Multilingual Features Based Keyword Search for Very Low-Resource Languages

Pavel Golik[1], *Zoltán Tüske*[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
{golik,tuske,schlueter,ney}@cs.rwth-aachen.de

Abstract

In this paper we describe RWTH Aachen's system for keyword search (KWS) with very limited amount of transcribed audio data available in the target language. This setting has become this year's primary condition within the Babel project [1], seeking to minimize the amount of human effort while retaining a reasonable KWS performance. Thus the highlights presented in this paper include graphemic acoustic modeling; multilingual features trained on language data from the previous project periods; comparison of tandem and hybrid DNN-HMM acoustic models; processing of large amounts of text data available on the web and the morphological KWS based on automatically derived word fragments.

The evaluation is performed using two training sets for each of the six current project period's languages – full language pack (FLP), consisting of 30 hours and very limited language pack (VLLP), comprising less than 3 hours of transcribed audio data. We put our focus on the latter of the two, which is clearly more challenging. The methods described in this work allowed us to exceed 0.3 MTWV on five out of six languages using development queries.

Index Terms: acoustic modeling, keyword search, graphemic, multilingual, neural networks, semi-supervised learning

1. Introduction

Constantly growing amounts of audio data have increased the demand for robust and scalable keyword search (KWS) systems. The goal of a KWS system is to detect occurrences of a keyword in the index computed from audio data. This index is usually obtained by a speech-to-text (STT) system and stored as a lattice. The search procedure thus can select hypotheses even if they are not part of the first-best path through the word graph. A hypothesis is selected if the score based on the word posterior probability exceeds a decision threshold. Given a set of hypotheses and the true transcription of the test data, the performance of a KWS system can be measured in Actual Term Weighted Value (ATWV), a quantity based on the negative average value lost per term [2]. The value loss is a linear combination of the probabilities of miss and false alarm errors at the actual detection threshold. The threshold is optimized on a development corpus with a development keyword set by maximizing the term weighted value (MTWV). In this work we report the MTWV results, which does not require tuning the decision threshold and allows to compare KWS systems easily.

One of the goals of the IARPA Babel program is to explore the operating point that minimizes the amount of human effort required to prepare language specific resources for setting up a KWS system in a new language. This year's evaluation includes the very limited language pack (VLLP), a scenario where only 3 hours of transcribed audio data and about 30-40 hours of untranscribed data is available. However, no pronunciation lexicon is provided with the VLLP. In contrast, the full language pack (FLP) contains about 30 hours of transcribed speech and a proper pronunciation lexicon is available.

Table 1: *Development languages of Babel's project period OP2.*

Id	Language	Acronym	Data set
205	Kurmanji Kurdish	KMR	IARPA-babel205b-v1.0a
207	Tok Pisin	TPI	IARPA-babel207b-v1.0a
301	Cebuano	CEB	IARPA-babel301b-v2.0b
302	Kazakh	KAZ	IARPA-babel302b-v1.0a
303	Telugu	TEL	IARPA-babel303b-v1.0a
304	Lithuanian	LIT	IARPA-babel304b-v1.0b

The VLLP scenario poses a challenge to building robust acoustic models. The goal of this paper is to describe and evaluate a set of methods for building KWS systems capable of dealing with the limited resources available in the target languages. These methods include training of multilingual DNNs using rectified linear units and training of hybrid DNN-HMM models on multilingual features, as well as graphemic pronunciation modeling. The evaluation is performed on six languages listed in Table 1.

This paper is structured such that the reader can follow the steps of KWS system development. We put special focus on the following topics. In case of VLLP, we start by creating a graphemic pronunciation lexicon (Sec. 3.1). For acoustic model training we make heavy use of the multilingual features (Sec. 3.2) that have been shown to outperform unilingual features especially in case of limited resources [5]. A semi-supervised training is performed using confidence weighting on HMM state level (Sec. 3.3). The speaker adaptation of both, tandem [7] and hybrid acoustic models [8] also takes the confidence weights into account (Sec. 3.4). We generate lattices for the KWS by using large lexica and language models (LMs) based on text resources downloaded from the web (Sec. 3.5). The KWS is performed on multiple levels: an in-vocabulary (IV) search of the keywords known at the decoding time can be performed directly on the lattice; the out-of-vocabulary (OOV) terms are first expanded by a weighted finite state transducer (WFST) that allows substitutions of graphemes; alternatively, the OOV terms are split into subword units using a morphological segmentation learned from the data and searched in a lattice that has been created using a subword lexicon and LM (Sec. 3.6). The results are presented and discussed in Sec. 4 and conclusions are drawn in Sec. 5.

