SEMI-SUPERVISED TRAINING IN LOW-RESOURCE ASR AND KWS

Florian Metze⋆*, Ankur Gandhe*⋆*, Yajie Miao*⋆*, Zaid Sheikh*⋆*, Yun Wang*⋆*, Di Xu*⋆*, Hao Zhang*⋆*,*
Jungsuk Kim†*, Ian Lane*†*, Won Kyum Lee*†*, Sebastian Stüker*‡*, and Markus Müller*‡

⋆ Language Technologies Institute, † Department of Electrical and Computer Engineering
Carnegie Mellon University; Pittsburgh, PA/ Moffett Field, CA; U.S.A
‡ Karlsruhe Institute of Technology; Karlsruhe; Germany
fmetze@cs.cmu.edu

ABSTRACT

In particular for "low resource" Keyword Search (KWS) and Speech-to-Text (STT) tasks, more untranscribed *test* data may be available than *training* data. Several approaches have been proposed to make this data useful during system development, even when initial systems have Word Error Rates (WER) above 70%. In this paper, we present a set of experiments on low-resource languages in telephony speech quality in Assamese, Bengali, Lao, Haitian, Zulu, and Tamil, demonstrating the impact that such techniques can have, in particular learning robust bottle-neck features on the test data. In the case of Tamil, when significantly more test data than training data is available, we integrated semi-supervised training and speaker adaptation on the test data, and achieved significant additional improvements in STT and KWS.

***Index Terms*—** spoken term detection, automatic speech recognition, low-resource LTs, semi-supervised training

1. INTRODUCTION

Low-resource speech and language technology has been a focus area for several research groups over the last several years. Semi-supervised training of (Deep Neural Network based) acoustic models can be used to exploit situations in which more audio data is available than has been transcribed [1, 2], despite baseline Word (or Token) Error Rates (WER/ TER) hovering around 70%.

Despite the existence of an active community that is developing pattern-matching approaches to keyword search [3, 4, 5], the predominant approach is still to develop a word-based STT system, and search for keywords in a symbolic index, that is often given by confusion networks (CN, [6]) generated by the recognition process, similar to the overall approach described in [7]. In order to be able to also detect out-of-vocabulary (OOV) keywords, we implement Probabilistic Phonetic Retrieval (PPR, [8]), which performs query expansion using observed phonetic confusions and then searches a phone lattice, similar to [9]. Because many low-resource languages also exhibit high morphological complexity, we explored the feasibility of automatically generating morphemes, and built a morpheme-based (rather than word-based) system.

In this paper, we describe our efforts to build STT and KWS systems for six low-resource languages, highlight the most important achievements, and report the final performance achieved on each of them in the Babel program [10] and OpenKWS [11] evaluations.

2. TASK AND DATA DESCRIPTION

For our task, the primary evaluation metric is "Actual Term Weighted Value" (ATWV), a weighted combination of precision and recall averaged over a given list of keywords, computed under the "NTAR" (No Target Audio Reuse) condition, i.e. audio data was processed without knowledge of the keywords, and keyword search was performed on an index, without accessing the original audio [11]. STT is evaluated using standard WER/ TER. We worked on the Babel Option Period 1 (OP1) "development" languages and the OpenKWS 2014 "surprise" language, Tamil.[1] The following conditions were defined for system training:

FullLP, the "Full Language Pack", consists of ˜60 h of annotated speech for each of the languages used

LimitedLP, the "Limited Language Pack", consists of ˜10 h of annotated speech, plus the remaining 50 h of FullLP audio without transcriptions, allowing for semi-supervised training

OtherLR allows the addition of external data. In our case, we downloaded and added text data from the Internet

Most of the data is telephone-quality speech, recorded in a number of different environments, including car kits and hands-free devices on cell-phones in the street. Tamil contained some data with room acoustics, which was not treated differently in our work.

