Improvements in RWTH LVCSR Evaluation Systems for Polish, Portuguese, English, Urdu, and Arabic

M. Ali Basha Shaik[1], *Zoltan Tüske*[1], *M. Ali Tahir*[1], Markus Nußbaum-Thom[1], Ralf Schlüter[1], Hermann Ney[1,2]

[1]Human Language Technology and Pattern Recognition – Computer Science Department
RWTH Aachen University, 52056 Aachen, Germany
[2]Spoken Language Processing Group, LIMSI CNRS, Paris, France
{ shaik, tuske, tahir, nussbaum, schlueter, ney }@cs.rwth-aachen.de

10.21437/Interspeech.2015-635

Abstract

In this work, Portuguese, Polish, English, Urdu, and Arabic automatic speech recognition evaluation systems developed by the RWTH Aachen University are presented. Our LVCSR systems focus on various domains like broadcast news, spontaneous speech, and podcasts. All these systems but Urdu are used for Euronews and Skynews evaluations as part of the EU-Bridge project. Our previously developed LVCSR systems were improved using different techniques for the aforementioned languages. Significant improvements are obtained using multilingual tandem and hybrid approaches, minimum phone error training, lexical adaptation, open vocabulary long short term memory language models, maximum entropy language models and confusion-network based system combination.

Index Terms: LVCSR, LSTM, open-vocabulary, EU-Bridge

1. Introduction

This paper describes Portuguese, Polish, English, Urdu and Arabic large vocabulary continuous speech recognition (LVCSR) systems in detail. These systems but Urdu are developed for the Euronews-2015 and Skynews-2015 evaluations as part of European Community's 7^{th} Framework Programme under the project EU-Bridge. These evaluations focus on transcribing news and spontaneous speech data. Transcription of these types of audio data is challenging due to varying acoustic conditions. Our LVCSR systems previously developed for the Quaero project (Portuguese, Polish, and English) [1] and the Gale project (Arabic) [2] are improved using a number of recently developed approaches. Urdu LVCSR is developed from scratch as part of an in-house project which mainly focuses on transcription of spontaneous speech data.

Recently, neural networks are considerably improving the acoustic level robustness of an ASR system, either with a tandem approach [3], or a hybrid approach [4, 5]. In the tandem approach, neural network output is used as input features to improve Gaussian Mixture Model (GMM) based emission probability estimates. On the other hand, in the hybrid approach neural network output is directly used as Hidden Markov Model (HMM) state posterior probability estimates. Significant gains in LVCSR performance can be achieved by combining multilingual neural network training for both tandem and hybrid approaches. It is demonstrated in [6, 7] that the multilingual bottleneck features could benefit from the additional non-target language data and thus outperforming unilingual bottleneck approach. One of the dominant advantage of these bottleneck features is their flexible portability on an new language. Further acoustical mismatch between the training and testing can be reduced for any target language, by exploiting matched data from other languages [8]. Alternatively, a few attempts have been made to incorporate GMM within a framework of deep neural networks (DNN). The joint training of GMM and shallow bottleneck features was proposed using the sequence MMI criterion, in which the derivatives of the error function are computed with respect to GMM parameters and applied the chain rule to update the GMM simultaneously with the bottleneck features [9]. On the other hand, a different GMM-DNN integration approach was proposed in [10] by taking advantage of the softmax layer, which defines a log-linear model. A GMM and a log-linear mixture model (LMM) are equivalent in the sense that log-linear model with quadratic features corresponds with a Gaussian model [11]. A GMM with pooled covariance matrix can be converted into an LMM by using neural network elements: linear, softmax, and sum- or max- pooling layer. Therefore, a GMM can easily be integrated into a DNN by substituting a softmax layer in a neural network with a LMM. The joint training of a bottleneck and GMM is performed using a generalized softmax layer.

