This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2879361, IEEE Transactions on Vehicular Technology



# Short Utterance based Speech Language Identification in Intelligent Vehicles with Time-scale Modifications and Deep Bottleneck Features

Zhanyu Ma, *Senior Member, IEEE,* Hong Yu, *Member, IEEE,* Wei Chen, and Jun Guo

***Abstract*—Conversations in the intelligent vehicles are usually short utterance. As the durations of the short utterances are small (*e.g.*, less than three seconds), it is difficult to learn sufficient information to distinguish the type of languages. In this paper, we propose an end-to-end short utterances-based speech language identification (SLI) approach, which is especially suitable for the short utterance based language identification. This approach is implemented with a long short term memory (LSTM) neural network, which is designed for the SLI application in intelligent vehicles. The features used for LSTM learning are generated by a transfer learning method. The bottle-neck features of a deep neural network (DNN) which are obtained for a mandarin acoustic-phonetic classifier are used for the LSTM training. In order to improve the SLD accuracy with short utterances, a phase vocoder based time-scale modification (TSM) method is utilized to reduce/increase the speech rate of the test utterance. By connecting the normal, speech rate reduced, and speech rate increased utterances, we can extend the length of the test utterances such that the performance of the SLI system is improved. The experimental results on the AP17-OLR database demonstrate that the proposed method can improve the performance of SLD, especially on short utterance. The proposed SLI has robust performance under the vehicular noisy environment.**

***Index Terms*—Speech language identification, time-scale modification, DNN-BN feature, LSTM**

## I. Introduction

One of the challenges in the real-life applications of smart vehicles is how to design a convenient and effective interactive method between drivers and vehicles. Using the voices to transfer information to smart vehicles is an effective solution for assistant and cooperative driving. Intelligent Vehicles need to install a mixed lingual intelligent speech recognition (SR) system to understand vocal commands [1], [2].

As the front-end of SR system, speech language identification (SLI) technology is needed to recognized language of the input speech utterance, firstly, and then the SR system can call the corresponding decoder to translate the input speech utterance into correct command text [3], [4].

In practical applications, the duration of vocal commands are usually very short. So to develop a accurate and effective SLI method which is suitable for short utterances can improve the performance of mixed lingual SR systems and enhance comprehension between the drivers and the intelligent vehicles.

There exist many languages in the world and each of these languages has different distinguishing features. Many researches have dedicated to work on developing a universal, quick responsive, and effective SLI system [5]. Generally speaking, researches on SLI mainly focus on two domains, namely the feature domain and the classifier domain. In the feature domain [6], [7], the key task is to find features which can express the difference between languages. Then in the classifier domain, effective classification schemes are required to distinguish diverse languages.

In the feature domain [8], raw acoustic features, *e.g.*, linear predictive coding (LPC), filter bank feature, and formation features are mainly considered [9]. Then, the performance of the dynamic features, which include temporal information, are also investigated [10]. The prosody information, such as the patterns of duration, pitch, and stress of languages, is usually considered as the additional knowledge to improve the performance of the raw acoustic features [11], [12]. Token based features, such as phone, syllables, and words sequences which contain hight-level character information, are also utilized to realize the SLI system [13], [14].

In the classifier domain, when using the acoustic or the prosody features as front-ends, strong statistical models are usually selected to build the SLI system. In [15], [16], [17], [18], different languages are modeled by Gaussian mixture models (GMMs) and the log-likelihood ratios are used to make the languages identification decision. In addition, Hidden Markov models (HMMs) trained by speaker and text independent acoustic feature sequences of different languages are also applied to construct a SLI system [19], [20]. Neural networks and support vector machines are also implemented as the back-end to classify different speech languages [21], [22], [23]. The i-vector based method, which have been successfully used in speaker verification tasks, are also used to express the distinguishable features of different languages, followed by a task-oriented probabilistic linear discriminant analysis (PLDA) scoring method. The i-vector-PLDA model achieved significant success in SLI tasks[24], [25]. When selecting the token based features, *e.g.*, phone sequences, as front-ends, n-gram language models (LM) are needed as back-end to evaluate the confidence value between the input speeches and the target language, which is called phone recognition and language modelling (PRLM) [5]. Different variants of

Z. Ma, and J. Guo are with Pattern Recognition and Intelligent System Lab., Beijing University of Posts and Telecommunications, Beijing, China.

H. Yu is with School of Information and Electrical Engineering, Ludong University, Yantai, China.

W. Chen is with Beijing Sogou Technology Development Co., Ltd., China.

The corresponding author is H. Yu (hy@ldu.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2879361, IEEE Transactions on Vehicular Technology



PRLM method based on parallel phone recognition and phone selection on multilingual phone set have been discussed in [2], [26], [27].

Recently, with the development of deep learning [28], [29], [30], many deep neural networks (DNN) based solutions are also involved into SLI tasks. In [31], a fully connected feed-forward neural network trained by 21 frames stacking perceptual linear prediction(PLP) features were used to classify the languages directly. In [32], [33], a convolutional neural network trained by Mel-frequency cepstral coefficient(MFCC) and PLP feature maps was applied to build a language classifiers. In [34], [35], [36], [37], a DNN has been trained to generate frame-level bottleneck (BN) features and these features are used to train an i-vector based SLI systems. Segment-level X-vectors which were built by mean and standard deviation of BN features are also applied in language identification [38]. Recurrent neural networks (RNN) which can model the temporal information of features are also widely used in SLI takes. In [39] and [40], the long Short-Term Memory(LSTM) and the bidirectional LSTM (BLSTM) neural networks are trained to recognize different languages.

Many published results show that the DNN based SLI methods perform better than the conventional statistic models based methods, such as GMMs and i-vector,especially on the short utterances which can not supply sufficient statistical information. However, when the length of input utterance is shorter than three seconds, even the performance of the DNN based SLI systems will decrease significantly [41], [4], [42]. In the verbal system of the intelligent vehicles, most of the conversations are short utterances less than three second [43]. Hence, it is very important to build a SLI system that is suitable for very short utterances, especially in the condition of conversations in intelligent vehicles.

