Articulatory Features Based TDNN Model for Spoken Language Recognition

Jiawei Yu, Minghao Guo, Yanlu Xie, Jinsong Zhang
Beijing Advanced Innovation Center for Language Resources
Beijing Language and Culture University
Beijing, China
vyujiawei@gmail.com, gmhgmh8000@163.com, {xieyanlu, jinsong.zhang}@blcu.edu.cn

***Abstract*—In order to improve the performance of the Spoken Language Recognition (SLR) system, we propose an acoustic modeling framework in which the Time Delay Neural Network (TDNN) models long term dependencies between Articulatory Features (AFs). Several experiments were conducted on APSIPA 2017 Oriental Language Recognition(AP17-OLR) database. We compared the AFs based TDNN approach to the Deep Bottleneck (DBN) features based ivector and xvector systems, and the proposed approach provide a 23.10% and 12.87% relative improvement in Equal Error Rate (EER). These results indicate that the proposed approach is beneficial to the SLR task.**

***Keywords*-spoken language recognition; articulatory features; time delay neural network; deep bottleneck features;**

I. Introduction

Spoken Language Recognition (SLR) technology is to identify or verify the language being spoken in a speech utterance [1]. It can be used as the front-end of the multi-lingual speech recognition systems and the automatic translation systems [2], [3]. The state-of-the-art SLR systems can use a variety of features to distinguish one language from another, such as: acoustic features, prosody, phonotactic structure, lexical knowledge, vocabulary [1].

Generally, the SLR approach can be classified into two types according to features used: spectral-based and token-based. The spectral-based one exploit different distributions in the acoustic space between different languages. State-of-the-art modeling method, such as ivector and xvector, project acoustics of different languages to different places in high dimensional space. The token-based approaches utilize phonotactic information which characterizes how these phonemes are combined in a language. One of the examples is Phone Recognition followed by Language Modeling (PRLM) which converts a speech utterance into a sequence of phones by a phone recognizer, then uses an n-gram language model produces a likelihood score [1].

Compared to the token-based approach, the spectral-based one is weak at modeling the temporal information such as phonotactics for SLR task. On the contrary, the token-based approach can not accurately exploit the acoustic differences between languages. Especially, they heavily rely on the accuracies of recognizers [1] , which is usually hard for cross-language tasks.

In view of this, AFs were introduced to the SLR task [4], [5], [6]. The AFs represent the articulatory specification in the vocal tract when pronouncing a phone. The combination of a few AFs can determine a specific phone. The finer granularity of AFs have, the better cross-language modeling power are got when compared to phonemes. So the recognition performance of AFs is generally better than phonemes [7], and consequently the AFs based SLR systems perform better. Besides, n-gram LM based on AFs is capable of modeling the phonotactics of different languages, and can bring about further improvements to SLR task. However, it still suffers from the data sparsity problem of n-gram, especially when contextual width is to be lengthened [1].

In this paper, we propose a scheme of AFs plus TDNN for SLR task. Our motivations include utilizing the advantage of cross-language modeling by AFs, and the effectiveness of TDNN in modeling temporal dependencies in the acoustic signal.

The rest of this paper is organized as follows. Section II presents the AFs based TDNN SLR system in detail. The experimental setup is presented in Section III. Finally Section IV and Section V show the experimental results and conclusion.

II. AFs Based TDNN SLR System

The AFs based TDNN SLR system diagram is shown in Fig. 1. The system include two part, the front-end is a feature extractor which processing spoken language utterances into sequence of AFs using ASR DNN model. Once we get the AFs, the TDNN back-end will classify these tokens to the specific language.

A. Articulatory Features

The International Phonetics Association (IPA) classifies the sounds of a language by means of AFs [8]. A sound is described by a bundle of articulatory features, and a unique symbol is used as a shorthand to represent this bundle. The AFs generally used to assist automatic speech recognition (ASR) [9], and several studies have proved that AFs can be recognized more robustly across languages than phonemes [7]. In the token-based and DBN based SLR approaches, the accuracy of phone recognizer is a critical factor. Specifically, if a phoneme of another language to be recognized is always recognized as the one in the phone set designed for the phone recognizer, it is fine to model it in the language model based on the assumption of similarity between them. If some phonemes are very different from the phonemes of the language for phone

