INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Semi-supervised and Cross-lingual Knowledge Transfer Learnings for DNN Hybrid Acoustic Models under Low-resource Conditions

Haihua Xu[1*], *Hang Su*[2,5], *Chongjia Ni*[4], *Xiong Xiao*[1], *Hao Huang*[3], *Eng-Siong Chng*[1], *Haizhou Li*[1,4]

[1]Temasek Laboratories, Nanyang Technological University, Singapore
[2]Electrical Engineering & Computer Science Department, UC Berkeley, USA
[3]School of Information Science and Engineering, Xinjiang University, Urumqi, China
[4]Institute for Infocomm Research (I^{2}R), A*Star, Singapore
[5]International Computer Science Institute, Berkeley, CA, USA

Abstract

Semi-supervised and cross-lingual knowledge transfer learnings are two strategies for boosting performance of low-resource speech recognition systems. In this paper, we propose a unified knowledge transfer learning method to deal with these two learning tasks. Such a knowledge transfer learning is realized by fine-tuning of Deep Neural Network (DNN). We demonstrate its effectiveness in both monolingual based semi-supervised learning task and cross-lingual knowledge transfer learning task. We then combine these two learning strategies to obtain further performance improvement.

Index Terms: Multilingual, cross-lingual, semi-supervised training, deep neural network

1. Introduction

To build an automatic speech recognition (ASR) system for a low-resource language, two additional resources apart from transcribed training data are often used, namely, which are unlabeled audio data of the same language, and multilingual resource from other languages. The strategy using unlabeled data is called semi-supervised learning (SSL), while the other one is called cross-lingual knowledge transfer learning.

Semi-supervised learning is an effective way to improve performance of ASR systems developed with limited transcribed data [1–4]. In this case, a preliminary acoustic model (seed model) and plenty of unlabeled acoustic data are available for a specific language. The task is to exploit unlabeled data to boost performance of the model. In the framework of Deep Neural Network (DNN), we can use either Bottle-neck feature (BNF) pipeline [1, 3, 5, 6] or DNN-HMM hybrid acoustic models [2, 4] to perform semi-supervised learning. In practice, we first use a seed model to transcribe unlabeled data. Then we select the transcribed data based on confidence score of ASR outputs [1, 2]. Finally these selected data are merged with human transcribed data to update BNF extractor or DNN-HMM acoustic model.

On the other hand, when human transcribed data in other languages (i.e. multilingual data) are available, it is natural to find ways to take advantage of these data. In this scenario, multitask (i.e. multilingual) learning [7] and cross-lingual knowledge transfer come into play. DNN based multilingual training and cross-lingual knowledge transfer learning have been widely studied under low-resource conditions [8–15], where both BNF pipeline and DNN-HMM hybrid approach are used. For bottleneck feature (BNF) approach [12–15], a BNF extractor is trained to extract BNFs for target low-resource language. Shared hidden layers (SHL) are trained with multilingual data while softmax layers of the BNF extractor are language dependent [8, 12, 14]. Once multilingual BNF extractor is trained, it can be tuned using target language to realize cross-lingual transfer learning [12, 13]. For DNN-HMM hybrid approach, cross-lingual transfer learning is done using SHL directly [8–11, 16]. The benefit of this recipe lies in its simplicity and shorter feature extraction pipeline.

While semi-supervised and cross-lingual knowledge transfer learnings[1] are widely studied for low-resource acoustic modeling, one rarely examines their relationship. In this paper, we view semi-supervised learning and cross-lingual knowledge transfer learning as the same problem. Both of them are done using simple DNN "fine-tuning" method in practice. Furthermore, we proposed a word segment based data selection method to improve semi-supervised learning. We also attempt to gain additional improvement by combining cross-lingual knowledge transfer and semi-supervised learnings.