The main problem of SLI on short utterance is the inadequate information extracted from input speeches. In order to overcome this, we build an end-to-end SLI system based on the transfer learning features and the time-scale modification (TSM) method. The structure of the proposed SLI system is shown in Fig. 1. We use a TSM method to expend the length of short input utterances. The speech rate of test short utterances are adjusted by the phase vocoder method. By splicing the original speech with the rate increased speech and the rate decreased speech, the lack of information in the short utterance can be overcome. Next, the PLP features concatenating with pitch features generated by the length extended speeches are used to train the language classifiers. As many researches shows that the DNN-BN feature generated by a phonetic classifier have more information than raw acoustic features, we use the PLP+pitch features to train an mandarin phoneme classifier firstly and the pre trained DNN is used to extract the DNN-BN features which including more useful information. For the purpose of adapting to the short duration speeches, the generated DNN-BN features frames are packed into small blocks with 100 frames. For the purpose of fitting short input utterance with frame length less than 100, we use repeatedly padding method to fill the gaps. In the next step, feature blocks are feeded into two layer LSTMs, which are suitable for modeling feature sequences. Because the output of the last frame contents information of the whole block, a softmax layer is connected with the last frame of the LSTM outputs to realize the language classification task.

In the following sections, we first introduce the TSM method used for the short utterance length extending in Section II. The structure of the neural networks used for the DNN-BN features extracting and the language classification are described in Section III. In Section IV, we introduce the experimental configurations including the database used for evaluating the SLI models and the parameter settings of the SLI models. In this section, we also make intensive comparisons between the proposed TSM-DNN-BN-LSTM SLI model and two baseline systems, with clean and vehicular noisy speeches. Finally, we draw the conclusions in Section V.

## II. Time-scale Modification

Many published results show that the accuracy of SLI systems will decreased heavily with short input utterances (*e.g.*, less than three seconds). In order to overcome this problem, in this paper, we apply a TSM technology to adjust the speech rate of the input utterance. By concatenating the speeches with different rates, we can extend the length of input utterances and enhance the information content of the test speeches. Therefore, the performance of the SLI system can be improved.

The TSM technology can adjust the speech rates by changing the length of speeches. In this section we will introduce a classic TSM method, namely the phase vocoder, which can modify the rate of the input speech without damage on the pitch and the prosody information. Speech rate usually refers to the speed of pronunciation. Irregular speech rates will decrease the accuracy of continuous SR systems [44], [45]. In a speaker verification (SV) system, the mismatch of speech rates between the enrollment and the verification speeches will degrade the performance of SV system as well [46]. During the acoustic features extraction, the speed rate information will be included into the extracted features, inevitably. Hence, in the SR and the SV tasks, we need to restrain these mismatches [47], [48]. Moreover, in the SLI task, the abundant combinations of the speech with different rates will improve the performance of the SLI system [49].

The TSM method can modify the speech rate without changing the spectral information, *e.g.*, the fundamental frequency and the formant. Recently, many different TSM methods have been proposed and in this paper we select the phase vocoder method [50]. As shown in Fig 2, this method contains three steps to modify the speech rate. Firstly, the input utterance is segmented into frames with duration $L$ and step size $S_i$. For each frame, a Hann window is applied to reduce the high frequency components. The short-time Fourier transform (STFT) is applied on each frame and the time-frequency information $X(\lambda, k)$ can be generated as

$$X(\lambda, k) = \sum_{n=0}^{L-1} x(\lambda S_i + n) h(n) e^{-j(2\pi kn/N)}, \qquad (1)$$

where $x$ stands for the input speech, $h$ is the window function, $\lambda$ is the frame index, $k$ means the frequency bin index and $N$ is the point number in the discrete Fourier transform.

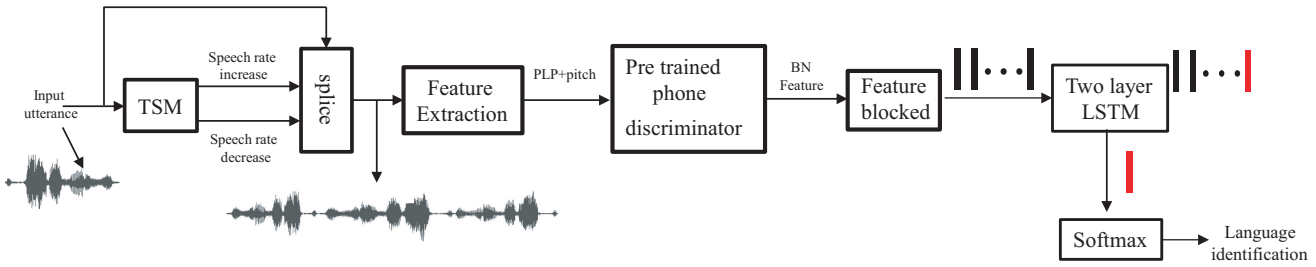



Fig. 1. The structure of the proposed SLI system.

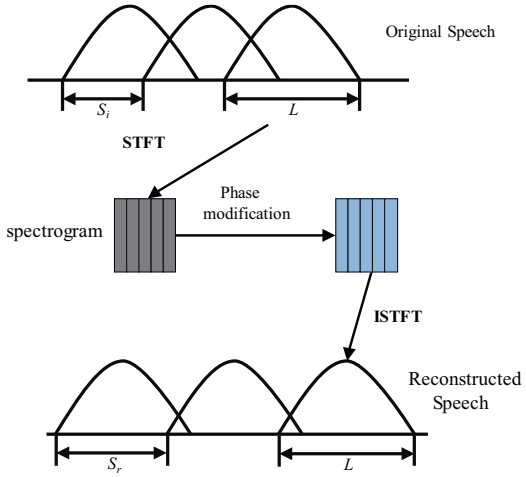



Fig. 2. The processing flow of the phase vocoder method.

