunity to compete in this evaluation campaign to compare the strength of their systems and advance the state of the art.

The German ASR task of the 2016 evaluation campaign employs data from the Microsoft Speech Language Translation (MSLT) task. The goal is to transcribe one side of bilingual Skype voice calls. Development and test data are audio files with a varying length between 1 and 30 seconds each containing one part of an informal conversation between two natural persons. This is challenging from multiple points of view. Informal conversations tend to be highly spontaneous which leads to an increased number of disfluencies and misarticulations. Conversational speech also covers a broader range of topics which makes it hard to train and fine tune appropriate language models.

To overcome these difficulties we propose a combination of two different acoustic models and two different language models. We used the LIUM auto-segmenter [1] to trim non-speech parts and to segment the longer of the audio files. From the audio files we extracted standard cepstral features: MFCC, PLP, Gammatone [2] and also the critical band energies (CRBE) of the corresponding pipelines. A hybrid LSTM acoustic model was trained on the Gammatone features to directly output tied-triphone state posterior probabilities. To generate a second, largely complementary acoustic model for German, we trained a Gaussian mixture model based HMM processing multilingually initialized deep bottleneck feed-forward NNs using the tandem approach [3, 4].

Both systems are complemented with two 5-gram language models for initial decoding [5]. In a second pass the generated lattices are rescored using LSTM recurrent NN based language models. In a final step all four lattices are combined using confusion network combination. The results were achieved using RETURNN, the RWTH extensible training framework for universal recurrent neural networks, in combination with RASR, the RWTH ASR toolkit[6, 7].

The remainder of this paper is organized as follows. Section 2 describes in detail the acoustic models while language models are presented in Section 3. The complete decoding setup including system combination is presented in Section 4. Our results are described in Section 5 and Section 6 concludes the paper.

2. Acoustic Models

For the training of acoustic models, no in domain audio data were provided by the organizers. This work utilizes training data from the Quaero project (2009 - 2013) to train two state-of-the-art neural network acoustic models. One is a bidirectional long short-term memory neural network (BLSTM) in a hybrid approach, and the other is a fine tuned multilingually initialized deep feed-forward network adapted to the German language.

2.1. Training Resources

For the German ASR task, acoustic training data was taken from German broadcast news (BN), speeches from European parliament plenary sessions (EPPS) held in German, and web domain [8]. Table 1 lists the amount of training data from each domain.

While the parliamentary speech and parts of the BN seg-

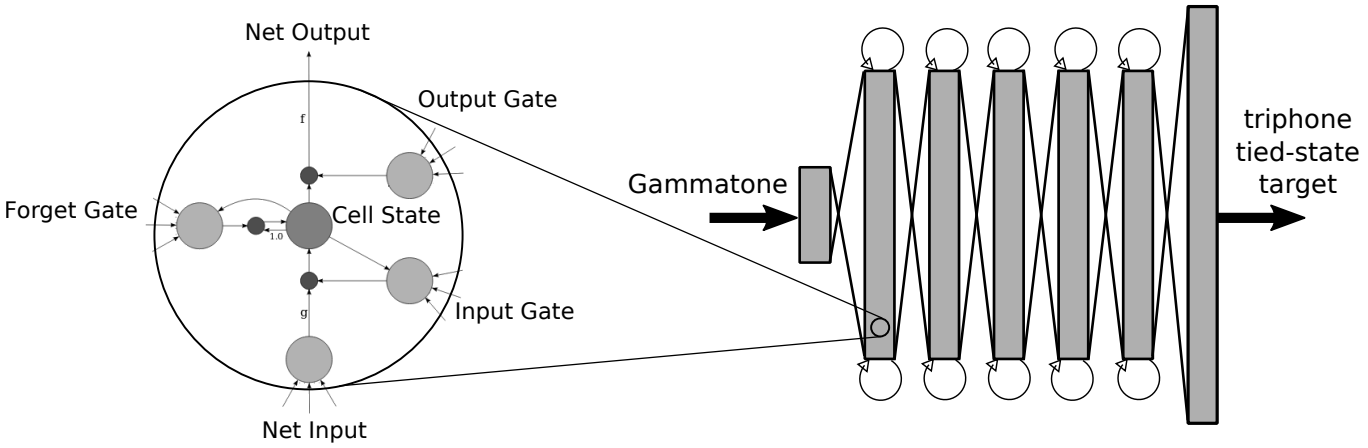

Figure 1: *Structure of an LSTM cell and the BLSTM acoustic model network. The Network consists of a 50 dimensional input layer, 5 hidden layers with 600 cells for both forward and backward direction and a 4500 dimensional output layer corresponding to triphone tied-state posteriors.*