2. Prior work and contributions

It was reported in [5, 6, 17] that semi-supervised and cross-lingual knowledge transfer learnings could be used simultaneously to address low-resource language acoustic modeling issue. However, only BNF pipeline is well-investigated in these works. In this paper, we study semi-supervised and cross-lingual knowledge transfer learnings using DNN hybrid framework instead. While we hope to improve performance of these two tasks in two separate attempts, we address them with the same knowledge transfer learning approach.

We note that data selection method plays a role in semi-supervised learning. This is because low resources ASR systems tend to give higher word error rate (WER). In [1, 17], utterance based data selection method was used for a BNF pipeline based semi-supervised learning. The confidence of each utterance is calculated as average of posteriors over all the words in the utterance. In [4], we tried a similar data selection method for DNN hybrid system, unfortunately, no improvement was obtained. Alternatively, a frame based data selection method was proposed in [2], and it was suggested that data selection at frame level is more effective than at utterance level. In this

*This work is supported by the DSO funded project MAISON DSOCL14045, Singapore

[1]In this paper, cross-lingual knowledge transfer learning is implicitly based on multilingual training.

http://dx.doi.org/10.21437/Interspeech.2016-1099

paper, we propose a word segment based data selection method using word confidences and time boundaries estimated with the method proposed in [18]. This strategy benefits from both utterance and frame based techniques.

As for cross-lingual knowledge transfer learning, [8] proposed a two-step framework including softmax training and entire network tuning. It was shown that the overall tuning with very limited data (3 hours) does not help. In this paper, we propose a one-step DNN fine-tuning approach to accomplish knowledge transfer learning, i.e. we conduct softmax training and shared-hidden layer tuning jointly. In this way, we don't need to change learning rate in a separate tuning step. We show it achieves better results even with very limited target language data (3 hours), in terms of both cross-entropy (CE) and state-level sequence based Minimum Bayesian Risk (sMBR) [19] trainings.

3. Resource description

The data for experiments in this paper are from NIST OpenKWS15 evaluation program[2]. Different from previous evaluation program [20, 21], OpenKWS15 allows participants to adopt multilingual training for acoustic modelling. Swahili is the target low-resource language in this program. To this end, Cantonese, Pashto, Turkish, Taglog, Vietnamese and Tamil data from previous OpenKWS challenges are released to participants. In addition to Limited Language Pack (LLP) and Full Language Pack (FLP), NIST also defines Very Limited Language Pack (VLLP) as low-resource task. In this work, low-resource refers to LLP data and VLLP data to align with previous works [2, 4, 16]. All results are reported on 10-hour Swahili *dev* data. Table 1 describes data statistics.

Apart from acoustic data, NIST also provides out-of-domain text data to build language model (LM) [22]. These data are collected from diversified sources include websites like Wikipedia, Wiktionary and open-subtitles of movies and TV shows. Overall, there are 84M words for background trigram LM training. The background LM is interpolated with various in-domain LMs built with different (VLLP, LLP, FLP) transcriptions during testing. We use trigram LMs for all experiments in this paper.

During evaluation, no manual lexicon is provided, and we use grapheme lexicons in all cases. Lexicons used for decoding contains both out-of-domain data mentioned above, and corresponding language pack transcriptions. The vocabulary is about 200k depending on specific language pack categories. We note that after evaluation NIST also released a manual lexicon in the FLP. We compared ASR performance using these two lexicons, and the WER difference is less than 1% absolute. This indicates that Swahili is very regular in terms of pronunciation.

Table 1: Overall experimental data distributions

Language (Babel Id)	Data set	Data length (hours)
Source language (Multilingual data)		
Cantonese (101)	FLP	141.3
Pashto (104)	FLP	78.4
Turkish (105)	FLP	77.2
Tagalog (106)	FLP	84.5
Vietnamese (107)	FLP	87.7
Tamil (204)	FLP	69.4
Target language (Swahili)		
Swahili (202)	FLP	55.4
	LLP	10.8
	VLLP	3.1
	dev	10.7

4. Semi-supervised learning

For effective semi-supervised learning, three factors are critical, namely, a good seed model, an appropriate data selection method, and an effective method for practical semi-supervised training.