Secondly, the amplitude $|X(\lambda,k)|$ and the phase $\theta(\lambda,k)$ of $X(\lambda,k)$ are calculated. Then we use the idea introduced in paper [51] to modify the phase into $\theta^{'}(\lambda,k)$.

Finally, the inverse short-time Fourier transform (ISTFT) is used to reconstruct the time-domain frame $y(\lambda)$ with the new phase information as

$$y(\lambda) = \text{ISTFT}\left(|X(\lambda,k)|e^{j\theta'(\lambda,k)}\right). \tag{2}$$

As shown in Fig. 2, by summing the reconstructed frames with $S_r$ as step size, we can modify the length of input speeches.

When the step size $S_r$ in the reconstruction procedure is shorter than the original step size $S_i$ in the frame segmentation procedure, the speech rates of new generated speeches is increased. On the contrary, we can reduce the speech rate of the input speech by setting $S_r$ longer than $S_i$ . Therefore, we can define the modification ratio of the original speech rate as

$$\alpha = \frac{S_i}{S_r} \tag{3}$$

and the length of the reconstructed speech $\widetilde{y}$ as

$$\text{length}(\widetilde{y}) = \frac{\text{length}(x)}{\alpha}. \tag{4}$$

Fig. 3 and Fig. 4 illustrate the waveform and spectrogram of original and time-scale modified signals where $\alpha$ is chosen as $0.8$ and $1.2$, respectively. The frame duration $L$ is set as $2048(128ms)$ and the discrete Fourier transform number $N$ is set as $2048$ as well. The step size of the reconstructed speech, $S_r$, is chosen as $512(32ms)$. From Fig. 4 we can observe that, when ignoring frame numbers difference before and after TSM and aligning the three spectrograms to the same size,

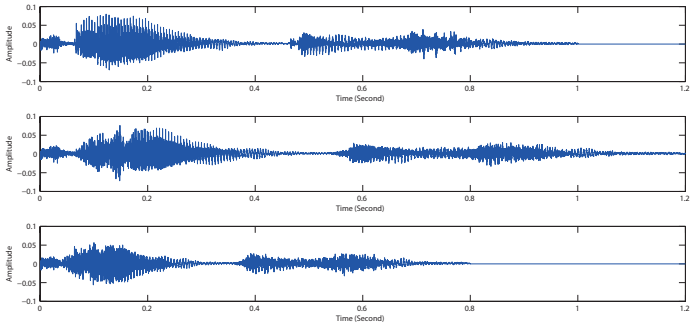



Fig. 3. From top to bottom: waveform of original, speech rate decreased ($\alpha = 0.8$) and speech rate increased signals ($\alpha = 1.2$).

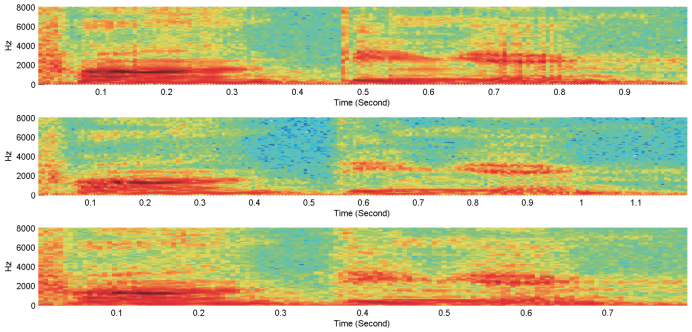



Fig. 4. From top to bottom: spectrogram of original, speech rate decreased ($\alpha = 0.8$) and speech rate increased signals ($\alpha = 1.2$).

the three aligned spectrograms are very similar. It indicates that using TSM method to extend or shorten the length of the speech will not yield the loss of the frequency domain information. Moreover, the rate modified speeches generated by the aforementioned TSM method have less distortion, which can supply more useful information. This modification is helpful to the SLI tasks.

## III. Neural Network based SLI model

In order to make the proposed SLI model suitable for short utterances, we use the TSM method to extend the length of input signals to increase the information which is helpful to language recognition in the front-end. At the meantime, in the back-end, we design a DNN based module to generate more meaningful feature to improve the accuracy of SLI.

### A. DNN-BN feature extractor

In the early stage of SLI realization, researches tended to train statistical models to present the underling information of different languages. Specially, i-vector based models trained only by raw acoustic features archived good performance [52]. The performance of the statistical model based method is

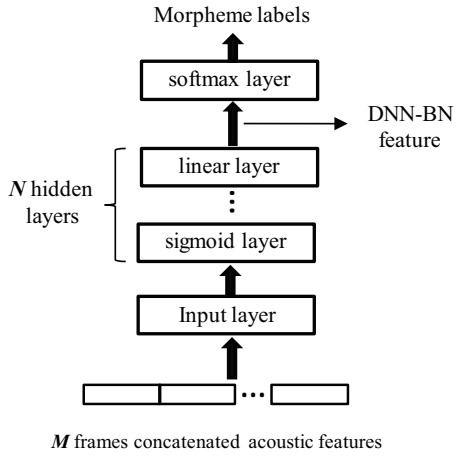



Fig. 5. The structure of the DNN-BN feature extractor.

closely related to the frame length of the test utterances. In short duration condition, limited amount of test feature frames cannot supply sufficient statistical information to the SLI models, which will severely affect the accuracy of language identification.

In order to make the trained model adaptive and robust to the short utterances, we need to adopt some other features to include more language discriminative information rather than the raw acoustic features only. Recall that the theoretical foundation of the PRLM model is that the languages are discriminated by their phonetic properties. This encourages us to build a feature extractor that can extract phonetic information. As shown in Fig. 5, we build a DNN using morpheme labels as the training target to extract the DNN-BN features.