Table 1: *Acoustic training data for the German LVCSR task.*

Corpus	Duration [h]	Segments	Running Words
EPPS08	5	1109	45,796
WEB08	14	3452	127,086
Quaero BN	123	25061	1,391,468

ments are usually well planned, a substantial part of the Quaero BN and the web corpus consists of talk shows and interviews where different people speak in an unplanned and spontaneous way. This is a good match for the MSLT Skype Task, which consist of spontaneous conversations. Nevertheless there is a mismatch between the usually high quality microphones used for talk show recordings and the cheaper consumer microphones used in private Skype conversations.

2.2. BLSTM Acoustic Model

Recently bidirectional long short-term memory (LSTM) recurrent neural networks have shown promising results in reducing the error rates of speech recognition systems [9]. The model was trained on alignments obtained by our previous best system, a multilingual hierarchical bottleneck FFNN [10].

2.2.1. Feature Extraction

For the BLSTM acoustic model Gammatone features [2] were extracted from the audio files. We used 50 Gammatone filters in the frequency range of 100 Hz to 7.5 kHz followed by full-wave rectification and temporal integration with a Hanning window of size 25 ms shifted by 10 ms. Then, cepstral decorrelation was performed, followed by 10th root compression and segment-wise mean and variance normalization. The resulting 50 dimensional features are directly input into the LSTM network.

2.2.2. Network Topology

The LSTM is a variant of recurrent neural networks. A recurrent neural network can be unfolded over time and such corresponds to a special deep neural net structure. The corresponding backpropagation algorithm is called *backpropagation through time*. Most implementations of recurrent neural networks share the deficiency, that the gradient either vanishes or grows exponentially during training. The LSTM architecture cures this deficiency by introducing a memory cell and gates which control the flow of information into and out of this cell. The structure of one LSTM cell is depicted in Fig. 1.

Let x_t be the input vector at time t, c_t the cell memory state, and s_t the output of the cell. Let A, R, $A_{\{i,f,o\}}$, $R_{\{i,f,o\}}$ be full matrices and $W_{\{i,f,o\}}$ be diagonal matrices which are all to be trained. Then the (recurrent) net input z_t to a cell is given by

$$z_t = \tanh\left(Ax_t + Rs_{t-1}\right)$$

The input gate i_t and forget gate f_t determine how the cell state should be updated.

$$\begin{aligned} i_t &= \sigma(A_i x_t + R_i s_{t-1} + W_i c_{t-1}) \\ f_t &= \sigma(A_f x_t + R_f s_{t-1} + W_f c_{t-1}) \\ c_t &= f_t \circ c_{t-1} + i_t \circ z_t \end{aligned}$$

The output gate o_t determines how much of the cell state should be outputted.

$$\begin{aligned} o_t &= \sigma(A_o x_t + R_o s_{t-1}) \\ s_t &= o_t \circ \tanh(c_t) \end{aligned}$$

where $\circ$ means element wise multiplication.