4.1. Seed model

In this work, we use DNN hybrid acoustic models trained with sMBR criterion as the seed model. Two types of seed models are used in the experiments. One is for monolingual semi-supervised training, which might be trained using VLLP or LLP; the other is for multilingual experiments, which contains models trained using human transcribed data set. Once the seed models are ready, they are used to decode unlabeled data, and generate ASR transcripts.

4.2. Data selection method

Since ASR transcripts always contain errors, it is necessary to select data that have higher confidence to a given acoustic models. We choose to use word-segment based data selection method. Technically, this is similar to the frame based data selection method proposed in [2] since we use the same method advocated in [18] to estimate word confidence and time boundaries in Kaldi[3]. However, our method does not affect the actual DNN training, and it is much simpler to implement. Besides, since our method is explicitly based on word segments, it is easier for us to filter out unwanted words like non-speech or noise. Finally, we merge those selected machine transcribed data and supervised data to form "semi-supervised" data.

4.3. Semi-supervised knowledge transfer learning

In this paper, we treat semi-supervised learning as a knowledge transfer learning process. In brief, we first use human transcribed data to train a DNN as the seed model, and then we perform semi-supervised learning using the seed model for initialization (with softmax layer randomly initialized). In other words, we don't do semi-supervised training from scratch, but rather start with a seed model that has already learnt from a limited amount of human transcribed data. We update the entire neural network during semi-supervised training phase. All semi-supervised data are used for cross-entropy DNN training, while for sMBR DNN training, only supervised data are used.

5. Cross-lingual knowledge transfer

5.1. Multilingual training

We use all 6 multilingual language data summarized in Table 1 to conduct multilingual training. Briefly, our training recipe is the same as those in [8, 16]. Practically, we first construct a multilingual DNN, which is topologically composed of two parts, a bottom part of many shared hidden layers, and a top part of parallel softmax layers with 6 sets of context-dependent (CD) tied-states as targets from 6 languages as mentioned. Training is complicated by the necessity of remembering which

[2]http://www.nist.gov/itl/iad/mig/openkws15.cfm

[3]https://github.com/kaldi-asr/kaldi

Table 2: Baseline WER (%) results on *dev* data with different monolingual supervised training methods

System	VLLP	LLP	FLP
GMM-HMM (SAT)	67.2	60.2	53.4
DNN (CE)	67.3	57.4	47.4
DNN (sMBR)	64.7	54.5	44.5

sample (frame and corresponding label) belongs to which language in each mini-batch of training data. Therefore, compared with monolingual DNN training, an extra utterance-to-language mapping file is needed to prepare beforehand.

5.2. Cross-lingual knowledge transfer learning

As the shared hidden layers are multilingually trained, they will learn the common characteristics of different languages. Such common characteristics are even shared by those languages unseen during training, realizing the possibility of cross-lingual transfer learning. The process of cross-lingual transfer learning is the same as that in Section 4.3. We notice that our recipe is different from what was advocated in [8], where softmax training and entire network tuning are conducted separately. Instead, we do both simultaneously.

6. Experimental setup

Experiments in this work are developed using the Kaldi toolkit. We use PLP+pitch features [23] as GMM-HMM front-end. The GMM-HMM system is trained up to the Speaker Adaptive Training (SAT) stage with 40 *dim* LDA+MLLT transformed features. Numbers of context-dependent states are set as 2k, 3k and 5k for the VLLP, LLP and FLP respectively (number of actual physical context-dependent states depends on language pack distribution).

All DNNs have 5 hidden layers with 2048 neurons in each layer. Input features are 25 *dim* composed of 22 *dim* filter-banks and 3 *dim* pitch features. They are first concatenated using a 21-frame window (10-1-10), and then passed through Hamming window and DCT transformation for DNN training.

