CROSS-LINGUAL AUTOMATIC SPEECH RECOGNITION EXPLOITING ARTICULATORY FEATURES

Qingran Zhan Shixuan Du Petr Motlicek
Yahui Shan Xiang Xie

Idiap-RR-05-2021

APRIL 2021

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Cross-lingual Automatic Speech Recognition Exploiting Articulatory Features

Qingran Zhan[1,3], *Shixuan Du*[1], *Petr Motlicek*[3],*Yahui Shan*[1],*Xiang Xie**[1,2]

[1]Information and Electronics Institute, Beijing Institute of Technology, Beijing, China
[2]Shenzhen Research Institute, Beijing Institute of Technology, Shenzhen, China
[3]Idiap Research Institute, Martigny, Switzerland

asrzhanqingran@gamil.com, jfhz4689@gmail.com, petr.motlicek@idiap.ch,yahui_shan@163.com,
xiexiang@bit.edu.cn

Abstract

Articulatory features (AFs) provide language-independent attribute by exploiting the speech production knowledge. This paper proposes a cross-lingual automatic speech recognition (ASR) based on AF methods. Various neural network (NN) architectures are explored to extract cross-lingual AFs and their performance is studied. The architectures include mutilayer perception(MLP), convolutional NN (CNN) and long short-term memory recurrent NN (LSTM). In our cross-lingual setup, only the source language (English, representing a well-resourced language) is used to train the AF extractors. AFs are then generated for the target language (Mandarin, representing an under-resourced language) using the trained extractors. The frame-classification accuracy indicates that the LSTM has an ability to perform a knowledge transfer through the robust cross-lingual AFs from well-resourced to under-resourced language. The final ASR system is built using traditional approaches (e.g. hybrid models), combining AFs with conventional MFCCs. The results demonstrate that the cross-lingual AFs improve the performance in under-resourced ASR task even though the source and target languages come from different language family. Overall, the proposed cross-lingual ASR approach provides slight improvement over the monolingual LF-MMI and cross-lingual (acoustic model adaptation-based) ASR systems.

Index Terms: deep neutral network, articulatory features, cross-lingual speech recognition.

1. Introduction

Deep neural networks (DNNs) have contributed to a significant increase of performance in automatic speech recognition (ASR) by replacing the traditional Gaussian Mixture Modeling (GMM) based approaches. However, due to the speaker variations and environmental noise, there is still an ongoing effort in developing robust feature extraction methods offering competitive performance for different conditions.

Articulatory features (AFs) are traditionally used to represent the movement of different articulators, such as lips and tongue, during speech production. AFs are known to offer additional robustness to noise and pronunciation variants compared to conventional acoustic features. There are several methods to extract the AFs: (i) use an X-ray radiometer to measure movements of vocal organs [1], (ii) use inverse filtering techniques on acoustic signal [2], or (iii) estimate the articulatory features form speech data using statistical classifiers based on linguistic knowledge. The first method is complex and time-consuming, while the second method requires high quality filters, technically hard to design. In order to do cross-lingual speech recognition, we choose the third method to extract the AFs. There are many approaches to generate the AFs (through so-called AF extractor), such as based on SVM [3], HMM [4], or DNN [**?**, 5, 6].

In recent years, new deep learning architectures for extracting articulatory features were studied. For instance in [7] and [**?**], they have found CNNs performing well in extracting AFs.

Unlike previous works, this paper proposes an AF-based approach, based on International Phonetic Alphabet (IPA) [8], for cross-lingual ASR from English to Mandarin. As the Mandarin comes from different language family than English, we introduce the AFs approach to the cross-lingual speech recognition task. We hypothesize that AFs can provide shared information between Mandarin and English which can be employed through a statistical modeling. The proposed method first focuses on building set of AF extractors which is then used to boost the ASR system. Various neural network (NN) architectures are investigated to extract AFs and their performance is analyzed, including multi-layer perception (MLP), convolutional NN (CNN) and long short-term memory recurrent NN (LSTM). More specifically, the set of AF extractors is first built using source (English, well-sourced) language. The developed AF extractors are then used to forward-pass AFs to the target (Mandarin, under-source) language. Finally, the AFs are combined with conventional speech features (MFCCs) and the stacked features are employed to train the ASR in target language. The ASR results show that the proposed approach outperforms both the monolingual ASR system as well as the adapted ASR system, where the parameters of DNN-based acoustic model are adapted to the target-language.

