# Improving Phonetic Recognition with Sequence-length Standardized MFCC Features and Deep Bi-Directional LSTM

Toan Pham Van
*Framgia Inc*
pham.van.toan@framgia.com

Hau Nguyen Thanh
*Framgia Inc*
nguyen.thanh.hau@framgia.com

Ta Minh Thanh
*Le Quy Don Technical University*
thanhtm@mta.edu.vn

***Abstract*—Phonetic recognition is one of the most challenging problems in the field of speech analysis. These applications can be mentioned such as dialect identification [1], mispronunciation detection [2], spoken document retrieval [3], and so on. There are different approaches to solve these problems such as improving the feature selection on input speech [4], applying deep learning technique [5][6][7] or combining both of them [8]. With the sequence data as the phonetics, the architecture which is based on recurrent neural network (RNN) is an appropriate approach [9]. It is even more powerful when combined with the improvement of features selection on input data. In our approach, we combine the Mel Frequency Cepstral Coefficients (MFCC) method with sequence-length to present the acoustic features of speech and use some RNN models to phonetic classification. Our experiments are implemented on the Texas Instruments Massachusetts Institute of Technology (TIMIT) [10] phone recognition dataset. Especially, our data processing and features selection method give consistently better results than other researches using the same neural network model. Currently, we have achieved the lowest error test rate (13.05%) by using Bidirectional LSTM, which is the best result in TIMIT dataset with the reduction of about 3.5% over the last best result [5][6].**

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

## I. Introduction

In the era of information, the communication between people and computers is increasingly narrowing the gap. There are a lot of computer science techniques to do that in which speech processing is one of the most indispensable techniques. In a general machine learning application, speech recognition technology is one of the most typical application. 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 learners in their pronunciation. The typical neural network model used for speech recognition system is recurrent neural network (RNN), an effective model in sequence-to-sequence problem.

In this paper, we introduce RNN model with some variants, along with 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 comparing to the original model. In the proposed method, we show the difference with other methods through two components: data processing and training models.

We evaluate the effectiveness of our models by using TIMIT dataset. The original data were converted to Mel Frequency Cepstral Coefficient (MFCC) features. MFCCs features were proved 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. When the feature scaling for input data is applied, the training process is not only more faster but also to get better results. We have achieved 13.5% PER, which is better than that of results from paper [5][6].

The rest of this paper is organised as follows. The related work is briefly introduced in section II. Section III introduces the proposed method and the experiment details. In section IV, we conclude the paper and present the future work.

## II. Related work

### A. TIMIT Speech Corpus

Phoneme recognition experiments are performed on the Texas Instruments (TI) and Massachusetts Institute of Technology (MIT) - TIMIT corpus [10]. TIMIT contains 6300 sentences in total, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. It was recorded at 16 kHz rate with 16 bits sample resolution. All sentences in this corpus were segmented at the phone level manually. The original corpus included 61 phonemes.

This dataset is commonly used in the speech recognition community because it is small enough to ensure that experiments are carried out easily, however it is also large enough to demonstrate the effectiveness of the method being used. The manual alignment and details of transcription can be found in the (Zue & Seneff, 1996) [11]. In many researches, some authors compact the origin phonemes to 48 phonemes [12] or 39 phonemes [13] in the training and testing phase. **Figure 1** describes this folding process and the resultant 39 phone set. In this paper, the experiment is carried out in two phoneme sets with the number of phonemes of the training and testing set are $61-61$ and $40-40$, respectively. For resultant 61 phonemes to 40 phonemes, we apply the similar method of (Lee & Hon, 1989) [13], however, we do not remove the **q** phonemes from the original dataset.

In the original TIMIT, the training set contains 3512 sentences of 462 speakers. The testing set contains 1344 sentences of 168 speakers. In our work, we use 184 sentences from training set to make the validation set.