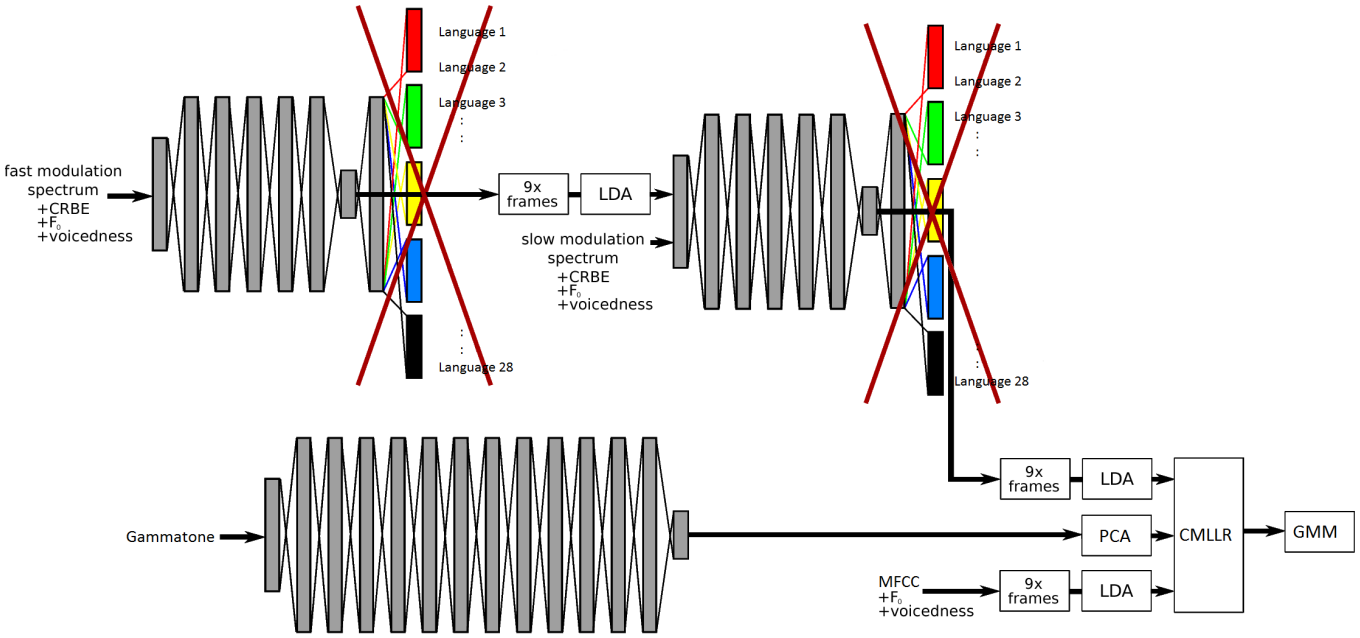

Figure 2: *Structure of the FFNN acoustic model network. A 2-step hierarchical MRASTA bottleneck network is combined with a 12 layer deep network and 9 frames of standard cepstral features in a tandem fashion.*

The LSTM network of this year's submission, which is depicted in Fig. 1, consists of an input layer, 5 hidden layers with 600 Cells for each direction, and a 4500 dimensional output layer corresponding to generalized triphones determined by a decision-tree-based clustering method (CART).

The network was trained with the framewise cross entropy loss criterion using Adam optimization with incorporated Nesterov momentum [11]. To prevent overfitting, we used an l_2 normalization parameter of 0.01. We also employed a dropout scheme with an empirically optimized dropout rate of $p = 0.05$. This means in every training step 5% of the hidden nodes outputs s_t were randomly set to zero. During recognition no cells were forced to zero but the inputs were rescaled by $(1-p)^{-1}$ to account for the higher number of inputs. This leads to a more robust network and further prevents overfitting.

2.3. FFNN Acoustic Model

With system combination in view, a second acoustic model was also trained. In order to make this acoustic model complementary to the first model, we modified the neural network structure and embedding, as well as the feature extraction pipeline, and weight initialization. Consisting of feed-forward neural network elements, two different models were trained. These were trained on the same corpus as the LSTM AM, but the audio files were resampled at 8 kHz.

On the one hand, for the evaluation a 12-layer network was prepared which contained 2000 nodes in each layer and used the rectified linear units (ReLU) non-linearity [12]. The last hidden layer was low-rank factorized by a linear 512-dimensional bottleneck (BN) [13]. On the other hand, a deep hierarchical MRASTA-BN feature extractor was also trained [14, 15]. The size of the BN layers was restricted to 62 nodes, whereas the other layers have 2000 sigmoidal neurons. The hierarchy consisted of 2 levels, and each level was built up from 6 layers of non-bottleneck hidden layers. The BNs were inserted before the last hidden layer. The two feed-forward networks were used in tandem with a Gaussian model [3].

The ReLU network was trained on the same Gammatone pipeline as the LSTM, however the input features had only 40 dimensions due to the reduced sampling rate. The hierarchical FFNN used three streams of critical band energies extracted from GT, PLP, MFCC pipelines. Each of them produced 15-dimensional feature streams. In addition to the MRASTA filtering of these streams, voicedness, F0 and the three current critical band energy (CRBE) frames were also fed into the NNs at each level of the hierarchy.