From a language modelling (LM) perspective, LMs generated using long short term memory (LSTM) approach [12] and maximum entropy approach [13, 14, 15] have shown better performance compared to count-based backoff language models in recognition. In the literature, sub-word language models (LMs) are experimented using recurrent neural network language models (RNNLMs) [16] and DNNs [17]. According to our knowledge, LSTMLMs have not been investigated for open vocabulary tasks in the literature, yet. Alternatively, for Urdu, the use of foreign word in place of a local word having a same semantic meaning is a common practice. It contains a significant fraction of loan words from Arabic, Turkish, Persian, Hindi and English languages either in original or in some modified form. But, the text is written in Urdu orthography in almost all the cases, except where there is no alternative. This causes an increase in the number of out-of-vocabulary (OOVs) words. We make an attempt to recognize OOVs using lexical adaptation.

In this paragraph, we give a brief summary of all investigations conducted in this paper. We compare and contrast state-of-the-art hybrid and tandem approaches, and integrate GMM within a framework of DNN on Polish task and also experiment tandem approach for Portuguese, English and Arabic. Due to sparse LM data conditions for Portuguese and Urdu languages, these systems are treated differently at a language model level. Further, Portuguese contains a significant frac-

tion of spoken words belonging to a foreign origin, resulting in high OOV rates. Therefore, morphemic sub-word LSTM approach is investigated for Portuguese task. Alternatively, we do lexical adaptation by adding a new canonical pronunciation (of the foreign word) to its corresponding Urdu word based on semantic similarity. In this approach, we select a fraction of the most-frequent words from the vocabulary. We then select English translations of these words using Urdu-English dictionary lookup. Pronunciations are extracted for these (English) translated words. These pronunciations are added as pronunciation variants to their corresponding Urdu words in the lexicon. These newly added pronunciation variants directly use the same probability mass of the existing Urdu words during recognition. Therefore, this method not only helps by reducing OOV rate to some extent, but also foreign-words are directly transcribed as Urdu text. Further, maximum entropy LMs are used for better generalization of word sequences followed by maximum a-posteriori LM adaptation. We combine advantages of all individual systems using confusion network combination for English and Portuguese [18]. We use ROVER for Polish [19].

2. Training Data

The amount of the data used for acoustic model training is shown in Table 1, multilingual bottleneck feature training is shown in Table 2, and language model training is shown in Table 3 for different languages.

Table 1: *Acoustic Training data (Lng.: Language, dur.: duration (hours), seg.:segments, PR: Portuguese, PL: Polish, EN: English, UR: Urdu and AR: Arabic)*

Lng.	**Corpus**	**#Dur.**	**#Segs**	**# words**
PR	Broadcast News	110	20 K	1.1 M
PL	Quaero +Broadcast News	110	29 K	1.0 M
AR	Broadcast News	110	52 K	850 K
EN	EPPS+Quaero Broadcast News	322	124 K	3.3 M
UR*	Broadcast News	99	29 K	800 K

* corpus provided by http://www.apptek.com

Table 2: *Multilingual bottleneck feature training data.*

language	German	English	French	Polish
Duration of speech [h]	142	232	317	110

Table 3: *Language model training data. Broadcast news data includes acoustic transcriptions.*

Lng.	**Type**	**# words**	**source**
PR	broadcast news	14 M	newspaper
PL	broadcast news blogs	688 M 710 M	newspaper web
AR	broadcast news wiki	1.20 B 50 M	newspaper web
EN	gigaword quaero	2.6 B 462 M	newspaper blogs+news
UR	broadcast news blogs	60 M 206 M	newspaper blogs+news

