BUT NEURAL NETWORK FEATURES FOR SPONTANEOUS VIETNAMESE IN BABEL

Martin Karafiát, František Grézl, Mirko Hannemann and Jan ”Honza” Černocký

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno, Czech Republic
karafiat,grezl,ihannema,cernocky@fit.vutbr.cz

ABSTRACT

This paper presents our work on speech recognition of Vietnamese spontaneous telephone conversations. It focuses on feature extraction by Stacked Bottle-Neck neural networks: several improvements such as semi-supervised training on untranscribed data, increasing of precision of state targets, and CMLLR adaptations were investigated. We have also tested speaker adaptive training of this architecture and significant gain was found. The results are reported on BABEL Vietnamese data.

Index Terms: speech recognition, discriminative training, bottleneck neural networks, adaptation of neural networks, region-dependent transforms

1. INTRODUCTION

This paper presents our recent effort to build an automatic speech recognition (ASR) system for Vietnamese spontaneous telephone conversations. The work was mainly driven by our participation in the BABEL project (“Babelon” consortium coordinated by BBN). Unlike the common style of ASR development on generous languages with almost infinite time and enough resources, BABEL aims at building keyword-spotting systems for languages with limited resources in limited amount of time.

In general, the data is split into two main conditions:

1. Full Language Pack (FullLP or FLP) - all Language Pack data (about 100h of clean speech)
2. Limited Language Pack (LimitedLP or LLP) - A subset of FullLP data (about 10h of clean speech), remaining part can be used for unsupervised training.

All BABEL teams had one year to build systems for four languages (Cantonese, Pashto, Tagalog and Turkish) and this effort culminated by one-month evaluation. Immediately after, a “surprise” language was released and all Participant teams had one month to build a system. This paper presents our efforts on this “surprise language” – Vietnamese (Language pack IARPA-babel107b-v0.7). It contains the description of our system as it was submitted to the evaluations, as well as post-evaluation experiments on speaker adaptive training.

2. STACKED BOTTLE-NECK NN FEATURE EXTRACTION

The main contributions of our paper are the improvements of Neural Network (NN) feature extraction used in our HMM-based system. The architecture used was Stacked Bottle-Neck (SBN) NN which were found to overcome standard Bottle-Neck features [1]. The scheme is given in figure 1. It contains two NNs: the BN outputs from the first one are stacked, down-sampled, and taken as an input vector for the second NN. This second NN has again a BN layer, of which the outputs are taken as input features for GMM/HMM recognition system.

Classically, the NNs are trained on transcribed data, where the training targets are obtained by forced alignment of the transcript. For Babel Vietnamese, three main topics were investigated:

1. Concerning the **NN training targets**, we have experimented with (1) Using less reliable data - the transcriptions contain many segments with un-intelligible or foreign speech without reliable phonetic conversion. (2) realigning the targets by current best system to get more precise training targets.
2. On LimitedLP domain we experimented with **unsupervised training of NN** and also with **adaptation of multilingual NN** trained on different languages into target domain (sections 4.2 and 4.3.
3. Finally, **Adaptation of Stacked Bottle Neck architecture** by Constrained Maximum Likelihood Linear Regression (CMLLR) [2], was investigated (section 4.4).

2.1. Semisupervised training of NN

Here we focused on bootstrapping approach of Semisupervised learning (SSL) for NNs used for feature extraction. The labeling was done using the “seeding” recognition system trained on LimitedLP data only. The data was selected based on a confidence measure of the most likely path through the segment (C_{max} confidence criterion was used). The dependence of NN on the quality of 1-best transcripts was examined. The classification error should be less harmful in BN architecture than in the HMM system as we consider BN more a feature extraction than the ultimate classification process. This approach is described in details in [3]; a similar approach was used by our colleagues [4] for hybrid (Deep)NN-HMM system.

2.2. Multilingual NN

An integral way of obtaining multilingual (or language independent) NN-based features was presented in [5, 6]. Here, the NN last – output – softmax layer is divided into language-specific parts which makes the main body of the NN language-independent.