The rest of the paper is organized as follows: Section 2 briefly discusses the previous work on articulatory features and cross-lingual ASR. Section 3 provides the definition of AFs. Experimental setup is given in section 4. Section 5 discusses the results. Finally, section 6 concludes the work and discusses the future steps.

2. Previous Work

2.1. Previous Work on Articulatory Features

In the past, combining the articulatory features (AFs) together with conventional acoustic features has shown to be beneficial for ASR, In [9], Mitra et al. first studied a DNN used to first estimate articulatory trajectories from speech signal and then apply the network to extract the articulatory trajectories for training and testing datasets for English ASR system. In [10], Emre et al. investigated combination of the articulatory features extracted from the labeled speech corpus with the traditional acoustic features for the the pathological speech recognition. Manjunath [11] et al. investigated a DNN based AF prediction methods applied in a fusion to enhance the acoustic features with

Table 1: *Statistic of speech corpora.*

Language	#Speakers		#Utterances		Duration (in hours)	
	M	F	Train	Test	Train	Test
Mandarin	30	10	10'893	2'496	27.2	6.2
English	125	126	28'539	2'620	100.6	5.4

various categories of AFs. Results showed that the AFs can be beneficial for the multilingual phone recognition.

2.2. Previous work on cross-lingual speech recognition

Overall, DNN based acoustic models have been shown to provide consistent advantages for multilingual speech recognition tasks [12, 13]. For cross-lingual ASR, [14] found that combining bottleneck features extracted using monolingual (English and Mandarin) ASR systems perform worse than the baseline ASR built from conventional acoustic (PLP) features. This suggests that the transfer knowledge from English to Mandarin is a difficult task. In [15], the authors have demonstrated the transfer knowledge from European languages to English and Chinese target languages. The trained DNN has the hidden layers made common across many languages while the softmax layers are made language dependent. The final ASR system has been shown to outperform the monolingual ASR systems by 3% - 5% relative in word error rate (WER).

3. Extraction of articulatory features

3.1. Phone set

English ASR developed in this paper uses the standard phone set with 39 phones defined in Worldset Symbols [16]. It comprises 15 vowels and 24 constants. There are also different articulatory features among languages, like *zero consonant* appearing for Mandarin, which do not exist for English. In order to simplify sharing the common units in IPA, we applied some modifications on the Mandarin phone set. For instance, in Mandarin dataset, AFs of two different vowels in a diphthongs are not the same, so diphthongs are transcribed using separate monophthongs[1]. After the modifications, we obtained 50 phones for Mandarin, comprising 25 vowels and 25 constants.

3.2. Definition of the articulatory features

In this work, we assume that AFs are similar across the languages and thus they can be viewed as language-independent sub-phonetic units. However, the languages tackled in this paper (English and Chinese Mandarin) belong to far different language families [17]. Based on the IPA and the linguistic knowledge [18, 19], we therefore define the AF types as described in Table 1. In this table, *Nil* means "not-specified", for example, *Height* of articulation does not exist for constants, so for constants, the AFs type in *Height* is *Nil*.

4. Experimental Setup

4.1. Speech corpora

For developing and analyzing the performance of the proposed cross-lingual ASR system, speech datasetes of English (source language) and Mandarin (target language) are considered. For Mandarin and English we choose 30 hours Thchs30 set [20] and 100 hours sub-set of Librispeech [21], respectively. The set of AF extractors is trained on English, while the acoustic and language models of the ASR system are developed and evaluated on Mandarin data. Thchs30 was developed by the Center for Speech and Language Technologies at Tsinghua University and involves about 30 hours of speech. Librispeech is a large scale corpus containing approximately 1'000 hours of speech aligned with their transcriptions. We used a 100 hour sub-set . The sampling rate of both datasets is 16 kHz. Table 3 shows the statistics of the two datasets.

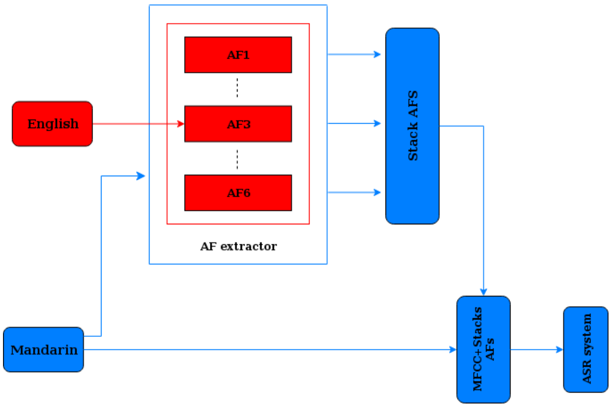


Figure 1: *Framework of the AF-based cross-lingual ASR. (i) Red block diagrams represent the set of AF classifiers developed using English data. (ii) Blue block diagrams represent the forward pass of target-language data through the set of AF classifiers for training the final ASR in Mandarin.*