Table 4: *Polish WERs (OOV=[dev12: 1.0, dev15: 1.1], PPL=[dev12: 656.1, dev15: 469.8], audio:[dev12: 3.0 hrs, dev15: 1.4 hrs, eval15: 1.4 hrs], vocab.: 600k, LM order: 5gm, SI: speaker independent, ML: maximum likelihood models, SAT-MPE: speaker adapted MPE models, CE: cross-entropy)*

sys.	AM	Criterion	dev12	dev15	eval15
project baseline		-	-	20.8	18.6
A	Multiling-BN + GMM	SI-ML	13.5	7.4	-
		SAT-MPE	11.8	6.8	7.9
B	DNN	SI-CE	13.0	6.9	7.6
C	Multiling-DNN	SI-CE	13.1	6.9	7.5
D	LMM-DNN	SI-CE	12.9	6.8	7.2
	ROVER A+B+C+D		-	**6.3**	**6.8**

3. Polish LVCSR

Three acoustic models were trained for Polish task. We reused the previous year's tandem system, a speaker adapted and Minimum Phone Error (MPE) trained acoustic model based on deep hierarchical multilingual MRASTA bottleneck features (i.e., **A** in Table 4) [1]. The MRASTA pipeline processed 20-dimensional MFCC based critical band energies. The GMM modelled 4500 position dependent tied triphone states. Besides the previous evaluation system, several speaker independent hybrid models were also trained on optimized cepstral feature extraction pipeline. Instead of using MFCC the AM was build on 17 frames of high-resolutional (50-dimensional) CRBEs of a Gammatone filter-bank output. The hybrid model estimates posterior probabilities of 12k states and contained 12 rectifier liner unit hidden layers containing 2000 nodes each. The matrix of the last layer was low-rank factorized by a linear 512-dimensional BN layer [20]. As can be seen in Table 4, the new speaker independent system without any multilingual boosting clearly outperforms the older model. In addition, we also applied the multilingual training on this hybrid model structure in the same way as with the MRASTA bottleneck features. The Polish data were included in the mutlilingual training and only the output layer was made language dependent. Additional fine-tuning step on the target language data was not performed. Furthermore, a recently proposed joint training of BN+GMM technique was also investigated [10]. First, a ML GMM with pooled covariance matrix was trained on the bottleneck output of the unilingual DNN described above. Instead of using 12k tied states, the GMM approximated the emission probability distribution of only 4.5k states. After performing 3 splits the GMM was transformed to Log Linear Mixture Model which can be considered as a softmax layer with hidden variables. Then the whole structure was fine-tuned according to the cross-entropy (CE) criterion. During the recognition maximum approximation was applied similar to [10]. Table 4 shows that a deep hybrid LMM with 4.5k outputs achieved slightly better result than the direct modelling of 12k outputs. Modelling less triphone states could results in smaller lexical prefix-tree for decoding. Although the multilingual DNN system resulted in very similar results to the unilingual DNN, the multilingual training could make the AM more robust. Modified Kneser-Ney smoothed 5gm domain-adapted language model is used during decoding. We combine all the advantages of the above mentioned systems using ROVER based system combination [21]. We achieved significant WER reductions of [dev:7.4%, eval:13.9% (rel.)] using ROVER, compared to system A. Technical details of the project baseline are not released by the project committee.

4. Urdu LVCSR

Two methods are used for acoustic training: Maximum Likelihood Estimation (MLE) and Minimum Phone Error using only short-term spectral features. Speaker adaptation is performed to reduce mismatches between the training data and the test data. The input features are sixteen mel-frequency cepstral coefficients with energy and a voicedness feature. Nine consecutive frames are then concatenated together to account for frame context dependencies, and then transformed by a linear discriminant analysis matrix to 45 dimensions. The acoustic model consists of Gaussian HMMs with a pooled diagonal covariance. The training data features are first linearly aligned to the transcriptions. Ten iterations of monophone alignment and subsequent single density accumulation are performed. The resulting alignments are then used to estimate classification and regression tree (CART) containing 4501 states, and the LDA matrix. These LDA transformed features are then (iteratively) further aligned and then split into mixture densities. This iterative process continues until a maximum of 256 densities per CART states is reached. The resultant model is then used to estimate the speaker-dependent Constrained Maximum Likelihood Linear Regression matrices (CMLLR), to normalize the effect of the speaker variations in the features [22]. After adding these matrices to the feature extraction pipeline, the mixture densities are re-estimated and split once again to the desired resolution (speaker adaptive training). Urdu lexicon contains

