SEMI-SUPERVISED BOOTSTRAPPING APPROACH FOR NEURAL NETWORK FEATURE EXTRACTOR TRAINING

Frantisek Grézl and Martin Karafiát *

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno, Czech Republic

ABSTRACT

This paper presents bootstrapping approach for neural network training. The neural networks serve as bottle-neck feature extractor for subsequent GMM-HMM recognizer. The recognizer is also used for transcription and confidence assignment of untranscribed data. Based on the confidence, segments are selected and mixed with supervised data and new NNs are trained. With this approach, it is possible to recover 40-55% of the difference between partially and fully transcribed data (3 to 5% absolute improvement over NN trained on supervised data only). Using 70-85% of automatically transcribed segments with the highest confidence was found optimal to achieve this result.

Index Terms— Semi-supervised training, bootstrapping, bottle-neck features

1. INTRODUCTION

One of the fundamental components in todays state-of-the-art LVCSR systems are neural networks (NNs) either in the role of feature extractor for subsequent GMM-HMM system [1] or in the role of probability distribution model for HMM in so called hybrid system [2]. To train such system, large amount of labeled data is needed. The demand for transcribed data is even more accented in case of NN training as they are trained on pairs of input feature vectors $\mathbf{x}_i$ and output labels y_i. Thus the transcription has to be done on frame level. This is achieved by forced alignment using a simpler system.

Since obtaining audio data is relatively easy, the burden of building a recognition system for a new language lies in the transcription of collected data. This process gets more demanding for more "exotic" languages and languages with small populations. If the scenario of a completely "untouched" language is considered, these costs are unavoidable since the transcription is needed not only for acoustic model training but also for language model training as the written and spoken forms may (and usually do) differ.

To save the costs, the use of untranscribed data together with transcribed one leads to development of *semi-supervised learning* (SSL) techniques. To accommodate the untranscribed data into the training, two main approaches can be considered: In the *bootstrapping approach*, the untranscribed data are automatically transcribed using model trained on transcribed data only. Then, the reliable segments are selected based on some measure and added to the next stage of training [3, 4, 5]. The second approach relies on an objective function which reflects reasonable *assumption about labeled and untranscribed data*. One such assumption is that the data belonging to the same class are close to each other after projection to a low-dimensional manifold. This is represented by graph-based methods [6]. Other assumption is that the decision boundaries lie in the regions where the data has lower density used in Transductive-SVM [7] or incorporating conditional entropy criterion [8].

It has been shown that increasing the amount of training data improves also the performance of NN [9, 10]. From this perspective, it would be highly desirable to accommodate the untranscribed data in the NN training especially when the amount of labeled data is low. Unfortunately, it seems that the extension of SSL to neural network training is less studied. The work [11] introduces graph-based SSL training objective together with entropy regularizer.

Although some authors disrespect the bootstrapping due to insufficient theoretical background, it performs similarly to graph-based methods - the data points nearby a labeled data point are given the same label. The labeling in bootstrapping can be done on several levels, depending on the knowledge about the data presented in the labeling process. Low-level includes frame by frame labeling, where individual vectors are labeled with respect to the closest transcribed one(s) by means of vector quantization, GMM, NN or other classifier. This approach does not assume any other knowledge about the data but it may cause problems with outliers and on label boundaries. Medium-level labeling, which incorporates some knowledge about the data, can label several frames together,

*This work 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 supported by the 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.

Table 1. Data analysis

Language	PA	TU	VI
FLP training speakers	1189	980	991
FLP training hours	194.3	192.7	181.0
LLP training speakers	126	121	121
LLP training hours	21.0	22.1	21.0
LM training sentences	9536	12025	11192
LM training words	108025	67706	110980
dictionary size	7025	12124	3119
dev speakers	121	121	119
dev hours	19.9	20.0	19.7
number of words	101803	9145	111957
OOV rate [%]	4.2	12.2	1.2

as, for example, phonemes. The high-level incorporates all knowledge about the data available - it forms words and sentences. In this case, whole recognition system is employed in the labeling process. The weakness of this approach lies in limited amount of transcribed data as there might be out-of-vocabulary (OOV) words and/or incomplete language model which will cause the wrong classification.

This paper focuses on bootstrapping approach of SSL for NNs used for feature extraction. The labeling is done using the "seeding" recognition system trained on transcribed data. The data are selected based on a confidence measure of the most likely path through the segment. Similar approach is used by our colleagues [12] in hybrid (Deep)NN-HMM system where they focus on the means of data selection together with training method of DNN.