Table 2: *Table with detailed set of articulatory features.*

AFs	AF types	Number
Backness	Front, Central, Back, Nil	4
Height	High, Middle-High, Second-highest, Half-high Medium,High, Half-low, Nil	8
Rounding	Round, Unround, Nil	3
Apical	High, Medium, Low, Interdental, Nil	5
Lingual sound	Front, Central, Back, Nil	4
Labial sound	Bilabial, Labiodental, Nil	3

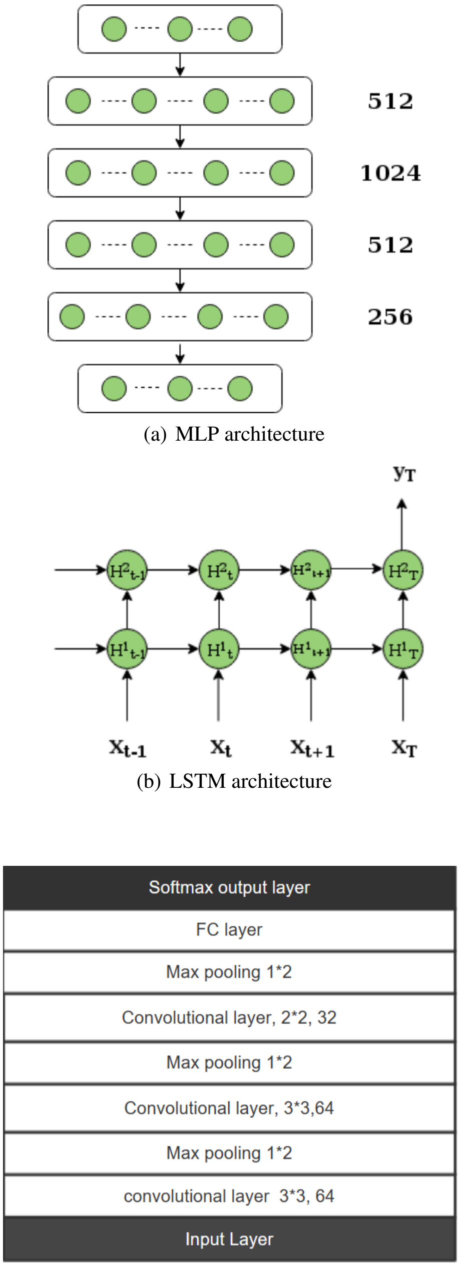
4.2. Proposed framework on extracting AFs

The proposed AF-based cross-lingual ASR system is implemented as shown in Figure 3. More specifically, the developed system can be split into two parts: (i) AF extractor (i.e. set of individual AF extractors), and (ii) cross-lingual ASR. The Set of AF extractors is trained on the source language using frame-level alignments obtained using previously developed monolingual GMM-HMM system. Different NN architectures are developed and evaluated to classify AFs on the source language. Further, the speech data of the target language are forward-passed through set of AF extractors and obtained features are combined with conventional MFCCs to train the final ASR system.

[1] https://en.wikipedia.org/wiki/Monophthong

Table 3: *Frame-level classification accuracy [%] of articulatory features. The classification results are provided for two tasks (a) mono-lingual - source-language test-set, and (b) cross-lingual - target-language test-set.*

architecture	Training set	Testing set	Backness	Height	Rounding	Apical	Lingual sound	Labial consonant
MLP	English	English	68.1	67.7	67.1	78.8	86.8	88.3
MLP	English	Mandarin	59.0	57.7	57.1	75.3	79.8	85.4
CNN	English	English	80.3	81.5	82.4	86.8	90.1	90.8
CNN	English	Mandarin	60.8	59.8	57.3	71.8	79.7	85.1
LSTM	English	English	82.5	82.4	84.8	85.4	95.1	94.7
LSTM	English	Mandarin	**69.7**	**70.0**	**67.3**	**82.0**	**84.0**	**89.1**