Instead of random initialization, both models were multilingually boosted reusing corresponding BN feature extractors from the BABEL project [16]. The cross lingually transferred models were trained on 1800 hours covering 28 languages from which none of them was German. Besides the 24 Babel languages the following resources were included in the multilingual training: subset of Fisher English part 1 and 2 (LDC2004S13 and LDC2005S13), approx. 213 hours, and the CMU pronunciation dictionary `cmudict-0.7b`; about 153 hours of Spanish from Fisher and CallHome audio (LDC2010S01 and LDC96S35) and CallHome pronunciation lexicon (LDC96L16); 145 hours of Mandarin Chinese, HKUST telephone speech corpus (LDC2005S15), CallHome

Table 2: *OOV rate and in-vocabulary-word (IV) character level PPLs are described. Comparative PPL is shown for the hybrid system (i.e., PPL normalized w.r.t. the full word (FW) system) for fair comparison. Vocabulary size is 377k for both the full word and hybrid systems.*

	character level PPL (IV word)					
	dev			eval		
	OOV	count	+ LSTM	OOV	count	+ LSTM
FW	0.9	2.97	2.84	1.0	2.98	2.84
Hyb	0.3	3.19	3.07	0.4	3.49	3.36

Table 3: *Total Perplexity - Interpolation of full word 5-gram count-based LM and LSTM-NNLM.*

LM	count	LSTM	interpolated
dev	275	261	210
eval	262	250	196

(LDC96S34), HUB5 (LDC98S69) along with the CallHome pronunciation lexicon (LDC96L15); and 172 hours of Levantine Arabic QT Training Data Set 5 (LDC2006S29). However, the FFNNs were language adapted to the target language following the same recipe of [17]. During this fine-tuning step the whole hierarchical structures were updated. Also, instead of using the previous best alignment for the language adaptation step, a new one was obtained from scratch using only the MRASTA-BN features.

For the tandem modeling the output of the FFNNs were further processed. Nine-frame window of the MRASTA-BN features were LDA transformed, whereas the ReLU network's 512-dimensional output was directly reduced by PCA. The final dimension of both FFNN streams was 64. All Parameters were empirically tuned. The ultimate tandem model was trained on the concatenation of both FFNN features and 45-dimensional LDA transformation of 9 frames of MFCC, voicedness, and F0 streams. On this acoustic model, speaker adaptive training (SAT) using constrained maximum likelihood linear regression (CMLLR) [18] was also applied. Furthermore, in the final training step, the minimum phone error (MPE) training criterion was also employed.

3. Language Modelling

The language modeling text contains sources from different domains like broadcast news, spontaneous speech transcripts, web data and audio transcriptions [8]. The text is normalized using a language dependent set of rules and semi-automatic methods. Vocabularies are generated based on word frequency. Domain adapted full word and hybrid count-based language models are generated. The hybrid vocabulary contains the most frequent 5k full words as preserved vocabulary. It has been observed that the news domain text is closer to the IWSLT-2016 skype audio corpus. In addition, both the full word and hybrid long short term memory (LSTM) LMs are generated using selected in-domain data containing 100M running words [19]. They are linearly interpolated with their respective count-based backoff LMs and are used for lattice rescoring [20, 21]. Table 3 compares full word perplexities for the count based, LSTM and interpolated language model. OOV rates and character level perplexities are shown in Table 2. As OOVs are mapped to an unknown token <unk> in both language models, total perplexities are not comparable. However they are shown for the full word system as an additional information. For this reason only character level perplexities measured on the in-vocabulary words for both systems are shown. However, a proper direct comparison of the full word and hybrid systems is difficult in terms of word/subword level PPLs due to different vocabularies, sizes, and OOV rates. Therefore, word/subword perplexities are renormalized to character level perplexities. This is also called comparative or projected PPL. The increase in in-vocabulary character perplexity can be attributed to the increased confusability introduced by subword modeling. Thus, a trade-off between a lowering of the OOV rate and an increase in-vocabulary word confus-