978-1-7281-5014-7/19/$31.00

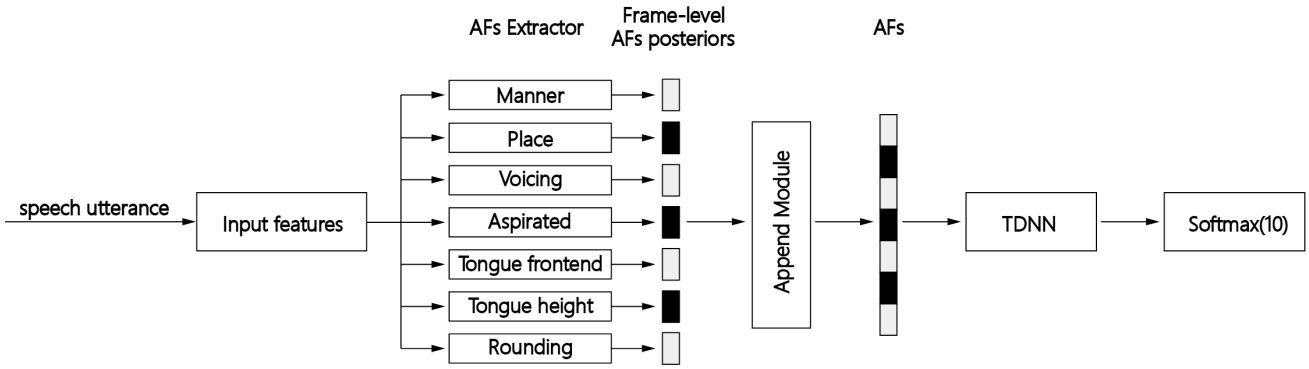

Figure 1. Block diagram of articulatory features based TDNN SLR system.

recognizer, they cannot be represented well in language modeling, which is quite common for spoken languages in different language families. So, we use AFs that are language-universal across all spoken languages as a front-end to obtain a more accurate recognizer and improve the performance of SLR system.

B. Articulatory Categories

In this paper, we choose 39 AFs, which belong to 7 categories listed in Table I according to the scheme of IPA. Each type of AFs having different items, e.g. "manner" has 6 items: affricate, fricative, nasal, liquid, sibilant and final labels. Except these AFs, We also use the "silence" token to represent the soundless segments.

TABLE I. OVERVIEW OF AFS TYPES USED

AFs	Numbers[a]	Description
Manner (MA)	9	*Manner of articulation*
Place (PA)	6	*Place of articulation*
Voicing (VO)	2	*Voicing*
Aspirated (AS)	3	*Aspiration*
Tongue frontend (TF)	8	*Tongue x position*
Tongue height (TH)	8	*Tongue y position*
Rounding (RO)	3	*Lips rounded*

a. The item number of each AF.

C. Articulatory Features Extraction

Since manual AF annotations of speech signals are rather difficult and costly to produce, one reasonable way of generating training material for the articulatory classifier is to convert phone-based training transcriptions to AFs transcriptions [10]. This can be achieved by using a canonically defined phone and AFs mapping table. In this study, we use Mandarin phone set converting AFs. Our mapping table is based on the [11], and we added three new mapping relationship between phone and AFs, as shown in Table II. In this paper, we used the posterior probabilities of the articulatory categories as the articulatory features. As shown in Fig. 1, the feature extraction module consists series of AFs' extractors according to 7 attribute categories described in Table I. A context dependent DNN-based AFs' extractor is separately built for each category. The current frame posteriors are linked to the possible class within that category. Subsequently, a group of the frame attribute posteriors will be fed into the append module, the append module stacks together with the attribute posteriors and generates a vector, which is AFs.

TABLE II. AFS AND THEIR ASSOCIATED PHONES IN MANDARIN

AFs	Category	Phone set
Tongue Frontend (TF)	Front 2	*ii*
	Front 1	*iii*
	Front	*i v*
	Half F	
	Central	*a*
	Half B	
	Back	*u*
Tongue Height (TH)	High	*i ii iii v u*
	Second H	
	Half H	
	Middle	
	Half L	
	Second L	
	Low	*a*
Rounding (RO)	Rounded	*u v*
	Unrounded	*a i ii iii*

D. TDNN Back-end

The TDNN structure is shown in Fig. 2. The architecture of TDNN is designed to work on sequential data. Specifically, A TDNN is formulated as a feedforward network but it has delays on the layer weights associated with the input weights. The data are represented at different time points by adding a set of delays to the input. This allows the TDNN to have a finite dynamic response to time series input data [12], [13], [14].

