QUALITATIVE ANALYSIS OF PLP IN LSTM FOR BANGLA SPEECH RECOGNITION

Nahyan Al Mahmud[1] and Shahfida Amjad Munni[2]

[1]Department of Electrical and Electronic Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh
[2]Cygnus Innovation Limited, Dhaka, Bangladesh

ABSTRACT

The performance of various acoustic feature extraction methods has been compared in this work using Long Short-Term Memory (LSTM) neural network in a Bangla speech recognition system. The acoustic features are a series of vectors that represents the speech signals. They can be classified in either words or sub word units such as phonemes. In this work, at first linear predictive coding (LPC) is used as acoustic vector extraction technique. LPC has been chosen due to its widespread popularity. Then other vector extraction techniques like Mel frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) have also been used. These two methods closely resemble the human auditory system. These feature vectors are then trained using the LSTM neural network. Then the obtained models of different phonemes are compared with different statistical tools namely Bhattacharyya Distance and Mahalanobis Distance to investigate the nature of those acoustic features.

KEYWORDS

LSTM, Perceptual linear prediction, Mel frequency cepstral coefficients, Bhattacharyya Distance, Mahalanobis Distance.

1. INTRODUCTION

Speech is the most effective way of communication among people. This is the most natural way of conveying information. Therefore, to interact vocally with computers, studies are being conducted in a number of fields. The objective is to simulate the humans' ability to talk, to carry out of simple tasks by computers through the means of machine-human interaction, to turning speech to text through Automatic Speech Recognition (ASR) systems. Recently, signal processing and detection have been applied in a variety of fields such as human activity tracking and recognition [1,2], computer engineering, physical sciences, health-related applications [3], and natural science and industrial applications [4].

In recent years, neural network technology has significantly enhanced the precision of ASR system, and many applications are being developed for smartphones and intelligent personal assistants. Many researches on end-to-end speech detection are being conducted to swap the hidden Markov model (HMM) based technique which has been used for many years. The end-to-end models with neural networks include connectionist temporal classification (CTC)-trained recurrent neural networks (RNN), encoder-decoder architectures [5], and RNN transducers [6,7]. Although HMM-based algorithms can be considered computationally efficient, they require many uneven memory accesses and a large memory foot-print. On the contrary, RNN based ASR has the advantage of low memory footprint; however, it demands many computationally expensive arithmetic operations for real-time inference.

DOI:10.5121/ijma.2020.12501

A lot of works has been done in ASR for a variety of major languages in the world. Regrettably, only a handful of works have been carried out for Bangla, which is among the most widely spoken languages in the world in terms of number of speakers. Some of these efforts can be found in [8]. However, majority of these studies mainly focussed on simple word-level detection worked on a very minor database. Also these works did not account for the various dialects of different parts of the country. The reasons for this can be pointed as the lack of proper speech database.

The primary target of this study is to examine the efficiency of various acoustic vectors for Bangla speech detection using LSTM neural network and assess their performances based on different statistical parameters.

2. DATA PREPARATION

2.1. Bangla Speech Database

Right now the real difficulty while experimenting on Bangla ASR is the absence of proper Bangla speech database. Either the existing database is not large enough or these databases are not considering the colloquial variations among the speakers. Also majority of these are not publicly available which is a massive hindrance in conducting research. Another difficulty is that the database must be properly segmented or labelled while using for supervised learning. To overcome this problem, we had to come up with our own solution [9].

The samples were first divided into two groups: training and testing. The male training corpus were prepared by 50 male and 50 female speakers. A separate set of 20 male and female speakers voiced the test samples. 200 common Bangla sentences were chosen for this database.

A medium sized room having styrofoam damper was used to record the samples. A dynamic unidirectional microphone was used to collect training samples. In order to mimic the real-world scenario for testing samples, a moderate grade smart phone was used.

2.2. Acoustic Feature Vectors

Linear predictive coding (LPC) was first introduced in 1966. It is widely used in audio signal processing, speech processing and telephony. LPC embodies the spectral envelope of speech in compressed form, using a linear predictive methodology. “These techniques can thus serve as a tool for modifying the acoustic properties of a given speech signal without degrading the speech quality. Some other potential applications of these techniques are in the areas of efficient storage and transmission of speech, automatic formant and pitch extraction, and speaker and speech recognition” [10].