In order to consider the temporal information, we use $M$ concatenated acoustic feature frames as the input. The DNN contains $L$ hidden layers, the activation functions of the bottom $L-1$ layers are all sigmoid functions and the top hidden layer is a linear layer which is connected with a softmax output layer. Since the outputs of the top hidden layer are closest to the morpheme classification outputs and they include abundant morpheme discriminative information, we use these outputs as the DNN-BN features to train a language classifiers. Comparing with unit-level token features, which also include phonetic information, the extracted frame-level DNN-BN features have higher temporal resolutions and are more suitable for short utterance SLI task.

In some papers, researchers tend to use the language label as the trained target to train the DNN classifiers [31]. Moreover, they use a DNN to learn the language discriminative information directly, while the language label is too coarse to supply enough supervisory information. Unlike language label, the morpheme labels can provide strong supervision information and lead the DNN to learn more useful discriminative information layer by layer. In addition, the morpheme information also has close correlation with the language identification task.

From another point of view, using the morpheme label to generate feature for language identification task can be considered as a variant of transfer learning, where a related task is used to pre-train a model for other tasks. In our work, the DNN-BN feature extractor supervised by the morpheme labels also has the benefit of cross-language learning, which means that the DNN trained by one language can also learn

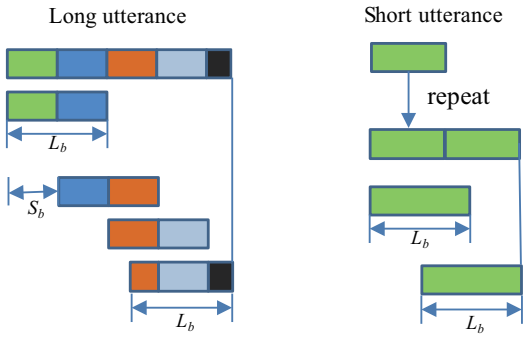



Fig. 6. Feature sequences partitioning method.

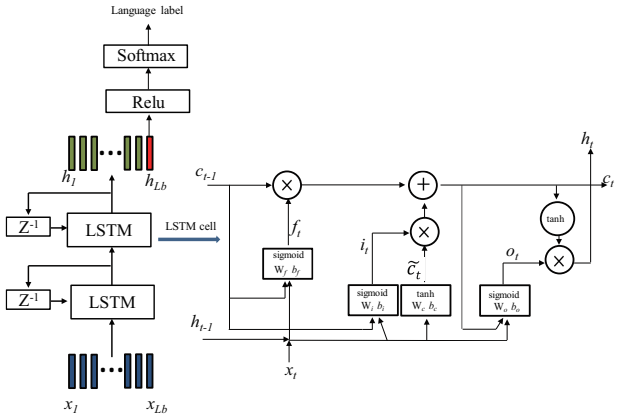



Fig. 7. Structure of the LSTM based language classifier.

the features to recognize another language. This is especially important for the the uncommon language recognition task, when we do not have enough training data.

### B. Feature Block Generation

In order to make the proposed SLI model suitable for short utterances, we pack the DNN-BN features into short-time blocks. In the test phase, the input utterances are recognized at the block-level and the average log-likelihood value of all blocks is used as the speech identification score.

The aforementioned packaging method is shown in Fig. 6. A DNN-BN feature sequence is segmented into blocks with block size $L_b$ and step size $S_b$, respectively. For the utterances with frame number larger than $L_b$, every $L_b$ frames are packed into a lock. If the remaining number of frames is smaller than $L_b$ , the last $L_b$ frames are packaged as a new block. For utterances with framed number less than $L_b$, repeating method is used to increase the frame number of the input features and then the extended features are packaged following the method used for partitioning the long utterances.

In the training step, packaging the extracted features into partial overlapped small blocks can make the utmost usage of the limited training data and adapt the trained model to the short utterance. When testing the system with short utterances, the feature repeating method can supply more effective information to trained a robust language classifiers.

### C. LSTM-based Language Classifier

The LSTM is a special type of RNN, which can learn temporal correlations. Moreover, the memory blocks and gates in a LSTM cells can avoid the long-term dependency problem [53].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2879361, IEEE Transactions on Vehicular Technology



TABLE I
DESCRIPTION OF THE AP17-OLR DATABASE.

| | Train/dev | | Test | |
|---|---|---|---|---|
| Language. | Speaker | Total utt. | Speaker | Total utt. |
| ka-cn | 86 | 4200 | 86 | 1800 |
| ti-cn | 34 | 11100 | 34 | 1800 |
| uy-id | 353 | 5800 | 353 | 1800 |
| ct-cn | 24 | 7559 | 6 | 1800 |
| zh-cn | 24 | 7198 | 6 | 1800 |
| id-id | 24 | 7671 | 6 | 1800 |
| ja-jp | 24 | 7662 | 6 | 1800 |
| ru-ru | 24 | 7109 | 6 | 1800 |
| ko-kr | 24 | 7196 | 6 | 1800 |
| vi-vn | 24 | 7200 | 6 | 1800 |

LSTMs are suitable for modeling feature sequences and have good performance on SLI tasks [54], [39]. In this paper, we train a LSTM model with two layers to carry out the language identification task.

As shown in Fig. 7, the language classifier is trained by DNN-BN feature blocks with $L_b$ frames. The LSTM based model can build a mapping from input feature sequences $(x_1, ..x_{L_b})$ to hidden layer outputs $(h_1, ..h_{L_b})$. Because $h_{L_b}$, the last output frame of the top hidden layer, is generated by all input features, it can present the information of the whole input feature block. We feed $h_{L_b}$ to a full connect layer with rectified linear units (Relu) as the activation function and use a softmax layer to classify different languages.