7. Results

7.1. Baseline

Table 2 reports our monolingual results of SAT based GMM-HMM, DNN with cross-entropy and sMBR criteria respectively. From Table 2, we see that both VLLP and LLP systems perform about 20% and 10% worse than those of FLP systems. Our goal is to bridge the performance gap with our proposed semi-supervised and cross-lingual knowledge transfer learning methods. We also notice from Table 2 that cross-entropy DNN system doesn't perform better than GMM-HMM system for VLLP data, and the best result comes from sMBR DNN system. This suggests that cross-entropy DNN system is rather sensitive to the size of training data.

7.2. Monolingual based semi-supervised learning

Table 3 reports our monolingual based semi-supervised learning results with different word confidence thresholds for unsupervised data selection. A threshold of zero for word confidence in the first row of Table 3 means all unsupervised data are selected, which is about 34.33 hours for VLLP and 26.60 hours for LLP

Table 3: WER (%) results on *dev* data for different DNN-HMM acoustic models with monolingual based semi-supervised learning

Word *conf.*	VLLP		LLP	
	CE	sMBR	CE	sMBR
- (Baseline)	67.3	64.7	57.4	54.5
0.0	69.0	63.8	54.3	52.2
0.5	68.5	63.5	53.8	51.9
0.7	67.7	63.4	53.5	51.7
0.9	66.2	**62.7**	**53.3**	**51.5**
1.0	**65.7**	**62.7**	53.7	51.9

(as is indicated in the first column in Table 4). We notice that less data are selected for the LLP case than for the VLLP case. This is because there are more unsupervised data available for VLLP (unsupervised data are FLP data that do not appear in LLP or VLLP).

In contrast to Table 2, we only see 2.0% or 3.0% absolute WER reduction for semi-supervised learning in the VLLP and LLP "sMBR" cases respectively, with 0.9 set as confidence threshold. Figure 1 summarizes the percentage of selected data as a function of word confidence thresholds, where the VLLP and LLP are denoted as "Mono (VLLP)" and "Mono (LLP)" respectively. When word confidence threshold is set to 0.9, about 40% and 50% of data are selected for these two cases respectively.

Comparing row one (threshold is 0) with row four (threshold is 0.9) in Table 3, we see that data selection method may contributes up to 1.1% and 0.7% absolute WER reductions in VLLP and LLP case respectively.

Figure 2 shows the actual averaged word accuracy versus word confidence threshold. It suggests that the bigger threshold we use, the smaller word accuracy gaps we have between VLLP and LLP systems. This indicates some word confidences are overestimated. When word threshold is set 1.0, there is only marginal WER gaps between the systems, regardless of monolingual or multilingual system (multilingual systems are marked with "ML (VLLP)" and "ML (LLP)"). This suggests that the raw word posteriors calculated from ASR lattice are far from optimal for effective word confidence estimate.

Table 4: Overall data size (hours) in different cases after all non-content words are removed from the decoded FLP data, when word confidence threshold is set zero

Data	Monolingual	Multilingual
VLLP	34.33	32.98
LLP	26.60	25.12

7.3. Cross-lingual knowledge transfer

Table 5 shows effectiveness of the proposed cross-lingual knowledge transfer learning method. The maximum absolute WER reductions for VLLP and LLP cases are 8.6% and 5.8%, respectively, which are pretty significant. Besides, we see the proposed cross-lingual knowledge learning method (represented as "Overall-tuning" in Table 5) consistently outperforms "Softmax-tuning". Even when there are only 3 hours of training data (VLLP case), "Overall-tuning" still works well. This is different from what was observed in [8], where an extra tuning on the entire network didn't help with 3 hours of training data.