978-1-5386-7983-8/18/$31.00 ©2018 IEEE

| | |
|---|---|
| aa, ao | aa |
| ah, ax, ax-h | ah |
| er, axr | er |
| hh, hv | hh |
| ih, ix | ih |
| l, el | l |
| m, em | m |
| n, en, nx | n |
| ng, eng | ng |
| sh, zh | sh |
| uw, ux | uw |
| pcl, tcl, kcl, bcl, dcl, gcl, h#, pau, epi | sil |
| q | - |

Fig. 1: Mapping from 61 phonemes to 39 phonemes

### B. Some works on TIMIT dataset

There are several researches worked on TIMIT dataset for speech recognizers. The trending of this field was based on the connectionist temporal classification (CTC) [20]. After that, CTC method is extended to RNN Transducer [21] for reducing the phone error rates (PER). Both of them focused on treating the alignment as a latent random variable over which MAP (maximum a posteriori) inference.

Application of attention-based models to speech recognition is proposed by Chorowski $et.\ al.$ [17] recently. It improved the PER comparing to [20] and [21] by using the attention mechanism deterministically aligns the input and the output sequences. It also proposed attention mechanism using both non-monotonic and monotonic alignment for a larger variety of tasks other than speech recognition. However, it may take a lager cost for training and testing phases.

A hybrid attention model to speech recognition [22] is also considered by using additional informations in the attention model such as the content-based, the location-based addressing, and so on. However, this model cannot work with long input sequences. That makes it cannot be applied on the real applications.

Recently, a method that was proposed the regularization post layer to improve a DNN generalization ability [6]. They proved that they can combine with other techniques, then that can be applied on many applications. Also, they can obtain better results than DNN and DBN pre-training. However, the main drawback of that is high computational requirements in the testing phase.

According to drawbacks of previous works, we focus on solving the technical issues by selecting the feature and using bidirectional LSTM to phone recognition task.

### C. Features Extraction

Features extraction is an important phase of the machine learning problems. With the phonetic classification problem, we try different speech features from extraction methods. Firstly, the speaker's speech samples are 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 the classifier. These allow the system to classify the mispronunciation by types. In our work, we use 26 **MFCC** features as input features.

### D. Other techniques in our work

*1) Sequence length:* In this paper, 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 allows us to use Tensorflow in training process. Also, with the use of sequence length, we 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 use is standardization. Feature vector is subtracted by the mean value (so standardized values always have a zero mean), and then is divided by the variance. So that, the resulting distribution has unit variance. In the 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 $\sigma$ is its standard deviation.

## III. The Proposed Method

The workflow of our methods have been shown in **Figure 2**. Number of phonemes are 40 phonemes or 61 phonemes.

Initially, we convert the raw audio signals to MFCC features with 26 MFCC features per 20ms. Then, we use the standardization technique to standardize MFCC feature vectors. We add the context MFCC features for each feature in standardized vectors. By doing so, we can improve the accuracy of our works. Then, we put them into bidirectional recurrent neural networks with the long short-term memory (LSTM) networks.

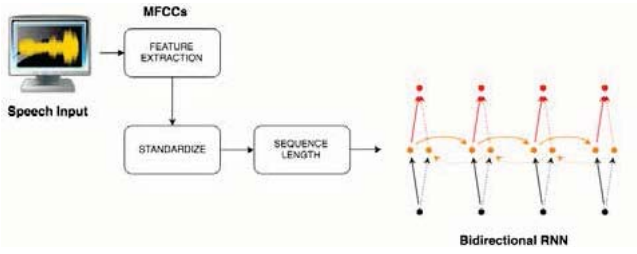


Fig. 2: Phonetic classification workflow.

After each layer, we use dropout for one. A fully connected layer is added after bidirectional recurrent neural networks. Finally, we add the output layer to get phonemes. The loss function we used here is cross-entropy, which was a convex function. This will help us more easily improve the performance of works.

## IV. Experiments

### A. Experimental environments

We train models by a computer in our research center, the detail of this configuration includes a CPU Intel Core I5-7500 3.40GHz x 4, a GPU GeForce GTX 1080 Ti, hard disk Samsung SSD 512 GB. The average training time is about 24 hours / 100 epochs for the RNN algorithm and about 2 days for the BRNN algorithm. The training time for the 61 phonemes dataset is a little bit longer than the dataset of 61 phonemes (about 1 hour / 100 epochs).