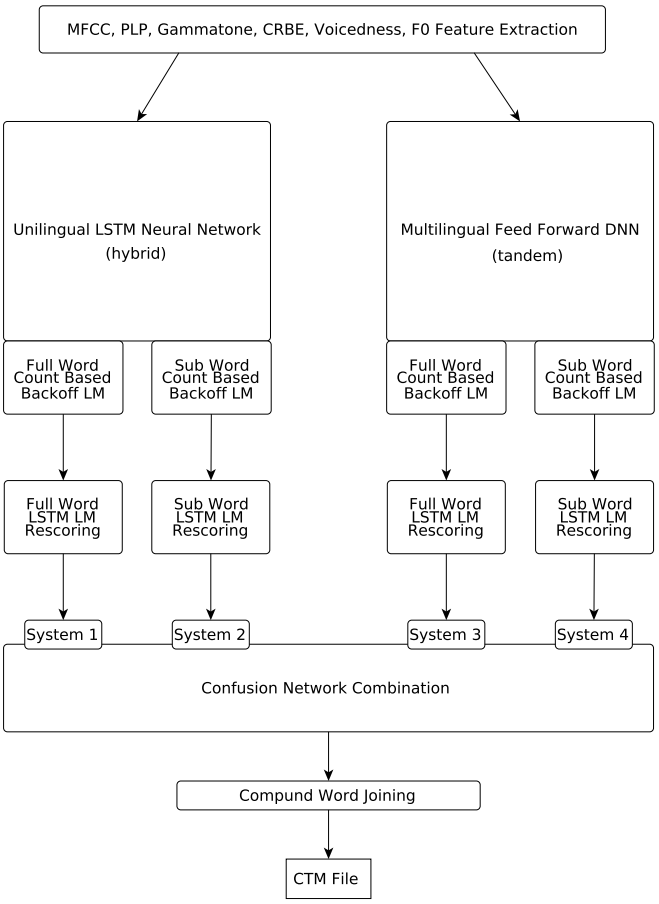

Figure 3: *Schematic diagram of our speech recognition system. The output of two acoustic models and two language models is combined using confusion network combination.*

Table 4: *Word and Character Error Rate [%] for each system separately and for different combination methods. Numbers in boldface indicate the systems used for combination. FW: full word, SW: subword, CN/CNC: Confusion Network Combination*

	dev2016				tst2016			
Pass	WER FW	WER SW	CER FW	CER SW	WER FW	WER SW	CER FW	CER SW
2013 system	27.3	27.3	15.0	15.1	24.4	24.7	14.6	14.8
LSTM AM	24.7	24.9	14.1	14.2	22.7	23.0	13.8	13.9
+ LSTM LM	**23.2**	**24.0**	13.0	13.5	**21.5**	**22.3**	13.3	13.7
FFNN AM	27.3	27.4	15.1	15.1	24.7	24.8	14.9	14.8
+ CMLLR	25.6	25.7	14.0	14.1	23.2	23.5	14.0	14.0
+ MPE	24.5	24.4	13.4	13.3	21.9	22.1	13.1	13.1
+ LSTM LM	**23.4**	**23.9**	13.0	13.1	**20.6**	**21.5**	12.6	12.8
Method	combined WER		combined CER		combined WER		combined CER	
ROVER	22.8		13.3		20.9		13.3	
CNC	**21.6**		12.9		**19.6**		13.1	

ability can be observed. After lattices have been created with both the full word and hybrid language models, confusion network combination is used to combine the advantages of both approaches.

4. Recognition Pipeline

In this section we describe the complete recognition pipeline and how the different acoustic and language models are combined to produce the final output. A schematic representation of our pipeline can be found in Figure 3.

We were given a set of separate audio files with a sampling rate of 16kHz. For the FFNN acoustic model these files were resampled at 8kHz. On both sets of files the LIUM auto segmenter was used to filter out non speech parts and to generate a segmentation. After this, MFCC, PLP and Gammatone features, along with the corresponding CRBE, voicedness and F0 frames were extracted.

These features were then used by both acoustic models to generate, in conjunction with both 5-gram language models, in total four different word hypothesis lattices. The LSTM language models are then used to rescore the corresponding lattices. The new LM score was obtained from an interpolation of both LSTM and 5-gram language model. The optimal interpolation weight was found by optimizing the resulting perplexity on the dev data.

To further improve the accuracy of the system, all systems were combined using ROVER or confusion network combination [22, 23, 24]. Confusion network combination (CNC) first transforms every lattice into a separate confusion network and then only joins the individual confusion networks.