Regarding the LSTM cell, a popular "peephole connections"[55] structure is selected. As shown in Fig. 7, the square icons stand for neural network layers and circular icons denote point-wise operation. The associated computations are given as fellows

$$f_t = sigmoid\left(W_f \cdot [c_{t-1}, h_{t_1}, x_t] + b_f\right), \tag{5}$$

$$i_t = sigmoid\left(W_i \cdot [c_{t-1}, h_{t_1}, x_t] + b_i\right), \tag{6}$$

$$\tilde{c}_t = tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right), \tag{7}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t, \tag{8}$$

$$o_t = sigmoid\left(W_o \cdot [c_t, h_{t-1}, x_t] + b_o\right), \tag{9}$$

$$h_t = tanh(c_t) * o_t, \tag{10}$$

where"$\cdot$" means matrix multiplication and "$*$" means points-wise multiplication.

## IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

### A. Database

The proposed SLI model is evaluated on the AP17-OLR databases which are used for the second oriental language recognition challenge [56]. The database is originally created by SpeechOcean and Multilingual Minorlingual Automatic Speech Recognition (M2ASR). In the databases, there are 10 languages in total, which includes Kazakh in China (ka-cn) , Tibetan in China (ti-cn), Uyghur in China (uy-id), Cantonese in China Mainland and Hongkong (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 dataset is divide into a train/dev part and a test part. The details about speaker number and total utterances of each language are listed in Tab. I. The male and female speakers and the utterances of each speaker are balanced. Speakers in train/dev and test subsets have no overlap.

All the utterances were recorded by mobile phones, with a sampling rate of 16kHz and a sample size of 16 bits. In the train/dev subset, each language has about 10 hours recordings for SLI model training. In order to investigate the performance of trained SLI model on short duration signals, the AP17-OLR database also supplies short duration subsets, which includes train-1s, train-3s, dev-1s, dev-3s, test-1s, and test-3s. These subsets were randomly segmented from the train/dev and the test sets, respectively.

### B. Experimental Configurations

In the front-end, the acoustic and prosody features are used for SLI model training. The input utterances are segmented into frames with 25ms length and 10ms step size. For each frame, 150 dimensional PLP coefficients (50 + $\Delta$ + $\Delta\Delta$ ) concatenating with 3 dimensional pitch features are extracted. A global mean and variance vectors are use to normalize the extracted features. A pre-trained DNN based voice activity detector (VAD) is used to remove the silence frames.

All the neural networks are trained by Tensorflow [57]. In the DNN-BN feature extractor training phase, as shown in Fig. 5, 11 frames concatenating the PLP+pitch feature are used. The phoneme discrimination DNN contains five hidden layers and the nodes number of each hidden layer are all set as 512, which means the dimension of DNN-BN features generated by the linear outputs of the top hidden layer is also 512. About 500 hours Mandarin speech collected from Sogou speech input method platform are used to train the phoneme discriminator. The input features are tagged to $6,294$ triphone labels by a trained acoustic model.

The DNN based phoneme classifier is trained by $1,683$ dimensional ($153\times11 = 1,683$ ) features and $6,294$ dimensional target labels, using cross entropy as the cost function and stochastic gradient descent (SGD) as the optimization method. The learning rate is set as $0.001$ empirically, the training epoch is set as 50, and the min-batch size is 256.

After the phoneme classifier training, the trained DNN is used as a feature extractor to generate the DNN-BN features. As described in Section III-B, the produced features with 512 dimensions are segmented into blocks with block size $L_b = 100$ (about one second) and step size $S_b = 50$.

As illustrated in Fig. 7, the packaged short-time feature blocks were send into a language classifier with two LSTM layers. The output nodes number of two LSTM layers were all set as 512. The nodes number of the Relu layer was set as $1,024$. The dimension of the softmax layer was 10, which stood for the number of languages to be identified.

In the language classifier, we also selected the cross entropy as the cost function and an Adam optimizer was used to update the parameters in the language classifier. The learning rate was chosen as $0.0002$ and the training epoch was 50. During training, only the parameters in the language classifier parts were updated while the parameters in the DNN-BN feature extractor part were fixed.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2879361, IEEE Transactions on Vehicular Technology



TABLE II
COMPARISON OF DIFFERENT SLI MODELS ON $C_{\text{AVG}}$ AND EER (IN %).

| Models | test-all | | test-3s | | test-1s | |
|---|---|---|---|---|---|---|
| | $C_{\text{avg}}$ | EER | $C_{\text{avg}}$ | EER | $C_{\text{avg}}$ | EER |
| i-Vector | 0.063 | 6.94 | 0.075 | 8.67 | 0.189 | 17.24 |
| LSTM | 0.092 | 9.64 | 0.121 | 11.05 | 0.136 | 14.3 |
| DNN-BN-LSTM | 0.017 | 1.92 | 0.024 | 2.55 | 0.128 | 13.26 |
| TSM-DNN-BN-LSTM | **0.007** | **0.86** | **0.011** | **1.14** | **0.067** | **6.95** |

### C. Baseline Systems

We built two baseline SLI systems base on the i-vector model and the LSTM model trained by the PLP + pitch features with 153 dimensions.

In the i-vector model, the universal background model (UBM) with 2048 Gaussian mixtures were trained by the utterances in AP17-OLR database. The dimension of the i-vectors was set as 400. The mean i-vector of one language in the train/dev subset can be used to represent that language. The score of a test utterance on a particular language can be computed by the cosine distance between the i-vector of the test speech and the language i-vector generated from the train/dev subset.

The structure of the baseline LSTM model is similar as the language classifier described in Section III-C. Instead of using the DNN-BN features, the packaged raw PLP + pitch feature blocks with 100 frame length and 50 frames step size were used for model training. The mean log-likelihood, computed by the outputs of the softmax layer, was used as the language identification scores.