This leads to truly multilingually trained weights in NN (except for language-specific parts of the output layer) which should be the

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. 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. The work was also partly supported by, Technology Agency of the Czech Republic grant No. TA01011328 and by IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070. M. Karafiat was supported by Grant Agency of the Czech Republic post-doctoral project No. P202/12/P604.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

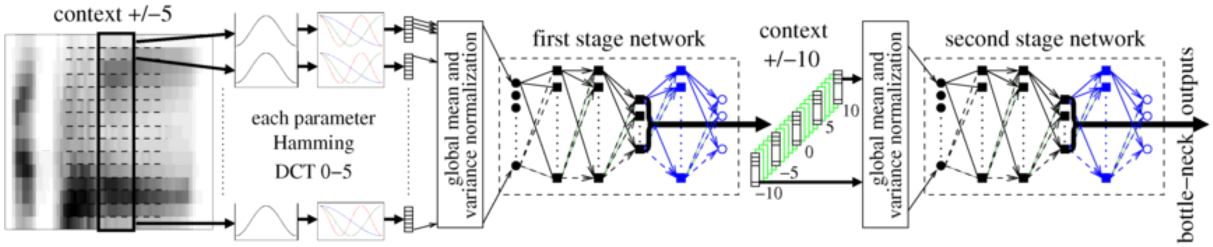

Fig. 1. *Stacked Bottle-Neck Neural Network feature extraction.*

best starting point for NN training on a new language especially if not enough training data is available.

The fine-tuning of NN trained on large amount of data from one language into target domain with little data by final fine-tuning was proposed in [7] and extended to multilingual NN in [8]. This approach eliminates the necessity of identification and approximation of new phonemes. Our fine-tuning procedure was adopted from [7] and extended to SBN structure. The approach should benefit from large amount of data available for other languages that allows for proper training of sufficiently big NN. This network serves as starting point for training a NN for the target language. As the amount of target data is small, the retraining is fast even for large NN which is the second benefit of the fine-tuning approach. [1]

2.3. CMLLR adaptation of SBN NN

The speaker adaptation of NN is a problem that has not been entirely solved. A general adaptation technique used with NNs is vocal-tract length normalization (VTLN) [9]. Hamid et al. [10] use a speaker-dependent code as part of the NN; this is an interesting approach but quite difficult to implement in unsupervised manner. A typical speaker adaptation uses standard cepstral coefficients (PLP or MFCC) together with CMLLR as NN input [11]; on the other hand, Mel-filter bank outputs (FBANK) are known to work better with NN [12], but their direct adaptation is difficult. Our trick to adapt the FBANKS is very simple: To estimate the adaptation matrix, it is necessary to train a GMM system on the NN input features, however, FBANKs are difficult to model by diagonal covariance models due to high correlations. This problem is solved by using Discrete Cosine Transform (DCT) in the same way as in MFCC computation. Next, the speaker independent GMM HMM system is estimated by single-pass retraining with FBANK-DCT appended by derivatives and acceleration coefficients. A block-diagonal CMLLR transform is estimated for each speaker s and only the first, spectrum corresponding, part of the transform is taken for further processing. New features for NN training are estimated simply by: $\hat{\mathbf{x}}(t) = \mathbf{x}(t)\mathbf{A}_{DCT}\mathbf{A}_{CMLLR}\mathbf{A}_{DCT^{-1}}$.

Next, we focused on adaptation of the SBN inner product – output of the first stage NN. The bottle-neck output is known not to be highly correlated, therefore the CMLLR can be applied easily. Moreover, according to our analysis, the first-stage NN is doing mainly acoustic feature extraction and only the second-stage NN is processing acoustic clues in wider context. Therefore, it makes sense to use speaker-specific layer in this part as it is common in classical speech recognition scenarios (feature extraction, speaker adaptation, acoustic modeling).