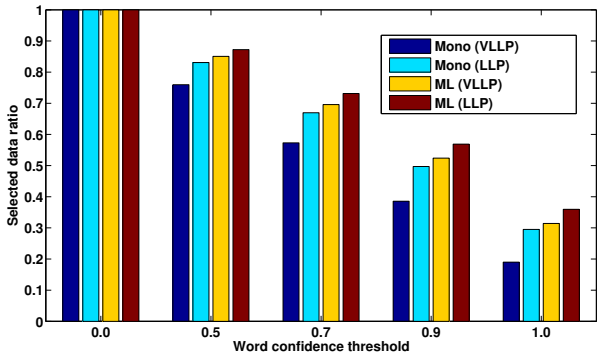

Figure 1: *Data selected ratio versus word confidence threshold for the selected unsupervised data under varieties of conditions, in which "Mono-VLLP" and "Mono-LLP" represent the cases where seed acoustic models are monolingual-trained. Similarly, "ML-VLLP" and "ML-LLP" represent the cases where seed acoustic models are trained with cross-lingual knowledge transfer learning*

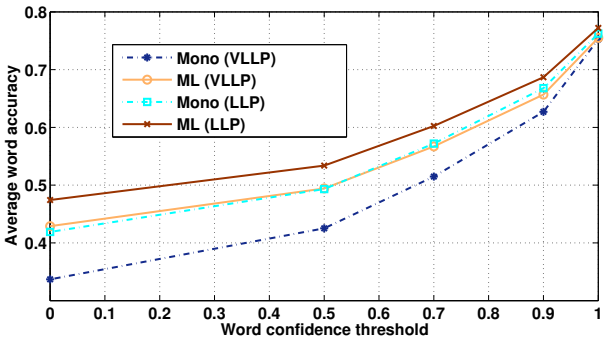

Figure 2: *Average word accuracy versus word confidence threshold for the selected unsupervised data under varieties of conditions, where an accuracy for a hypothesis word means the hypothesis and reference words should be the same and their time overlap should be over 85% of the reference word duration as well. "Mono-VLLP" and "Mono-LLP" represent the cases where seed acoustic models are monolingual-trained, for the VLLP and LLP respectively. Similarly, "ML-VLLP" and "ML-LLP" represent the cases where seed acoustic models are trained with cross-lingual knowledge transfer learning respectively*

While we understand that we cannot directly compare Table 5 and Table 3, it is quite clear that multilingual training works better in these scenarios. This can be attributed to two factors. Firstly. multilingual training benefits from more data. We only have about 30 hours of unsupervised monolingual data (see Table 4) while there are about 600 hours of multilingual data available (see Table 1). Secondly, data selection method is actually ineffective as is shown in Figure 2. Even with 1.0 threshold, our average word accuracy is less than 80%.

Table 5: WER (%) results on *dev* data for the DNN-HMM hybrid acoustic models with different cross-lingual knowledge transfer learning methods

System	VLLP		LLP	
	CE	sMBR	CE	sMBR
Monolingual	67.3	64.7	57.4	54.5
Softmax-tuning	58.0	56.1	51.3	50.0
Overall-tuning	57.6	55.9	50.5	48.7

7.4. Combination

As shown in Sections 7.2 and 7.3, a better seed model is critical for effective knowledge transfer learning. In this section we adopt the best systems in Table 5 as the seed systems for semi-supervised learning. We would like to see if further performance improvement can be achieved over Table 5. We summarize the results in Table 6. From Table 6, we only observe marginal improvements, particularly in the VLLP case. VLLP DNN cross-entropy system even gets worse. We notice that an obvious improvement is reported in [5] by combining these two methods with more unsupervised data.

Now let's turn to Figure 1 again, where four different seed models are used, we observe that a better seed model leads to more selected unsupervised data. We also observe from Figure 2, better seed models offer higher performance. However, when looking at the threshold segment between 0.9 and 1.0 in Figure 2, we see the difference of the average word accuracy of the selected data is dramatically diminished. This explains why our semi-supervised training has limited benefit from the data selection method even for the best seed model.