Table 5: Urdu WERs (OOV:[dev: 0.9, eval: 5.1], audio:[dev:1.0 hrs, eval: 0.5 hrs], vocab.: 79K, LM order: 5gm)

Experiment	**dev**		**eval**	
	PPL	WER (%)	PPL	WER(%)
Speaker independent	114	30.5	182	44.1
+Speaker adaptation		29.8		34.8
+ MPE		27.9		33.6
+ BO∽Maxent LM	107	27.7		33.6
+ BO∽Adap. Maxent LM			72	33.2
+ Lexical Adaptation				**32.6**

∽: represents linear interpolation, BO: count-based LM

Table 6: *LM perplexities on dev-15 Portuguese corpus (FW: full-word LM, SW: sub-word LM, LSTM training: 40 iterations)*

LM	**# classes**	**dev-15**		**eval-15**	
		FW	**SW**	**FW**	**SW**
BO	-	286.3	308.4	226.1	250.7
LSTM	20	210.0	338.1	-	-
	50	201.2	309.2	-	-
	100	**200.9**	**308.1**	170.8	193.0
BO∽LSTM	-	**178.9**	**207.5**	153.8	167.1

∽: represents linear interpolation, BO: count-based LM

79k words as vocabulary along with hesitation and noise related special tokens. It consists of 57 basic phonemes, inclusive of silence phoneme and pronunciations are generated using linguistic rules. However, Urdu language contains a lot of loan words, used sometimes either in their foreign word form or in some modified form. For instance, word like حقوق (h u q U q) in Urdu means 'Rights' in English. These pairs are generally used alternatively in the same context at spoken level. But, the same phenomenon is not observed in written text as purists prefer to write only Urdu orthography in almost all the cases, except where there is no alternative. Therefore, an English word like Rights is an unseen word carrying zero probability mass. To solve this problem, we investigate a supervised lexical adaptation using Urdu-English dictionary look-up, where, we add the pronunciation of the word Rights to the word حقوق as a pronunciation variant in the lexicon. Therefore, an OOV word like Rights could still use the same probability mass as the word حقوق and thus not an OOV anymore. For our experiments, we apply this approach only for a selected fraction of the most-frequent existing words based on word frequency.

In recognition, after the first pass, CMLLR and MLLR parameters are estimated. The speaker labels for adaptation are assigned by clustering the segments in the feature space. The Gaussian means, variance and the features are modified accordingly and speaker adaptation pass is performed. In this work, N-gram features are used to generate 5-gram maximum entropy LMs [13]. Supervised LM adaptation is also performed to minimize the domain mismatch at LM level. We perform maximum a-posteriori (MAP) adaptation using Gaussian priors over trained maximum entropy LMs. The maximum entropy LM is trained on background text along with the features of an in-domain text. The priors computed from background text are used to learn parameters from the in-domain text. During adaptation, the regularized log-likelihood of the adaptation data is maximized. As a supervised adaptation, development corpus is used as an in-domain text to generate 5-gram adapted maximum entropy LMs [14]. To capture the advantages of both the count-based backoff N-gram and the (adapted or non-adapted) maximum entropy based language models, linear interpolation is performed between them [23]. Using the above techniques, we achieved significant WER reductions compared to the baseline system, as shown in Table 5.

5. Portuguese, English and Arabic LVCSRs

Multilingual MRASTA features concatenated with short-term spectral features are applied for Portuguese, English, and Arabic languages.