### D. Experimental Results

Following the recommendation introduces in 2015 NIST language recognition evaluation plane, the performances of different SLI systems were evaluated by $C_{\text{avg}}$ and the equal error rate (EER). The pair-wise loss that composes the missing and false alarm probabilities for a particular target/non-target language pair is defines as

$$C(L_t, L_n) = P_{\text{Target}} P_{\text{Miss}}(L_t) + (1 - P_{\text{Target}}) P_{\text{FA}}(L_t, L_n), \quad (11)$$

where $L_t$ and $L_n$ are the target and non-target languages, respectively; $P_{\text{Miss}}$ and $P_{\text{FA}}$ are the missing and false alarm probabilities, respectively. $P_{\text{Target}}$ is the prior probability for the target language, which is set to 0.5 in the evaluation. $C_{avg}$ is defined as the average of the above pair-wise performance as

$$\begin{aligned} C_{\text{avg}} = \frac{1}{N}\{[P_{\text{Target}} \cdot \sum_{L_t} P_{\text{Miss}}(L_t)] \\ + \frac{1}{N(N-1)}[(1 - P_{\text{Target}}) \cdot \sum_{L_T} \sum_{L_N} P_{\text{FA}}(L_t, L_n)]\}, \end{aligned} \quad (12)$$

where $N$ is the number of languages.

The performances of different SLI models were evaluated on full test subset, test-all and two short duration subset, test-3s and test-1s. The utterances level $C_{\text{avg}}$ and the EER of different models are shown in Tab. II.

Firstly, we compare the performance of two baseline systems. It can be observed that on the full data set and test-3s data set, the statistical i-vector model performs slightly better than the neural network based LSTM model. However, in the very short duration data set (test-1s), as the test utterance can not provide sufficient statistical information, the LSTM model performs better than the i-vector model.

Secondly, if we changed the training feature from the raw acoustic features to the DNN-BN features (DNN-BN-LSTM model), the performance of SLI has been significantly improved. It indicates that the phoneme distinguishing features are very useful for the SLI task. When we have sufficient amount of training data, more complex targets labels can help the neural networks to learn features with richer information.

Then, as described in SectionII, the phase vocoder based TSM method was applied to extend the length of test utterances (TSM-DNN-BN-LSTM model). Here, we set the speech rate changing parameter $\alpha$ as 0.8 and 1.2, respectively. This means the original speech are concatenate with the speed increased speech and the speed decreased speech.

From the results in Tab. II, we can observe that, the TSM based length expending method can improve the accuracy of speech identification. Without changing the parameters of the trained neural networks and just by simple preprocessing of input waveform signals, the error rate of SLI system can decrease about 50% on the long and short duration data set.

In order to investigate the effect of the speech rate modification to the SLI accuracy, we tried different speed rate modification combinations and evaluated their performances on the test-1s data set. From the results in Tab. III it can be observed that, concatenating some speed modified utterances together can improve the accuracy of SLI, comparing with the original short signals. Splicing the original speech with a speed increased and a speed rate decreased signal can make the performance better than splicing two speech rate increased speech or two speech rate decreased speech only. Moreover, the speech rate modification should be moderate, too large range of speech rate modification will decrease the SLI accuracy.

TABLE III
COMPARISON OF SPEED RATE MODIFICATION COMBINATIONS ON TEST-1 DATA SET.

| $(\alpha_1, \alpha_2)$ | (0.8,1.2) | (1.1,1.2) | (0.8,0.9) | (0.7,1.3) |
|---|---|---|---|---|
| $C_{\text{avg}}$ | **0.067** | 0.075 | 0.071 | 0.067 |
| EER (in %) | **6.95** | 7.02 | 7.17 | 7.01 |

It is important to evaluate the performance of proposed SLI model in real driving environment. To this end, the Street and Car noises were added to the test data set. These noises are recoded by the voice interaction technology center at Sogou in the real scenarios with the total recording time about 80 hours. The noisy speech was made by randomly selecting a segment, which has the same length as the relevant speech signal, from the noise recording. The selected noise was then scaled to the desired SNR level and then added to the speech. The noise level of each test utterance is randomly selected from 0dB to 20dB. With many noisy speech generations, the noise level is controlled at 10dB on average. From the experimental results shown in Table IV, it can be observed that the TSM method can also improve the robustness of the SLI model significantly under car noisy condition.

TABLE IV
COMPARISON OF DIFFERENT SLI MODELS IN NOISY CONDITION.

| Models | test-all | | test-3s | | test-1s | |
|---|---|---|---|---|---|---|
| | $C_{avg}$ | EER | $C_{avg}$ | EER | $C_{avg}$ | EER |
| DNN-BN-LSTM | 0.15 | 15.01 | 0.22 | 20.32 | 0.24 | 23.44 |
| TSM-DNN-BN-LSTM | **0.10** | **10.04** | **0.13** | **12.95** | **0.21** | **20.51** |

## V. CONCLUSION

In this paper we proposed an end-to-end speech language identification (SLI) model. Three measurements were used to make the trained model more suitable to short utterances. In the waveform domain, we used a time-scale modification (TSM) method to extend the length of input short utterances. In the feature domain, we used the transfer learning idea to train a deep phoneme classifier. The bottleneck features of the phoneme classifier, which included the phoneme discriminative information, were used to train the language classifiers. In the language classifier domain, a LSTM based classifier was trained by the short time feature blocks which can make the trained model adaptive to short duration inputs. The experimental results on AP17-OLR database demonstrated that, comparing with the i-vector model and simple LSTM model, the proposed method can significantly enhance the classification accuracy of the SLI, especially on short duration utterance. In addition, the structure of proposed SLI model is very simple, which only occupied about 20M storage.

## ACKNOWLEDGEMENT

This work was supported by the National Key R&D Program of China No. 2018YFC0807205, by National Natural Science Foundation of China No. 61773071, 61872170, by the Beijing Nova Program No. Z171100001117049, by the Beijing Nova Program Interdisciplinary Cooperation Project No. Z181100006218137, by the Beijing Natural Science Foundation No. 4162044, and by the Fundamental Research Funds for the Central Universities No. 2018XKJC02.

