# ARTICULATORY FEATURES BASED DILATED LSTM MODEL FOR SHORT UTTERANCE LANGUAGE RECOGNITION

*Jiawei Yu, Minghao Guo, Jinsong Zhang*

Beijing Advanced Innovation Center for Language Resources
Beijing Language and Culture University, Beijing, China
vyujiawei@gmail.com, gmhgmh8000@163.com, jinsong.zhang@blcu.edu.cn

## ABSTRACT

The performance of Spoken Language Recognition (SLR) systems is often significantly degraded when the test utterance duration is as short as 3 seconds or less. To overcome this, we present an acoustic modeling system that the Dilated Long Short-Term Memory Networks (DLSTM) modeled Articulatory Features (AFs). The motivations of using AFs and DLSTM include utilizing the advantage of cross-language modeling of AFs and the effectiveness of DLSTM in modeling temporal dependencies in the acoustic signal for short utterance. The experiments were conducted on the AP17-OLR database. We compare the AFs based DLSTM approach and the commonly used approaches in short utterance SLR tasks in the feature and model domains. The proposed approach provides a 9.04% relative improvement to our best baseline system (deep bottleneck features based LSTM system) in terms of Equal Error Rate (EER) for 1 second utterance. Moreover, the fusion of the proposed system and baseline systems further enhanced the performance. These results indicate that the proposed approach is beneficial to the short utterance SLR task.

***Index Terms***— Spoken language recognition, short utterance, articulatory features, bottleneck features, dilated LSTM

## 1. INTRODUCTION

Significant efforts have been made to remedy the performance degradation in short utterance SLR tasks. It can be divided into two domains, namely the feature domain and the model domain.

In the feature domain, the most advanced SLR systems often use phonetic features, such as phone posteriors and deep bottleneck features (DBFs), extracted from the phonetic features recognizer [1, 2]. The performance of these phonetic features based SLR systems heavily relies on the accuracy of recognizers [3]. If the recognizers are more accurate, the SLR system will reach better performance consequentially. In general, there are two methods to extract phonetic features in the SLR tasks. The first one is to use a recognizer developed for one language like Hungarian and applied it to all language utterances [4]. However, this approach does not perform well because the recognizers developed for one language cannot accurately recognize other languages' phonetic features. The second one is to build a parallel recognizer so that each language has a recognizer separately [2], but this approach will consume a significant amount of time and computational resources.

In the model domain, since the application of Deep Neural Networks (DNN) to SLR tasks, DNN related methods have started to achieve better performance than i-vector based methods, especially in short utterance SLR tasks [5]. Then methods such as Time Delay Neural Network (TDNN) [6] and Long Short-Term Memory (LSTM) [7] based SLR systems bring further performance improvements by capturing robust sequential information from the given input features.

In this paper, we make new attempts at both feature domain and model domain to propose an Articulatory Features (AFs) based Dilated LSTM (DLSTM) model for short utterance SLR. In the feature domain, AFs are introduced to SLR tasks. 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 [8]. There are three advantages of adopting AFs: 1) The AFs can be defined universally across all languages. Therefore, adopting AFs avoids poor recognizer accuracy caused by using a single recognizer developed for one language. 2) The number of AFs is typically smaller than the phones, and one AF is usually shared by multiple phones. For example, English phonemes */m/* and */n/* are both *nasal* sounds of AFs. Therefore, more training material is available by using AFs, which means the AFs' recognizer can be trained more robustly. 3) AFs can reflect subtle differences in articulatory level between languages. For example, in Vietnamese, the *fricative* sounds */f/* and */v/* does not occur in a word's final position, but this phenomenon can happen in English, such as *beef*. So even if test utterances are short, subtle differences between languages can be captured.

One the other hand, We use the DLSTM model for the first time in a short utterance SLR task. Compared to TDNN and LSTM, which have been successfully applied to short ut-

terance SLR tasks [6, 7], DLSTM can capture longer discriminative information over the input sequence. In summary, the scheme of AFs plus DLSTM can build a robust short utterance SLR system.