2. EXPERIMENTAL SETUP

2.1. Data

The IARPA Babel Program data[1] simulate a case of what one could collect in limited time from a completely new language: it consists of two parts: scripted (speakers read text through telephone channel) and conversational (spontaneous telephone conversations). The *dev* data contains conversational speech only. Two training scenarios are defined for each language – Full Language Pack (FLP), where all collected data are available for training; and Limited Language Pack (LLP) — only one tenth of FLP. For the LLP condition, the rest of the audio material can be used as untranscribed (blind) audio data. Vocabulary and language model training data are also defined with respect to the scenario.

To evaluate the bootstrapping SSL approach, three language collections were selected: Pashto language collection release babel104b-v0.4aY (PA), Turkish language collection release babel105-v0.6 (TU) and Vietnamese language collection release babel107b-v0.7. The overview of training material for these languages is given in Tab. 1. The selection of languages covers one with almost 100% complete dictionary - Vietnamese. This is a syllable-based language and the OOVs are mainly foreign words. Pashto is language with reasonable OOV rate which one can expect when having small amount of training data. Turkish is extensively agglutinative which increases number of words in the language. This leads to high OOV rate. Thus it is possible to compare the effect of imperfect labeling on the SSL procedure.

[1]Collected by Appen `http://www.appenbutlerhill.com`

Table 2. Baseline results for individual Language packs.

Lang. Pack		num. targ.	hours	WER [%] HLDA-PLP	WER [%] BN
PA	FLP	216	64.7	76.9	62.0
	LLP	126	7.1		71.4
TU	FLP	126	56.6	77.4	61.9
	LLP	126	7.3		69.3
VI	FLP	303	52.3	79.9	63.6
	LLP	273	6.5		72.1

Note, that the amounts of the raw audio are given, which in case of conversational speech, includes one recording for each side of the conversation. Thus the data contains huge portion of silence useless for training. The amounts of data used for training are given in Tab. 2.

2.2. NNs for feature extraction

The features obtained using Neural Networks are the Bottle-Neck (BN) features. A structure of two 6-layer NNs is employed according to [13]. It is depicted in Fig. 1.

The NN input features are based on critical band energies (squared FFT magnitudes binned by Mel-scaled filter-bank and logarithmized) concatenated with estimates of F0 and probability of voicing. The estimation of F0 is based on normalized cross-correlation function. The maximum of this function indicates F0 value. Dynamic programming is used for smoothing the estimates. It is implemented according to [14]. Although it might seem not necessary to use the F0 and probability of voicing parameters for non-tonal languages, it turns out that these features are useful and their incorporation brings nice improvement of the final systems [15].

The conversation-side based mean subtraction is applied on the whole feature vector. 11 frames are stacked together. Hamming window followed by DCT consisting of 0th to 5th base are applied on the time trajectory of each parameter resulting in 102 coefficients on the first stage NN input.

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 downsampled 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

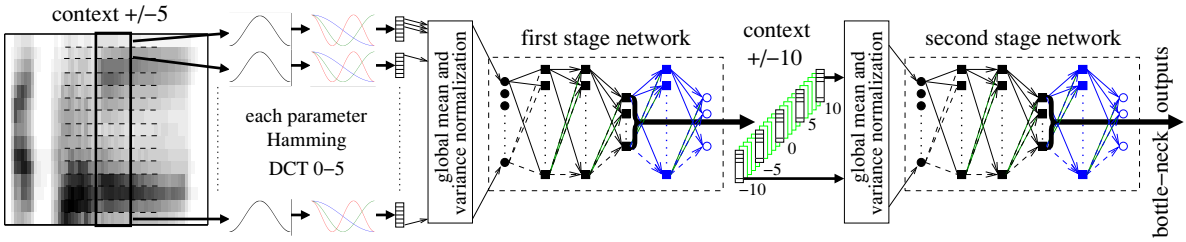

Fig. 1. Block diagram of Bottle-Neck feature extraction. The blue parts of NNs are used only during the training. The green frames in context gathering between the NNs are skipped. Only frames with shift -10, -5, 0, 5, 10 form the input to the second stage NN.

outputs are 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]. Before the features enter the NNs' input layer, global mean and variance normalization is performed.

The NN targets are phoneme states obtained by forced alignment of training data. The numbers of targets for individual language packs are given in Tab. 2. The forced alignments were generated with provided segmentations, however it was found that they still contain large portion of silence (50%–60%). Therefore, new segmentation, which reduced the amount of silence to 15%-20%, was generated. The final amounts of data used for NN training are also given in Tab. 2.