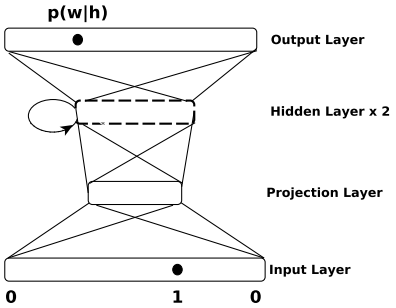

Figure 1: *Architecture of a Neural Network LM*

$$p(w_m|w_1^{m-1}) = p(w_m|c(w_m), w_1^{m-1})p(c(w_m)|w_1^{m-1}) \quad (1)$$

Short-term features include either MFCCs or PLPs, in-addition to a voiced feature. Warping factors are estimated using vocal tract length normalization (VTLN). All feature vectors within a sliding window are concatenated and projected to a 45 dimensional feature space using linear discriminative analysis (LDA). Multilingual approach is used to extract robust MLP features instead of uni-lingual approach. Speaker independent and the speaker adapted acoustic models are generated using the similar recipe, i.e., tandem approach as described in [1]. We generated full-word count-based backoff language models for Portuguese, English and Arabic LVCSR systems [24]. Alternatively, we generated sub-word level count-based and LSTM language models for Portuguese. To generate sub-words, word decomposition is performed using *Morfessor* [25], a statistical tool to generate decomposition of words using minimum

Table 7: *Results (count-based: count-based backoff LM, CN: confusion network, audio: [dev15:1.4 hrs, eval15: 1.4 hrs] for Arabic & Portuguese, [dev: 3.0 hrs, eval: 1.0 hrs] for English, LM order : 4gm for Arabic and English, 5gm for Portuguese)*

Task	**vocab.**	**method**	**dev15**					**eval15**				
			OOV	**PPL**		**WER (%)**		**OOV**	**PPL**		**WER (%)**	
				word	**char**	**mfcc**	**plp**		**word**	**char**	**mfcc**	**plp**
Arabic‡	project baseline		-	-	-	35.5		-	-	-	33.7	
	475k	count-based	4.8	534.0	2.718	**23.5**	-	2.6	614.4	2.816	**19.7**	-
English†	170k	count-based	0.7	124.5	2.411	22.2	21.3	0.8	177.4	2.503	19.5	17.8
		CN combination (count-based)				**19.1**		-	-	-	**17.5**	
Portuguese‡, full-word	project baseline		-	-	-	34.7		-	-	-	35.1	
	171k	count-based	1.0	286.3	2.250	21.2	21.8	1.4	226.1	2.218	17.9	18.2
		LSTM		200.9	2.128	19.2	19.8		170.8	2.128	16.0	16.2
		count-based∞LSTM		178.9	2.103	19.0	19.5		153.8	2.096	15.9	**16.1**
sub-word	160k	count-based	0.9	308.4	2.274	21.4	21.8	1.4	250.7	2.252	17.9	18.2
		LSTM		308.1	2.274	19.6	19.9		193.0	2.167	16.2	16.2
		count-based∞LSTM		207.5	2.148	19.5	19.9		167.1	2.122	16.2	**16.0**
full-word, sub-word		CN combination (count-based)				**20.4**		-	-	-	**17.1**$^{\blacktriangle}$	
		CN combination (count-based∞LSTM)				**18.6**		-	-	-	**15.2**	

‡: Official scoring, †: Skynews system, ∞ : Linear interpolation, $^{\blacktriangle}$: Evaluation submitted system

description length (MDL) principle. As the LM text for Portuguese is sparse (14M running words), decomposition model is trained using all the words from the corpus. This model is still capable of decomposing unseen words. The resulting decomposed sequences are post-processed to generate a clean and not very-short morphemes, which are generally difficult to recognize. For easy reconstruction of a full-word from sub-word sequences, each non-boundary sub-word is marked with a symbol '+'. Pronunciations for sub-word units are generated using word pronunciation alignment [26]. The recognition vocabulary contains a fraction of most-frequent full-words preserved in their original form without decomposition, followed by sub-word units [27]. This is necessary, as count estimates of most-frequent full-words are more reliable compared to remaining words. This helps to recognize most frequent words as full-words instead of sub-word sequences to some extent. Otherwise, a mis-recognition of a single sub-word for most frequent words could have a negative impact on the WER. However, the optimal point is found using 160k hybrid vocabulary, where 145k are full-words, and the remaining are sub-word units. This kind of optimal point is expected due to data sparsity. Since sub-word LMs need more contextual information, LSTM language models are investigated using hybrid vocabulary of 160k size. Our recognition setup contains speaker independent recognition pass followed by recognition using speaker adapted models. LM rescoring followed by system combination is performed in additional passes.