Table 6: WER (%) results on *dev* data with semi-supervised learning using cross-lingual knowledge transferred systems in Table 5 as seed systems

Word *conf.*	VLLP		LLP	
	CE	sMBR	CE	sMBR
- (Baseline)	57.6	55.9	50.5	48.7
0.0	59.6	56.6	50.0	48.7
0.5	59.3	56.2	49.9	48.5
0.7	58.8	56.0	49.6	48.3
0.9	58.2	**55.4**	**49.3**	**48.1**
1.0	**57.7**	55.7	49.5	48.3

8. Conclusions

In this paper, we proposed to use knowledge transfer learning framework for both semi-supervised and cross-lingual knowledge transfer learnings. The transfer learning is done by tuning DNN parameters using semi-supervised data or supervised data respectively. We demonstrated its effectiveness in these two kinds of learning tasks. We also attempted to combine these two techniques to get better performance improvement. Due to limited unsupervised data available, semi-supervised learning makes limited performance improvement over baseline system trained with cross-lingual transfer learning in both VLLP and LLP cases respectively.

9. References

[1] F. Grézl and M. Karafiát, "Semi-supervised bootstrapping approach for neural network feature extractor training," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2013.

[2] K. Veselý, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2013.

[3] H. Xu, H. Su, E.-S. Chng, and H. Li, "Semi-supervised training with bottle-neck feature based DNN-HMM hybrid modeling framework," in *INTERSPEECH 2014*, 2014.

[4] H. Su and H. Xu, "Multi-softmax deep neural network for semi-supervised training," in *Proceedings of INTERSPEECH*, 2015.

[5] J. Cui, B. Kingsbury, B. Ramabhadran, A. Sethy, K. Audhkhasi, X. Cui, E. Kislal, L. Mangu, K. Nussbaum-Thom, M. Picheny, Z. Tüske, P. Golik, R. Schlüter, H. Ney, M. J. Gales, K. M. Knill, A. Ragni, H. Wang, and P. Woodland, "Multilingual representations for low resource speech recognition and keyword search," in *Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015.

[6] M. Cai, Z. Lv, C. Lu, J. Kang, L. Hui, Z. Zhang, and J. Liu, "High-performance Swahili keyword search with very limited language pack: the THEE system for the OpenKWS15 evaluation," in *Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015.

[7] R. Caruana, "Multitask learning," *Machine Learing*, vol. 28, no. 1, pp. 41–75, 1997.

[8] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[9] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *ICASSP*, 2013.

[10] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *ICASSP*, 2013.

[11] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schulta, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *ICASSP*, 2014.

[12] Z. Tüske, D. Nolden, R. Schlüter, and H. Ney, "Multilingual mrasta features for low-resource keyword search and speech recognition systems," in *ICASSP*, 14.

[13] F. Grézl, M. Karafiát, and K. Veselý, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *ICASSP*, 2014.

[14] K. Veselý, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *IEEE Workshop on Spoken Language Technology (SLT)*, 2012.

[15] Y. Zhang, E. Chuangsuwanich, and J. Glass, "Language ID-based training of multilingual stacked bottleneck features," in *Proceedings of INTERSPEECH 2014*, 2014.

[16] H. Xu, V. H. Do, X. Xiao, and E.-S. Chng, "A comparative study of BNF and DNN multilingual training on cross-lingual low-resource speech recognition," in *Proceedings of INTERSPEECH*, 2015.

[17] F. Grézl and M. Karafiát, "Combination of multilingual and semi-supervised training for under-resourced languages," in *Proceedings of INTERSPEECH*, 2014.

[18] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance," *Computuer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

[19] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proceedings of INTERSPEECH*, 2013.

[20] N. F. Chen, S. Sivadas, B. P. Lim, H. G. Ngo, H. Xu, V. T. Pham, B. Ma, and H. Li, "Strategies for Vietnamese keyword search," in *ICASSP*, 2014.

[21] N. Chen *et al.*, "Low-resource keyword search strategies for Tamil," in *ICASSP*, 2015.

[22] N. F. Chen, V. T. Pham, H. Xu, X. Xiao, V. H. Do, C. Ni, I.-F. Chen, S. Sivadas, C.-H. Lee, E. S. Chng, B. Ma, and H. Li, "Exemplar-inspired strategies for low-resource spoken keyword search in Swahili," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[23] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Jrmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recogniton," in *ICASSP*, 2014.