Classifying Dysarthria Patients with Long Short-Term Memory Networks

Alex Mayle
School of EECS
Ohio University
am218112@ohio.edu

Abstract—**This paper presents a recurrent neural network architecture for the binary classification of Mandarin speaking individuals into two classes, those who are afflicted with some form of Dysarthria, and those who are not. Specifically, a series of Long Short-Term Memory (LSTM) networks are evaluated on the task using accuracy, and the rate of both false positives and negatives as metrics. A double layer, one directional LSTM is shown to slightly outperform the others, and significantly improve upon a baseline Multi-layer Perceptron employed for the same task. While the results are not indicative of a practical replacement for a medical diagnosis, we show that the LSTM's ability to leverage temporal information from within its inputs makes for an effective step in the pursuit of accessible Dysarthria diagnoses.**

I. Introduction

There are approximately 7 million individuals in China suffering from various speech disabilities. One such disorder, Dysarthria, results in an increased difficulty to articulate phonemes, or that which distinguishes one word from another. The impact of Dysarthria is exacerbated in Mandarin speaking individuals due to the fact that variations in tone have the potential to carry different meanings. Given the amount of Chinese speakers suffering from this particular disease, and the challenges it poses to effective communication, accessible means to a diagnosis is paramount. To this end, we present a collection of Recurrent Neural Network (RNN) architectures capable of discerning those who suffer from Dysarthria when given Mandarin character pronunciations as input.

While there are established medical practices regarding the diagnosis of Dysarthria, such as the Frenchay Dysarthria Assessment [1], such techniques require the patient be physically present and undergo a series of examinations. In contrast, the system presented here increases accessibility by merely relying on speech as input. While our system is not currently robust enough to replace the services of a medical practitioner, it has the potential to provide a less invasive, preliminary step in seeking care.

We present an architecture primarily relying on a Long Short-Term Memory (LSTM) network [2] and evaluate it using three variants of the model. While we do not know of any studies whose methods are directly comparable, the results suggest recurrent neural networks, and LSTM networks in particular, may be a fruitful method for Dysarthria classification.

II. Model

Given an audio clip containing the pronunciation of a Mandarin character X from a single speaker, the model is to produce a label Y indicating whether or not the speaker suffers from Dysarthria. We refer to a positive result as a diagnosis of Dysarthria. The raw waveform X is first transformed into an MFCC feature vector $X' = \{x_1, x_2, ...x_t\}$, where t is dependent on the length of X. Each element of X' is then fed into a LSTM network sequentially. After x_t has been input, the LSTM network produces the vector h_t, which is then used as input to logistic regression. Finally, the regression layer outputs a label Y. Figure 1 illustrates a single training example's path through the network.

A. Pre-processing

We began by transforming the raw audio X into an MFCC feature vector X' using a sliding window of 25 milliseconds and a 10 millisecond stride. Each MFCC in $X' = \{x_1, x_2, ...x_t\}$ consists of 13 coefficients $x_i = \{\theta_1 ... \theta_{13}\}$. In practice, the network was trained on many such inputs X. These were collected and normalized such that the kth coefficient θ_k had zero-mean and unit variance with respect to θ_k across all training examples. The validation and testing set were processed in the same way.

Since each input X' contains a varying amount of MFCC's, each mini-batch fed into the network is 0-padded such that each X' in the mini-batch has the same length t_{max}, where t_{max} is the largest t value in the mini-batch. However, we do keep track of the lengths of each X' and instruct the LSTM network not to process the padded portions of each input X'. That is, the LSTM runs t time steps for each input $X' = \{x_1, x_2, ...x_t, 0_{t+1}, ..., 0_{t_{max}}\}$ in the mini-batch.

B. LSTM Architecture

After pre-processing, X' is then fed into a LSTM network. As noted, because X' is a time-series of MFCC's, we do not input them concurrently. Instead, we feed in one MFCC each time-step. The network produces an output h_i at each time step, but only the last output h_t is used as input to the logistic regression layer. The model is implemented to copy the output from the last non-zero-padded input h_t and copy it to $h_{t_{max}}$. In doing so, we guarantee that the output sent to the logistic regression layer is not affected by the zero-padding.

We experimented with several variants of the LSTM model, including adding layers and using a bidirectional LSTM. For the models with one layer, $L2$ regularization was used. The bidirection model employed dropout regularization between the two LSTM layers and between the output of the LSTM and the input to logistic regression. Dropout is never applied