## 2. PROPOSED METHOD

Our proposed short utterance SLR system diagram is shown in Figure 1. In the following part, we describe the individual components in detail.

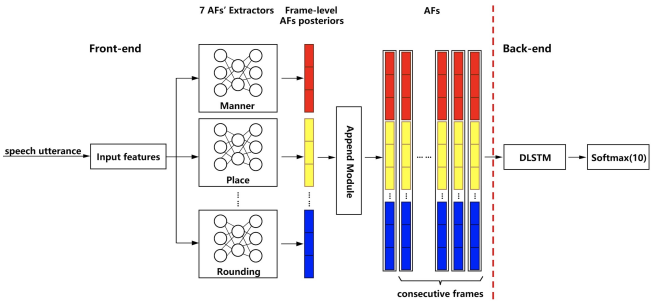


**Fig. 1**. Block diagram of articulatory features based DLSTM model for short utterance SLR system.

### 2.1. Articulatory features

AFs represent the target of the articulators in the vocal tract when pronouncing a specific phone. The identity of specific phone spoken can be linked to the combination of AFs, i.e., phones are just a short-hand for series of AFs. The AFs generally used to assist ASR systems [9], and several studies have proved that AFs can be recognized more robustly across languages than phone[10, 11].

Here we use 30 AFs which belong to 7 categories: place of articulation (PA), manner of articulation (MA), aspiration (AS), voicing (VO), tongue front-end (TF), tongue height (TH) and rounding (RO) listed in Table 1. Except above AFs, We also use the "silence" token to represent the soundless segments.

### 2.2. 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 AFs' extractors is to convert phone-based training transcriptions to AFs transcriptions [11]. This process can be achieved by using a canonically defined phone and AFs mapping table. Here we use Mandarin phone set converting AFs. The mapping table is based on [12].

Previously studies on AFs based SLR systems, the AFs' extractors were usually built using HMM or shallow neural networks [8]. Here we want to test deeper architectures. A TDNN-based AFs' extractor is separately built for each

**Table 1**. Overview of AF categories used

| Categories | Items |
|---|---|
| Manner (MA) | Stop, Fricative, Affricate, Nasal, Lateral, Approximant, Tap or Flap |
| Place (PA) | Bilabial, Labiodental, Alveolar, Dental, Retroflex, Palatal, Velar, Palatal/Front, PA-Central, Velar/Back |
| Voicing (VO) | Voiced, Unvoiced |
| Aspirated (AS) | Aspirated, Unaspirated |
| Tongue frontend (TF) | High, Middle, Low |
| Tongue height (TH) | Front, Central, Back |
| Rounding (RO) | Rounded, Unrounded |

AFs' category [13]. In TDNN, hidden layers are usually constructed by sigmoid units, and the output layer is a softmax layer. The values of the nodes can, therefore, be expressed as:

$$x^i = \begin{cases} W_1 o_t + b_1 & i = 1 \\ W_i y^{i-1} + b_i & i > 1 \end{cases} \tag{1}$$

$$y^i = \begin{cases} sigmoid(x^i) & i < L \\ softmax(x^i) & i = L \end{cases} \tag{2}$$

where $W_1$ and $W_i$ are the weight matrices, $b_1$ and $b_i$ are the bias vectors, $o_t$ is the input frame at time $t$, $L$ is the total number of the hidden layers, and both sigmoid and softmax functions are element-wise operations. The vector $x^i$ corresponds to pre-nonlinearity activations, and $y^i$ and $y^L$ are the vectors of neuron outputs at the $i^{th}$ hidden layer and the output layer, respectively. The softmax outputs were considered as an estimate of AFs posterior probabilities according to the categories of AFs that we want to model:

$$p(C_j|o_t) = y_t^L(j) = \frac{exp(x_t^L(j))}{\sum_i exp(x_t^L(i))} \tag{3}$$

where $C_j$ represents the $j^{th}$ AFs (e.g., Manner, Place, Aspirated etc.) and $y^L(j)$ is the $j^{th}$ element of $y^L$. The TDNN is trained by maximizing the log posterior probability over the training frames $x$.