2.3. Recognition system

First, a system based on standard Mel-PLP features is created. 13 PLP coefficients are generated together with first, second and third order derivatives. HLDA is estimated with Gaussian components as classes to reduce the dimensionality to 39. Then the conversation-side based mean and variance normalization is applied. Based on these features, baseline HLDA-PLP speech recognition system system is trained using LLP data only. It is HMM-based cross-word tied-states triphone system, with approximately 4500 tied states and 18 Gaussian mixture components per state for all languages. It is trained from scratch using mix-up maximum likelihood training. The HLDA-PLP system is used for alignment of training data for NN training.

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. Next, 12 maximum likelihood iterations follow to better settle new HMMs in the new feature space.

Final word transcriptions are decoded using 3gram Language Model (LM) trained only on the transcriptions of LLP training data[2].

The results obtained with baseline HLDA-PLP systems are given in Tab. 2. The rather poor performance is given by the limited amount of data for acoustic as well as language model. Also note, that Turkish has quite high OOV rate.

2.4. Supervised NN systems

First, two systems with NNs trained on transcribed data were evaluated. The first one used only the LLP data for NN training. This one is the baseline for the semi-supervised systems trained later. The second one uses FLP for NN training, but the rest of the system is trained with LLP data only. This system serves as upper bound for the SSL experiments - the goal is to get as close as possible to its performance. Both systems use only BN features (no PLP ones). The WER obtained by these systems are given in Tab. 2.

It is important to state that forced alignments for training of NNs were obtained with a HLDA-PLP system trained on respective language pack. Since some phonemes are underrepresented in LLP, they are merged with acoustically closest ones to form more compact phoneme set. That is why there are different numbers of NN targets (phoneme states) for LLP and FLP.

2.5. Seeding system

To obtain the best labels for the SSL, more sophisticated seeding system was constructed. PLP-HLDA (39dim.), BN features (after the MLLT – 30dim.) and F0 (see sec. 2.2) with delta and acceleration coefficient (3dim.) are concatenated to form 72 dimensional feature vector. Next, two iterations of speaker-based Constrained Maximum Likelihood Linear Regression (CMLLR) transforms estimation and retraining the HMM in speaker adaptive training (SAT) scheme were done [17].

Finally, two sets of Region Dependent Transform (RDT) [18] were estimated, both performed dimensionality reduction from 72 to 69 dimensions: RDT_{concat} on top of original

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

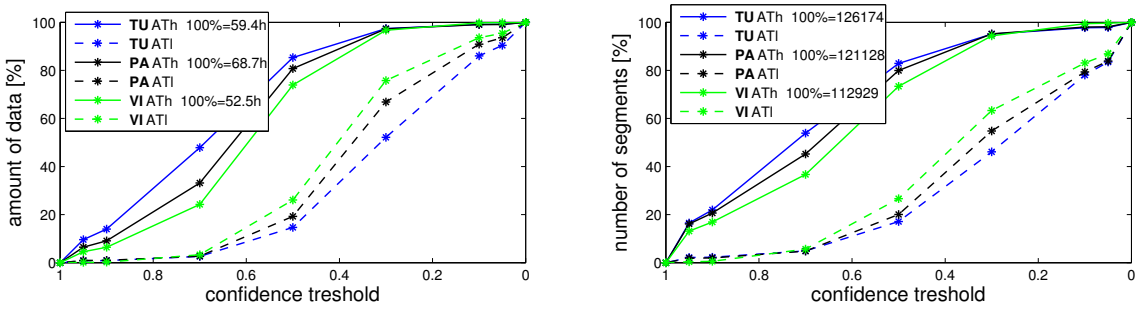

Fig. 2. The amount of data and number of segments in individual sets for all languages.

72dimensional features and RDT_{SAT} on top of CMLLR rotated features. RDT parameters and ASR acoustic model parameters are alternately updated in several iterations. RDT parameters are updated using discriminative MPE criterion, ML update is used for acoustic model parameters. For detailed information, see [15].

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

Utterance level confidence is weighted average of non-silence words in the segment:

$$C_{utt} = \frac{1}{T} \sum_{w=1}^{W} t^w C_{max}^w, \qquad (1)$$

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.

3. EXPERIMENTS AND ANALYSIS

The goal of the experiments is to minimize the difference in performance of system trained in semi-supervised manner and one trained on FLP data. 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 with high confidence.