The most efficient acoustic feature extraction methods are mainly based on models that are similar to human hearing system. There are some limiting features of the human auditory system such as nonlinear frequency scale, spectral amplitude compression, reduced sensitivity of at lower frequencies etc. Mel Frequency Cepstrum Coefficients (MFCC) and Perceptual Linear Predictive (PLP) to some extent adopts the behaviour of human auditory system. MFCC was first introduced in [11]. PLP introduced in [12] is based on ideas similar to the MFCCs. But the computational steps of PLP are a bit elaborate and thus produce marginally better result which is illustrated in Figure 1.

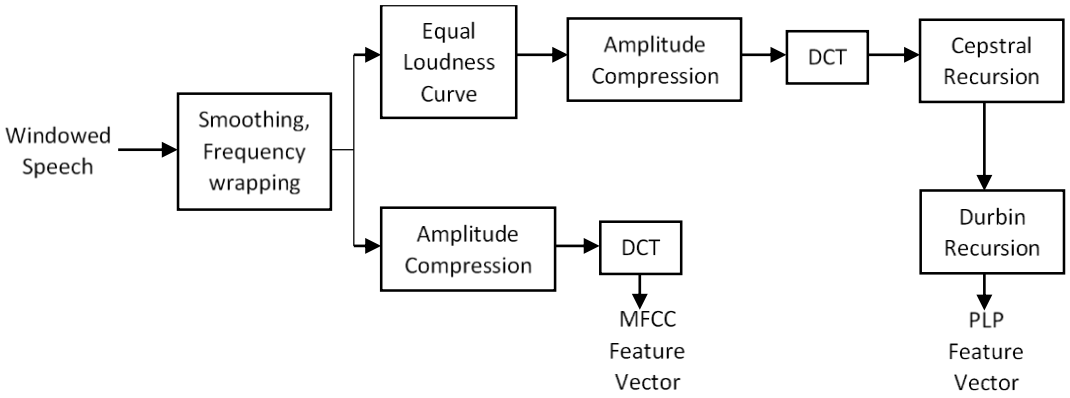

Figure 1. The steps of PLP and MFCC feature vector extraction.

3. EXPERIMENTAL SETUP

3.1. LSTM Neural Network Structure

LSTM, and in general, RNN based ASR systems [13,14,15] trained with CTC [16] have recently been shown to work extremely well when there is an abundance of training data, matching and exceeding the performance of hybrid DNN systems [15]. In the conventional DNN if the time is long, the amount of remaining vector reduces that needs to be looped back. This process inherently breaks down the network-updating scheme [17]. LSTM is naturally immune to this problem as it can employ its comparatively long-term memory structure [5].

LSTM and RNN are very prominent in sequence prediction and labelling at the same time, LSTM proved to have the better performance than RNN and DNN in context [18]. Recently, according to the successful application of DNN to acoustic modelling [19,20,21] and speech enhancement [22], a CTC output layer, is imposed on the acoustic model [23,24] in conjunction with a deep LSTM network. Figure 2 shows the LSTM speech recognition block.

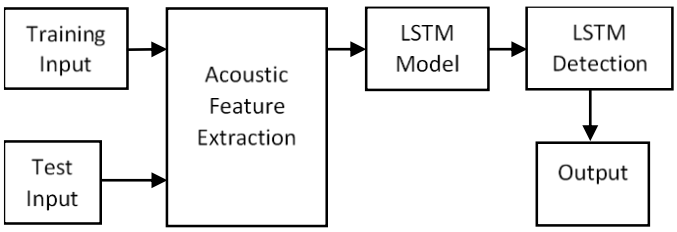

Figure 2. LSTM speech detection block

A standard LSTM neural network structure is comprised of an input layer, a recurring LSTM layer and an output layer. In this study the input layer is a Time Delay Neural Network (TDNN) layer and this layer is connected to the hidden layer.

The new LSTM structure proposed in [23] has been used in this work. As this method is based on CTC training, there is a significant performance growth in speech detection tasks. This technique has been adopted by Google. This typical LSTM cell structure is illustrated in Figure 3. The LSTM cell consists of three gates: input, forget, and output, which control, respectively, what fraction of the input is passed to the "memory" cell, what fraction of the stored cell memory is retained, and what fraction of the cell memory is output.