Thus, the current frame posteriors are related to the possible items within that category. Subsequently, a group of the frame AFs' posteriors will be fed into the append module. The append module stacks the posterior probabilities delivered by each AFs' extractor into a supervector of AFs' detection scores, as indicated in Figure 1.

### 2.3. Dilated LSTM back-end

Dilated LSTM (DLSTM) can extend the range of temporal dependencies with fewer parameters because of its dilated recurrent skip connection and its use of exponentially increasing dilation [14].

#### 2.3.1. *Dilated recurrent skip connection*

Denote $c_t^{(l)}$ as the cell in layer $l$ at time $t$. The dilated skip connection can be represented as:

$$c_t^{(l)} = f\left(x_t^{(l)}, c_{t-s^{(l)}}^{(l)}\right) \tag{4}$$

where $s^{(l)}$ is the skip length, or dilation of layer $l$; $x_t^{(l)}$ as the input to layer $l$ at time $t$; and $f()$ denotes LSTM cell and output operations.

#### 2.3.2. *Exponentially increasing dilation*

DLSTM extract long temporal dependencies by stack dilated recurrent layers, and the dilation increases exponentially across layers. Denote $s^{(l)}$ as the dilation of the $l$-th layer. Then,

$$s^{(l)} = M^{l-1}, l = 1, \ldots\ldots, L \tag{5}$$

The Figure 2 depicts an example of DLSTM with $L = 3$ and $M = 2$. On one hand, stacking multiple dilated recurrent layers increases the model capacity. On the other hand, exponentially increasing dilation brings two benefits. First, it makes different layers focus on different temporal resolutions. Second, it reduces the average length of paths between nodes at different timestamps, which improves the ability of LSTM to extract long-term dependencies [14].

Once the AFs are extracted, they will be fed into the DLSTM. Then, the probability of a given utterance belonging to one of the languages is computed by respectively averaging the log of the softmax output of all its frames.

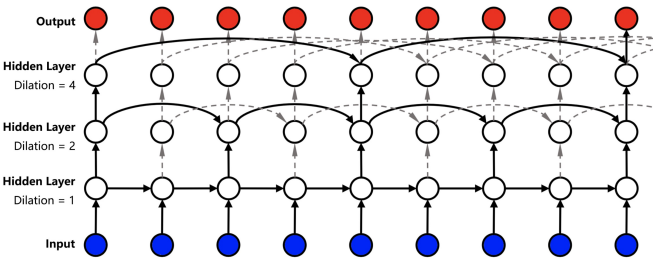


**Fig. 2**. An example of a three-layer DLSTM with dilation 1, 2, and 4. The picture is reproduced from [14].

## 3. EXPERIMENTAL SETUP

For comparison with AFs, we used MFCC, phone posteriors, and DBFs on the front-end; On the back-end, we used the classic i-vector, TDNN and LSTM methods to establish the baseline systems to compare with DLSTM. All the experiments were conducted with Kaldi toolkit [1].

### 3.1. Databases

Both phone posteriors and DBFs are extracted from the same ASR DNN trained on two mandarin corpora [15, 16]. A total of 250,000 utterances spoken by 1800 speakers (300 hours) were used for acoustic modeling.

All SLR systems evaluated on the APSIPA 2017 Oriental Language Recognition (AP17-OLR) database [17], which is used for second oriental language recognition challenge. This database consists of 10 oriental languages. 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 test condition called "test_1s" which means the duration of test utterance is 1 second, and the total number of test utterances are 22051.

### 3.2. Baseline systems

The i-vector system follows the procedure described in [18]. It is based on the GMM-UBM, and the UBM is a 2048 component full-covariance GMM. The system uses a 400-dimensional i-vector extractor and 9-dimensional Linear discriminant analysis (LDA) for scoring.

