Improvement the phonetic recognition with sequence-length MFCC features and deep bidirectional LSTM

Toan Pham Van
Framgia Inc R&D Group
pham.van.toan@framgia.com

Hau Nguyen Thanh
Framgia Inc R&D Group
nguyen.thanh.hau@framgia.com

Thanh Ta Minh
Le Quy Don Technical University
ta.minh.thanh@framgia.com

***Abstract*—Phonetic recognition is an meaningful problem in the many fields of speech analysis. This applications can mention as dialect identification [1], mispronunciation detection [2], spoken document retrieval [3] etc. Some main approaches to solving this problem are improve the feature selection on input speech [4], apply some deep learning techniques [5][6][7] or combination both of them [8]. With the sequence data as the phonetics, the architecture based on recurrent neural network is an appropriate approach [9]. It is more powerful when combined with the improvement of features selection in input data. In our approach, we combining the Mel Frequency Cepstral Coefficients (MFCC) method with sequence-length for present the acoustic features of speech and using some RNN models to phonetic classification. All the experiments are implement on the Texas Instruments Massachusetts Institute of Technology (TIMIT) [10] phone recognition dataset. Especially, our data processing and features selention method give consistently better results with other researches when applied the same neural network model. Currently, we achieved the lowest error test rate - 15.24% by using Bidirectional LSTM, which is currently the best result in TIMIT dataset and reduction of about 1.3% over the last best result [5][6].**

***Keywords*—Phonetic Recognition, MFCC features, sequence-length, bidirectional LSTM, TIMIT**

I. Introduction

In a general machine learning application, speech recognition technology is one of the most typical application recently. In speech recognition technology, given a sequence of acoustic observations, this technology decodes the corresponding sequence of words or phonemes. From that, we can use it for helping language learner in pronunciation. The typical neural network model is used for speech recognition system is recurrent neural network (RNN), an effective model in sequence-to-sequence problem.

In this paper, we introduce recurrent neural network model and some variants, along with some techniques to improve the accuracy for phonetic classification problem such as: sequence length, feature scaling, deep long-short-term memory (deep LSTM), bidirectional LSTM. With these techniques, the phonetic classification problem is greatly improved compared to the original model.

We evaluate the effectiveness of models using TIMIT dataset. The original data were converted to Mel Frequency Cepstral Coefficient (MFCC) features. MFCCs features were said to have better results in speech recognition problem. In each generated output of sequence, we use the results of previous and next steps by using bidirectional LSTM. In others, when applying feature scaling for input data, the training process will be more faster, and it gets better results. We get achieved 13.5% PER, the best result in TIMIT dataset until now.

II. Related works

A. Baseline

A simple approach to solve this problem is using phoneme-based recognition and identifying pronunciation errors in the input speech of non-native speakers. But in actually, the accuracy for detection words is much higher than phonetic detection even for native speakers. It also mean that we can not directly apply **Automatic Speech Recognition** system for mispronunciation detection. Instead we add to the ASR system a pronunciation model with possible faulty pronunciation variations are used to recognize the most likely phone sequences when it knew previous phones. Finally, the mispronunciation detection system is worked by executing the forced-alignment of ASR with extended pronunciation recognizer based on possible phonetic confusions. The figure below is a simulation of workflow in this system:

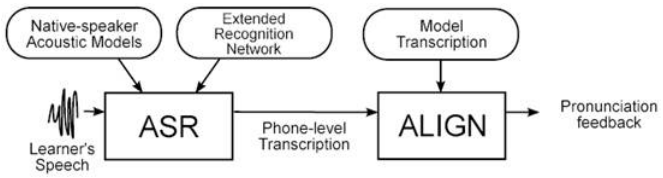

Fig. 1: *ASR-based system to detect and diagnose L2 learners mispronunciation*

B. Feature Extraction

As discussed above, phone-level mispronunciation detection can detect mispronunciations in units of phones, words or sentences. Firstly, the speaker's speech samples are first converted to certain types of features such as Linear predictive cepstral coefficients (**LPCC**), Mel-frequency cepstral coefficients (**MFCC**), Power spectral analysis (**FFT**) or Mel scale cepstral analysis (**MEL**) etc. These features are used as the input of classifier. These allow the system to classify the mispronunciation by types. In our works, we have used 26 **MFCC** features as input features.