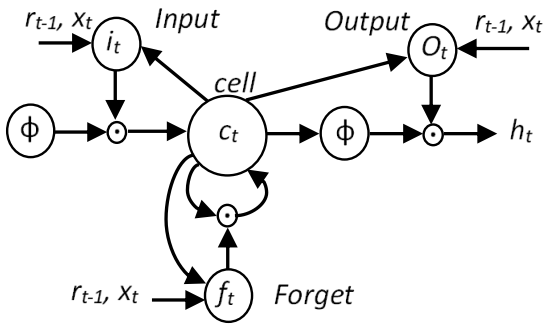

Figure 3. LSTM cell model

$i_t = \sigma(W_{ix}x_t + W_{ir}x_{t-1} + b_i)$.....(1)

$f_t = \sigma(W_{fx}x_t + W_{fr}x_{t-1} + b_f)$.....(2)

$C_t = f_t \odot C_{t-1} + i_t \odot g(W_{cx}x_t + W_{cr}r_{t-1} + b_c)$.....(3)

$o_t = \sigma(W_{ox}x_t + W_{or}r_{t-1} + W_{oc}c_t + b_o)$.....(4)

$h_t = o_t \odot \Phi(c_t)$.....(5)

Equations (1)-(5) are vector formulas that describe the LSTM cell.

Typically, LSTM parameters are initialized to small random values in an interval. However, for the forget gate, this is a suboptimal choice, where small weights effectively close the gate, preventing cell memory and its gradients from flowing in time. In this work this issue has been addressed by initializing the forget gate bias to a large value.

3.2. Bhattacharyya Distance

The Bhattacharyya Distance [25,26] measures the similarity of two discrete or continuous probability distributions which can be expressed by the following expressions.

$$D_B = \frac{1}{8}(\mu_1 - \mu_2)^T \sum(\mu_1 - \mu_2)^{-1} + \frac{1}{2} ln\left(\frac{\det \Sigma}{\sqrt{\det \Sigma_1 \det \Sigma_2}}\right)$$....(6)

where μi and Σi are the means and covariances of the distributions, respectively and

$$\Sigma = \frac{\Sigma_1 + \Sigma_2}{2}$$....(7)

3.3. Mahalanobis Distance

The Mahalanobis distance [26] is a measure of the distance between a point and a distribution, introduced by P. C. Mahalanobis in 1936.

$$D_M(x) = \sqrt{(x - \Sigma)^T S^{-1}(x - \Sigma)}$$....(8)

Where μ and Σ are the means and covariances of the distributions respectively and x is the point whose distance is to be measured. In this work, the Mahalanobis distance between each frame is measured for different phonemes.

4. PERFORMANCE ANALYSIS

The next steps are training the LSTM model using the recorded training set of data for Bangla speech. Three popular feature extraction methods, namely, LPC, MFCC and PLP has been tried while training by 5000 male samples and testing this model using 1000 different male samples. This combination has been chosen carefully as it performed reasonably well in [9].

Statistical distance quantifies the distance between two statistical objects, which can be two random variables, or two probability distributions or samples. The distance can be between an individual sample point and a population or a wider sample of points.

In this work both Bhattacharyya distance and Mahalanobis distance of various phoneme coefficients has been computed for LPC, MFCC and PLP.

Table 1. Bhattacharyya distance between similar sounding phonemes

Phonemes to be compared		**LPC**	**MFCC**	**PLP**
'"r"'	'"er"'	1.268892	1.609877	1.935885
'"b"'	'"v"'	0.602648	1.019706	1.176067
'"b"'	'"bh"'	0.674364	1.99346	2.228276
'"d"'	'"dh"'	0.091483	0.732092	0.630423
'"aa"'	'"ax"'	1.398714	0.257079	0.278304
'"ch"'	'"chh"'	0.477755	0.288733	0.318336
'"ey"'	'"y"'	6.73841	1.88483	2.006171
'"n"'	'"ng"'	1.090365	2.900765	2.830959
'"uh"'	'"ih"'	1.399962	3.502966	3.659523
'"uh"'	'"eh"'	9.00648	3.199748	3.829925
'"uh"'	'"u"'	7.451423	1.100306	1.12853
'"dh"'	'"d"'	0.091483	0.732092	0.630423
'"s"'	'"sh"'	5.384669	0.149458	0.150676
'"v"'	'"bh"'	0.513366	1.427321	1.206649
'"oy"'	'"ay"'	3.930717	1.342863	1.604389
'"hh"'	'"h"'	1.363757	2.696534	2.499212
'"ah"'	'"ax"'	8.392236	0.288995	0.300063
'"ay"'	'"oy"'	3.930717	1.342863	1.604389
'"z"'	'"jh"'	3.993217	0.83873	0.956689
'"z"'	'"j"'	7.540232	0.810661	0.860811
'"zh"'	'"jh"'	0.934331	1.320837	1.200627
'"j"'	'"z"'	7.540232	0.810661	0.860811
'"j"'	'"zh"'	1.583787	1.722148	1.625498
'"j"'	'"jh"'	3.002899	0.511258	0.524364