(a) MLP architecture

(b) LSTM architecture

(c) CNN architecture

Figure 2: *Analysed NN architectures for AF extractors.*

4.3. Set of AF extractors using frame alignments

This sections provides a detailed analysis on different neural network architectures for building set of individual articulatory classifiers. We conducted experiments with three different NN architectures to estimate the frame-level AFs. Individual articulatory extractors were built specifically for each AF class. The GMM-HMM model built on the source-language was used to generate the phone-level alignments, further converted into AFs values using the phone-to-AF mapping[2]. Further, MLP, CNN and LSTM NN classifiers were built. The input for the AF classifiers is represented by 39-dimensional conventional MFCCs, enlarged by the context of 5 left and 5 right consecutive frames with the middle frame being the frame to classify. The size of the output layers of the individual AFs classifiers are specified in Table 1. All the NN architectures apply a ReLU non-linearity. The batch normalization [22] is applied before the ReLU non-linearity. More details are given in Figure 2.

4.4. ASR system building

This section describes both monolingual and cross-lingual ASR works. All ASR systems built in this work are developed using Kaldi [23] from a GMM-HMM architecture exploiting conventional 39-dimensional MFCCs including the deltas and delta-deltas. The monophone (both English and Mandarin) acoustic models have 3 states to model non-silence phones) and 5 states to model silence phones. The context-dependent (tri) acoustic models are trained using MFCCs, followed by speaker adaptive training. The TDNN-HMM acoustic models (using CE loss) comprise 6 hidden layers, each consisting of 625 units, trained from 11 consecutive frames using the alignment from the GMM-HMM. We also trained a TDNN-HMM with LF-MMI objective function [24]. As a cross-lingual ASR baseline, we exploit an usual approach where TDNN-HMM (CE) acoustic model is first built using source-language data (i.e. 100 h English corpora). Then, the output layer is replaced with a new output layer corresponding to the Mandarin target units and the whole TDNN-HMM model is retrained (i.e. the transferred layers use a smaller learning rate while the output layer use a larger learning rate). For AFs based cross-lingual ASR, we combined the AF and conventional MFCCs. The combined features are deployed in the acoustic model (TDNN-HMM (CE) built using target-language data.

For all ASR experiments on target-language, we employ standard (3-gram) language model, distributed with the dataset.

[2] https://github.com/iezhanqingran

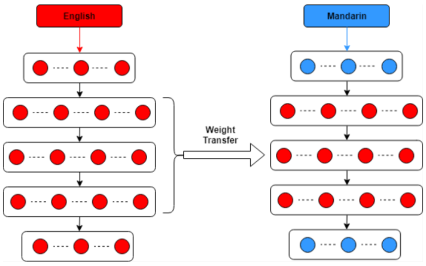


Figure 3: *Cross-lingual adaptation form English to Mandarin.*

5. Results and discussions

5.1. AF classification results

Table 3 illustrates the frame-based accuracies on the AF classification task for three different types of AF extractors. The evaluation is performed on source-language as well as target-language datasets. From the table we can see that the LSTM-based classifier provides the best performance to generate the AFs, both for monolingual (English test-set) and cross-lingual (Mandarin test-set) tasks. Based on obtained results, the following ASR experiments exploit the LSTM classifier for extracting the articulatory features.

5.2. ASR results

This section compares the ASR results obtained on the target-language test-set.

First, the conventional acoustic models are built (monophone and triphone GMM-HMMs, TDNN-HMM (CE) and TDNN-HMM (LF-MMI)) using standard MFCC features. Training data from the target-language are used. Then, the cross-lingual AFs, combined with MFCCs, are used to train the same types of acoustic models. More specifically, as described above, the set of AF classifiers is trained on source-language. AFs are therefore obtained by forward-passing the Mandarin (training and testing) data through these classifiers. The ASR results (in terms of WER) are given in Table 4 for different types of acoustic models. From the table we can find that the AFs-based ASR approach provides the best performance across all ASR systems (both GMM-HMM and TDNN-HMM based), although the improvements are not large.

Finally, the proposed AF-based ASR system is compared with the conventional cross-lingual ASR baseline (i.e. the HMM-TDNN trained on source language, further adapted to the target language as described in Section 4.4). The results are given in Table 5. AFs-based ASR system slightly outperforms the adapted ASR which indicates the effectiveness of our proposed method. As the Mandarin language is considered as an under-resource language in this paper, another observation which can be made from the results is that the AF-LSTM classifiers are able to capture the common information (i.e. on articulatory feature level) which can be used to improve the performance of low-resource ASR.