For development purposes, 10 h of test data were available for every language, while the unseen test data consisted of 75 h in Tamil, and 15 h each for all the other languages. In Tamil, an 15 h "Eval-Part1" subset (of the 75 h set) has also been defined. Hence, in the "LimitedLP" condition, and for Tamil in particular, more data is available to test systems, than to train them. Table 1 lists the characteristics of the data sets. Even though various keyword lists were used during development, results shown in this paper have been computed with the official "evaluation" keywords that were distributed with the evaluation data, unless noted otherwise.

Supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ ARL) contract number W911NF-12-C-0015. 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. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. Specifically, it used the Blacklight system at the Pittsburgh Supercomputing Center (PSC).

[1] IARPA-babel102b-v0.4 (Assamese), IARPA-babel103b-v0.3 (Bengali), IARPA-babel201b-v0.2b (Haitian Creole), IARPA-babel203b-v3.1a (Lao), IARPA-babel206b-v0.1e (Zulu), and IARPA-babel204b-v1.1b (Tamil).

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

FullLP	Assam.	Beng.	Haitian	Lao	Zulu	Tamil
Vocab.	22 k	24 k	13 k	6.2 k	54 k	52 k
OOV	3.4%	3.8%	1.6%	0.5%	11.9%	8.6%
PPL	245	295	137	112	420	414
LimitedLP (10 h subset of 60 h “FullLP”)						
Vocab.	7.7 k	7.9 k	4.9 k	3.2 k	14 k	14 k
OOV	8.0%	8.7%	3.8%	2.0%	20.2%	15.1%
PPL	243	291	157	134	299	343

Table 1. Characteristics of the data sets used in this work. OOV rate and perplexity (PPL) have been computed against the 10 h development set, using the training data’s vocabulary.

WER/ ATWV	MFCC	lMEL	PLP
ML		70.9%/ .332	
SAT bMMIE	69.1%/ .350	69.0%/ .339	68.4%/ .349
DNN	69.1%/ .352	68.5%/ .359	68.2%/ .357
7-way Combo		64.7%/ .433	

Table 2. Performance of the individual components of the Tamil FullLP system on development data.

3. SYSTEM OVERVIEW

The STT systems were built using the Janus ASR toolkit using the Ibis decoder [12]. Retrieval is based on a confusion network-based index that has been generated from a union of lattices from several decoding passes, as described in Section 4. Posting lists from multiple retrieval systems are combined together using CombMNZ [13]. Compared to our earlier work [14], the main novelties are:

- Automatic segmentation of the test data is now performed using Kaldi [15], and shared within the RADICAL team’s Kaldi and Janus systems.
- The code for lattice and confusion network generation has been re-written in order to retain more variability in the resulting index [16]. Pruning thresholds have been optimized across all available Babel languages.
- Retrieval (Keyword Search) is now using Probabilistic Phonetic Retrieval (PPR) in order to be able to detect out-of-vocabulary (OOV) words.
- Automatic (pseudo-)morphological decomposition of the words of a language has been employed to further improve retrieve out-of-vocabulary (OOV) words, c.f. Section 4.5.
- The “recipes” for training systems under “FullLP” and “LimitedLP” conditions have been streamlined, and several new techniques have been incorporated, including semi-supervised training and Maxout DNNs for LimitedLP, c.f. Section 4.3.

4. EXPERIMENTS

All systems used in these experiments are based on three different feature sets, which are used as inputs to DNN bottle-neck feature extractors [17], namely BNF_MFCC+Janus_Pitch [18] (“MFCC”), BNF_lMEL+FFV [19] (“lMEL”), and BNF_PLP+Kaldi_Pitch [20] (“PLP”), with per-speaker CMS/ CVN normalization of features. All systems use trigram language models trained on the appropriate data set, unless otherwise noted. The lexicon was used as provided, with some rare phones merged, and tones integrated as tags, if given in a particular language.

4.1. FullLP System

The FullLP systems are a 7-way combination of individual systems: a first-pass unadapted and ML trained GMM model in a lMEL feature space, and then three hybrid DNN and three bMMIE-trained GMMs in three BNF feature spaces generated as described above, then adapted by cMLLR [21]. The GMMs have been trained with semi-tied covariances and LDA, stacking 7 frames, the hybrid models use a context window of 11 frames.