Table 1. Data analysis

FLP training hours	181
LLP training hours	21
LM training words	110980
dictionary size	3119
dev hours	19.7
OOV rate [%]	1.2

3. SYSTEM DESCRIPTION

3.1. Data, ASR system and baseline features

Table 1 summarizes the available data. The IARPA Babel Program simulates a case of what one could collect in limited time from a completely new language: the data consists of two parts: scripted (speakers read text through telephone channel) and conversational (spontaneous telephone conversations). The *dev* data contains conversational speech only.

Speech recognition system is HMM-based on cross-word tied-states triphones, it is trained from scratch using standard maximum likelihood training. Final word transcriptions are decoded using 3-gram Language Model (LM) trained only on the transcriptions of training data[2].

Mel-PLP features are generated in classical way, the resulting number of coefficients is 13. Deltas, double- and in HLDA system [13] also triple-deltas are added, so that the feature vector has 39, respectively 52, dimensions. Cepstral mean and variance normalization is applied with the means and variances estimated per conversation side. HLDA is estimated with Gaussian components as classes to reduce the dimensionality to 39.

3.2. SBN feature extraction

The NN input features are 15 critical band energies (squared FFT magnitudes binned by Mel-scaled filter-bank and logarithmized) concatenated with estimates of F0 and probability of voicing. It makes 17 dimensional feature stream. The estimation of F0 (implemented according to [14]) is based on normalized cross-correlation function. Dynamic programming is used for smoothing the estimates. This configuration was found useful and incorporation of F0 into NN systems gave good improvement even for non-tonal languages [15].

The conversation-side based mean subtraction is applied on the speaker basis and 11 frames are stacked together. Hamming window followed by DCT consisting of 0th to 5th base are applied on the

[1] The fine-tuning is often called "NN adaptation" although it is done by standard cross-entropy training with small learning rate. Therefore, the term *fine-tuning* is used in this paper as it better reflects the approach used and can not be confused with adaptation of the HMM systems.

[2] This is coherent to BABEL rules, where *the provided data only* can be used for system training in the primary condition

time trajectory of each parameter (17x6) resulting in 102 coefficients at the first stage NN input (see fig. 1).

The first-stage NN has four hidden layers with 1500 units each except the BN layer. The BN layer is the third hidden layer and its size is 80 neurons. Its outputs are stacked over 21 frames and down-sampled before entering the second-stage NN. This NN has the same structure and sizes of hidden layers as the first one. The size of BN layer is 30 neurons and its outputs are the final outputs forming the BN features for GMM-HMM recognition system.

Neurons in both BN layers have linear activation functions as they were reported to provide better performance [16]. The NN targets are phoneme states obtained by forced alignment of training data.

To train the system on Bottle-Neck features, the BN outputs are transformed by Maximum Likelihood Linear Transform (MLLT), which considers HMM states as classes. Then, new models are trained by single-pass retraining from HLDA-PLP baseline system. 12 Gaussian components per state were found to be sufficient for BN features trained from single-pass retraining. Most of the tests run on SBN system directly. The best found configuration was further trained with advanced techniques described in the following.

3.3. Final system

The final system is based on feature level fusion by Region Dependent Transform (RDT) [17]. The 3 feature streams PLP-HLDA (39 dimensions), SBN features (30 dim.) and F0 with delta and acceleration coefficients (3 dim.) are concatenated and adapted using speaker-based CMLLR.[3] This feature stream was fed to RDT performing dimensionality reduction to 69 dimension.

In RDT framework, an ensemble of linear transformations is trained with the discriminative Minimum Phone Error (MPE) criterion. Each transformation corresponds to one region in partitioned feature space by a GMM. Each feature vector is then transformed by a linear transformation corresponding to the region the vector belongs to. According to our previous experiments, GMM with 125 components was chosen. From our experience, incorporation of contextual information leads to significantly better results compared to the RDT style proposed in [17], where feature vectors of multiple frames were stacked at the RDT input. For detailed information of used configuration, see [15].

In our setup, two sets of RDTs were trained, both performed dimensionality reduction from 72 to 69 dimensions: RDT_{concat} on top of the original 72-dimensional features (for 1^{st} pass decoding) and RDT_{SAT} on top of CMLLR rotated features. The final GMM system was trained using MPE [18] on top of SAT RDT features.