In a final step the resulting hypotheses are post-processed with a compound joining script that identifies and joins compound words and numbers in the target language [25]. The error rates of each step are reported in the next section.

5. Results

In this section we describe in detail the results and relative improvements of each step in our ASR pipeline. The word error rates are shown for the dev2016 and tst2016 set computed using an unofficial scoring script. Official error rates can be found in [26], where the combined system is our primary system and the other bold systems are contrastive systems. Since this is the first year with a Skype video conference transcription task, we give the results of our previous best German system [8] on this task as a progress report. As the previous language models were adapted to the lecture domain, we used the old acoustic model together with the newly estimated language models to give a fair comparison.

Table 4 shows the results for both acoustic and language models. The single best system is the combination of LSTM acoustic model and full word language model. For all acoustic models we see a performance degradation when using the subword language model. When we rescore with the LSTM language model, the difference increases to up to 0.9% absolute (4% relative). Due to the fairly low OOV rates, this might be expected, as subwords also increase confusability. However, combining full word and subword LMs lead to improvements.

Comparing both acoustic models we see that the LSTM acoustic model baseline error rates are lower than the FFNN baseline. CMLLR speaker adaption reduces the error rates by 6% relative and with MPE training we gain an additional 4% relative improvement. The final multilingually initialized and combined FFNN acoustic model performs then on par with the LSTM AM.

There are two methods of system combination proposed in this paper. ROVER only combines the single best recognition output of each system and achieves an improvement of 0.4% absolute (1.7% relative) compared to the single best system. Confusion network combination combines multiple hypotheses of all systems and can further improve the recognition results. The error rate after CNC is 1.6% absolute (7% relative) below the single best system.

This year's submission performs significantly better than our last submission to the IWSLT from 2013. We can report an improvement of 5.7% absolute (20.8% relative).

On the test set we see similar results. Error rates for the test set are about 10% lower compared to the dev set. Comparing the different systems we see that the full word language model still performs better than the subword model. While on the dev set the LSTM acoustic model was sightly better than the FFNN AM, on the test set the FFNN outperforms the LSTM by about 1% absolute (5% relative).

6. Conclusions

We presented the RWTH ASR 2016 system for the German MSLT Skype transcription task [26]. We made heavy use of different deep and/or recurrent neural network architectures in acoustic modeling.

A deep multilingual hierarchical bottleneck MRASTA FFNN trained on narrow-band audio was successfully combined with a very deep monolingual FFNN in a tandem approach to accompany a deep hybrid bidirectional LSTM NN acoustic model. Full word and hybrid 5-gram language models were used in decoding. The Results were further refined using LSTM language models for rescoring of hypothesis lattices. Although the hybrid LM did not outperform the traditional full word approach, the advantages of both language models and both acoustic models were sucessfully combined using confusion network based system combination.

We were able to achieve a WER of 21.6% on the dev set and 19.6% on the test set. We also reported a relative improvement of 20% WER compared to our previously submitted IWSLT German LVCSR system [8].

7. Acknowledgements

This effort used the following IARPA Babel Program language collection releases: IARPA-babel{101-v0.4c,105b-v0.4,106-v0.2f,107b-v0.7,102b-v0.5a,103b-v0.4b,201b-v0.2b,203b-v3.1a,204b-v1.1b,206b-v0.1e,202b-v1.0d,205b-v1.0a,207b-v1.0a,301b-v2.0b,302b-v1.0a,303b-v1.0a,304b-v1.0b,104b-v0.bY,305b-v1.0b,306b-v2.0c,307b-v1.0b,401b-v2.0b,402b-v1.0b,403b-v1.0b}

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract no. W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

8. References

[1] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An Open-source State-of-the-art Toolbox for Broadcast News Diarization," in *Interspeech*, Lyon, France, August 2013, pp. 1477–1481.

[2] 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, pp. 649–652.

[3] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, June 2000, pp. 1635–1638.

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

[5] M. A. B. Shaik, Z. Tüske, M. A. Tahir, M. Nussbaum-Thom, R. Schlüter, and H. Ney, "RWTH LVCSR Systems for Quaero and EU-Bridge: German, Polish, Spanish and Portuguese," in *Interspeech*, Singapore, September 2014, pp. 973–977.