The MFCC and lMEL hybrid acoustic models use i-vector speaker normalization [22, 23], while the DNNs were trained with ReLU nonlinearities [24], mostly for speed reasons.

Table 2 lists the performance of the individual sub-systems, as well as the final performance in Tamil. Figure 1 shows the performance on evaluation data, and for the other languages. All subsystems perform similarly for all languages, i.e. there is a slight advantage of hybrid over BNF-GMM models. Greedy system combination using ROVER or CombMNZ [13] almost always mandated the inclusion of all seven systems (“7-way Combo”) on development data, so that it was used in all evaluation submissions.

For simplicity, no OOV retrieval was performed for the KWS task in the “FullLP” condition.

4.2. “Single” System

For fast and simple indexing, we investigated the performance of a single, unadapted hybrid acoustic model, trained on FullLP data in lMEL feature space, using a FullLP language model. On average, this system performed best of all individual systems. The performance of these systems is shown in Figure 1 in Section 5. Adding OOV retrieval [8] improved ATWV by about .01 for Zulu and Tamil. It was however only included in a Virtual Machine-based prototype, which can be used to easily analyze a set of data in any of the six languages covered in this work, but not the evaluation submissions described here.

4.3. LimitedLP System

For the LimitedLP case, we developed three sub-systems for each language, one for each of the MFCC, lMEL, and PLP feature spaces. Initial supervision for adaptation is again generated by an unadapted and ML trained GMM model with lMEL features. In contrast to FullLP, every sub-system uses four acoustic models: two hybrid DNNs and two bMMIE-GMMs, which were trained in a standard sigmoid BNF feature space as well as a Maxout [25, 26] BNF feature space. Each subsystem’s output is either the consensus word hypothesis generated on four individual lattices, or the retrieval result generated from the combined consensus network of these four individual models. As in the “FullLP” case, the hierarchical use of DNNs trained on top of BNF features creates a TDNN-like [27], almost-convolutional [28] structure with fast training time [17]. Attempts at using i-vector speaker normalization [22] and data augmentation techniques [29] did not consistently improve results at the time of the evaluation, and were therefore not included. The three sub-systems are then combined using Rover (for STT) or CombMNZ (for KWS), in order to produce the final output.

Table 3 shows the results of our semi-supervised training experiments. In each case, we select the 50 % of training segments with highest averaged word confidence score, and retrained BNF features

WER (%)	Ass.	Ben.	Haitian	Lao	Zulu	Tamil
LimitedLP	62.4	65.5	61.9	60.5	67.9	74.0
SS1	60.3	63.0	57.7	56.6	65.8	72.2
SS2	58.7	61.3	-	-	-	70.9
ATWV						
LimitedLP	.255	.223	.318	.316	.273	.218
SS1	.289	.264	.369	.371	.305	.251
SS2	.312	.284	-	-	-	.274

Table 3. Semi-supervised training experiments: "SS1" lists semi-supervised training on the training data, while "SS2" shows performance after re-training on the evaluation data (15 h, except Tamil, which has 75 h). ATWV values are reported using "development" keywords, which for most languages are "harder" than the evaluation keywords, and using word-based retrieval only (i.e. no morphemic systems, no PPR).

WER/ ATWV	MFCC	lMEL	PLP
CNC SS1		72.2%/ -	
CNC SS2	72.1%/ .253	71.9%/ .268	71.8%/ .281
3-way Combo		70.5%/ .313	

Table 4. Performance of the different training strategies for the Tamil LimitedLP system. Results are given using evaluation keywords on evaluation data, without morphemic systems or PPR.

with the expanded training data in a cMLLR adapted feature space [21]. GMM training, GMM adaptation and hybrid acoustic model training are performed using only 10 h of supervised data.

Table 4 breaks down the results of the individual sub-systems in the case of the SS2-Tamil system. The other languages show similar relative performance levels. SS2 systems were not built for Haitian, Lao, and Zulu, because performance goals could be achieved without the extra computation required to perform training of the SS2 system using the SS1 hypotheses, while SS1 was itself trained using a pure LimitedLP system's hypotheses.