Table 2: *Statistics of VLLP training and testing corpora.*

Language	Set				Lexicon size
	Training		Test		
	amount of speech [h]	running words		OOV [%]	
205 KMR	2.5	30k	100k	9.4	3.7k
207 TPI	2.7	35k	99k	4.6	1.9k
301 CEB	2.6	33k	99k	10.9	3.8k
302 KAZ	2.6	26k	81k	14.5	5.3k
303 TEL	2.4	25k	81k	23.2	7.1k
304 LIT	2.7	27k	102k	18.1	5.5k

2. Task description

The participants of the Babel project are given about 100 hours of audio data in each language. This year, less than 50% of the data is transcribed. After segmentation and discarding large silence intervals, the amount of transcribed training data totals to about 30 hours per FLP. A subset of less than 3 hours is used for training on the VLLP task. Table 2 shows the corpus statistics for the VLLP task. The test set for both language packs is the same and consists of approx. 8 hours. The tuning of meta-parameters for the VLLP task was performed on a 3 hours tuning set `vllp_tune`. In the VLLP track, all of the untranscribed data can be used for unsupervised training. A list of about 2000 development query terms is available in each of the six languages. As shown in Table 5, 24 to 45% of these keywords are OOV. The KWS system is required to find all occurrences of each query term in the test audio. The higher the resulting MTWV score, the better. The Babel program considers an ATWV of 0.3 to be the goal on each of six languages and each of the two language packs (VLLP and FLP).

3. Experimental setup

For acoustic training and recognition we used the RASR toolkit [9, 10]. The WFST based software for KWS was provided by our project partner IBM and integrated into the RWTH Aachen infrastructure. In the basic feature extraction pipeline we concatenate Gammatone [11] features, probability of voicedness [12] and an F_0 estimate [13]. We always use Kneser-Ney smoothed LMs with optimized discounts [14]. The STT performance is measured in term error rate (TER), which is equivalent to word error rate for this year's languages.

3.1. Graphemic pronunciation modeling

The set of graphemes was derived from the training transcriptions after removing non-word events and punctuation and converting the text to lower case (where casing is applicable). The generation of the pronunciation lexicon is then done by a character lookup. This lexicon was used to train an initial context-independent (CI) GMM acoustic model (AM) and to force-align the training transcriptions. In order to train a context-dependent (CD) model with a CART based state-tying, we usually need a set of phonetic questions that is used to group phonemes in some linguistically meaningful clusters. In case of graphemic modeling, we follow methods described in [3, 4] in order to obtain "graphemic questions" in a data-driven way. The idea is to perform a greedy bottom-up clustering of the mean vectors of the CI single Gauss model. Then, every intermediate cluster is considered a "question" to be used during CART estimation. E.g. on Tok Pisin (207), this procedure automatically discovered reasonably looking clusters like $\{c, k, p, t\}$ and $\{l, m, n\}$.

3.2. Multilingual features

Since the VLLP evaluation condition allows to re-use the transcribed audio data of other languages from former project periods, we continued our effort on multilingual training of neural networks. The effectiveness of this approach has been shown earlier on the limited language pack [5] (approx. 10h of transcribed data).

The multilingual feature extraction is done by a hierarchical processing of MRASTA features [15, 16]. Two multilingual bottleneck (BN) DNNs are trained on a combination of several features: the first network is trained on the *fast* modulation part of MRASTA filtering, critical band energies (CRBE), F_0 and voicedness features. The input to the second network consists of the *slow* modulation part of MRASTA features, CRBE, F_0 and voicedness features as before, as well as the BN features from the first DNN. These BN features are stacked to accumulate a temporal context of 9 frames. The time delay in the feature stream introduced by the hierarchical processing can be interpreted as a convolution in time [17], if the training of both DNNs is performed jointly. The CRBE features are a concatenation of three single CRBE streams, derived from different short-term feature extraction pipelines (MFCC, PLP and Gammatone) [5].