### B. Model Hyperparameters

All the RNN models are implemented in **Tensorflow** - a open-source framework for training machine learning models. With LSTM and bidirectional LSTM network, we use the batch size of 32, number of LSTM layers is 3 with 256 units each layer. In both models above, we use **Adam Optimizer** [14] for optimization with learning rate is 0.001. The max length of a feature sequence is 776 for feature encoding.

### C. Implementation

In our experiments, we do not use traditional methods that are often used in classification problems such as SVM, KNN, Naive Bayes. We conduct our experiments with artificial neural network models including LSTM model and Bidirectional LSTM model. By combining two models above with the two dataset splits described in Section II, we conduct four experiments to compare the other algorithms in the most objective way.

Firstly, we experimented with the original TIMIT data set (61 phonemes) and the LSTM model. We run this algorithm with 100 epochs (about 20 hours). Our results in the train and validation set are shown in **Figure 3**. Secondly, we try the Bidirectional LSTM for the same datasets. It achieves the better accuracy than first method. In this experiment, due to the limited hardware, we only train the model within 100 first epochs. The results are shown in **Figure 4**. We archive the best result of the error in test set is **0.162** at epoch 98 and in the validation set is **0.143** at the same epoch. Thirdly, we try the LSTM network with similar architecture as above with the dataset of 40 phonemes. The results are shown in **Figure 5**. And the last one, we try the bidirectional LSTM network with the dataset of 40 phonemes. And the last one, we try the bidirectional LSTM network with the dataset of 40 phonemes. The best result of the error in test set is **0.185** at epoch 98 and in the validation set is **0.147** at the same epoch.

### D. Comparison with other methods

In Table I, we show the summary of our results together with a couple state-of-the-art recently published results. Comparing with the best result trained on TIMIT from the paper

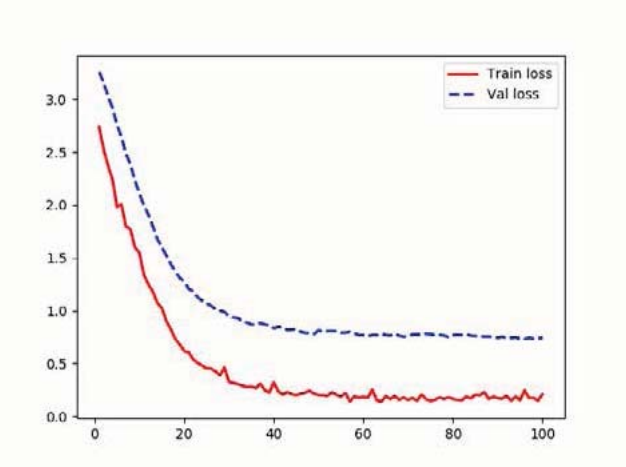


Fig. 3: *Training results with LSTM - 61 phonemes*

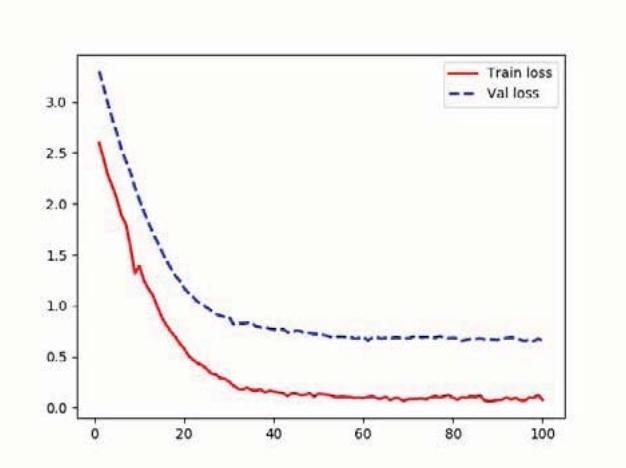


Fig. 4: *Training result with bidirectional LSTM - 61 phonemes*

[5] used hierarchical maxout CNN and dropout, our result is obviously improved. Our phone error rates (PER) is lower than that of all previous researches.

## V. Conclusion and future works

We have used the combination of deep bidirectional LSTM with end-to-end training which gives state-of-the-art results in phoneme recognition on the TIMIT database. Our plan is to extend the system for larger 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 Inc***.