By using cross-adaptation with a Kaldi-based system [15], results could typically be improved further by 0.5 % WER and 0.01 in ATWV, exploiting the diversity between these systems, however evaluation timing constraints did not permit integration.

4.4. "OtherLR" Systems based on Webtext

One way to reduce the OOV rate of a recognizer is to download additional text from the Internet or other data sources, from which an expanded vocabulary and additional language model training text is gathered [30]. Using such resources places a system submission into the "OtherLR" category. We generated queries to language-specific Google servers from the LimitedLP training text, which we then used to download additional text documents. We also downloaded Wikipedia in each of these languages, and created an (almost) language independent text normalization and filtering process, which we applied to these corpora.

Table 5 shows the amount of data crawled, and the resulting OOV rate and ATWV. After filtering HTML tags, English words, and non-text parts, the merged data was sorted according to perplexity and vocabulary overlap, and the fraction that empirically best matched the development data was retained.

No improvements could be achieved on Assamese, presumably because the data crawled was too similar to Bengali, and Lao. The crawls were conducted in two parts: one part was done in October/ November 2013, and a second part was done in January/ February 2014. On average, the text from both crawls performed about the same, and merging the results did not improve overall results much, so that we believe that the presented results represent the optimum that can be achieved with our technique, and no significant dependency on "time of crawl" or other random factors exists.

Lines of text	Ass.	Ben.	Haitian	Zulu	Tamil
Queried	95 k	181 k	333 k	318 k	204 k
Wikipedia	274 k	1.6 M	1.3 M	48 k	5.4 M
OOV rate					
KWS	24.8 %	13.4 %	25.6 %	33.6 %	30.2 %
TXT	6.4 %	7.9 %	2.9 %	14.3 %	13.0 %
ATWV					
LimitedLP	.311	.318	.370	.305	.302
OtherLR	.310	.340	.394	.362	.313

Table 5. Characteristics of the web data downloaded, and resulting benefit. 67 k lines of text from a government web-site (ZA-Corp) were added to the Zulu Wikipedia data (48 k lines), as they were empirically found to be helpful. OOV rates, measured against evaluation keywords ("KWS") and against the development test set ("TXT"), as well as ATWV (on dev data) are also reported, c.f. Table 7. For a comparison with the morphemic strategy, see Table 6. The Lao baseline OOV rate was already low, so no experiments were performed.

ATWV	Ass.	Ben.	Haitian	Zulu	Tamil
Baseline	.311	.318	.370	.305	.302
OOV ATWV	.084	.083	.164	.132	.074
Fused ATWV	.330	.322	.388	.326	.312
OOV	1.2 %	3.2 %	1.2 %	0.3 %	0.1 %

Table 6. OOV rates (with respect to keywords) and ATWV for morphemic systems on evaluation data. For a comparison with the "OtherLR" webtext-based strategy, c.f. Table 5.

4.5. Influence of Morphemic Systems

We performed several experiments training a segmentation model on LimitedLP data using Morfessor-Cat (version 0.9.2, [31]), Fast_umorph (https://github.com/vchahun/fast_umorph) and also a data-driven segmentation developed at JHU. Overall, Morfessor performed best, so we will not present further details on the other approaches. The The trained model was used to segment the lexicon, the language model text and the keywords. The pronunciations for pseudo-morphemes were obtained using G2P [32], trained on the LimitedLP lexicon. Acoustic Models were used as-is. In Tamil, a small gain was realized by re-training the morpheme acoustic models using the morpheme vocabulary. The final system output is generated by combining the posting lists generated by the word-based and morpheme-based retrieval systems.

Parameters were optimized on Zulu (fixing the perplexity threshold $b = 50$), and then applied to the other languages, c.f. Table 6.

4.6. "Radical" Systems

For both LimitedLP and FullLP system, system combination was performed with the Kaldi systems described in [15], using ROVER and CombMNZ between all available systems respectively. Figure 1 shows that a significant gain was achieved in every condition.