Neural network based LM like LSTMLM takes advantage of a long range word context, which is a highly desirable property, unlike backoff LM. Neural network topology is shown in Fig. 1, where LSTM units are plugged into the second recurrent layer. We created a network with a linear layer with identity activation function and two subsequent LSTM layers, each comprising of 300 neurons. To speed up the training process, words are split into predefined number of hard classes using singular value decomposition (SVD) and k-means algorithm [28]. If $c(w_m)$ is a class of w_m, then word probability is computed using Eq. 1. The number of classes are optimized and perplexities for various language models are shown in Table 6. For direct comparison, character perplexities are also shown for full-word and sub-word LMs in Table 7. As shown in Table 7, LSTM sub-word LMs performed marginally better compared to the full-word LMs on eval corpus interms of word error rate, under sparse data conditions. Similar reductions are observed in-terms of the word and character-level perplexities. Alternatively, significant reduction in WER is observed by combining full-word and sub-word systems using confusion network combination [18]. Confusion network combination helped to unify complementary information from both the full-word and sub-word systems. As LSTM LMs were not ready at the time of evaluation, only count-based LMs were used for experiments. On the other hand, it is observed that MFCC system gave better WER performance compared to PLP system for Portuguese, unlike English. We achieved WER reductions of {[Portuguese-dev: 8.8%, eval: 11.1%] (rel.)} using CN combination on the count-based∞LSTM system, compared to the baseline count-based CN combination system. Technical details of the project baselines are not released by the project committee, yet.

6. Conclusions

Improved LVCSR systems for various languages developed by the RWTH Aachen University are presented. These systems are used for Euronews-2015 and Skynews-2015 evaluations as part of the EU-Bridge project. Multilingual bottleneck features are investigated. The performance of hybrid and tandem approaches are compared. Integrated Gaussian mixture models with deep neural network models performed better than the other investigated approaches for Polish. We achieved significant reductions in WER using language model adaptation and lexical adaptation for Urdu. We made an attempt to recognize OOVs using lexical adaptation for Urdu, and sub-word language models for Portuguese. We achieved significant reductions both in-terms of the perplexities and WERs using open-vocabulary LSTM language models, along with confusion network combination for Portuguese. The RWTH LVCSR systems for the languages described in this work all ranked first in both the Euronews-2015 and Skynews-2015 evaluation campaigns.

7. Acknowledgements

This work was funded by the European Community's 7^{th} Framework Programme (FP7: n° 287658) under the project EU-Bridge (http://www.eu-bridge.eu). Hermann Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Ile-de-France.

8. References

[1] 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, Sep. 2014, pp. 973–977.

[2] D. Vergyri, A. Mandal, W. Wang, A. Stolcke, J. Zheng, M. Graciarena, D. Rybach, C. Gollan, R. Schlüter, K. Kirchhoff, A. Faria, and N. Morgan, "Development of the SRI/Nightingale Arabic ASR system," Brisbane, Australia, Sep. 2008, pp. 1437–1440.

[3] 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, Hawaii, USA, Apr. 2007, pp. 757 – 760.

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

[5] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual Acoustic Models using Distributed Deep Neural Networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013.

[6] K. Veselý, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *IEEE Workshop on Spoken Language Technology*, Miami, Florida, USA, Dec. 2012, pp. 336–341.

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

[8] Z. Tüske, J. Pinto, D. Willett, and R. Schlüter, "Investigation on cross- and multilingual MLP features under matched and mismatched acoustical conditions," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 7349–7353.