Both DNNs have language specific output layers [18] with 1500 outputs per language corresponding to the tied triphone states. The error is calculated w.r.t. the cross entropy (CE) criterion and the backpropagation is carried out only in the output layer from the same language as the input vector. The DNNs have five hidden layers before and one after the BN layer. The large hidden layers consist of 2000 rectified linear units (ReLUs) [19] while the BN layer of size 62 uses the sigmoid activation function. During the final feature extraction, the activation function of the BN layer is replaced by identity.

The multilingual training is performed on FLPs of the 11 languages of the first two project periods, totaling about 600 hours of data. In addition, the multilingual features can be adapted to the target language in a supervised and a semi-supervised manner. Given the small amount of transcribed data, the supervised adaptation of the multilingual features [5] was done on the VLLP corpus augmented with replicas generated by the Vocal Tract Length Perturbation (VTLP) approach [20, 21]. The semi-supervised training (cf. Sec. 3.3) of the transferred BN features was carried out on the union of the untranscribed and supervised corpora. The training concludes with a supervised fine-tuning step of the DNN [22].

3.3. Semi-supervised learning

Our basic semi-supervised learning (SSL) procedure is the following:

1. train a supervised AM on 3h VLLP data
2. transcribe unsupervised data
3. filter by confidence on state level
4. down-weight the unsupervised frames
5. merge supervised and unsupervised data
6. train the final AM

The filtering in Step (3) is performed in order to remove the frames whose labels are likely to be wrong [6]. We found a threshold of 0.75 to work quite well, which leads to removing about 1/4 of all recognized data.

The down-weighting in Step (4) is meant to keep a balance between the supervised and unsupervised data. The selected 3/4 of approx. 50h of untranscribed data (about 37h) is down-weighted by multiplying the frame weights by 0.25 similar to [23]. This makes the automatically transcribed data to contribute about 75% to what is used in Step (6).

3.4. Hybrid DNN-HMM acoustic models

The input features for the hybrid DNN models are the same as for the GMM: LDA transformed Gammatone+voicedness+F_0 features concatenated with LDA transformed multilingual bottleneck features (115 dimensions). In contrast to GMM, the DNN can be easily trained on multiple stacked frames. Thus, the input to the DNN has $17 \cdot 115 = 1955$ dimensions. The output layer corresponds to approx. 1.8k CD states obtained from a CART trained on the VLLP data. There are 6 hidden layers with 2048 rectified linear units each. This topology has been optimized on the tuning set. The speaker adaptation is performed by multiplying the input features with CMLLR matrices obtained from the GMM setup [24, 16]. The models were trained according to the CE criterion using L_2 regularization factor of 10^{-4}.

Further, semi-supervised training of the DNN is performed based on the alignment obtained from the GMM system. As described in Sec. 3.3, the DNN training can take frame weights into account when performing backpropagation using possibly erroneous transcriptions. In order to reduce the effect of wrong labels on the model quality, we ran a few training epochs on the supervised VLLP data only in order to "fine-tune" the semi-supervised model.

The CE model can be further improved by performing sequence-discriminative training. We trained the output layer of the DNN according to the minimum phone error (MPE) criterion [25] in a setup similar to that in [10]. The optimization was done using the Rprop algorithm [26].

3.5. Lexicon and LM extension with web data

The web data compiled for all project participants by IBM and BBN can be used in the VLLP track to increase the lexicon size and improve the LMs. We have integrated the data provided by both teams into the STT and KWS pipelines. First, a list of 100k most frequent words is selected from all available text data and merged with the original lexicon built from VLLP audio transcriptions. The list is restricted to words with characters that are already present in the audio training data since the graphemic acoustic model needs observations for every character. Then this lexicon is used to estimate two separate LMs (one on BBN data, one on Wikipedia from IBM's release [27]). The resulting LMs are then interpolated with the original LM by optimizing perplexity on `vllp_tune`. It is worth mentioning that the original LM estimated on approx. 25k running words has been assigned the largest interpolation weight (about 0.7), while the web data (50M to 100M running words) has been assigned a weight of about 0.25. We use a 4-gram LM for STT experiments and a bigram LM for KWS lattice generation.

