# Data Augmentation, Feature Combination, and Multilingual Neural Networks to Improve ASR and KWS Performance for Low-resource Languages

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

## Abstract

This paper presents the progress of acoustic models for low-resourced languages (Assamese, Bengali, Haitian Creole, Lao, Zulu) developed within the second evaluation campaign of the IARPA Babel project. This year, the main focus of the project is put on training high-performing automatic speech recognition (ASR) and keyword search (KWS) systems from language resources limited to about 10 hours of transcribed speech data. Optimizing the structure of Multilayer Perceptron (MLP) based feature extraction and switching from the sigmoid activation function to rectified linear units results in about 5% relative improvement over baseline MLP features. Further improvements are obtained when the MLPs are trained on multiple feature streams and by exploiting label preserving data augmentation techniques like vocal tract length perturbation. Systematic application of these methods allows to improve the unilingual systems by 4-6% absolute in WER and 0.064-0.105 absolute in MTWV. Transfer and adaptation of multilingually trained MLPs lead to additional gains, clearly exceeding the project goal of 0.3 MTWV even when only the limited language pack of the target language is used.

**Index Terms**: ASR, KWS, MTWV, MLP, rectified linear units, multilingual, low-resource

## 1. Introduction

Speech technologies are applied to a growing number of languages. Thus, there is a large interest for methods which ease the training and improve the quality of the models, especially in the first steps of the development phase where only limited amount of data is available. The Babel project funded by IARPA is addressing these goals by the development of robust speech technologies, focusing on spoken term detection, which can be applied to any language with a limited amount of transcription in a limited time [1]. With the progress of the project the main focus is moved, this year the participants have to accomplish the desirable keyword search (KWS) performance using limited (about 10 hours of speech) transcription on the following languages: Assamese, Bengali, Haitian Creole, Lao, and Zulu.

As has been already shown, Neural Networks (NN) play a key role in achieving the project goals [2, 3] either by of the tandem [4] or the hybrid acoustic modeling approach [5]. Applying multilingual training of e.g. [6] to deep Multilayer Perceptrons (MLP), [7, 8, 9, 10] demonstrated that borrowing orders of magnitude more data from other languages improves the ASR and KWS performance enormously if only limited amount of data is available in the target language.

In previous years, novel non-linearities which are biologically more plausible than sigmoid have been proposed for NN. For instance, maxout and Rectified Linear Units (ReLU) have been successfully applied to machine learning problems [11, 12]. In [13] the authors showed significant improvement using ReLU on a Large Vocabulary Continuous Speech Recognition (LVCSR) task. Besides the different activation units introduced recently, label preserving data augmentation techniques – widely used in image recognition tasks – also show consistent improvement for low-resource speech recognition [14, 15]. Furthermore, exploiting multiple representation of the speech signal, neural network based feature combination resulted in considerable improvement on Spanish broadcast news and conversation LVCSR task [16].

Therefore, in this paper we investigate the application of data perturbation, feature combination, and ReLU activation units to improve low-resourced ASR and KWS systems for a diverse set of languages. The experiments are carried out with the tandem approach. According to the primary goal of the Babel project we concentrate on unilingual systems, however, the best MLP architectures and techniques are also tested with multilingual approaches as well.

The paper is organized as follows, Section 2 gives a short corpus description and the overview of the keyword search task of the Babel Program. We give a summary of the investigated methods and the details on our experimental setups in Section 3. The ASR and KWS results are presented in Section 4. The paper closes with conclusions in Section 5.

## 2. Task description

One of the main goals of the IARPA-Babel Program is to reduce the performance gap of speech applications between high-resource well-studied languages (like English) and low-resource languages which have not yet been researched extensively. The participants compete in keyword search evaluations. The performance is measured in Actual Term Weighted Value (ATWV) based on the average value lost per term [17]. The loss is a weighted 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). The performer should achieve a minimum of 0.3 ATWV on the evaluation set with the evaluation keyword set.

On each language more than 100 hours of data are collected – full language pack (FLP) –, however, a considerable portion is non-speech, and in the current period about 75% of the corpus is transcribed. The limited language pack (LLP) comprises only about 10 hours of speech. The ASR and KWS tasks are challenging because the data sets contain speech originating from various sources covering several environments and dialects, and most of the data is narrow-band telephony speech. The provided pronunciation lexicons use a variant of X-SAMPA and cover the words appearing in transcribed training data [18]. In the base period (BP) of the project, Cantonese (CAN), Tagalog (TAG), Pashto (PAS), Turkish (TUR), and Vietnamese (VIE) as surprise language pack were released. In the second period (OP1) five other languages – Assamese (ASM), Bengali (BEN), Haitian Creole (HAI), Lao (LAO), Zulu (ZUL) – were given to the performers. Table 1 shows the phone set overlap between the languages. Cantonese, Vietnamese, Lao are tonal languages and Zulu phoneset has the distinctive attribute of having click consonants.