[6] D. Rybach, S. Hahn, P. Lehnen, D. Nolden, M. Sundermeyer, Z. Tüske, S. Wiesler, R. Schlüter, and H. Ney, "RASR - The RWTH Aachen University Open Source Speech Recognition Toolkit," in *Proc. IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, Waikoloa, HI, USA, December 2011.

[7] P. Doetsch, A. Zeyer, P. Voigtlaender, I. Kulikov, R. Schlüter, and H. Ney, "RETURNN: The RWTH extensible training framework for universal recurrent neural networks," 2016. [Online]. Available: https://arxiv.org/abs/1608.00895

[8] M. A. B. Shaik, Z. Tüske, S. Wiesler, M. Nussbaum-Thom, S. Peitz, R. Schlüter, and H. Ney, "The RWTH Aachen German and English LVCSR systems for IWSLT-2013," in *Proc. 10th Int. Workshop on Spoken Language Translation (IWSLT 2013)*, Heidelberg, Germany, Dec. 2013, pp. 120–127.

[9] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid Speech Recognition with Deep Bidirectional LSTM," in *Proc. IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, Olomouc, Czech Republic, December 2013, pp. 273–278.

[10] Z. Tüske, R. Schlüter, and H. Ney, "Multilingual Hierarchical MRASTA Features for ASR," in *Interspeech*, Lyon, France, August 2013, pp. 2222–2226.

[11] T. Dozat, "Incorporating Nesterov momentum into Adam," Stanford University, Tech. Rep., 2015. [Online]. Available: http://cs229.stanford.edu/proj2015/054_report.pdf

[12] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Neural Networks," in *Proc. 14th Int. Conf. Artificial Intelligence and Statistics*, Ft. Lauderdale, FL, USA, April 2011, pp. 315–323.

[13] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, "Low-rank matrix factorization for Deep Neural Network training with high-dimensional output targets," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 6655–6659.

[14] F. Valente and H. Hermansky, "Hierarchical and parallel processing of modulation spectrum for ASR applications," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Las Vegas, NV, USA, March 2008, pp. 4165–4168.

[15] Z. Tüske, R. Schlüter, and H. Ney, "Deep hierarchical bottleneck MRASTA features for LVCSR," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 6970–6974.

[16] "http://www.iarpa.gov/Programs/ia/Babel/babel.html."

[17] Z. Tüske, D. Nolden, R. Schlüter, and H. Ney, "Multilingual MRASTA Features for Low-resource Keyword Search and Speech Recognition Systems," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 7904–7908.

[18] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[19] M. A. B. Shaik, Z. Tüske, M. A. Tahir, M. Nussbaum-Thom, R. Schlüter, and H. Ney, "Improvements in RWTH LVCSR evaluation systems for Polish, Portuguese, English, Urdu, and Arabic," in *Interspeech*, Dresden, Germany, September 2015, pp. 3154–3157.

[20] M. Sundermeyer, R. Schlüter, and H. Ney, "rwthlm - The RWTH Aachen University Neural Network Language Modeling Toolkit ," in *Interspeech*, Singapore, September 2014, pp. 2093–2097.

[21] M. Sundermeyer, H. Ney, and R. Schlüter, "From Feedforward to Recurrent LSTM Neural Networks for Language Modeling," *Proc. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 517–529, 2015.

[22] J. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, Santa Barbara, CA, December 1997, pp. 347–354.

[23] G. Evermann and P. Woodland, "Posterior Probability Decoding, Confidence Estimation and System Combination," in *NIST Speech Transcription Workshop*, College Park, MD, USA, May 2000.

[24] B. Hoffmeister, "Bayes Risk Decoding and its Application to System Combination," Ph.D. dissertation, RWTH Aachen University, Computer Science Department, RWTH Aachen University, Aachen, Germany, Jul. 2011.

[25] M. Nußbaum-Thom, A. El-Desoky Mousa, R. Schlüter, and H. Ney, "Compound Word Recombination for German LVCSR," in *Interspeech*, Florence, Italy, August 2011, pp. 1449–1452.

[26] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, R. Cattoni, and M. Federico, "The IWSLT 2016 Evaluation Campaign," in *Proc. 13th Int. Workshop on Spoken Language Translation (IWSLT 2016)*, Seattle, WA USA, December 2016.