4. EXPERIMENTS

4.1. Realigning and un-intelligible speech segments

The NN training targets are obtained by forced alignment of transcriptions by a simple PLP system. Since the system performance significantly improves by employing the new NN features, it makes sense to rebuild the alignments, and re-train the whole network. Also, originally, we omitted sentences containing the un-intelligible speech from the training. This was to make sure that we were training on clean speech only. Instead of excluding the data, we introduced a new word with "<unk>" phoneme transcription and mapped the problematic data on it. Table 2 presents the effect of NN target realignment and addition of more training data (containing "<unk>") on Vietnamese LimitedLP. It shows 1.4% absolute improvement on LLP and 1.6% on FLP by more precise timing of NN targets (+ Realignment). In additions, we show 0.8% improvement by adding more segments containing unintelligible speech and 0.1% degradation on FLP probably due to noise coming from "<unk>" phoneme.

Table 2. *Effect of realignment and sentences with "<unk>".*

System	LLP WER[%]	FLP WER [%]
PLP	81.3	72.5
InitSystem	72.9	55.8
+ Realignment	71.5	54.2
+ UNK	70.7	54.3
+ UNKDROP	70.5	53.5

The NN training is a discriminative process, therefore "<unk>" phoneme should be discarded as it introduces noise in the data. The last line (+UNKDROP) of table 2 presents 0.2% absolute improvement on LLP and 0.8% on FLP by removing this phoneme from the training data.

4.2. Semisupervised training of NN

To obtain the baselines, the NNs were first trained on both, LLP and FLP. Training on LLP will provide us with the lower bound - if the evaluated technique does not exceed this threshold, it is not worth the effort. The FLP result will serve as upper bound. The closer the results will be to this value, the more the technique benefits from other resources. All HMMs were trained on LLP clean data only.

The automatic transcription are naturally erroneous due to many reasons such as imperfect acoustic model, OOVs or poor language model. Thus it is important to select sentences with reasonable transcription. We use utterance-level confidence defined as a weighted average of non-silence words in the segment: $C_{utt} = \frac{1}{T}\sum_{w=1}^{W} t^w C_{max}^w$, where W is number of words, C_{max}^w is word confidence measure [19], t^w is length of the word in frames and T is length of all non-silence words.

The untranscribed data were decoded twice: the "final" MPE SAT RDT with RDT_{concat} transforms and models was used to produce 1-best output for CMLLR adaptation and RDT_{SAT} system was used to generate lattices for confidence measure.

Table 3 shows results with different thresholds on the confidence measure going from "reliable" data ($C_{max} > 0.6$) (67% of all untranscribed data) to "dirty" data ($C_{max} > 0.1$) (93% of all untranscribed data). Finally, we also made experiments with balancing of the training data. The data with true transcripts were cloned to reach the same amount as data with unsupervised transcripts. Afterwards, the data were shuffled. This procedure gave a nice 0.4% improvement. Note, number of hours counts NN training data where silence was limited therefore the data sizes do not correspond to Table 1.

4.3. Multilingual fine-tuning into LLP domain

The multilingual NN was trained on on FullLP of all previous BABEL languages (Cantonese, Pashto, Tagalog and Turkish). This NN guarantees a good starting point for fine-tunning to the target language.

Our approach to NN fine-tuning has two phases: First, only the last layer is trained. Since our initial NN is multilingual, the output layer has large number of units. We need to initialize the output

[3] Note that our experiments showed a marginal effect of VTLN on the PLP feature stream, therefore, VTLN was not applied for simplicity.