In our experiments we concentrate on LLPs of OP1. According to the primary submission condition, the acoustic and language model training is based on the transcription of the limited corpus. Not surprisingly, this result in a high out-of-vocabulary (OOV) rate and also in large number of OOV terms (at least one word of the query is OOV).

For the multilingual experiments the data is borrowed from the BP FLPs, a total of approx. 350 hours of speech. Using non-target language data from the project corresponds to the "BabelLR" condition, where all Babel resources of non-target languages can be used.

Table 2 summarizes the corpus statistics, and shows the amount of speech retained for acoustic model and MLP training after segmentation and silence removal steps. Our training corpus is based on the reference segmentation. The segmentation of the 20 hours of test set was prepared by our project partner IBM. As can be seen, Zulu is the most difficult one of the OP1 languages. It has the highest number of unique phones (Table 1) and due to the rapid vocabulary growth almost 2/3 of the queries are OOV terms (Table 2).

Since evaluation keywords and data were not available, the spoken term detection efficiency was measured in terms of MTWV on the test set using the development keyword list [17]. The ASR performance is reported in WER on the development sets of the Babel corpora.

Table 1: *Phone set overlap across the Babel languages without splitting di- or triphthongs and without differentiating between tones of the phonemes*

| Lang. | CAN | PAS | TAG | TUR | VIE | ASM | BEN | HAI | LAO | ZUL | Unique |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CAN | 37 | 19 | 21 | 19 | 13 | 15 | 15 | 15 | 18 | 12 | 8 |
| PAS | | 45 | 29 | 29 | 22 | 28 | 29 | 24 | 22 | 20 | 5 |
| TAG | | | 49 | 27 | 21 | 29 | 29 | 25 | 21 | 22 | 8 |
| TUR | | | | 42 | 21 | 27 | 27 | 23 | 24 | 20 | 6 |
| VIE | | | | | 68 | 23 | 22 | 22 | 23 | 20 | 35 |
| ASM | | | | | | 51 | 43 | 29 | 23 | 25 | 4 |
| BEN | | | | | | | 53 | 29 | 22 | 25 | 5 |
| HAI | | | | | | | | 32 | 20 | 21 | 1 |
| LAO | | | | | | | | | 42 | 20 | 7 |
| ZUL | | | | | | | | | | 47 | 19 |

# 3. Experimental setup

## 3.1. Vocal tract length normalization and perturbation

The differing vocal tract sizes of different speakers lead to frequency shift of the formants. In order to reduce the influence of these shifts on the Gaussian Mixture Model (GMM) based acoustic models Vocal Tract Length Normalization (VTLN) could be applied [19]. As was shown in [20], using a generic text-independent speech model the normalization step can be performed already before the first pass recognition. Speaker normalized features could also improve the performance of MLP [21, 22]. In our experimental setups the factor of piecewise linear warping is quantized between 0.88 and 1.12 with 0.02 step.

Table 2: *Statistics of IARPA-Babel training and testing corpora and development keyword lists*

| Lang. | Set | | | | Lexicon size | OOV KW [%] |
|---|---|---|---|---|---|---|
| | Training | | Test | | | |
| | amount of speech [h] | running words | | OOV [%] | | |
| ASM | 9.2 | 73k | 66k | 8.6 | 8.7k | 28 |
| BEN | 9.4 | 82k | 70k | 8.9 | 9.5k | 31 |
| HAI | 10.0 | 103k | 95k | 5.8 | 5.9k | 16 |
| LAO | 9.3 | 101k | 90k | 1.8 | 4.0k | 10 |
| ZUL | 10.0 | 68k | 59k | 16.0 | 16.2k | 61 |

Recently, it has been shown that data augmentation using label preserving transformations could further improve the neural network training [14]. We implemented the Vocal Tract Length Perturbation (VTLP) in a similar way as [15]. Instead of choosing the warping factor randomly, the perturbation was performed in a deterministic manner around the estimated warping factor. The data augmentation was then carried out by creating other replicas. Based on initial experiments the warping factor perturbation step was set to 0.04, and 4 additional slightly different copies of the original feature streams were generated.