In TDNN structure, a narrow temporal context is provided to the first layer and increasingly wide contexts are available to the subsequent hidden layers, i.e., each layer in a TDNN operates at a different temporal resolution, in this way, the higher layers of the network are able to learn longer temporal relationships.

To explain in more detail how TDNN learns long term temporal dependencies between AFs, we use the following example to illustrate. as shown in Fig. 2, Suppose t is current frame, at the input layer (layer1), frames [t-2,t+2] are spliced together. Layers 2, 3 and 4 we splice together frames [t-1, t+2], [t-3, t+3] and [t-7, t+2] respectively. In

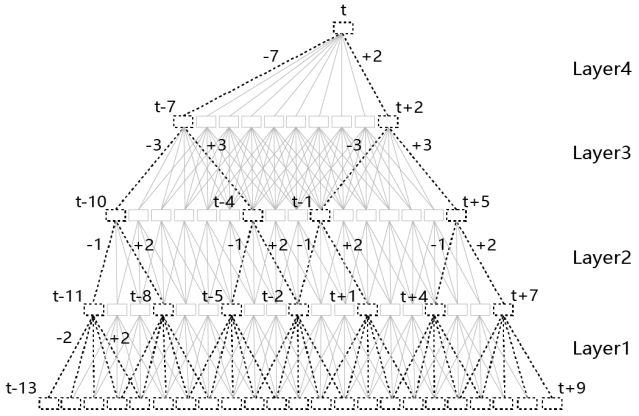

Figure 2. TDNN architecture.

total, the DNN has a left-context of 13 and a right-context of 9.

III. Experimental Setup

To establish a baseline framework, we built a classical ivector and xvector system, the feature we use include: MFCC, DBN features and AFs. At the same time, we also establish a n-gram LM based back-end system for comparing the performance with TDNN. Baseline systems are summarized below. All the experiments were conducted with Kaldi toolkit [15].

A. Description of Database

The DBN features are extracted from an ASR DNN trained on two mandarin corpus. The first one is from the Chinese National Hi-Tech Project 863 for Mandarin LVCSR system development [16], and the second corpus is an open-source Mandarin speech corpus called AISHELL-1 [17]. A total of 250,000 utterances spoken by 1800 speakers (300 hours) were used for acoustic modeling.

All experiments were evaluated on AP17-OLR databases which used for second oriental language recognition challenge [18]. The database consists of 10 different languages: Kazakh in China (ka-cn) , Tibetan in China (ti-cn), Uyghur in China (uy-id), Cantonese in China Mainland and Hong Kong (ct-cn), Mandarin in China (zh-cn), Indonesian in Indonesia (id-id), Japanese in Japan (ja-jp), Russian in Russia (ru-ru), Korean in Korea (ko-kr), and Vietnamese in Vietnam (vi-vn). The duration of training data for each language is about 10 hours and the speeches were recorded with mobile phones, at a sampling rate of 16 kHz and 16 bits resolution. Our systems evaluated on one of databases' condition called "test all".

B. Features Extraction

The acoustic MFCC features are 40-dim without cepstral truncation and with a frame-length of 25ms. These features are equivalent to filter bank coefficients, but are more compressible.

The phonetic DBN features are from an ASR DNN acoustic model. The DNN is a time-delay acoustic model with p-norm nonlinearities, and the DNN model was trained on the AISHELL-1 and 863 mandarin corpus described on Section III. The DNN has 6 hidden layers, and the dimension of each DNN layer is set to be 650, except for the last hidden layer is replaced with a 100 dimensional linear bottleneck layer. Its input features are 40-dimentional MFCC. Excluding the softmax output layer, which is not needed to compute DBN features. No fMLLR or ivectors are used for speaker adaptation.