3.6. G2G and morphological KWS

The keywords that were out-of-vocabulary during the lattice generation, cannot be found by iterating over the arcs. The regular approach to solve this problem is to expand the query and the lattice on the grapheme level and to allow for substitutions and deletions by composing the query with a *grapheme-to-grapheme* (G2G) WFST whose weights are estimated on the training data [28, 29]. We train the G2G WFST by calculating a grapheme confusion matrix of two alignments – one obtained by force-aligning the training data to the true transcriptions, and one obtained by recognition with a unigram LM.

An alternative approach is to segment the in-vocabulary words automatically in order to obtain some kind of *morphological* decomposition, where the known words share multiple fragments. The lattice generation is then performed with a new lexicon and a new LM estimated on the decomposed words. Finally, the OOV queries are decomposed in the same manner and the KWS is done on the new lattice. We used the Morfessor2 toolkit [30] to obtain a morphological segmentation of the text data. Finally, the hit lists from both IV and morph KWS are merged for scoring. It is also possible to merge hit lists from different KWS systems for system combination purposes.

The average lattice density exceeds 10000 arcs/second, the G2G expansion of the graphemic form of the queries was limited to 5000 best paths. A sum-to-one score normalization was also applied for each keyword separately [28].

4. Experimental results

Following the same structure as in the previous section, we start by giving a brief summary on the comparison between phonetic and graphemic systems. The graphemic systems (cf. Sec. 3.1) have shown an STT performance competitive with the phonetic baseline systems and even outperformed them on some of the VLLP tasks. We will omit the detailed results for this topic, since the comparison was done on an early stage of system development and the differences tended to become smaller as the acoustic models improved. Hence all results reported in this work are based on graphemic pronunciation lexica.

As mentioned earlier, this year's primary evaluation condition allows to use multilingual features in the VLLP track. The detailed description of the multilingual DNN can be found in Sec. 3.2. We first compared different methods of adapting these features transferred from the multilingual DNN to the target language. The baseline setup shown in Table 3 corresponds to the case where no further processing of the BN features is performed, which is the fastest approach if the multilingual DNN has been already trained before. The next row shows the results obtained after fine-tuning the multilingual DNN on the target language. The amount of in-domain data has been increased fivefold by adding replica of the VLLP set artificially perturbed using VTLP. Given the small amount of training data, this still can be done very fast. If we now use this AM to recognize the untranscribed data, we can perform the semi-supervised learning (SSL) on two levels (cf. Sec. 3.3): fine-tuning the bottleneck (BN) features and training the GMM. As the last two rows in Table 3 suggest, the improvement of updating both models is rather small on 2 out of 3 languages.

Table 3: *Supervised and semi-supervised adaptation of multilingual BN features transferred to the target language. Results in TER [%] after SAT+MPE training of the tandem AM.*

AM	KMR	TPI	TEL
baseline	71.6	47.4	76.7
+ VTLP (BN)	71.5	47.1	76.4
+ SSL (GMM)	70.3	45.5	76.1
+ SSL (BN+GMM)	70.2	44.8	76.0

In the next set of experiments we compare the effect of speaker adaptive training (SAT), semi-supervised learning (SSL) and sequence discriminative training w.r.t. the MPE criterion on a tandem GMM-HMM and a hybrid DNN-HMM acoustic model in parallel (cf. Sec. 3.4). Table 4 shows the STT results on 3 most difficult languages of the current project period. While the GMM benefits greatly from the SAT using CMLLR transforms, the DNN gets its largest improvement by including more training data in form of SSL. Even though the difference in TER after the MPE training becomes small, the KWS results show that tandem AM can still outperform the hybrid model. Presumably, the word posterior scores in a DNN lattice follow a different distribution than in a lattice generated by a GMM, which is less beneficial for the KWS purposes. We can benefit from this difference by combining the posting lists of the two KWS systems. A strong increase in MTWV shows that the lattices capture hypotheses that are complementary to some extent.