Only the MLPs were trained on the perturbed data, the GMMs were always estimated on the original feature set. It should be noted that application of perturbation in the described way could become infeasible with larger amount of data due to the five times longer MLP training.

## 3.2. Feature extraction

### 3.2.1. Short-term features

In this paper, we investigate three different short-time speech representations for MLP based combination of multiple feature streams. The standard feature pipelines were slightly modified to extract the critical band energies (CRBE) for MRASTA filtering, see Section 3.2.2.

Similar to our previous study [10], we use the Gammatone features (GT) [23] when no feature combination is performed. The pipeline is based on an audiologically motivated Gammatone filterbank implemented as a cascade of infinite impulse response filters. In the post-processing step the features are smoothed by a Hanning window in time and in frequency followed by 10th root compression.

In addition, we also extracted PLP features [24], the pipeline applies cubic-root compression and all-pole model fitting before restoring the critical band energies. Our third feature extraction method is the standard MFCC features [25].

In short, the main differences between the features refer to the shape of the critical band filters: *gammatone*, *trapezoid*, or *triangular*. Further difference concerns the distinction how the decreasing frequency resolution of the human ear is modeled with higher frequencies: *Greenwood*, *Bark*, or *Mel-scales*.

Moreover, fundamental frequency (F0) and voicedness features [26, 27] were always extracted independent of the language property.

#### 3.2.2. Hierarchical BN-MLP features

The 101 frames of the CRBEs were smoothed by two-dimensional bandpass filters to cover the relevant modulation frequency range (MRASTA) [28]. Following the work of [29, 30], the modulation spectrum is processed by hierarchical bottleneck (BN) MLPs and concatenated with LDA transformed GT features. The input of the first MLP contains the fast modulation part of the MRASTA filtering. The second MLP is trained on the slow modulation components and the windowed BN output of the first MLP. In a third step, the joint training of the two-level NN was performed, similar to [31]. Due to windowing function in the middle the hierarchical MLP corresponds to a time-delay NN [32]. The modulation features were augmented by the single frame of the actual CRBE and 17 frames of F0 and voicedness features. In order to have a consistent feature extraction pipeline – as required by multilingual MLP training – F0 features were always used.

The MLPs are initialized by discriminative pretraining [21] and trained using the frame-wise cross-entropy criterion only. The MLPs estimated 1500 tied-triphone states posterior probabilities per language. We adjusted the learning rate parameter on 10% of the training corpus.

The BN layer consisted of 60 units and was placed before the last hidden layer. The other hidden layers contained 2000 neurons. Due to the limited amount of training data mostly shallow networks are trained for unilingual acoustic modeling (see Section 4.1). The deep BN-MLP experiments were carried out with MLPs with 6 non-BN hidden layers. Based on prior experiments, the BN layer was kept sigmoidal when we switched from sigmoid activations to ReLU of [12]. The ReLU MLPs were trained using $L_2$ regularization and momentum.

The multilingual training of the BN features was performed on fully randomized feature vector set of the joint corpora of the languages. Although the transcriptions are available in X-SAMPA format, this paper applies language dependent output layers [6] instead of using a single softmax layer and the joint phone set [33, 34]. According to [31] the former results usually in lower word error rates. All hidden layers were shared between the languages. The multilingual training was carried out on the BP languages with deep MLPs. Then the networks was transferred and adapted to OP1 language. Depending on the experimental setup, adaptation was done on the original or the augmented corpus.

To carry out the training of different MLPs we extended the Quicknet toolkit with the following features [35]: arbitrary deep MLP structure, availability of input features at each level, $L_2$ regularization, time-delay element, ReLU activation function, and multilingual training of [6].

### 3.3. Acoustic and language modeling

The acoustic model uses up to 2500 triphone context-dependent states clustered by decision trees and GMM with a globally pooled diagonal covariance matrix and less than 800k densities. In the following, the training procedure of the baseline acoustic models is outlined. First, monophone models are trained based on a linear alignment, without MLP features. In the second step, the triphone acoustic model training is performed on the same features. Then the MLP features and the speaker independent/adapted GMM models are trained iteratively including a realignment in each step until the WER converges. Unless stated otherwise, the alignment of the last iteration step was used in our experiments. Speaker adaptive training was applied using constrained maximum likelihood linear regression [36, 37]. The initial GMMs were trained according to the maximum likelihood criterion. Minimum Phone Error (MPE) training was performed on the final speaker adapted acoustic models [38]. The speech recognition was carried out with 4-gram language models (LM). According to the ”BaseLR” submission condition, no additional resources were used during the estimation. In order to smooth the language models, the discount parameters were optimized [39].