TABLE I: *Japanese phonetic classifier performance of four ANN architectures*

	CNN 1D	*Normal RNN*	*LSTM*	*CNN + LSTM*
FAR	21.92	17.21	15.93	11.73
FRR	11.29	9.32	9.02	6.23

C. Dataset

In our works for phonetic classification and detection, we used the dataset named **Japanese Phonetic Database - JPD** [1]. The JPD provides IPA phonetic transcriptions that accurately indicate how Japanese names and words are pronounced in actual speechs, as well as accent codes, for each entry. It includes detail descriptions of 130,000 entries in Japanese.

The advantage of this dataset are: it includes the pitch accent position for each phonetic and it is so meaningful for many speech processing tasks. An example of phonetics collection in **JPD** is shown below.

Ortho-graphic	Kana	Phonetic	Accent	Pitch Pattern	Remarks
井川	イカワ	[ikawa]	1	HLL(L)	first mora accented
井田	イダ	[ida]	0	LH(H)	accentless
磯貝	イソガイ	[isoŋai]	2	LHLL(L)	second mora accented
鏡	カガミ	[kaŋami]	3	LHH(L)	last mora accented
形	カタチ	[katatɕi]	0	LHH(H)	accentless

Fig. 2: *JPD phonetics dataset with pitch accent*

D. Phonetic classification methods

Deep learning is a technique used a lot recently. There are several methods of deep learning that we use there. Recurrent Neural Network (RNN) is a class of artificial neural network where connections between units form a directed graph along a sequence. This allows exhibitting dynamic temporal behavior for a time sequence. We use RNN with 3 layers of LSTM cell. Another deep neural network class we use there is Convolutional Neural Network (CNN), it also find the connectivity between features and help us to improve the accuracy. The architecture of CNN we use there is combination of convolutional layers, maxpool and fully connected layers. Finally we use both RNN and CNN in a combination called deepspeech, they help us increasing the accuracy in this works.

E. Result

We tried some models of deep learning architecture mentioned above to training the extended pronunciation recognizer. Following our previous works, two metrics false acceptance rate **(FAR)** and false rejection rate **(FRR)** are used to measure the system performance. The comparison of our Neural Network accuracy with some algorithms is shown in ***Table 1***

III. Methods

A. Recurrent Neural Network

Recurrent Neural Networks (RNNs) are popular models that have shown great promise in many Natural Language Processing, Artificial Speech Recognition, Time Series, Sequence-to-sequence tasks. The idea behind RNNs is to make use of sequential information. Here is what a typical RNN look like:

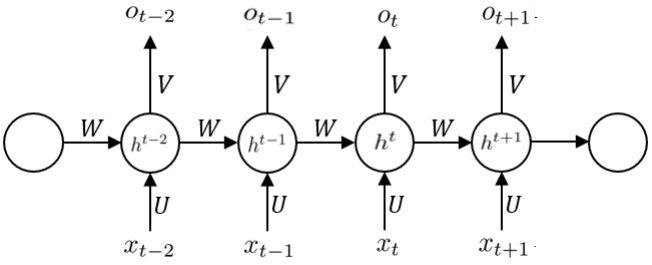

Fig. 3: *Recurrent Neural Network*

$$h_t = Ux_t + Wh_{t-1}$$
$$o_t = Vh_t$$

B. Long Short Term Memory

In theory RNNs can make use of information in arbitrarily long sequences, but in practice they are limited to looking back only a few steps. From that, LSTM cell have been designed to get around this problem. Long Short Term Memory Network (LSTMs) are a special kind of RNN, capable of learning long-term dependencies.

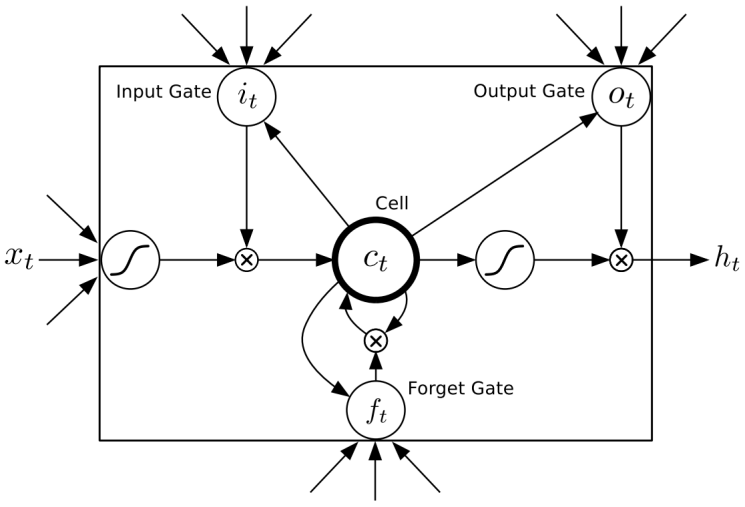

Fig. 4: *Long Short Term Memory Cell*

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
c_t &= f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \\
h_t &= o_t tanh(c_t)
\end{aligned}
$$

where σ is the logistic sigmoid function, and i, f, o and c are respectively the *input gate*, *forget gate*, *output gate* and *cell memory* vectors.