The TDNN system follows the procedure described in [6]. And The TDNN model is 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 related to the number of languages in the AP17-OLR database.

The LSTM system follows the procedure described in [7]. It has two hidden layers, and each layer contains 512 nodes. The configuration of the output layer is the same as the above TDNN system.

### 3.3. Proposed DLSTM system

The DLSTM system follows the procedure described in [14]. It has nine hidden layers and 50 nodes per layer. The configuration of the output layer is the same as the above TDNN system. The dilation increases exponentially across layers, and the max dilations are 256. Moreover, default nonlinearities and RMSProp optimizer with learning rate 0.001 and decay rate of 0.9 are used.

[1] Kaldi toolkit, http://kaldi-asr.org

**Table 2**. System performance in different methods in terms of percentage of EER and $minCavg \times 100$(reported within parenthesis) for 1 second condition

| Number | Systems | MFCC | Posteriors | DBFs | AFs | DBFs+AFs |
|---|---|---|---|---|---|---|
| 1 | $ivector + LDA$ | 18.04(18.43) | 14.78(15.03) | 13.29(14.38) | 12.95(13.47) | 12.11(12.66) |
| 2 | $TDNN$ | 14.04(13.76) | 10.75(10.32) | 9.92(9.82) | 9.32(9.02) | 8.62(8.44) |
| 3 | $LSTM$ | 13.62(13.53) | 10.22(10.08) | 9.18(9.06) | 9.04(9.00) | 8.22(8.21) |
| 4 | $DLSTM$ | 13.18(13.47) | 9.23(9.12) | 8.65(8.32) | 8.35(8.18) | **8.02(7.88)** |

## 4. RESULTS AND ANALYSIS

### 4.1. Proposed system vs baseline systems

The results in terms of the equal error rate (EER) and $minCavg$ metric are shown in Table 2.

In the feature domain, we have the following conclusions. Firstly, the performance of DBFs based systems is better than phone posteriors based systems. It indicates that the DBFs are a more compact representation of the phonetic content. Secondly, AFs based systems are better than phone posteriors and DBFs based system. These results emphasize the importance of the accuracy of the recognizers. Since AFs are more fundamental units, this makes AFs' recognizer can be trained more robustly. At the same time, AFs can reflect subtle differences in articulatory level between languages. These advantages are beneficial to improving the performance of a short utterance SLR task. Finally, when we combine DBFs and AFs (DBFs+AFs), SLR system performance can be further enhanced.

In the model domain, it can be seen that the DLSTM based systems outperform all the baseline systems. AFs based DLSTM system, which has a 7.6% relative improvement, performs better than AFs based LSTM in EER. The DBFs+AFs based DLSTM system gets the best result, and the EER is 8.02%. It reveals that a DLSTM back-end is effective in the SLR task.

### 4.2. Analysis of feature space

To better understand the ability to differentiate between languages of different features learned by the back-end model. We used $J$-measure in the test sets' feature space to assess the capacity to distinguish between languages [19]. The $J$-measure is the ratio between inter-class scatter to intra-class scatter and larger the value of $J$-measure, the better the discrimination of the classes in the feature space.

Table 3 shows the $J$-measure values. It can be seen that AFs based systems have a higher $J$-measure compared to DBFs based systems. It indicates that AFs have a better ability to distinguish between languages compared to DBFs. Besides, DLSTM based systems perform better than LSTM and i-vector systems. It shows that DLSTM is more sensitive to capture language-specific information.

**Table 3**. Comparison of J-measure on the proposed DLSTM system with the baseline systems based on DBFs and AFs.

| J-measure | DBFs | AFs |
|---|---|---|
| $ivector + LDA$ | 4.68 | 4.82 |
| $LSTM$ | 6.58 | 6.69 |
| $DLSTM$ | 7.24 | **7.86** |

### 4.3. Fusion system

The fusion results shown in Table 4. As observed, the combination of the DLSTM and i-vector (Fusion1) gets a $> 8.2\%$ gain of performance in terms of EER to our best individual DLSTM system. This fact shows that i-vector and DLSTM based systems produce uncorrelated information that can be successfully combined. Furthermore, although Fusion2 fused more systems than Fusion1, the performance of the Fusion2 achieved is only slightly better than the Fusion1. It may be because the three systems used in Fusion2 are all DNN-based, and the information they explored is homogenous. Finally, we present the fusion of all the developed systems in Fusion3. The EER of this fusion system is 6.92%.