### 3.4. Keyword search systems

The speech recognition experiments are conducted with the publicly available RASR toolkit [40]. The weighted finite state transducer based keyword search system was provided by our project partner IBM. On most of the languages, the spoken term detection task is carried out on word lattices. The in-vocabulary queries are searched in the graph directly, whereas the OOV terms are looked for in the phonetic form of the lattice. After the grapheme-to-phoneme (G2P) conversion of the query, its phonetic form was further expanded with a transducer modeling phone confusions [41]. The phone confusion estimation is derived from the Viterbi alignment of the reference transcription and the first best hypothesis of unigram recognition of the training corpus [42].

Due to the high OOV query rate on Zulu, an alternative morpheme-graph based OOV search was also implemented. The morphologically segmented queries, training and development data were provided by the Columbia University morphology team. By the help of sublexical modeling 43% of the OOV queries are covered. These were searched as in-morph-vocabulary queries in the morph graph similar to the search of in-vocabulary queries in a word-graph. The OOV queries not covered by morphological segmentation are searched in the morpheme lattice after the G2P and P2P expansion steps.

In all cases, the lattices were obtained by decoding with a bigram word/morpheme LM. The lattices contained over 10000 arcs/sec on average, the P2P expansion of the phonetic form of the queries was limited to 5000 best paths. A sum-to-one score normalization was also applied per keyword.

## 4. Experimental results

### 4.1. Improving unilingual BN features

In the first set of experiments we investigated the optimal structure of BN-MLP on Assamese. Because of the great success of deep neural networks in acoustic modeling and feature extraction, hierarchical MRASTA BN-MLP structure optimized previously for LVCSR tasks was revisited. Besides the MRASTA features we also experimented with 17 neighboring CRBE frames directly. The BN features were extracted in hierarchical and classical, furthermore, deep and shallow structures. The effect of VTLN normalization of the input features was only partially tested. Finally, rectified linear units were applied to the best structure.

Results in Table 3 indicate that if only limited data is available for BN-MLP training both VTLN and phychoacoustically motivated MRASTA processing are crucial. Thus, in the following experiments the BN features were fixed to hierarchical MRASTA. It can also be seen that training a hierarchy of shallow BN features – which is a deep time delay NN – outperformed the classical deep structures (6-7th rows). Application of ReLU improved the best results by 2.1% absolute.

In the second set of experiments the effect of VTLN, ReLU, MRASTA processing of multiple streams, and VTLP were

Table 3: *Optimization of NN features for Assamese LLP. Word Error Rate (WER) was measured after SAT w/o MPE.*

| Features | VTLN | NN structure | | | WER |
|---|---|---|---|---|---|
| | | hier. | deep | activation | |
| MRASTA | no | yes | no | sigmoid | 67.4 |
| | yes | | | | 66.4 |
| | | | yes | | 68.1 |
| | | | no | ReLU | 64.3 |
| CRBE | | no | | sigmoid | 68.9 |
| | | | yes | | 68.8 |
| | | yes | no | | 67.5 |

tested on several languages (Table 4). We consider the recognition and KWS results obtained by a discriminatively trained MRASTA based GMM as the baseline. The results show that successive application of VTLN, ReLU, and feature combination significantly improves the acoustic models on all languages. In the next step, we investigated whether the better acoustic model leads to a better alignment. As can be seen, on Assamese, Bengali, and Haitian the better alignment resulted in measurable improvement in WER. Additional experiments revealed, that VTLN, ReLU and feature combinations are not fully additive, and the same recognition performance could be achieve without VTLN (6th row in Table 4). Finally, the data augmentation technique was applied. Replicating the training data consistently improved the results on all languages. Experiments (results not presented here) with deep hierarchical MLP structures and the artificially increased VTLP corpus did not show further improvement. After the MPE training of the best acoustic model we also measured KWS performance (Table 5). In summary, the optimized BN-MLP features show 20-37% relative MTWV and 5-8% relative WER improvement.