Table 4: *Comparison of the two strategies: language adapted multilingual BN tandem GMM-HMM vs. hybrid DNN-HMM based on multilingual features. Term error rates in [%]. MTWV measured on dev queries.*

	AM	KMR		CEB		TEL	
		GMM	DNN	GMM	DNN	GMM	DNN
TER [%]	baseline	73.3	75.5	68.7	67.1	78.3	78.3
	+ SAT	71.8	74.5	65.9	66.4	76.8	78.6
	+ SSL	70.5	71.7	64.4	62.5	76.3	75.6
	+ MPE	70.2	70.8	63.9	62.1	76.0	75.2
MTWV		0.234	0.211	0.283	0.321	0.215	0.194
combination		0.246		0.332		0.224	

In the following experiment, we make use of the web data to increase the lexicon size to approx. 100k words and to estimate a new LM as described in Sec. 3.5. As shown in Table 5, this step consistently improves the STT results (especially on Lithuanian and Kazakh) and also strongly reduces the OOV rate among keywords. Please note that the perplexity values cannot be compared directly, since they are calculated on different vocabularies. The numbers are included in the table in order to illustrate the quality of text resources available in each language.

Table 5: *Comparison of STT performance using the base lexicon and LM derived from audio transcriptions with webdata based systems.* $(^*)$ *GMM tandem acoustic model*

id	Language	TER [%]		PPL		KW OOV [%]	
		base	web	base	web	base	web
205	KMR*	70.2	69.6	148.8	227.2	29.2	11.9
207	TPI*	44.8	44.3	67.3	82.3	23.8	10.9
301	CEB	62.1	60.3	111.9	175.2	29.3	12.4
302	KAZ	61.7	59.9	185.1	374.7	35.3	10.6
303	TEL	75.2	74.0	261.7	547.3	43.6	19.3
304	LIT	58.4	52.9	153.8	388.1	41.6	12.4

Since the KWS of OOV queries is known to be very difficult, we compared the performance of different OOV handling approaches described in Sec. 3.6. Thus, our baseline systems employ G2G WFSTs for OOV keywords. In addition to the web data based lattices, we have set up KWS systems based on morphologically decomposed vocabularies. And finally, in order to see if the combination of these two methods can further improve the KWS performance, we learn the morphological decomposition and estimate the subword LM on all available texts including web data. Table 6 shows the results obtained with these four methods on all development languages. Clearly, using web data to boost the lexicon size and the LM is the most simple and effective method to improve the MTWV.

Table 6: *Comparison of different KWS pipelines on VLLP tasks. Results in MTWV.*

OOV KWS	KMR	TPI	CEB	KAZ	TEL	LIT
baseline	0.234	0.395	0.321	0.335	0.215	0.415
morph	0.239	0.403	0.331	0.353	0.262	0.472
webdata	**0.249**	0.404	0.349	0.409	0.290	**0.549**
web-morph	0.245	**0.406**	**0.355**	**0.411**	**0.296**	0.542

For reference, we include the overall best results obtained on current period's development languages. Table 7 shows the STT and KWS results for both language packs in each language. We only report TERs of single systems in order to illustrate the quality of acoustic models. Some of the KWS results were obtained by combining hit lists from hybrid and tandem systems.

Table 7: *Overall results. VLLP results are boosted with multilingual features and text resources from the web.*

id	Language	TER [%]		MTWV	
		VLLP	FLP	VLLP	FLP
205	KMR	69.6	65.6	0.283	0.289
207	TPI	44.3	40.5	**0.419**	**0.458**
301	CEB	60.3	58.1	**0.355**	**0.408**
302	KAZ	59.9	57.5	**0.411**	**0.409**
303	TEL	74.0	70.1	**0.328**	**0.353**
304	LIT	52.9	52.2	**0.549**	**0.549**

5. Conclusions

In this paper we have shown experimentally, that even without a phonetic pronunciation lexicon and with less than 3 hours of transcribed audio data in the target language, it is possible to achieve a reasonable KWS performance. The most essential requirements are bottleneck features extracted by a multilingual DNN trained on a large amount of non-target language data, as well as the access to text resources from the web. Several semi-supervised learning methods have been evaluated and applied for training of tandem GMM-HMM, and hybrid DNN-HMM models. We compared the effect of speaker adaptation and MPE training on both types of acoustic models. We found that both models achieve a very similar performance on both STT and KWS tasks and a combination of posting lists obtained by two different models can further improve the results. These methods allowed us to exceed 0.3 MTWV on five out of six languages using development queries.

6. Acknowledgements

H. Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Île-de-France. 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.

7. References

[1] "IARPA Babel Program," http://www.iarpa.gov/index.php/research-programs/babel, accessed: 2015-02-27.

