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 APSIPA 2017 Oriental Language Recognition (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

In recent years Spoken Language Recognition (SLR) tasks have made great progress. However, the SLR systems' performance significantly decreases when dealing with short test utterances (less than 3 seconds) [1, 2]. The main problem of short utterance SLR systems is that the duration of test utterance is too short to provide reliable language-specific information. 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 critical task is to find features with better robustness. Spectral features, such as MFCC and Shifted Delta Coefficients (SDC), etc., are the most commonly used features [3, 4]. Then the most advanced SLR systems often use phonetic features, such as phone posteriors and deep bottleneck features (DBFs), extracted from the phonetic features recognizer [5, 6, 7, 8]. The performance of these phonetic features based SLR systems heavily relies on the accuracy of recognizers [9]. 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 [10]. 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 [8], but this approach will consume a significant amount of time and computational resources.

In the model domain, effective models are required to distinguish diverse languages. 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 [11, 12, 13]. Then methods such as Time Delay Neural Network (TDNN) [14] and Long Short-Term Memory (LSTM) [15] 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 improve the performance of the short utterance language recognition system 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

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

[16, 17, 18]. There are three advantages of adopting AFs:

1) The AFs are fundamental units and can be defined universally across all languages. Therefore, adopting AFs avoids the problem of poor recognizer accuracy caused by using a single recognizer developed for one language. At the same time, It avoids the problem of using a parallel recognizer that consumes more time and computational resources.

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 in duration, 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 utterance SLR tasks [14, 15], 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.

The rest of this paper is organized as follows. Section 2 presents the AFs based DLSTM for short utterance SLR system in detail. The experimental setup is presented in Section 3. After presenting the experimental results in Section 4, the conclusions are made in Section 5.

2. PROPOSED METHOD

Our proposed short utterance SLR system diagram is shown in Figure 1. The system includes two parts. The front-end is a series of feature extractors that process spoken language utterances into a sequence of AFs using the Automatic Speech Recognition (ASR) DNN model. Then a DLSTM back-end will use the AFs to model target languages. In the following part, we describe the individual components in detail.

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 [19], and several studies have proved that AFs can be recognized more robustly across languages than phone[18, 20, 21].

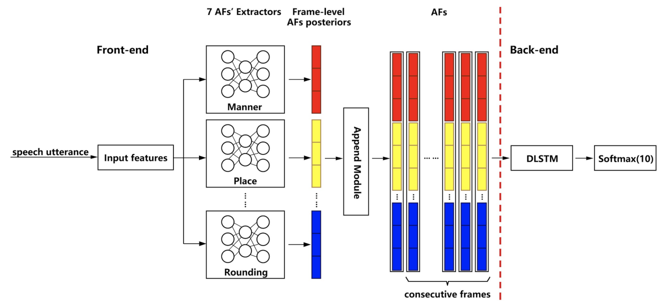

Fig. 1. Block diagram of articulatory features based DLSTM model for short utterance SLR system. Spectral features are fed into the time-delay neural network (TDNN), and the TDNN will generate the articulatory posterior probabilities. Then the append module will merge all posterior probabilities to form the final articulatory features. The back-end is the DLSTM that will model the target languages and make the final decision.

In the phonetic features such as phone posteriors and DBFs based SLR approaches, the accuracy of recognizers of phonetic features is a critical factor [9]. Specifically, If the phone set of a language to be recognized differs from the phone used to train the phone recognizer, then the accuracy of the phone recognizer will be reduced. This phenomenon is quite common for spoken languages in different language families. As discussed in the introduction, AFs can be obtained from a particular language and shared across other languages. So AFs can be used to derive a universal set of speech units. In this study, we use AFs as a front-end feature to obtain a more accurate recognizer and then improve the SLR systems' performance.

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 [21]. 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 [22].

In this study, we used the posterior probabilities of the articulatory categories as the articulatory features. As shown in Figure 1, the feature extraction module consists of a bank of AFs' extractors. Since we have 7 AFs' categories (shown in Table 1), correspondingly we have 7 AFs' extractors. 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' extractor is used to extract one specific articulatory feature.

Previously studies on AFs based SLR systems, the AFs' extractors were usually built using HMM or shallow neural networks [17, 23]. Here we want to test deeper architectures. A TDNN-based AFs' extractor is separately built for each AFs' category [24, 25]. 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 linked 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 [26].

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 [27] 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 [26].

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.

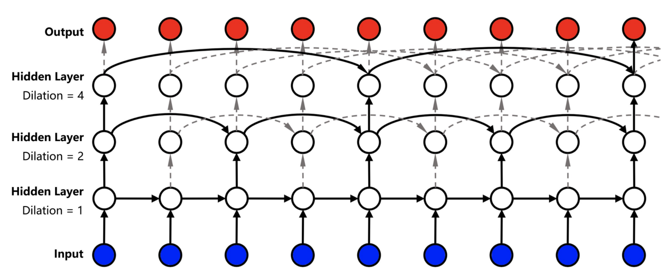

Fig. 2. An example of a three-layer DLSTM with dilation 1, 2, and 4. The picture is reproduced from [26].

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. The first one is from the Chinese National Hi-Tech Project 863 for Mandarin LVCSR system development[28], and the second corpus is an open-source Mandarin speech corpus called AISHELL-1[29]. A total of 250,000 utterances spoken by 1800 speakers (300 hours) were used for acoustic modeling.