However, the question of optimal threshold setting is more complicated. Lowering the threshold will increase the amount of data used for the training so that the degrading effect of erroneous part of the data might diminish or disappear. Improperly set threshold can cause degradation in further processing.

Thus it is also interesting to see how does adding high-confidence data compare with adding low-confidence data. To investigate this behavior with only one threshold (taking into account that by decreasing the threshold, we always want to include more data), we introduce the following notation:

- the complete set of automatically transcribed segments is denoted AT.
- the confidence threshold is denoted ct, in our experiments, we use the following values: $ct = \{1.1; 0.95; 0.9; 0.7; 0.5; 0.3; 0.1; 0.05; -0.1\}$.
- for a given value of ct
  - the set of *high-confidence* data ATh_{ct} is given by selecting all segments with confidence value higher than ct.
  - the set of *low-confidence* data ATl_{ct} is given by selecting all segments with confidence value lower than $1 - ct$.

Obviously,

$$ATh_{1.1}{=}\emptyset \qquad ATl_{1.1}{=}\emptyset$$
$$ATh_{-0.1}{=}AT \qquad ATl_{-0.1}{=}AT$$

Note, that in figures, the boundary confidence thresholds 1.1 and -0.1 will be replaced by points at 1 and 0 respectively. The amounts of data and the numbers of segments in sets are depicted in Fig 2. Note also, that the amount of AT data is not equal FLP - LLP as the segmentation is based on voice activity detection and not on provided transcriptions.

It can be seen that the amount of data with confidence below 0.3 ($ATl_{0.7}$) is relatively small for all languages (less than 1.9 hours). On the other hand, the amount of data with the highest confidence is relatively large. There are 3.3 to 8.2 hours of data with confidence higher then 0.9 ($ATh_{0.9}$) and 2.4 to 5.7 hours of data with confidence above 0.95 ($ATh_{0.95}$) depending on language. That means that segments with high confidence transcription are able to double the amount of training data.

For the training of neural network, the selected set of automatically transcribed data is mixed with the training part of LLP data. The cross-validation part (about 10% of LLP data) is kept unchanged which will allow to observe the behavior also on the CV frame level accuracy.

For each set, new NNs are trained from random initialization without any seeding by NN trained on LLP data only. The differences in achieved CV frame accuracy are given in Fig. 3. The behavior is somewhat unexpected. Adding data

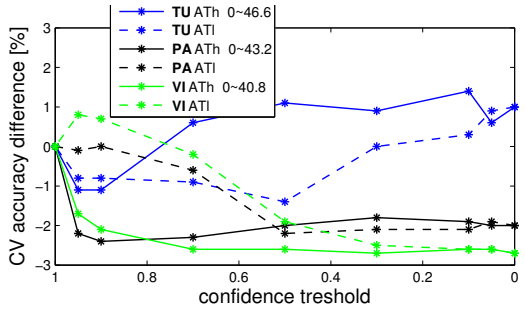

Fig. 3. Absolute difference in cross-validation frame accuracy. 0 corresponds to the accuracy obtained with LLP data.

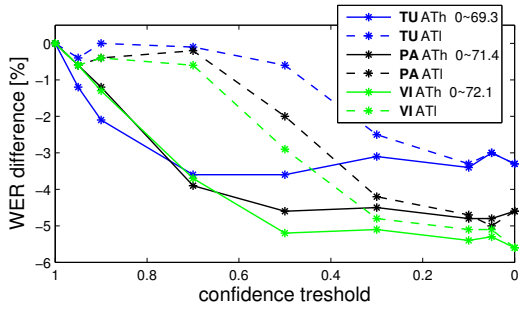

Fig. 4. Absolute difference in WER. 0 corresponds to the WER with LLP data.

with high confidence transcripts leads to a drop in frame accuracy. But when segments with low confidence transcripts are added to training data, the frame accuracy improves. Also, only the NNs for Turkish language finally achieve better CV accuracy.

The features provided by each new NN were also evaluated in terms of WER. The results are shown in Fig. 4. Although a degradation on frame accuracy was observed, the WER shows improvement when segments with high confidence transcription are added. The reason for the drop in frame accuracy might be that these sentences are not the easy ones and diverge the NN parameters far from LLP subset from which the CV part is selected. The WER does not change much when low confidence segments are added to NN training. Note, that the amount of such data is low (see above).

It can be also seen that adding data with confidence lower than 0.5 does not improve the system performance. It stays about the same, or, in case of Turkish, degrades slightly. Although the degradation is not dramatic compared to overall improvement from LLP NNs, the automatically transcribed data are used only for feature extractor training here and might hurt more in further processing. Thus the optimal threshold can be set to 0.5.