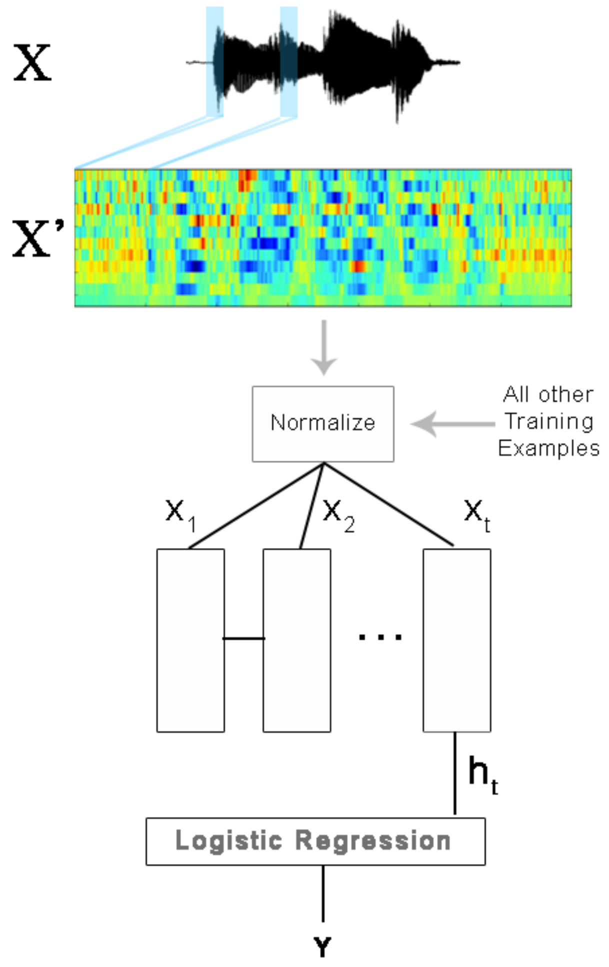

Fig. 1. A single training example, X, as it flows through the network: The input is first converted to X', a variable length MFCC vector of length t. It is then normalized with respect to the entire training dataset. Finally, each MFCC in X' is input to the LSTM network sequentially. The final output h_t is fed to the logistic regression layer to produce the label Y, indicating whether or not the speaker is afflicted with Dysarthria.

applied between the time-steps, as suggested by Zaremba et al [3].

The standard LSTM model is able to use inputs from arbitrarily distant time steps to change its output at the current time step [4]. However, it cannot use information from subsequent time steps to affect previous ones. Bidirectional LSTM networks overcome this limitation by performing two concurrent passes on the data. One pass starts from step 0 to t as normal, while another pass starts at time step t and ends at step 0. Each pass produces an output, which we handle by concatenating them together and feeding them to the logistic regression layer.

III. Experiments

A. Data

The data consists of solely Mandarin speaking individuals divided into four groups, as depicted in Table I. These were first grouped and shuffled, then split into a training, validation, and test set with a ratio of 2:1:1. Each data point represents a spoken Mandarin character recorded under the supervision of a medical professional.

TABLE I. Data Set

	Female	Male	Ratio
Healthy	1600	1605	53.4%
Patient	1001	1792	46.6%
Total	2601	3397	100%

B. Methodology

All models were trained using Adam gradient descent [5] to minimize the cross entropy between the predictions of the network and the ground truth provided by the medical practitioners who collected the data. Training occurred for 40 epochs (with early stopping) on mini-batches of size 32.

We employ the accuracy, precision, and recall [6] as metrics to judge the performance of each model. Precision and recall are considered due to the medical nature of these experiments. That is, most people do not suffer from Dysarthria, but it is the instances in which one does that are important to classify correctly. We therefore define a metric to measure the chance that a Dysarthria patient will receive a negative result $FN = 1 - recall$. In an analogous fashion, we define a metric for false-positives $FP = 1 - precision$. Because individuals who receive a negative prediction (i.e., they do not suffer from Dysarthria) are less likely to seek a second opinion, we are especially interested in the minimization of FN.

Of course, we could hard code the model to output positive predictions for every input, thus achieving $FN = 0$; however, this medical test would certainly be without merit. Alternatively, we implemented early stopping [7] using the combined F1 score metric, $F1 = 2 * (precision * recall)/(precision + recall)$. Specifically, after each training epoch j, $F1_j$ is evaluated on the validation set. Further, the maximum value seen thus far, $F1_{max}$, is stored in memory. If the ratio between $F1_{max}$ and $F1_j$ is greater than a threshold a, a five epoch grace period is given to see if the F1 score rebounds. If it does not, the training is cut short and the weights that achieved $F1_{max}$ on the validation set are used.

C. Experiments

TABLE II. Experimental results

	Accuracy	FP	FN
Baseline	80.1%	-	-
LSTM-1	88.7%	11.5%	8.8%
LSTM-2	88.7%	12.0%	8.3%
Bi-LSTM-1	87.8%	13.4%	8.2%

In total, four models are tested, including a baseline model for comparison. A value of 1.25% was used for the early stopping parameter a. The LSTM models use a hidden state of size 200, resulting in a total of 800 parameters. These are initialized to a truncated normal distribution with $\mu = 0$ and $\sigma = 0.5$. Values generated beyond two standard deviations are discarded and re-picked. The logistic regression layer is also of size 200, except in the case of the bidirectional LSTM, where it is of size 400 to accommodate the concatenation of the two passes' outputs.

1) Baseline: multi-layer Perceptron (MLP) consisting of one hidden layer that is fed into the logistic regression classifier.
2) LSTM-1: Single layer, one-directional LSTM starting from time step 0
3) LSTM-2: Double layer, one-directional LSTM starting from time step 0
4) Bi-LSTM-1: Single layer, bi-directional LSTM involving two concurrent passes of the training example. One starting from time step 0, and the other starting at time step t.