In the case of Tamil, the fused Kaldi keyword list was rescored using a Poisson Point Process Model [33], before this list was fused with a fused Janus list, which resulted in a boost of about .015 in ATWV alone. Also, the class-based language model described in [15] was also used in the Janus LimitedLP Tamil system, however the expanded lexicon technique did not improve performance of Janus-based systems with respect to either ATWV or WER.

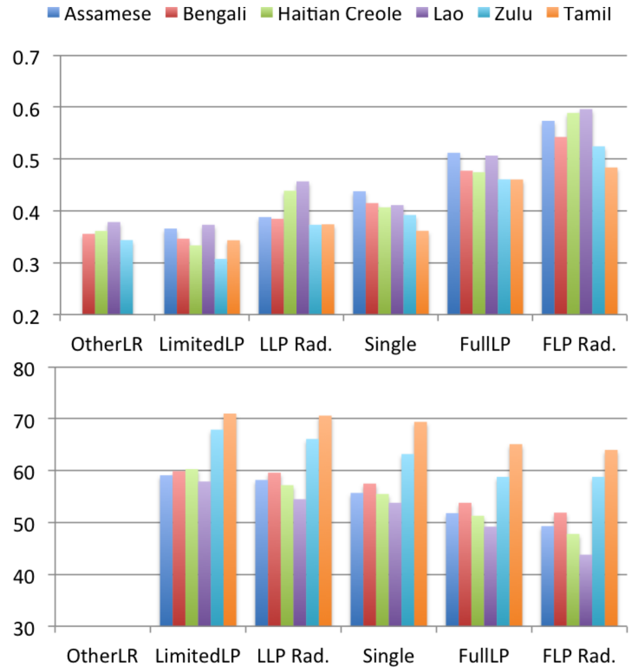

Fig. 1. ATWV (top) and WER (bottom) achieved on evaluation data. The "LimitedLP Rad." and "FullLP Rad." conditions represent the "Radical" team's combination of the systems presented in this paper with the Kaldi systems described in [15]. No WERs were computed for "OtherLR", as the web-text language models were often mismatched, and did not improve WER.

WER (%)	Ass.	Ben.	Haitian	Lao	Zulu	Tamil
FullLP	49.7	52.2	45.8	43.4	54.8	63.9
LimitedLP	58.0	60.3	55.7	54.4	65.1	70.2
ATWV						
FullLP	0.50	0.50	0.59	0.57	0.53	0.47
LimitedLP	0.35	0.34	0.44	0.44	0.37	0.32

Table 7. Summary of results obtained on the evaluation sets of the OpenKWS 2014/ Babel OP1 languages, using evaluation keywords, and the evaluation data. Development focused on the "LimitedLP" condition; Tamil results were achieved within a four-week window for training a system and processing the evaluation data.

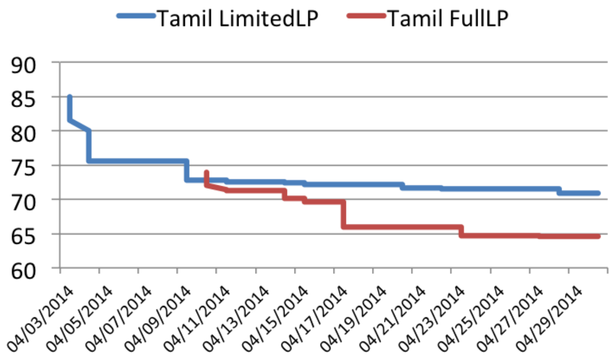

Fig. 2. Best Word Error Rate (%) of Janus Tamil systems over time during OpenKWS 2014, on development data. Effort was almost linear over time, although work on the "FullLP" systems started only once the initial "LimitedLP" systems had been built. "SS1" systems became available on Apr 9, and were only tuned slightly after. Morphological LMs were introduced on April 20, while "SS2" systems were introduced on April 28, but did not improve performance on development data quite as much as they helped on evaluation data (hence the smaller "drop" in WER shown here). System combination became standard for FullLP experiments on April 17.