## REFERENCES

[1] O. Yürür, C. H. Liu, C. Perera, M. Chen, X. Liu, and W. Moreno, "Energy-efficient and context-aware smartphone sensor employment," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 9, pp. 4230–4244, Sept 2015.

[2] I. Bisio, C. Garibotto, A. Grattarola, F. Lavagetto, and A. Sciarrone, "Smart and robust speaker recognition for context-aware in-vehicle applications," *IEEE Transactions on Vehicular Technology*, pp. 1–1, 2018.

[3] M. Sun, A. Schwarz, M. Wu, N. Strom, S. Matsoukas, and S. Vitaladevuni, "An empirical study of cross-lingual transfer learning techniques for small-footprint keyword spotting," in *IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2017, pp. 255–260.

[4] Y. Jang, J. Ham, B. Lee, and K. Kim, "Cross-language neural dialog state tracker for large ontologies using hierarchical attention," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2072–2082, Nov 2018.

[5] Z. Tang, D. Wang, Y. Chen, L. Li, and A. Abel, "Phonetic temporal neural model for language identification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 134–144, 2018.

[6] S. Ganapathy, S. H. Mallidi, and H. Hermansky, "Robust feature extraction using modulation filtering of autoregressive models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 8, pp. 1285–1295, Aug 2014.

[7] S. Ranjan, G. Liu, and J. H. L. Hansen, "An i-vector PLDA based gender identification approach for severely distorted and multilingual DARPA RATS data," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 331–337.

[8] M. Sarma and K. K. Sarma, "Long-term critical band energy-based feature set for dialect identification using a neuro-fuzzy approach," *IEEE Intelligent Systems*, vol. 33, no. 1, pp. 40–52, Jan 2018.

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

[10] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *International Conference on Spoken Language Processing*, 2002.

[11] R. W. M. Ng, T. Lee, C. Leung, B. Ma, and H. Li, "Spoken language recognition with prosodic features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1841–1853, Sept 2013.

[12] H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, and K. Kashino, "Generative modeling of voice fundamental frequency contours," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1042–1053, June 2015.

[13] A. Sizov, K. A. Lee, and T. Kinnunen, "Direct optimization of the detection cost for i-vector-based spoken language recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 588–597, March 2017.

[14] F. Chen, L. Wang, H. Chen, and G. Peng, "Investigations on Mandarin aspiratory animations using an airflow model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2399–2409, Dec 2017.

[15] W. Kim, T. Song, T. Kim, H. Park, and S. Pack, "VoIP capacity analysis in full duplex WLANs," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 12, pp. 11 419–11 424, Dec 2017.

[16] J. Willmore, R. Price, and W. Roberts, "Comparing Gaussian mixture and neural network modelling approaches to automatic language identification of speech," in *Aust. Int. Conf. Speech Sci. & Tech*, 2000, pp. 74–77.

[17] M. A. Pathak and B. Raj, "Privacy-preserving speaker verification and identification using Gaussian mixture models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 397–406, Feb 2013.

[18] L. Ferrer, Y. Lei, M. McLaren, and N. Scheffer, "Study of senone-based deep neural network approaches for spoken language recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 105–116, Jan 2016.

[19] K. Wong and M.-h. Siu, "Automatic language identification using discrete hidden markov model," in *Eighth International Conference on Spoken Language Processing*, 2004.

[20] S. Nakagawa, Y. Ueda, and T. Seino, "Speaker-independent, text-independent language identification by hmm," in *Second International Conference on Spoken Language Processing*, 1992.

[21] Y. K. Muthusamy, "A segmental approach to automatic language identification," 1993.

[22] S. C. Kwasny, B. L. Kalman, W. Wu, and A. M. Engebretson, "Identifying language from speech: An example of high-level, statistically-based feature extraction," in *Proceedings of Annual Conference of the Cognitive Science Society*, 1992.

[23] W. M. Campbell, E. Singer, P. A. Torres-Carrasquillo, and D. A. Reynolds, "Language recognition with support vector machines," in *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.

[24] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Annual conference of the international speech communication association*, 2011.

[25] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language recognition in ivectors space," in *Annual Conference of the International Speech Communication Association*, 2011.

[26] R. Kazemi, M. Boloursaz, S. M. Etemadi, and F. Behnia, "Capacity bounds and detection schemes for data over voice," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 11, pp. 8964–8977, Nov 2016.

[27] P. Matejka, L. Burget, P. Schwarz, and J. Cernocky, "Brno university of technology system for nist 2005 language recognition evaluation," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*. IEEE, 2006, pp. 1–7.

[28] Z. M. Fadlullah, F. Tang, B. Mao, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems," *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2432–2455, Fourthquarter 2017.

[29] N. Kato, Z. M. Fadlullah, B. Mao, F. Tang, O. Akashi, T. Inoue, and K. Mizutani, "The deep learning vision for heterogeneous network traffic

control: Proposal, challenges, and future perspective," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 146–153, June 2017.

[30] X. Wang, L. Gao, S. Mao, and S. Pandey, "Csi-based fingerprinting for indoor localization: A deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 763–776, Jan 2017.

[31] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 5337–5341.

[32] A. Lozano-Diez, R. Zazo-Candil, J. Gonzalez-Dominguez, D. T. Toledano, and J. Gonzalez-Rodriguez, "An end-to-end approach to language identification in short utterances using convolutional neural networks," in *Annual Conference of the International Speech Communication Association*, 2015.

[33] M. Jin, Y. Song, I. Mcloughlin, L.-R. Dai, and Z.-F. Ye, "Lid-senone extraction via deep neural networks for end-to-end language identification," in *Proc. of Odyssey*, 2016.

[34] P. Matejka, L. Zhang, T. Ng, H. S. Mallidi, O. Glembek, J. Ma, and B. Zhang, "Neural network bottleneck features for language identification," in *Proceedings of Odyssey*, vol. 2014, 2014, pp. 299–304.