Table 4: *Improving unilingual acoustic models (AM) by rectified linear units (ReLU), feature combination, and data augmentation (VTLP). Recognition results are measured after speaker adaptive (SA) training optionally followed by discriminative training (DT) and are in Word Error Rate (WER [%])*

| | AM | ASM | BEN | HAI | LAO | ZUL |
|---|---|---|---|---|---|---|
| Baseline | DT | 66.7 | 70.5 | 63.2 | 60.9 | 73.8 |
| +VTLN | SA | 66.4 | 69.0 | 61.8 | 60.6 | 73.9 |
| +ReLU | | 64.3 | 66.7 | 59.9 | 58.8 | 72.7 |
| +fea.comb. | | 63.8 | 66.1 | 58.9 | 58.0 | 72.0 |
| +realign. | | 63.1 | 65.8 | 58.4 | 58.0 | 72.0 |
| +VTLP | | 62.3 | 65.0 | 58.2 | 57.3 | 71.4 |
| | DT | 61.9 | 64.4 | 58.1 | 56.5 | 70.1 |

Table 5: *Keyword search performance comparison of the baseline and the improved systems. Results are measured in MTWV (more is better) after MPE training of the acoustic models.*

| | ASM | BEN | HAI | LAO | ZUL |
|---|---|---|---|---|---|
| Baseline | 0.288 | 0.295 | 0.461 | 0.408 | 0.180<br>0.244* |
| +ReLU, +fea.comb.,<br>+realign., +VTLP | 0.358 | 0.400 | 0.552 | 0.496 | 0.234<br>0.316* |

*using morpheme-graph based keyword search for OOV queries

### 4.2. Initialization with multilingual MLP

As can be seen in Table 5, Assamese and Zulu are the two most difficult languages wrt. spoken term detection performance. In the next experiment we investigated the effect of transfer and adaptation of multilingually trained MLPs for these languages. The first multilingual MLP was trained on GT pipeline based MRASTA features of five BP languages, and was adapted with only 10 hours of speech of the LLP corpus without using VTLN, this experimental setup is similar to the baseline in Table 4. In the second multilingual experiment, the transferred MLP was trained on three streams of MRASTA features (GT, MFCC, PLP) and was adapted to the target languages using the perturbed LLP corpora. These features could be compared to the best features in Table 4, except for the application of the sigmoidal non-linearities in multilingual case. As can be seen in Table 6, the first type of multilingual MLP resulted in 16% and 12% relative KWS improvement for Assamese and Zulu over the best unilingual systems. Comparison to the baseline results in Table 5 shows even larger, over 40% relative MTWV increase which agrees with our previous investigations in [10] regardless of word- or morpheme-graph based keyword search pipeline.

Application of the novel techniques on multilingual MLP features improves the Assamese results even further resulting in 20% relative MTWV improvement over the best unilingual system. On Zulu we observed only slight gains related to data augmentation and feature combination methods if the OOV query search was performed on word lattice.

Table 6: *Boosting speech recognition (WER) and keyword search results with multilingually trained BN-MLP features for Zulu and Assamese. MTWV$_W$ and MTWV$_M$ indicate word- and morpheme-graph based OOV query search performance*

| | ASM | | ZUL | | |
|---|---|---|---|---|---|
| | WER [%] | MTWV$_W$ | WER [%] | MTWV$_W$ | MTWV$_M$ |
| best unilingual | 61.9 | 0.358 | 70.1 | 0.234 | 0.316 |
| multilingual | 59.1 | 0.416 | 68.2 | 0.263 | 0.334 |
| +fea.comb. +VTLP | 58.4 | 0.433 | 68.3 | 0.265 | 0.344 |

## 5. Conclusions

We showed experimentally that acoustic models for low-resource speech recognition task can be significantly improved by MLPs with rectified linear activation units, MLP based feature combination, and vocal tract length perturbation. With the help of these approaches the unilingual query term detection performance for all IARPA-Babel OP1 languages increased by 20-30% relative. And as has also been shown, adaptation of multilingual, multistream MLP with artificially augmented corpus to the target language resulted in 20% and 9% relative gain over the best unilingual system for Assamese and Zulu, respectively.

## 6. Acknowledgement

The study was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. This effort used the following IARPA Babel Program language collection releases: IARPA-babel{101-v0.4c, 102b-v0.5a, 103b-v0.4b, 104b-v0.4aY, 105b-v0.4, 106-v0.2f, 107b-v0.7, 201b-v0.2b, 203b-v3.1a, 206b-v0.1d} 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. H. Ney was partially supported by a senior chair award from DIGITEO.

# 7. References

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

[2] B. Kingsbury *et al.*, "A High-Performance Cantonese Keyword Search System," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 8277–8281.