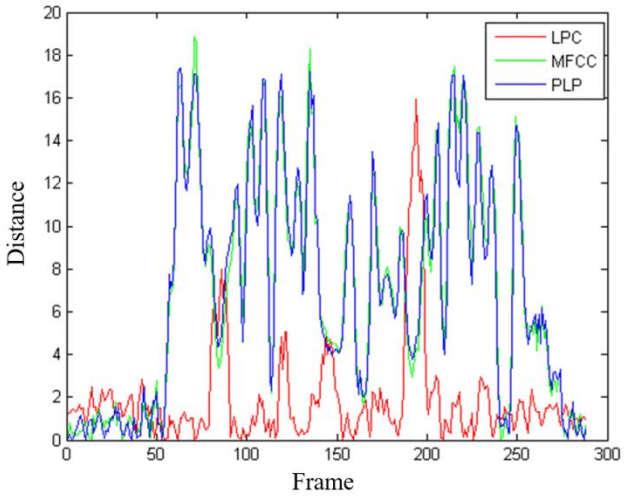

Figure 4. Mahalanobis distance between the phoneme 'sil' and the word "bideshi bonduk amdani"

Table 2. Timeframe of the word "bideshi" in Figure 4.

Frame start	Frame end	Phoneme
0	53	sil
53	63	bi
63	69	d
69	74	ey
74	78	eh
78	89	sh
89	94	ih
94	100	sp

5. DISCUSSION

From Table 1 it is evident that both PLP and LPC show some benefits over MFCC for similar sounding phonemes. However, in most cases LPC shows some abnormally large distances which should not be that large, as these phonemes sound similar.

If the timeframe of the word "bideshi" in Table 2 is matched with Figure 4, both PLP and MFCC manages to distinguish 53th frame from the phoneme 'sil', while LPC detects something on the onset of the phoneme 'sh' at frame no 78.

6. CONCLUSION

From Table 1, Table 2 and Figure 4 it is clear that, PLP and MFCC shows significantly better result than LPC in terms of statistical distance. Combining the findings in [9] it can be said clearly that PLP performs better than MFCC. So for large data set or for large sets of training and testing, PLP seems to be a better alternative. However, RNN and LSTM mainly relies on sequential processing. Which means they are inherently slow. The performance of PLP shines through by combining it with a more advanced neural network like the transformer network [27]. As the transformer network can process faster, PLP should perform more aggressively with this types of framework.

REFERENCES

[1] Uddin, M.T.; Uddiny, M.A. "Human activity recognition from wearable sensors using extremely randomized trees". In Proceedings of the International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, Bangladesh, 13–15 September 2015; pp. 1–6.

[2] Jalal, A. "Human activity recognition using the labelled depth body parts information of depth silhouettes". In Proceedings of the 6th International Symposium on Sustainable Healthy Buildings, Seoul, Korea, 27 February 2012; pp. 1–8.

[3] Ahad, M.A.R.; Kobashi, S.; Tavares, J.M.R. "Advancements of image processing and vision" in healthcare. J. Healthcare Eng. 2018.

[4] Jalal, A.; Quaid, M.A.K.; Hasan, A.S. "Wearable sensor-based human behaviour understanding and recognition in daily life for smart environments". In Proceedings of the International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 18–20 December 2017.

[5] C. Chiu, T. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R.Weiss, K. Rao, E. Gonina, et al. "State-of-the art speech recognition with sequence-to-sequence models". In Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference, pages 4774–4778. IEEE, 2018.

[6] Kanishka Rao, Ha¸sim Sak, and Rohit Prabhavalkar. "Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer". In Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE, pages 193–199. IEEE, 2017.

[7] Eric Battenberg, Jitong Chen, Rewon Child, Adam Coates, Yashesh Gaur Yi Li, Hairong Liu, Sanjeev Satheesh, Anuroop Sriram, and Zhenyao Zhu. "Exploring neural transducers for end-to- end speech recognition". In Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE, pages 206–213. IEEE, 2017.

[8] Kishore, S., Black, A., Kumar, R., and Sangal, R. "Experiments with unit selection speech databases for Indian languages," National Seminar on Language Technology Tools: Implementation of Telugu, Hyderabad, India, October 2003.