[2] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddingtion, "Results of the 2006 spoken term detection evaluation," in *Proc. of ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, 2007, pp. 51–57.

[3] K. Beulen and H. Ney, "Automatic question generation for decision tree based state tying," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, USA, May 1998, pp. 805–808.

[4] S. Kanthak and H. Ney, "Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Orlando, FL, USA, May 2002, pp. 845–848.

[5] Z. Tüske, P. Golik, D. Nolden, R. Schlüter, and H. Ney, "Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages," in *Proc. Interspeech*, Singapore, Sep. 2014, pp. 1420–1424.

[6] C. Gollan, S. Hahn, R. Schlüter, and H. Ney, "An improved method for unsupervised training of LVCSR systems," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 2101–2104.

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

[8] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Norwell, MA, USA: Kluwer Academic Publishers, 1993.

[9] 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 Workshop (ASRU)*, Hawaii, USA, Dec. 2011.

[10] S. Wiesler, P. Golik, R. Schlüter, and H. Ney, "Investigations on sequence training of neural networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Brisbane, Australia, Apr. 2015.

[11] 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, Hawaii, USA, Apr. 2007, pp. 649–652.

[12] A. Zolnay, R. Schlüter, and H. Ney, "Robust speech recognition using a voiced-unvoiced feature," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 2, Denver, CO, USA, Sep. 2002, pp. 1065–1068.

[13] X. Lei, M. Siu, M.-Y. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for Mandarin broadcast news speech recognition," in *Proc. Interspeech*, Pittsburgh, PA, USA, Sep. 2006, pp. 1237–1240.

[14] M. Sundermeyer, R. Schlüter, and H. N. Ney, "On the estimation of discount parameters for language model smoothing," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 1433–1436.

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

[16] 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.

[17] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 3, pp. 328–339, Mar. 1989.

[18] S. Scanzio, P. Laface, L. Fissore, R. Gemello, and F. Mana, "On the use of a multilingual neural network front-end," in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008, pp. 2711–2714.

[19] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. of the 27th Int. Conf. on Machine Learning*, Haifa, Israel, Jun. 2010, pp. 807–814.

[20] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Int. Conf. on Machine Learning (ICML), Workshop on Deep Learning for Audio, Speech and Language Processing*, Atlanta, GA, USA, Jun. 2013.

[21] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Florence, Italy, May 2014, pp. 5582–5586.

[22] F. Grézl and M. Karafiát, "Combination of multilingual and semi-supervised training for under-resourced languages," in *Proc. Interspeech*, Singapore, Sep. 2014, pp. 820–824.

[23] R. Hsiao, T. Ng, F. Grézl, D. Karakos, S. Tsakalidis, L. Nguyen, and R. M. Schwartz, "Discriminative semi-supervised training for keyword search in low resource languages," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Olomouc, Czech Republic, Dec. 2013, pp. 440–445.

[24] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Hawaii, USA, Dec. 2011, pp. 24–29.

[25] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Orlando, FL, USA, May 2002, pp. 105 – 108.

[26] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *Proc. of the Int. Conf. on Neural Networks*, San Francisco, CA, USA, Mar. 1993, pp. 586–591.

[27] G. Mendels, E. Cooper, V. Soto, J. Hirschberg, M. Gales, K. Knill, A. Ragni, and H. Wang, "Improving speech recognition and keyword search for low resource languages using web data," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015, submitted.

[28] L. Mangu, H. Soltau, H.-K. Kuo, B. Kingsbury, and G. Saon, "Exploiting diversity for spoken term detection," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vancouver, Canada, May 2013, pp. 8282–8286.

[29] M. Saraclar, A. Sethy, B. Ramabhadran, L. Mangu, J. Cui, X. Cui, B. Kingsbury, and J. Mamou, "An empirical study of confusion modeling in keyword search for low resource languages," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Olomouc, Czech Republic, Dec. 2013, pp. 464–469.

[30] S. Virpioja, P. Smit, S.-A. Grönroos, and M. Kurimo, "Morfessor 2.0: Python implementation and extensions for Morfessor Baseline," Department of Signal Processing and Acoustics, Aalto University, Helsinki, Finland, Report 25/2013 in Aalto University publication series SCIENCE + TECHNOLOGY, 2013.