Table 3. *SBN system, semi-supervised training on <unk> sentences.*

data [C_{max}]	length [h]	SSL WER [%]	SSL+bal. WER [%]
FLP(upper bound)	63.0	-	-
LLP only	9	71.5	71.5
LLP + 0.6	9+45.8	67.8	67.8
LLP + 0.3	9+61.2	67.4	67.0
LLP + 0.1	9+62.1	67.4	67.5

Table 4. *Multilingual training and fine-tuning into Vietnamese domain.*

NN System	LLP WER [%]	SSL WER [%]
LLP	71.5	67.0
Multiling - no fine-tuning	72.8	-
Multiling - fine-tuning	67.7	66.3

Table 5. *Effect of using a CMLLR adaptation in NN training.*

System	WER[%]
No NN adapt	52.3
CMLLR on FBANK	51.1
CMLLR on 1stageNN+MLLT	48.9
CMLLR on 1stageNN	48.9

Table 6. *Final system results on FLP.*

PLPHLDA+NN+F0 SAT RDT MPE System	WER[%]
NN - (clear - nounk)	48.9
NN - realign unkdrop	47.2
NN - +CMLLRon1stageNN	45.8
DNN - sMBR	45.5

layer randomly with the proper number of outputs matching the target language phoneme set. If the whole NN was retrained now, the error caused by the random weights in the last layer could be propagated deeper in the NN and the training could drift apart from the optimum. This is why the rest of NN is fixed and only the last layer is trained.

In the second phase, the other layers are released and the whole NN is retrained once more. Since this retraining starts from an already trained network, the learning rate for this phase is set to one tenth of its original value[4].

Only the second stage NN is tuned to keep the training process simple and fast. Detail description of fine-tuning multilingual NN can be found in [20].

The "Multiling - no fine-tuning" in table 4 shows that initial multilingual NN is giving slightly worse accuracy than NN trained on LLP data only even if no Vietnamese data was used in training. The First column presents 3.8% absolute improvements by using multilingual NN with fine-tuning instead of plain LLP NN. It is obviously caused by increasing of NN training data. Consequently, we were interested in using multilingual NN together with semi-supervised data from the previous section. The best configuration from the previous section (SSL $C_{max} > 0.3$+data balancing) was used for this experiments. The improvement is not so huge as without SSL but 0.7% absolute improvement is still good.

4.4. Adaptation of NN by CMLLR

The experiments with using CMLLR adaptation of NN were running on FullLP. First, 15-dimensional FBANK features were decorrelated by DCT, and expanded by delta and acceleration coefficients. Block diagonal CMLLR transform was estimated in order to keep a part of CMLLR corresponding to plain FBANKs independent on feature derivatives. After the adaptation, the first 15 coefficient were projected back into original by space and SBN was trained on adapted features. Table 5 presents a 1.2% absolute improvement in comparison to non-adapted NN.

The CMLLR adaptation of first stage NN from SBN structure is more straightforward, as no de-correlation is necessary. Table 5 shows no-effect when the first stage BN output (80 dimension) was de-correlated by MLLT. 2.4% absolute improvement from CMLLR adaptation of the first stage NN is very nice and could actually be due to better speaker characterization from more informative BN features (bigger dimensionality, wider contextual information).

Note, NNs and HMMs were both trained on all data (including "<unk>" segments), therefore the baseline number do not correspond to Table 2 where HMMs were trained on clean data only.

4.5. The ultimate system

Finally, all partial improvements coming from data handling (realignment and "<unk>") were put into a final discriminatively trained system (MPE SAT RDT). Table 6 presents 1.6% absolute improvement from retraining the NN on better treated data. Next, 1.4% absolute improvement was coming from NN adaptation.

Just for comparison, we also trained Deep NN (DNN) system where HMM posteriors are estimated directly by NN. The NN was trained on top of SAT PLP using Sequence Minimum Bayes Risk criterion [21]. It was very successful single system in this evaluations. Similar performance and structure different from the described SBN system made both systems well complementary [22].

5. CONCLUSIONS

The paper deal with multiple facets of NN feature extraction training. Not surprisingly, we found that data preparation is crucial for the success of NN training. In case we dispose of data from other (well represented) languages, we should go for it as we have shown that multilingual fine-tuning outperforms unsupervised training.

The outcomes of our work were important for "Babelon" submission to the BABEL evaluation as our advanced feature extraction (1) allowed a simple Maximum-Likelihood system to have state-of-the-art performance, so that the team could concentrate on issues related to pronunciation dictionaries, keyword-spotting. etc, and also (2) the feature-level fusion with BBN system was very successful [22].