[3] D. Karakos *et al.*, "Score normalization and system combination for improved keyword spotting," in *Proc. of ASRU*, 2013, pp. 210–215.

[4] H. Hermansky *et al.*, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, 2000, pp. 1635–1638.

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

[6] S. Scanzio *et al.*, "On the Use of a Multilingual Neural Network Front-End," in *Proc. of Interspeech*, 2008, pp. 2711–2714.

[7] K. Knill *et al.*, "Investigation of multilingual deep neural networks for spoken term detection," in *Proc. of ASRU*, 2013, pp. 138–143.

[8] J. Gehring *et al.*, "DNN acoustic modeling with modular multilingual feature extraction networks," in *Proc. of ASRU*, 2013, pp. 344–349.

[9] J.-T. Huang *et al.*, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 7304–7308.

[10] Z. Tüske *et al.*, "Multilingual MRASTA Features for Low-resource Keyword Search and Speech Recognition Systems," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2014.

[11] I. J. Goodfellow *et al.*, "Maxout networks," in *ICML*, 2013.

[12] X. Glorot *et al.*, "Deep Sparse Rectifier Neural Networks," in *Proc. of AISTATS*, 2011, pp. 315–323.

[13] G. E. Dahl *et al.*, "Improving Deep Neural Networks for LVCSR using Rectified Linear Units and Dropout," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 8609–8613.

[14] N. Jaitly and G. E. Hinton, "Vocal Tract Length Perturbation (VTLP) improves speech recognition," in *ICML*, 2013.

[15] X. Cui *et al.*, "Data Augmentation for Deep Neural Network Acoustic Modeling," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2014.

[16] C. Plahl *et al.*, "Feature combination and stacking of recurrent and non-recurrent neural networks for LVCSR," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 6714–6718.

[17] J. G. Fiscus *et al.*, "Results of the 2006 spoken term detection evaluation," in *Proc. of ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, 2007, pp. 51–57.

[18] "www.phon.ucl.ac.uk/home/sampa."

[19] L. Welling *et al.*, "Speaker adaptive modeling by vocal tract normalization," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 6, pp. 415–426, Sep. 2002.

[20] S. Wegmann *et al.*, "Speaker normalization on conversational telephone speech," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, 1996, pp. 339–341.

[21] F. Seide *et al.*, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011, pp. 24–29.

[22] Z. Tüske *et al.*, "A study on speaker normalized MLP features in LVCSR," in *Proc. of Interspeech*, 2011, pp. 1089–1092.

[23] R. Schlüter *et al.*, "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2007, pp. 649–652.

[24] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[25] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[26] X. Lei *et al.*, "Improved tone modeling for Mandarin broadcast news speech recognition," in *Proc. of Int. Conf. on Spoken Language Processing*, 2006, pp. 1237–1240.

[27] A. Zolnay *et al.*, "Robust Speech Recognition Using a Voiced-Unvoiced Feature," in *Proc. of Int. Conf. on Spoken Language Processing*, 2002, pp. 1065–1068.

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

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

[30] C. Plahl *et al.*, "Hierarchical Bottle Neck Features for LVCSR," in *Proc. of Interspeech*, 2010, pp. 1197–1200.

[31] K. Veselý *et al.*, "The language-independent bottleneck features," in *Proc. of IEEE Workshop on Spoken Language Technology*, 2012, pp. 336–341.

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

[33] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, Aug. 2001.

[34] N. T. Vu and T. Schultz, "Multilingual Multilayer Perceptron For Rapid Language Adaptation Between and Across Language Families," in *Proc. of Interspeech*, 2013, pp. 515–519.

[35] "http://www.icsi.berkeley.edu/Speech/qn.html."

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

[37] G. Stemmer *et al.*, "Adaptive training using simple target models," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, 2005, pp. 997–1000.

[38] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, 2002, pp. I–105–I–108.

[39] M. Sundermeyer *et al.*, "On the Estimation of Discount Parameters for Language Model Smoothing," in *Proc. of Interspeech*, 2011, pp. 1433–1436.

[40] D. Rybach *et al.*, "RASR - The RWTH Aachen University Open Source Speech Recognition Toolkit," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.

[41] L. Mangu *et al.*, "Exploiting diversity for spoken term detection," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2013, pp. 8282–8286.

[42] M. Saraclar *et al.*, "An empirical study of confusion modeling in keyword search for low resource languages," in *Proc. of ASRU*, 2013, pp. 464–469.