6. Conclusions

In this paper, we propose a framework for cross-lingual ASR for Mandarin. The approach exploits the source (English) data to train set of AFs classifiers. The generated AFs obtained by forward-passing the Mandarin data through the AF extractors are then combined with conventional MFCCs and used to build the ASR system on target language. The frame-level AF classification accuracies demonstrate that the LSTM are most effective to generate the AFs. The cross-lingual ASR results indicate that AFs can eventually improve the speech recognition performance, over other cross-lingual baselines. Our future work will focus on multi-task work, extending the proposed approach to more languages.

Table 4: *ASR systems (WER%) evaluated on Mandarin test set. We evaluate different types of acoustic models, trained on the target-language. Different type of features are used, either conventional MFCCs (i.e. monolingual ASR system) or AF-LSTM based features (estimated on source-language) (i.e. cross-lingual ASR system).*

	Features	
Acoustic Model	MFCC	Cross-lingual AF-LSTM
Mono GMM-HMM	53.5	**52.6**
Tri GMM-HMM	35.6	**34.7**
TDNN-HMM (CE)	32.9	**32.3**
TDNN-HMM (LF-MMI)	31.8	**31.5**

Table 5: *TDNN-HMM (CE) ASR systems evaluated on the target-language (Mandarin) test set.*

Acoustic model	Features	WER [%]
Mono-lingual	MFCC	32.9
Cross-lingual (DNN adaptation)	MFCC	32.4
Cross-lingual (AF)	MFCC+AF-LSTM	32.3

7. Acknowledgements

The authors gratefully acknowledge partial support from the China Scholarship Council (CSC). This work has been supported by National Nature Science Foundation of China (No.11590770-11590774), Science and Technology Innovation Foundation of Shenzhen (JCYJ20180504165826861).

8. References

[1] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *The Journal of the Acoustical Society of America*, vol. 92, no. 2, pp. 688–700, 1992.

[2] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 133–150, 1994.

[3] O. Scharenborg, V. Wan, and R. K. Moore, "Capturing fine-phonetic variation in speech through automatic classification of articulatory features," *Vaccine*, vol. 25, no. 16, pp. 3101–4, 2006.

[4] D. T. F. A. Der, K. Kirchhoff, and B. Juni, "Robust speech recognition using articulatory information," *Ph.d.dissertation University of Bielefeld*, 1999.

[5] B. Abraham, S. Umesh, and N. M. Joy, "Articulatory feature extraction using ctc to build articulatory classifiers without forced frame alignments for speech recognition." 2016.

[6] B. Abraham, "Cross-lingual techniques and use of articulatory features in acoustic modeling of low-resource languages."

[7] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, and zgr etin, "Articulatory feature classifiers trained on 2000 hours of telephone speech," *Proc Interspeech*, pp. 2485–2488, 2007.

[8] Wikipedia contributors, "Ipa — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=IPA&oldid=888604190, 2019, [Online; accessed 1-April-2019].

[9] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, and E. Saltzman, "Articulatory features from deep neural networks and their role in speech recognition," in *IEEE International Conference on Acoustics*, 2014.

[10] E. Ylmaz, V. Mitra, C. Bartels, and H. Franco, "Articulatory features for asr of pathological speech," in *Interspeech*, 2018.

[11] K. Manjunath, D. B. Jayagopi, and V. Ramasubramanian, "Indian languages asr: A multilingual phone recognition framework with ipa based common phone-set, predicted articulatory features and feature fusion."

[12] Z. Tüske, J. Pinto, D. Willett, and R. Schlüter, "Investigation on cross-and multilingual mlp features under matched and mismatched acoustical conditions," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7349–7353.

[13] N. T. Vu and T. Schultz, "Multilingual multilayer perceptron for rapid language adaptation between and across language families." in *Interspeech*, 2013, pp. 515–519.

[14] J. Li, R. Zheng, B. Xu *et al.*, "Investigation of cross-lingual bottleneck features in hybrid asr systems," 2014.

[15] 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 *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7304–7308.

[16] J. L. Hieronymus, "Ascii phonetic symbols for the worlds languages: Worldbet."

[17] Wikipedia contributors, "Language family — Wikipedia, the free encyclopedia," 2019, [Online; accessed 27-March-2019]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Language_family&oldid=882233459

[18] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[19] I. Yuen, M. H. Davis, M. Brysbaert, and K. Rastle, "Activation of articulatory information in speech perception," *Proceedings of the National Academy of Sciences*, vol. 107, no. 2, pp. 592–597, 2010.

[20] D. Wang and X. Zhang, "Thchs-30: A free chinese speech corpus," *arXiv preprint arXiv:1512.01882*, 2015.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[22] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 648–656.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[24] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." 2016.