5. DISCUSSION

Table 7 lists the results achieved by the Janus-based systems in the OpenKWS/ Babel OP1 evaluation using development data and evaluation keywords. Figure 1 shows the ATWV and WER achieved by the submitted systems on unseen evaluation data. It can be seen that all systems comfortably achieve ATWV>0.3 in the LimitedLP condition, the goal for Babel program performers. Generally, performance on development and evaluation data is very similar, and the languages behave as one would have predicted by looking at their vocabulary growth and OOV characteristics, i.e. Tamil and Zulu are hardest, followed by Assamese and Bengali.

It is interesting to note that the FullLP (60 h trained) "Single" systems data are generally just slightly better than the (10 h) "LimitedLP" systems w.r.t. WER, and comparable in terms of ATWV. The "LimitedLP" systems consist of a combination of 12 acoustic models in three different feature spaces. Transcribing an additional 50 h of data therefore facilitates KWS greatly, even if comparable performance can be achieved using only 10 h of transcribed data.

This paper analyzed the wealth of data generated by submitting competitive STT and KWS systems to the 2014 OpenKWS and Babel evaluations. Even with a large team, and a month's time, system performance seems to be already in an asymptotical regime, and only small gains can be achieved towards the end, as shown in Figure 2.

The systems presented here were developed with a focus on the "LimitedLP" condition, and the STT performance is on par with the top-performing teams in the OpenKWS evaluation. KWS performance however is relatively lacking for OOV words. Probabilistic phonetic retrieval, web-text based vocabulary and language model expansion, and the generation of a purely data-driven morpheme-based index present three approaches that ameliorated this problem to some extent only. Semi-supervised training of BNF features was applied successfully on training *and* test data in a low-resource setting, with little initial training data and high error rates.

6. ACKNOWLEDGMENTS

Dietrich Klakow provided the class-based language model that was used for Tamil, and Anjali Menon extracted the PLP and Kaldi-Pitch features that were used for some systems. Susanne Burger downloaded the text data for the web-text experiments.

7. REFERENCES

[1] F. Grézl and M. Karafiat, "Semi-supervised bootstrapping approach for neural network feature extractor training," in *Proc. ASRU*, Dec 2013, pp. 470–475.

[2] R. Hsiao, T. Ng, et al., "Discriminative semi-supervised training for keyword search in low resource languages," in *Proc. ASRU*, Dec 2013, pp. 440–445.

[3] N. Rajput and F. Metze, "Spoken web search," in *Proc. MediaEval Workshop*, Pisa; Italy, Sept. 2011.

[4] X. Anguera, L. J. Rodríguez-Fuentes, et al., "Query by example search on speech at Mediaeval 2014," in *Proc. MediaEval Workshop*, Barcelona; Spain, Oct. 2014, http://ceur-ws.org/Vol-1263.

[5] K. Kintzley, A. Jansen, and H. Hermansky, "Featherweight phonetic keyword search for conversational speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[6] L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks," *Computer, Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[7] D. R. H. Miller, M. Kleber, et al., "Rapid and accurate spoken term detection," in *Proc. INTERSPEECH*, Antwerpen; Belgium, Aug. 2007, ISCA.

[8] D. Xu and F. Metze, "Word-based probabilistic phonetic retrieval for low-resource spoken term detection," in *Proc. INTERSPEECH*, Singapore, Sept. 2014, ISCA.

[9] Y.-C. Li, W.-K. Lo, et al., "Query expansion using phonetic confusions for chinese spoken document retrieval," in *Proceedings of the Fifth International Workshop on on Information Retrieval with Asian Languages*, New York, NY, USA, 2000, IRAL '00, pp. 89–93, ACM.

[10] Intelligence Advanced Research Projects Activity, "IARPA-BAA-11-02," `http://www.iarpa.gov/index.php/research-programs/babel`, 2011, Last accessed July 7, 2014.

[11] The National Institute of Standards and Technology, "NIST Open Keyword Search 2014 Evaluation (OpenKWS14)," `http://www.nist.gov/itl/iad/mig/openkws14.cfm`, Apr. 2014, Last accessed: July 3, 2014.