The AFs are also from an ASR DNN acoustic model, but they were extracted from the posteriors of softmax output layer instead of the bottleneck layer. The DNN is a chain time-delay acoustic model with p-norm nonlinearities, and the DNN model was trained on the AISHELL-1 and 863 mandarin corpus described on Section III. The DNN has 6 hidden layers, and the dimension of each DNN layer is set to be 625. Its input features are 40-dimentional MFCC. Furthermore, the input features are at the original frame rate of 100 per second and the output frame rate is reduced by 3-fold.

C. I-vector Baseline

The ivector system follows the procedure described in [19]. It is based on the GMM-UBM and the UBM is a 2048 component full-covariance GMM. The system uses a 400 dimensional ivector extractor and cosine for scoring. The input features to ivector system separately incorporates mentioned above three features: MFCC, DBN and AFs.

D. X-vector Baseline

The xvector system follows the procedure described in [20]. The network of xvector is a 5 layer TDNN. The input of each layer is the sliced output of the previous layer. The sliced indices in the consecutive layers were [t-2, t-1, t, t+1, t+2; t-2, t, t+2; t-3, t, t+3; t; t]. The dimension of 1-4 layer is 512, and 5th layer is 1500. The segment-level part is a 2-layer fully connected network with 512-dim per layer and the nonlinearities are rectified linear units (ReLUs). The output is a softmax layer and the size is 10 (the number of languages).

E. TDNN SLR Classifier

The detail of TDNN back-end is described in section II. The configuration of TDNN shows below. The TDNN model was composed of 6 layers and the dimension of each layer is 650. The activation function was p-norm and the spliced indices in the consecutive layers were [t-2, t-1, t, t+1, t+2; t-1, t, t+1; t-1, t, t+1; t-3 ,t ,t+3; t-6, t-3, t]. The output is a softmax layer and the size is 10 (the number of languages).

IV. Results

A. AFs based TDNN vs Baseline System

Table II shows the performance of different features in ivector, xvector and the proposed SLR system. We use All_AFs to denote 7 attributes combined together. As we can see, AFs based systems achieve lower EER than DBN based system and MFCC based system. Overall, AFs

based TDNN outperforms DBN based TDNN by about 46%, AFs based ivector outperform DBN based ivector by about 9% and AFs based xvector outperform DBN based xvector by about 20%. It is evident that AFs are beneficial to improving the performance of SLR task.

Moreover, the result of AFs based TDNN is impressive. AFs based TDNN which has a 15% relative improvement perform better than AFs based ivector. At the same time, the result of AFs based TDNN and AFs based xvector is close. It reveals that a simple TDNN back-end with AFs is effective to SLR task.

TABLE III. SYSTEM PERFORMANCE IN DIFFERENT METHODS IN TERMS OF PERCENTAGE OF EER AND MINCAVG (REPORTED WITHIN PARENTHESIS)

Feature	TDNN	Ivector+cosine	Xvector+cosine
$MFCC$	11.29(12.09)	6.22(6.87)	5.76(5.13)
DBN	7.17(6.88)	5.02(4.76)	4.43(4.53)
All_AFs	**3.86 (3.56)**	4.56(4.32)	**3.52 (3.22)**

B. Fusion

The fusion results shown in table III, we evaluate the fusion system using the Focal toolkit [21], and we fuse AFs based TDNN system to different approach. As we can see, the fusion systems get significant performance gain for SLR. Especially, The AFs_ivector+AFs_xvector+AFs_TDNN fusion system delivered a relative improvement of 45% in EER to the best AFs based xvector approach alone.

TABLE IV. THE PERFORMANCE OF DIFFERENT FUSION SYSTEM IN TERMS OF PERCENTAGE OF EER AND MINCAVG (REPORTED WITHIN PARENTHESIS)

Fusion	EER(minCavg)
$MFCC\_TDNN + AFs\_TDNN$	3.46(3.76)
$DBN\_TDNN + AFs\_TDNN$	2.95(3.21)
$DBN\_ivector + AFs\_TDNN$	2.56(2.32)
$DBN\_xvector + AFs\_TDNN$	2.21(2.36)
$AFs\_ivector + AFs\_TDNN$	2.27(2.53)
$AFs\_xvector + AFs\_TDNN$	2.14(2.01)
$AFs\_ivector + AFs\_xvector + AFs\_TDNN$	**1.92 (1.84)**