[4] Learning rate of 0.004 is used for training from random weights, and 0.0004 for the fine-tuning.

6. REFERENCES

[1] Frantisek Grezl, Martin Karafiat, and Lukas Burget, "Investigation into bottle-neck features for meeting speech recognition," in *Proc. Interspeech 2009*, 2009, number 9, pp. 2947–2950.

[2] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," 1997.

[3] Grezl F., Karafiat M., and Vesely K., "Adaptation of neural network feature extractor for new language," in *in Proceedings of ASRU 2013*, Olomouc, Czech Republic, 2013.

[4] Karel Veselý, Mirko Hannemann, and Lukáš Burget, "Semi-supervised training of deep neural networks," in *Proc. of ASRU 2013*, Dec 2013.

[5] Stefano Scanzio, Pietro Laface, Luciano Fissore, Roberto Gemello, and Franco Mana, "On the use of a multilingual neural network front-end," in *Proceedings of INTERSPEECH-2008*, 2008, pp. 2711–2714.

[6] Karel Veselý, Martin Karafiát, František Grézl, Miloš Janda, and Ekaterina Egorova, "The language-independent bottle-neck features," in *Proceedings of IEEE 2012 Workshop on Spoken Language Technology*. 2012, pp. 336–341, IEEE Signal Processing Society.

[7] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4269–4272.

[8] Samuel Thomas, Michael Seltzer, Kenneth Church, and Hynek Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, Vancouver, Canada, may 2013, IEEE.

[9] Frank Seide, Gang Li, Xie Chen, and Dong Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription.," in *ASRU*. 2011, pp. 24–29, IEEE.

[10] Ossama Abdel-Hamid and Hui Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *Proc. ICASSP 2013*, Vancouver, Canada, 2013.

[11] Y.-Q. Wang and M. J. F. Gales, "Tandem system adaptation using multiple linear feature transforms," in *Proc. ICASSP 2013*, Vancouver, Canada, 2013.

[12] Abdel-Rahman Mohamed, Geoffrey E. Hinton, and Gerald Penn, "Understanding how deep belief networks perform acoustic modelling.," in *ICASSP*. 2012, pp. 4273–4276, IEEE.

[13] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, pp. 283–297, 1998.

[14] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. Paliwal, Eds., New York, 1995, Elseviever.

[15] Martin Karafiát, František Grézl, Mirko Hannemann, Karel Veselý, and Jan "Honza" Černocký, "BUT BABEL System for Spontaneous Cantonese," in *Proceedings of Interspeech 2013*, 2013, pp. 2589–2593.

[16] Karel Veselý, Martin Karafiát, and František Grézl, "Convolutive bottleneck network features for LVCSR," in *Proceedings of ASRU 2011*, 2011, pp. 42–47.

[17] Bing Zhang, Spyros Matsoukas, and Richard Schwartz, "Recent progress on the discriminative region-dependent transform for speech feature extraction," in *Proc. of Interspeech 2006*, Pittsburgh, PA, USA, Sep 2006, pp. 2977–2980.

[18] D. Povey, *Discriminative training for large vocabulary speech recognition*, Ph.D. thesis, University of Cambridge, 2003.

[19] Frank Wessel, Ralf Schlüter, Klaus Macherey, and Hermann Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 288–298, 2001.

[20] Martin Karafiát Frantisek Grézl and Karel Veselý, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *Accepted for: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, Florence, Italy, May 2014, IEEE.

[21] Karel Veselý, Arnab Ghoshal, Lukáš Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks," in *Proceedings of Interspeech 2013*. 2013, pp. 2345–2349, International Speech Communication Association.

[22] Karakos D., Schwartz R., Tsakalidis S., Zhang L. Ranjan S., Ng T., Hsiao R., Saikumar G., Bulyko I., Nguyen L., Makhoul J., Grezl F., Hannemann M., Karafiat M., Szoke I. Vesely K., Lamel L., and Le V.-B., "Score normalization and system combination for improved keyword spotting," in *Proc. of ASRU 2013*, Dec 2013.