[12] H. Soltau, F. Metze, et al., "A One-pass Decoder based on Polymorphic Linguistic Context Assignment," in *Proc. Automatic Speech Recognition and Understanding (ASRU)*, Madonna di Campiglio, Italy, Dec. 2001, IEEE.

[13] E. A. Fox and J. A. Shaw, "Combination of multiple searches," in *Proc. 2nd Text REtrieval Conference (TREC-2)*, Gaithersburg, MD; U.S.A., 1994, pp. 243–252.

[14] F. Metze, Z. A. W. Sheikh, et al., "Models of tone for tonal and non-tonal languages.," in *Proc. ASRU*, 2013, pp. 261–266.

[15] J. Trmal, G. Chen, et al., "A keyword search system using open source software," in *Proc. IEEE Workshop on Spoken Language Technology*, South Lake Tahoe, NV; USA, Dec. 2014, IEEE.

[16] Y. Wang and F. Metze, "An in-depth comparison of keyword specific thresholding and sum-to-one score normalization," in *Proc. INTERSPEECH*, Singapore, Sept. 2014, ISCA.

[17] J. Gehring, Y. Miao, et al., "Extracting deep bottleneck features using stacked auto-encoders," in *Proc. ICASSP*, Vancouver, BC; Canada, May 2013, IEEE.

[18] K. Schubert, "Grundfrequenzverfolgung und deren Anwendung in der Spracherkennung," M.S. thesis, Universität Karlsruhe (TH), Germany, 1999, In German.

[19] K. Laskowski, M. Heldner, and J. Edlund, "The Fundamental Frequency Variation Spectrum," in *Proc. 21st Swedish Phonetics Conference (Fonetik 2008)*, Gothenburg, Sweden, June 2008, pp. 29–32.

[20] P. Ghahremani, B. BabaAli, et al., "A pitch extraction algorithm tuned for automatic speech recognition," in *Proc. ICASSP*, May 2014, pp. 2494–2498.

[21] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Tech. Rep., Cambridge University, Cambridge; UK, May 1997, CUED/F-INFENG/TR 291.

[22] G. Saon, H. Soltau, et al., "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. ASRU*, Dec 2013, pp. 55–59.

[23] Y. Miao, H. Zhang, and F. Metze, "Towards speaker adaptive training of deep neural network acoustic models," in *Proc. INTERSPEECH*, Singapore, Sept. 2014, ISCA.

[24] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11)*, G. J. Gordon and D. B. Dunson, Eds. 2011, vol. 15, pp. 315–323, Journal of Machine Learning Research - Workshop and Conference Proceedings.

[25] I. J. Goodfellow, D. Warde-Farley, et al., "Maxout networks," *CoRR*, vol. abs/1302.4389, 2013.

[26] Y. Miao, F. Metze, and S. Rawat, "Deep maxout networks for low-resource speech recognition," in *Proc. ASRU*, Olomouc; Czech Republic, Dec. 2013, IEEE.

[27] A. Waibel, T. Hanazawa, et al., "Phoneme recognition using time-delay neural networks," *Acoustics, Speech and Signal Processing, IEEE Trans. on*, vol. 37, no. 3, pp. 328–339, Mar 1989.

[28] O. Abdel-Hamid, A.-r. Mohamed, et al., "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *Proc. ICASSP*, March 2012, pp. 4277–4280.

[29] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. Deep Learning for Audio, Speech and Language Processing Workshop at ICML*, Atlanta, GA; U.S.A., June 2013.

[30] I. Bulyko, M. Ostendorf, et al., "Web resources for language modeling in conversational speech recognition," *ACM Trans. Speech Lang. Process.*, vol. 5, no. 1, pp. 1:1–1:25, Dec. 2007.

[31] M. Creutz and K. Lagus, "Inducing the morphological lexicon of a natural language from unannotated text," in *Proc. AKRR*, Espoo, Finland, June 2005.

[32] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, May 2008.

[33] K. Kintzley, A. Jansen, and H. Hermansky, "Featherweight phonetic keyword search for conversational speech," in *Proc. ICASSP*, May 2014, pp. 7859–7863.