**Table 4**. The performance of fusion systems for 1 second condition

| Name | Fusion systems | EER(minCavg $\times$ 100) |
|---|---|---|
| $Fusion1$ | $1+4$ | 7.41(7.31) |
| $Fusion2$ | $2+3+4$ | 7.24(7.21) |
| $Fusion3$ | $1+2+3+4$ | **6.92(6.84)** |

## 5. CONCLUSIONS

In this work, we have explored using AFs based DLSTM model for short utterance SLR tasks. This approach took advantage of two domains. In the feature domain, adopting AFs can avoid poor recognizer accuracy compared to phonetic-based features (phone posteriors and DBFs). In the model domain, DLSTM has a better capability of capturing long term dependencies between input features than TDNN and LSTM. The experiments' results revealed the effectiveness of the proposed approach.

## 6. REFERENCES


[1] Pavel Matejka, Le Zhang, Tim Ng, Ondrej Glembek, Jeff Z Ma, Bing Zhang, and Sri Harish Mallidi, "Neural network bottleneck features for language identification.," in *Odyssey*, 2014.

[2] Radek Fér, Pavel Matějka, František Grézl, Oldřich Plchot, and Jan Černockỳ, "Multilingual bottleneck features for language recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[3] Haizhou Li, Bin Ma, and Kong Aik Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[4] Kong Aik Lee, Chang Huai You, Ville Hautamäki, Anthony Larcher, and Haizhou Li, "Spoken language recognition in the latent topic simplex," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[5] Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez, and Pedro Moreno, "Automatic language identification using deep neural networks," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 5337–5341.

[6] Daniel Garcia-Romero and Alan McCree, "Stacked long-term tdnn for spoken language recognition.," in *INTERSPEECH*, 2016, pp. 3226–3230.

[7] Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Haşim Sak, Joaquin Gonzalez-Rodriguez, and Pedro J Moreno, "Automatic language identification using long short-term memory recurrent neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] Sabato Marco Siniscalchi, Jeremy Reed, Torbjørn Svendsen, and Chin-Hui Lee, "Exploring universal attribute characterization of spoken languages for spoken language recognition," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

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

[10] Sebastian Stuker, Florian Metze, Tanja Schultz, and Alex Waibel, "Integrating multilingual articulatory features into speech recognition," in *Eighth European Conference on Speech Communication and Technology*, 2003.

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

[12] Wei Li, Sabato Marco Siniscalchi, Nancy F Chen, and Chin-Hui 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.

[13] Vijayaditya Peddinti, Daniel Povey, and Sanjeev 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.

[14] Shiyu Chang, Yang Zhang, Wei Han, Mo Yu, Xiaoxiao Guo, Wei Tan, Xiaodong Cui, Michael Witbrock, Mark A Hasegawa-Johnson, and Thomas S Huang, "Dilated recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 77–87.

[15] Sheng Gao, Bo Xu, Hong Zhang, Bing Zhao, Chengrong Li, and Taiyi Huang, "Update progress of sinohear: advanced mandarin lvcsr system at nlpr," in *Sixth International Conference on Spoken Language Processing*, 2000.

[16] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao 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.

[17] Zhiyuan Tang, Dong Wang, Yixiang Chen, and Qing 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.

[18] Alan McCree, Gregory Sell, and Daniel Garcia-Romero, "Augmented data training of joint acoustic/phonotactic dnn i-vectors for nist lre15," *Proc. of IEEE Odyssey*, 2016.

[19] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava, "Selecting the right interestingness measure for association patterns," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 32–41.