[9] Nahyan A. M. "Performance analysis of different acoustic features based on LSTM for Bangla speech recognition." The International Journal of Multimedia & Its Applications (IJMA) Vol.12, No. 1/2/3/4, August 2020, DOI :10.5121/ijma.2020.12402 17

[10] Rafal J., Wojciech Z., and Ilya S. "An empirical exploration of recurrent network architectures". In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, pages 2342–2350, 2015.

[11] Atal, B. S., "Speech analysis and synthesis by linear prediction of the speech wave." The Journal of The Acoustical Society of America 47 (1970) 65.

[12] Davis, S. B and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-28, no. 4, pp. 357 – 366, August 1980.

[13] Alex Graves and Navdeep Jaitly. "Towards end-to-end speech recognition with recurrent neural networks". In Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, pages 1764–1772, 2014.

[14] Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. "Deep speech: scaling up end-to-end speech recognition". CoRR, abs/1412.5567, 2014.

[15] Hagen Soltau, Hank Liao, and Hasim Sak. "Neural speech recognizer: acoustic-to-word LSTM model for large vocabulary speech recognition". CoRR, abs/1610.09975, 2016.

[16] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". In Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006, pages 369–376, 2006.

[17] Sepp Hochreiter and Jurgen Schmidhuber, "Long short-term memory", Neural Computation, vol.9, no.8,pp.1735 780,Nov.1997

[18] Mike Schuster and Kuldip K.Paliwal, "Bidirectional recurrent neural networks," Signal Processing, IEEE Transactions, vol. 45, no. 11, pp.2673-2681,1997.

[19] Abdel Rahman Mohamed, George E. Dah1, and Geoffrey E.Hinton, "Acoustic modeling using deep belief networks," IEEE Transactions on Audio, Speech & Language Processing, vol. 20, no. 1, pp. 14-22, 2012.
[20] George E. Dah1, Dong Yu, Li Deng, and Alex Acero, "Contextdependent pre-trained deep neural networks for large-vocabulary speech recognition," IEEE Transactions on Audio, Speech & Language Processing, vol. 20, no. 1, pp. 30-42, Jan. 2012.
[21] Navedeep Jaitly, Patrick Nguyen, Andrew Senior, and Vincent Vanhoucke, "Application of pre-trained deep neural networks to large vocabulary speech recognition," in Proceedings of INTERSPEECH, 2012.
[22] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "An experiment study on speech enhancement based on deep neural networks", IEEE Signal Processing, vol. 21, no. 1, pp. 65-68, Nov. 2013.
[23] Hasim Sak, Andrew Senior, and Francoise Beaufays, "Long Short-Term memory based recurrent neural network architectures for large vocabulary speech recognition", ArXiv e-prints, Feb.2014.
[24] Z. Chen, Y. Zhuang, Y. Qian, K. Yu, et al. "Phone synchronous speech recognition with CTC lattices" IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 25(1): 90–101, 2017
[25] S. Dubuisson. "The computation of the Bhattacharyya distance between histograms without histograms" 2nd International Conference on Image Processing Theory Tools and Applications (IPTA'10), Jul 2010, Paris, France. pp.373-378,
[26] W. F. Basener and M. Flynn "Microscene evaluation using the Bhattacharyya distance", Proc. SPIE 10780, Multispectral, Hyperspectral, and Ultraspectral Remote Sensing Technology, Techniques and Applications VII, 107800R (23 October 2018); https://doi.org/10.1117/12.2327004
[27] Wang, Alex; Singh, Amanpreet; Michael, Julian; Hill, Felix; Levy, Omer; Bowman, Samuel (2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Stroudsburg, PA, USA: Association for Computational Linguistics: 353–355.

AUTHORS

Nahyan Al Mahmud

Mr. Mahmud graduated from Electrical and Electronic Engineering department of Ahsanullah University of Science and Technology (AUST), Dhaka in 2008. Mr. Mahmud has completed the MSc program (EEE) from Bangladesh University of Engineering & Technology (BUET), Dhaka. Currently he is working as an Assistant Professor of EEE Department in AUST. His research interests include system and signal processing, analysis and design.

Shahfida Amjad Munni

Shahfida Amjad Munni completed her master of engineering degree at the Institute of Information and Communication Technology (IICT) in Bangladesh University of Engineering and Technology (BUET), Bangladesh in March 2018. Currently she is working as a software engineer at Cygnus Innovation Limited, Bangladesh. Her research interests include optical wireless communication, data science, wireless communication, OFDM modulation and LiFi.