Table 3. Summary of results

Language	PA	TU	VI
FLP WER	62.0%	61.9%	63.6%
LLP WER	72.1%	69.3%	71.4%
LLP + $ATh_{0.5}$	66.8%	65.7%	66.9%
absolute WER red.	4.8%	3.8%	3.3%
LLP to FLP dif. red.	47.5%	51.3%	42.3%
AT hours	68.7	59.4	52.5
$ATh_{0.5}$ hours	55.4	50.6	38.8
$ATh_{0.5}/AT$	80.6%	85.2%	73.9%

4. CONCLUSIONS

The suitability of bootstrapping SSL approach for neural networks training was evaluated. The untranscribed data were handled at segment level and automatically transcribed by sophisticated seeding system which includes Bottle-Neck NN based features, HLDA-PLP features, speaker dependent transform and region dependent transform. The system also provided a confidence measure used for segment selection.

Selected segments were mixed with supervised data and new NNs were trained. The BN features obtained with new NNs were evaluated in a simplified recognition system. It was observed that adding data with confidence less than 0.5 does not significantly improve the system performance and on contrary, can be harmful.

The overview of achieved results is given in Tab 3. It can be seen that the absolute WER improvement is 3.3 to 4.8% over the situation when only supervised data is used for NN training. This improvement corresponds to 40-50% recovery of the gap between LLP and FLP training while using 70-85% of the available untranscribed data. Note that it is not possible to recover part of the performance difference due the incomplete dictionary and poor LM which bring errors into automatic transcriptions.

When the full recognition system (see the description of the seeding system in sec.2.5) is used for recognition of Vietnamese dev set, 66.0% WER is achieved with NN trained on LLP data only. The WER decreases to 60.5 when NN is trained also on automatically transcribed data.

Our future work will focus on the improvement of NN training procedure to get more advantage from supervised part of the data. Other direction we would like to investigate is the combination of multilingual and semi-supervised training.

5. REFERENCES

[1] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP 2000*, Turkey, 2000.

[2] N. Morgan and H. Bourlard, "Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 25–42, 1995.

[3] Lori Lamel, Jean-Luc Gauvain, and Gilles Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–129, 2002.

[4] Frank Wessel and Hermann Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 1, pp. 23–31, 2005.

[5] Lan Wang, M. J. F. Gales, and P. C. Woodland, "Unsupervised training for mandarin broadcast news and conversation transcription," in *Proc. ICASSP*. IEEE, Apr 2007, vol. 4.

[6] Amarnag Subramanya and Jeff Bilmes, "The semi-supervised switchboard transcription project," in *Proc. INTERSPEECH 2009*, Sep 2009.

[7] Thorsten Joachims, "Transductive inference for text classification using support vector machines," in *machine learning-international workshop then conference*. Morgan Kaufmann Publishers, INC., 1999, pp. 200–209.

[8] Jui-Ting Huang and Mark Hasegawa-Johnson, "Semi-supervised training of Gaussian mixture models by conditional entropy minimization," *Optimization*, vol. 4, pp. 5, 2010.

[9] D. Ellis and N. Morgan, "Size matters: An empirical study of neural network training for large vocabulary continuous speech recognition," in *Proc. ICASSP 1999*, Phoenix, Arizona, USA, Mar. 1999, pp. 1013–1016.

[10] Petr Schwarz, *Phoneme recognition based on long temporal context*, Ph.D. thesis, Brno University of Technology, Czech Republic, 2009.

[11] Jonathan Malkinn, Amarnag Subramanya, and jeff Bilmes, "On the semi-supervised learning of multi-layered perceptrons," in *Proc. INTERSPEECH 2009*, Sep 2009.

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

[13] F. Grézl, M. Karafiát, and L. Burget, "Investigation into bottle-neck features for meeting speech recognition," in *Proc. Interspeech 2009*, Sep 2009, pp. 294–2950.

[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, number 8, pp. 2589–2593, International Speech Communication Association.

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

[17] Langzhou Chen, Mark J. F. Gales, and K. K. Chin, "Constrained discriminative mapping transforms for unsupervised speaker adaptation," in *Proc. ICASSP*, Prague, Czech Republic, 2011.

[18] Bing Zhang, Spyros Matsoukas, and Richard Schwartz, "Recent progress on the discriminative region-dependent transform for speech feature extraction," in *Proc. of Interspeech2006*, Pittsburgh, PA, USA, Sep 2006, pp. 2977–2980.

[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.