All SLR systems evaluated on the AP17-OLR database [30], 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. Features extraction

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

The phone posteriors and DBFs are extracted from an ASR acoustic model. The training used AISHELL-1 and 863 mandarin corpus described in section 3.1. The model is a TDNN with six hidden layers, and each layer contains 650 nodes. The output layer contains 463 units, corresponding to the number of GMM senones. Its input features are 40-dimensional MFCC. Once the model was trained, 463-dimensional phone posteriors were derived from the output layer. For the extraction of DBFs, the configuration of TDNN is the same as above, except that the number of nodes in the last hidden layer is 100. And they are Excluding the softmax output layer, which is not needed to compute DBFs.

The AFs are also extracted from an ASR acoustic model. The database, the configuration of the TDNN, and the input features are the same as above, except that the number of nodes in the output layer of TDNN is related to the number of items (shown in Table 1) in the categories of AFs.

3.3. Baseline systems

The i-vector system follows the procedure described in [31]. 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 150-dimensional Linear discriminant analysis (LDA) for scoring [32]. The input features to the i-vector system separately incorporate mentioned above four features: MFCC, phone posteriors, DBFs, and AFs.

The TDNN system follows the procedure described in [14]. 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 [15]. 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. Activations are forward propagated for a fixed step time of 20 over a subsequence of an input utterance, the cross entropy gradients are computed for this subsequence and backpropagated to its start.

3.4. Proposed DLSTM system

The DLSTM system follows the procedure described in [26]. 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. The standard normal distribution initializes all weight matrices. The batch size is set to 128.

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. The best performance is marked in boldface.

In the feature domain, we have the following conclusions.

- Firstly, phone posteriors, DBFs, and AFs based systems achieve lower EER than spectral features MFCC based system. And 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

[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)**

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 [33]. 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

In this section, we aim to analyze the score correlation among DLSTM based system and baseline systems and, in particular, how that can lead to a good combination strategy. For this purpose, we defined three different groups of systems and combined them using the multi-class logistic regression framework [34]. The features we used are DBFs+AFs because it achieved the best performance in the above experiments. This fusion and calibration procedure was conducted through the FoCal toolkit [35].

The fusion results shown in Table 4. As observed, the combination of the DLSTM and i-vector based system (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 compared to 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 different fusion systems in terms of percentage of EER and $minCavg \times 100$(reported within parenthesis) 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 the problem of poor recognizer accuracy by using phonetic-based features (phone posteriors and DBFs). And using AFs can capture subtle differences in articulatory level between languages. In the model domain, DLSTM has a better capability of capturing long term dependencies between input features than TDNN and LSTM for short utterances. The experiments were performed on the AP17-OLR database revealed the effectiveness of the proposed approach. Specifically, The experimental results show that our proposed approach provides a 37.17% and 3.47% relative improvement in terms of EER to DBFs based i-vector and DLSTM approaches. Also, we evaluated the fusion system. The fusion of AFs based DLSTM system with different baseline systems got significant performance gain for SLR tasks. These results are strong support that the AFs based DLSTM approach is beneficial to the short utterance SLR system.

6. REFERENCES

[1] Ruchir Travadi, Maarten Van Segbroeck, and Shrikanth S Narayanan, "Modified-prior i-vector estimation for language identification of short duration utterances," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[2] Meng Ge Wang, Yan Song, Bing Jiang, Li Rong Dai, and I. Mcloulghlin, "Exemplar based language recognition method for short-duration speech segments," in *IEEE International Conference on Acoustics, Speech & Signal Processing*, 2013.

[3] Pedro A Torres-Carrasquillo, Elliot Singer, Mary A Kohler, Richard J Greene, Douglas A Reynolds, and John R Deller Jr, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Seventh international conference on spoken language processing*, 2002.

[4] Jerry Foil, "Language identification using noisy speech," in *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1986, vol. 11, pp. 861–864.

[5] Yan Song, Bing Jiang, YeBo Bao, Si Wei, and Li-Rong Dai, "I-vector representation based on bottleneck features for language identification," *Electronics Letters*, vol. 49, no. 24, pp. 1569–1570, 2013.

[6] 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.

[7] Fred Richardson, Douglas Reynolds, and Najim Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE signal processing letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Alicia Lozano-Diez, Javier Gonzalez-Dominguez, Ruben Zazo, Daniel Ramos, and Joaquin Gonzalez-Rodriguez, "On the use of convolutional neural networks in pairwise language recognition," in *Advances in Speech and Language Technologies for Iberian Languages*, pp. 79–88. Springer, 2014.

[13] Alicia Lozano-Diez, Ruben Zazo-Candil, Javier Gonzalez-Dominguez, Doroteo T Toledano, and Joaquin Gonzalez-Rodriguez, "An end-to-end approach to language identification in short utterances using convolutional neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

[15] 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.

[16] Tanja Schultz and Katrin Kirchhoff, *Multilingual speech processing*, Elsevier, 2006.

[17] 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.

[18] Sebastian Stuker, Tanja Schultz, Florian Metze, and Alex Waibel, "Multilingual articulatory features,"

in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*. IEEE, 2003, vol. 1, pp. I–I.

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

[20] 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.

[21] 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.

[22] 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.

[23] Hamid Behravan, Ville Hautama, Sabato Marco Siniscalchi, Tomi Kinnunen, Chin-Hui Lee, et al., "Introducing attribute features to foreign accent recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5332–5336.

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

[25] 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.

[26] 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.

[27] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] Najim Dehak, Pedro A Torres-Carrasquillo, Douglas Reynolds, and Reda Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Twelfth annual conference of the international speech communication association*, 2011.

[33] 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.

[34] Niko Brummer and David A Van Leeuwen, "On calibration of language recognition scores," in *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*. IEEE, 2006, pp. 1–8.

[35] N. Brummer, "Focal multi-class toolkit," `https://sites.google.com/site/nikobrummer/focal`, 2014.