[9] M. Paulik, "Lattice-based training of bottleneck feature extraction neural networks," in *Interspeech*, Lyon, France, 2013, pp. 89–93.

[10] Z. Tüske, M. A. Tahir, R. Schlüter, and H. Ney, "Integrating Gaussian Mixtures into Deep Neural Networks: Softmax layer with hidden variables," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Brisbane, Australia, Apr. 2015, pp. 4285–4289.

[11] G. Heigold, R. Schlüter, and H. Ney, "On the Equivalence of Gaussian HMM and Gaussian HMM-like Hidden Conditional Random Fields," in *Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 1721–1724.

[12] M. Sundermeyer, H. Ney, and R. Schlüter, "From Feedforward to Recurrent LSTM Neural Networks for Language Modeling," vol. 23, no. 3, Mar. 2015, pp. 517–529.

[13] T. Alumäe and M. Kurimo, "Efficient Estimation of Maximum Entropy Language Models with N-gram features: an SRILM extension," in *Interspeech*, Chiba, Japan, September 2010.

[14] M. A. B. Shaik, A. El-Desoky, R. Schlüter, and H. Ney, "Investigation of Maximum Entropy Hybrid Language Models for Open Vocabulary German and Polish LVCSR," in *Interspeech*, Portland, OR, USA, Sep. 2012.

[15] M. A. B. Shaik, A. El-Desoky Mousa, R. Schlüter, and H. Ney, "Feature-rich sub-lexical language models using a maximum entropy approach for German LVCSR," in *Interspeech*, Lyon, France, Aug. 2013, pp. 3404–3408.

[16] T. Mikolov, S. Ilya, D. Anoop, L. Hai-Son, K. Stefan, and C. Jan, "Subword Language Modeling with Neural Networks," website: http://www.fit.vutbr.cz/ imikolov/rnnlm/, Tech. Rep., 2012.

[17] A. El-Desoky, H. Jeff Kuo, L. Mangu, and H. Soltau, "Morpheme-based Feature-rich Language Models Using Deep Neural Networks for LVCSR of Egyptian Arabic," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 8453–8439.

[18] B. Hoffmeister, "Bayes risk decoding and its application to system combination," Ph.D. dissertation, RWTH Aachen University, Aachen, Germany, july 2011.

[19] 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 Workshop*, Santa Barbara, CA, USA, December 1997, pp. 347 – 354.

[20] 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 *ICASSP*, Vancouver, Canada, May 2013, pp. 6655–6659.

[21] J. G. 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 Workshop*, Santa Barbara, CA, 1997, pp. 347 – 352.

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

[23] C. Samuelsson and W. Reichl, "A class-based language model for large-vocabulary speech recognition extracted from part-of-speech statistics," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Phoenix, AZ, USA, Mar. 1999, pp. 537 – 540.

[24] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 2, Denver, Colorado, USA, Sep. 2002, pp. 901 – 904.

[25] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0," Computer and Information Science Helsinki University of Technology, Finland, Tech. Rep., Mar. 2005.

[26] R. I. Damper, Y. Marchand, J. D. Marsters, and A. Bazin, "Aligning letters and phonemes for speech synthesis," in *5th ISCA Speech Synthesis Workshop*, Pittsburg, PA, USA, Jun. 2004, pp. 209 – 214.

[27] M. A. B. Shaik, A. El-Desoky, R. Schlüter, and H. Ney, "Hybrid Language Models Using Mixed Types of Sub-lexical Units for Open Vocabulary German LVCSR," in *Interspeech*, Florence, Italy, Aug. 2011, pp. 1441 – 1444.

[28] A. El-Desoky, M. A. B. Shaik, R. Schlüter, and H. Ney, "Morpheme Based Factored Language Models for German LVCSR," in *Interspeech*, Florence, Italy, Aug. 2011, pp. 1445 – 1448.