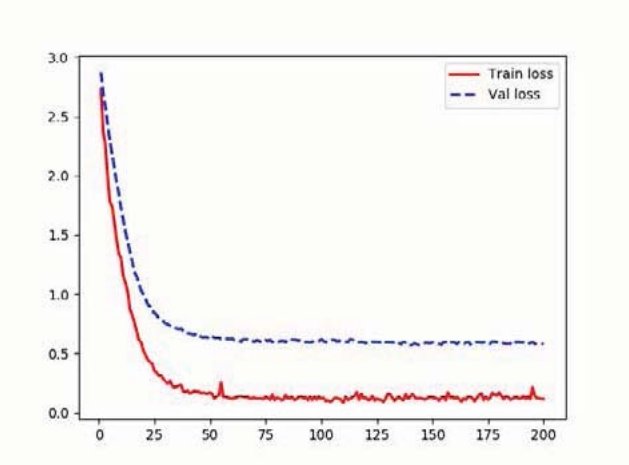


Fig. 5: *Training results with LSTM - 40 phonemes*

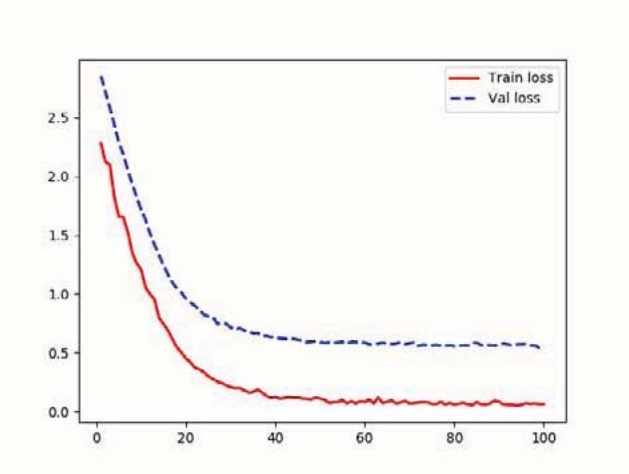


Fig. 6: *Training results with bidirectional LSTM - 40 phonemes*

TABLE I: Phone error rate trained on TIMIT

| Methods | PER |
|---|---|
| (first, modern) HMM-DBN [19] | 23% |
| Wavenet architecture [16] | 18.8% |
| Complex ConvNets on raw speech w melfbanks init [4] | 18.0% |
| Bi-LSTM + skip connections w CTC [9] | 17.7% |
| Bi-RNN + Attention [17] | 17.6% |
| RNN-CRF on 24(x3) MFSC [18] | 17.3% |
| CNN in time and frequency + dropout, 17.6% wo dropout [8] | 16.7% |
| DBN with last layer regularization [6] | 16.5% |
| Hierarchical maxout CNN + Dropout [5] | 16.5% |
| **BRNN + LSTM + 40 phonemes – (Our method)** | **13.0**% |

## 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).

[11] Zue, V. & Seneff, S. (1996). Transcription and alignment of the TIMIT database. In Hiroya Fujisaki (Ed.), Recent research toward advanced man-machine interface through spoken language. Amsterdam: Elsevier, pp 464-447, 1996.

[12] Lopes, Carla, and Fernando Perdigao. "Phoneme recognition on the TIMIT database." Speech Technologies. InTech, 2011.

[13] Lee, K. & Hon, H. (1989). Speaker-independent phone recognition using hidden Markov models. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.37 (11), November 1989, pp. 1642-1648, ISSN: 0096-3518.

[14] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

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

[16] 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.

[17] Chorowski, Jan K., et al. "Attention-based models for speech recognition." Advances in neural information processing systems. 2015.

[18] 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.

[19] Lu, Liang, et al. "Segmental recurrent neural networks for end-to-end speech recognition." arXiv preprint arXiv:1603.00223 (2016).

[20] Alex Graves, et al., "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," In ICML-06, 2006.

[21] Alex Graves, "Sequence transduction with recurrent neural networks," In ICML-12, 2012.

[22] Sainbayar Sukhbaatar, et al., "Weakly supervised memory networks," arXiv preprint arXiv:1503.08895, 2015.