C. Bidirectional Recurrent Neural Networks

Bidirectional Recurrent Neural Networks (BRNNs) are based on the idea that the output at time t may not only depend on the previous elements in the sequence, but also future elements. Bidirectional RNNs are quite simple. They are just two RNNs stacked on top of each other. The output is then computed based on the hidden state of both RNNs.

D. Deep Bidirectional Recurrent Neural Networks

Deep (Bidirectional) RNNs are similar to Bidirectional RNNs, only that we now have multiple layers per time step. In practice this gives us a higher learning capacity (but we also need a lot of training data). <picture and explaination>

E. Other techniques

1) Sequence length: In this paper, we use the input sequence is a 2D array representing each utterance of the sentence. Where each row is a feature vector with 26 MFCC values. Number of columns is the number of feature vectors of the sentence which have the longest MFCC features, called "max length". If the sentence has the number of features less than "max length", vectors with a value of 0 will be added to fit with "max length". This will allow us to use Tensorflow in training. Also, with the use of sequence length, we will ignore the dependence between sentences when we connect the sentences together, obviously for greater efficiency because sentences completely independent.

2) Feature Scaling: Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization/standardization and is generally performed during the data preprocessing step. By using feature scaling with input data before training, we get the better results and faster training. The technique we used there is standardization, first it subtracts the mean value (so standardized values always have a zero mean), and then it divides by the variance so that the resulting distribution has unit variance. In others, standardization is much less affected by outliers.

$$
x' = \frac{x - \bar{x}}{\sigma} \tag{1}
$$

Where x is the original feature vector, $\bar{x}$ is the mean of that feature vector, and σ is its standard deviation.

IV. Experiments

Phoneme recognition experiments were performed on the TIMIT corpus. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States.

V. Conclusion and future works

We have used the combination of deep, bidirectional Long Short-term Memory RNNs with end-to-end training gives state-of-the-art results in phoneme recognition on the TIMIT database. An our plan is extend the system to large vocabulary speech recognition. Another plan would be use some another techniques in deep learning such as convolutional neural networks (CNNs), gated recurrent unit in RNNs to improve the accuracy.

Acknowledgment

This research was partially supported by ***Framgia Vietnam***. We are thankful to our colleagues who provided expertise that greatly assisted the research, although they may not agree with all of the interpretations provided in this paper.

References

[1] Zissman, Marc A., et al. "Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech." Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on. Vol. 2. IEEE, 1996.

[2] Harrison, Alissa M., et al. "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training." International Workshop on Speech and Language Technology in Education. 2009.

[3] Ng, Kenney, and Victor W. Zue. "Phonetic recognition for spoken document retrieval." Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on. Vol. 1. IEEE, 1998.

[4] Zeghidour, Neil, et al. "Learning Filterbanks from Raw Speech for Phone Recognition." arXiv preprint arXiv:1711.01161 (2017).

[5] Tóth, László. "Phone recognition with hierarchical convolutional deep maxout networks." EURASIP Journal on Audio, Speech, and Music Processing 2015.1 (2015): 25.

[6] Vaněk, Jan, et al. "A Regularization Post Layer: An Additional Way How to Make Deep Neural Networks Robust." International Conference on Statistical Language and Speech Processing. Springer, Cham, 2017.

[7] Mohamed, Abdel-rahman, George Dahl, and Geoffrey Hinton. "Deep belief networks for phone recognition." Nips workshop on deep learning for speech recognition and related applications. Vol. 1. No. 9. 2009.

[8] Tóth, László. "Combining time-and frequency-domain convolution in convolutional neural network-based phone recognition." Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014.

[9] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." Acoustics, speech and signal processing (icassp), 2013 ieee international conference on. IEEE, 2013.

[10] Garofolo, John S., et al. "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1." NASA STI/Recon technical report n 93 (1993).