[35] H. Yu, Z.-H. Tan, Z. Ma, and J. Guo, "Adversarial network bottleneck features for noise robust speaker verification," in *Proceedings of INTERSPEECH*, 2017.

[36] H. Yu, Z.-H. Tan, Y. Zhang, Z. Ma, and J. Guo, "DNN filter bank cepstral coefficients for spoofing detection," *IEEE Access*, vol. 5, pp. 4779–4787, 2017.

[37] H. Yu, Z.-H. Tan, Z. Ma, R. Martin, and J. Guo, "Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2018.

[38] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors," in *Odyssey: The Speaker and Language Recognition Workshop, Les Sables d'Olonne*, 2018.

[39] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, and P. J. Moreno, "Automatic language identification using long short-term memory recurrent neural networks," in *Annual Conference of the International Speech Communication Association*, 2014.

[40] S. Fernando, V. Sethu, E. Ambikairajah, and J. Epps, "Bidirectional modelling for short duration language identification," in *Proc. Interspeech 2017*, 2017, pp. 2809–2813.

[41] R. Zazo, P. S. Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak, "Age estimation in short speech utterances based on LSTM recurrent neural networks," *IEEE Access*, vol. 6, pp. 22 524–22 530, 2018.

[42] B. Mao, Z. M. Fadlullah, F. Tang, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "Routing or computing? the paradigm shift towards intelligent computer network packet transmission based on deep learning," *IEEE Transactions on Computers*, vol. 66, no. 11, pp. 1946–1960, Nov 2017.

[43] T. Taniguchi, K. Furusawa, H. Liu, Y. Tanaka, K. Takenaka, and T. Bando, "Determining utterance timing of a driving agent with double articulation analyzer," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 3, pp. 810–821, March 2016.

[44] J. Yuan, M. Liberman, and C. Cieri, "Towards an integrated understanding of speaking rate in conversation," in *International Conference on Spoken Language Processing*, 2006.

[45] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

[46] C. J. van Heerden, E. Barnard, E. Van Heerden *et al.*, "Speech rate normalization used to improve speaker verification," in *Proceedings of the Symposium of the Pattern Recognition Association of South Africa*. Citeseer, 2007, pp. 2–7.

[47] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, 2007.

[48] Y. Nejime, T. Aritsuka, T. Imamura, T. Ifukube, and J. Matsushima, "A portable digital speech-rate converter for hearing impairment," *IEEE transactions on rehabilitation engineering*, vol. 4, no. 2, pp. 73–83, 1996.

[49] X. Miao, J. Zhang, H. Suo, R. Zhou, and Y. Yan, "Expanding the length of short utterances for short-duration language recognition," *Journal of Tsinghua University (Science and Technology)*, vol. 58, no. 3, pp. 254–259, 2018.

[50] E. S. Ottosen and M. Dörfler, "A phase vocoder based on nonstationary gabor frames," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2199–2208, Nov 2017.

[51] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

[52] W. Lin, M. Mak, and J.-T. Chien, "Multisource i-vectors domain adaptation using maximum mean discrepancy based autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2412–2422, Dec 2018.

[53] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. L. Roux, and K. Takeda, "Duration-controlled LSTM for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2059–2070, Nov 2017.

[54] Y. Tian, L. He, Y. Liu, and J. Liu, "Investigation of senone-based long-short term memory rnns for spoken language recognition," in *Proc. Odyssey*, 2016, pp. 89–93.

[55] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN)*, vol. 3. IEEE, 2000, pp. 189–194.

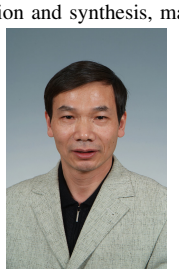
[56] Z. Tang, D. Wang, Y. Chen, and Q. Chen, "Ap17-olr challenge: Data, plan, and baseline," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 749–753.

[57] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

**Zhanyu Ma** has been an Associate Professor at Beijing University of Posts and Telecommunications, Beijing, China, since 2014. He is also an adjunct Associate Professor at Aalborg University, Aalborg, Denmark, since 2015. He received his Ph.D. degree in Electrical Engineering from KTH (Royal Institute of Technology), Sweden, in 2011. From 2012 to 2013, he has been a Postdoctoral research fellow in the School of Electrical Engineering, KTH, Sweden. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in multimedia signal processing, data mining, biomedical signal processing, and bioinformatics.

**Hong Yu** received the B.Sc. and M.Sc. degrees inelectronic information engineering from Shandong University, Jinan, China, in 2003 and 2006, respectively, and the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications, Beijing, China, in 2018. He has been an lecture at Ludong University, Yantai, Shandong, China, since 2006. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in speech processing, image processing, data mining, biomedical signal processing, and bioinformatics.

**Wei Chen** has been the chief scientist of the voice interaction technology center at Sogou, an innovator in search and a leader in China's internet industry since 2016. He received his Ph.D. degree in signal and information processing from BUPT (Beijing University of Posts and Telecommunications) , China, in 2011. Over the past six years, he led his team to establish the company's voice-based multimodal human-computer interaction operating system, the Zhiyin OS. His research interests include multi-modal interaction technology, speech recognition and synthesis, machine translation.



**Jun Guo** received B.E. and M.E. degrees from Beijing University of Posts and Telecommunications (BUPT), China in 1982 and 1985, respectively, Ph.D. degree from the Tohuku-Gakuin University, Japan in 1993. At present he is a professor and a vice president of BUPT. His research interests include pattern recognition theory and application, information retrieval, content based information security, and bioinformatics. He has published over 200 papers on the journals and conferences including SCIENCE, Nature Scientific Reports, IEEE Trans. on PAMI, Pattern Recognition, AAAI, CVPR, ICCV, SIGIR, etc.