TABLE V. TDNN BASED SYSTEM PERFORMANCE IN DIFFERENT AFS AND AF'S COMBINATION IN TERMS OF PERCENTAGE OF EER AND MINCAVG (REPORTED WITHIN PARENTHESIS)

Feature	EER(minCavg)
$MFCC$	11.29(12.09)
$Manner(MA)$	9.51(10.21)
$Place(PA)$	**6.53 (6.14)**
$Voicing(VO)$	10.45(10.23)
$Aspirated(AS)$	11.66(11.54)
$Tonguefrontend(TF)$	9.82(10.28)
$Tongueheight(TH)$	10.15(10.53)
$Rounding(RO)$	10.98(10.68)
$MA + VO + AS$	7.12(7.45)
$MA + PA + VO + AS$	5.25(4.78)
All_AFs	3.86(4.10)

C. Performance of different AFs

To investigate the performance based on the different AFs, we did several experiments to evaluate the system performance for different AFs and AFs' combination. Table V shows the SLR results, the description of AFs is described in Table I of section II. The features shown in Table V were evaluated using the TDNN system (see Section II). The results shows that place of articulatory (PA) can significantly improve the performance of SLR. The PA based system show the highest performance among all single AFs system, and this indicates that different AFs have different effects on the SLR system. Furthermore, the performance will improve when combine different AFs.

V. CONCLUSION

In this paper, we have explored using AFs based TDNN modeling for SLR task. This approach took advantage of the cross-lingual characters of AFs and the capability of TDNN capturing long term dependencies between input features . The experiments were performed on AP17-OLR database revealed effectiveness of the proposed approach. Specifically, The experimental results show that our proposed approach provides a 23.10% and 12.87% relative improvement in EER to DBN features based i-vector and xvector approach. The AFs based i-vector or x-vector approach also achieved performance gain to DBN based approach. Furthermore, we evaluated the fusion system. The fusion of AFs based TDNN approach with different baseline approach got significant performance gain for SLR task. Finally, We evaluated the effectiveness of different AFs. The result shows that the place of articulation is the most effective feature compared to other AFs. These results is a strong support that the AFs based TDNN approach is beneficial to the SLR task.

ACKNOWLEDGMENT

This study was supported by Advanced Innovation Center for Language Resource and Intelligence (KYR17005), the Fundamental Research Funds for the Central Universities (16ZDJ03, 18YJ030006, 19YCX113), and the project of "Intelligent Speech technology International Exchange". Jinsong Zhang is the corresponding author.

REFERENCES

[1] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

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

[3] A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1297–1313, 2000.

[4] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Exploring universal attribute characterization of spoken languages for spoken language recognition," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[5] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Exploiting context-dependency and acoustic resolution of universal speech attribute models in spoken language recognition," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[6] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer Speech & Language*, vol. 27, no. 1, pp. 209–227, 2013.

[7] S. Stuker, T. Schultz, F. Metze, and A. Waibel, "Multilingual articulatory features," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–I.

[8] I. P. Association, I. P. A. Staff *et al.*, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[9] F. Metze and A. Waibel, "A flexible stream architecture for asr using articulatory features," in *Seventh International Conference on Spoken Language Processing*, 2002.

[10] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3-4, pp. 303–319, 2002.

[11] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with dnn-based speech attribute modeling," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6135–6139.

[12] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *Backpropagation: Theory, Architectures and Applications*, pp. 35–61, 1995.

[13] A. Waibel, "Modular construction of time-delay neural networks for speech recognition," *Neural computation*, vol. 1, no. 1, pp. 39–46, 1989.

[14] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] 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," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[16] S. Gao, B. Xu, H. Zhang, B. Zhao, C. Li, and T. Huang, "Update progress of sinohear: advanced mandarin lvcsr system at nlpr," in *Sixth International Conference on Spoken Language Processing*, 2000.

[17] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[18] Z. Tang, D. Wang, Y. Chen, and Q. Chen, "Ap17-olr challenge: Data, plan, and baseline," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 749–753.

[19] A. McCree, G. Sell, and D. Garcia-Romero, "Augmented data training of joint acoustic/phonotactic dnn i-vectors for nist lre15," *Proc. of IEEE Odyssey*, 2016.

[20] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 105–111.

[21] N. Brummer, "Focal multi-class tookit," *https://sites.google.com/site/nikobrummer/focal*, 2014.