Table II depicts the results for each experiment. LSTM-1 and LSTM-2 achieve similar performance, clearly beating the baseline and making a marginal improvement upon the Bi-LSTM-1 model.

D. Discussion

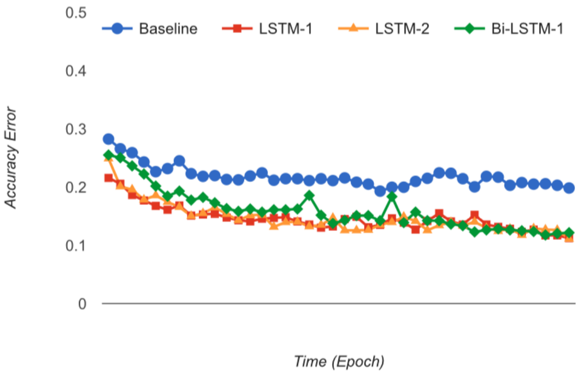

Fig. 2. Error rate over the 40 epochs of training. The three LSTM models start training with slightly different error rates, but all eventually converge to a rate superior to the baseline.

Figure 2 shows the accuracy convergence of each model. The LSTM's clearly outperform the baseline; however, increasing the expressive capacity through adding layers and passes over the data did not result in significant performance improvements. LSTM-1 and LSTM-2 achieved the same accuracy, with the latter improving upon the false negatives metric. Because adding a layer resulted in the same performance, and adding another pass was slightly detrimental, we did not consider a two-layer bidirectional LSTM model.

LSTM-2's 88.7% accuracy and 8.2% false negative rate constitute a promising attempt at classifying Dysarthria among both afflicted and healthy speakers, but it is not medically reliable. Of course, in such a setting, an accuracy of 99.9% is desired. While this is certainly hard to achieve and a slightly lower performance may be acceptable in practice, the results of these experiments are not. Specifically, the rate of false-negatives must be decreased by an order of magnitude.

We also observed that each network is not prone to over-fitting. While we implemented early stopping, it never triggered training to end early. Furthermore, neither $L!$, $L2$, or dropout regularization significantly improved performance, with the latter being detrimental to LSTM-1 and Bi-LSTM-1.

Instead of a bottleneck in the network, it may be the data that prevents an increase in performance. For example, a medical doctor would never diagnose someone with Dysarthria based off of how a patient pronounces a single character. If the input to the network was instead a series of speech from the the same speaker, the network may be given a better opportunity to discern the condition of the speaker. Moreover, the effects of Dysarthria may not present themselves to the same degree across different character pronunciations. If the input to the network is one with little variation between healthy and afflicted individuals, the network will be at a disadvantage to make an accurate diagnosis.

IV. Related Work

Carmichael et al [8] employed a multilayer perceptron to classify the different forms of Dysarthria using human speech. Unlike our work, however, the network inputs are assumed to come from a distribution of people known to have some form of Dysarthria. Prior to this, an effort was made to classify speakers into one of the categories of Dysarthria using the Frenchay Dysarthria Assessement of each patient as input [1] [9]. The more advanced topic of recognizing speech produced from someone with Dysarthria using RNN networks has also been investigated recently [10] [11].

V. Future Work

ZCA whitening is commonly employed as a preprocessing step in many audio classification tasks [12] [13] [14]. Unfortunately, this proved to be intractable on our implementation machine as the co-variance matrix of the data does not fit in memory. Given the ease of computing the transformation on an appropriate machine, it is a compelling next step in an effort to improve performance. Another technique, batch normalization has also been shown to improve performance [15]. While we considered implementing this, the level of abstraction with which we define the LSTM model did not provide the necessary level of granularity.

Barring minor improvements made possible by more foresight, we consider architectural additions which may increase performance. As already mentioned, changing the input to a series of character pronunciations from the same speaker may give the network an increased ability to classify Dysarthria patients. Another option is to use a recursive network structure similar to the one employed in [8]. For example, one network may classify the speakers' gender or rate of speech first, providing more information to the next layer to use. This pattern would culminate in a final Dysarthria classification layer.

VI. Conclusion

This paper investigates the effectiveness of LSTM networks in the classification of Dysarthria among both afflicted and healthy Mandarin speakers. When presented with a single character pronunciation, we found that single and double layer, one-directional LSTM networks slightly outperform their bidirectional single layer counterpart. Further, the double-layer LSTM regularized with dropout between layers exhibit an improvement in the rate of false negatives. While these methods are not yet practical as a standalone medical test, they do suggest that LSTM networks may provide a fruitful avenue for the realization of autonomous Dysarthria classification.

References

[1] P Enderby. Frenchay dysarthria assessment. *British Journal of Disorders of Communication*, 15(3):165–173, 1980.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[3] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *CoRR*, abs/1409.2329, 2014.

[4] Felix A Gers, Nicol N Schraudolph, and Jürgen Schmidhuber. Learning precise timing with lstm recurrent networks. *Journal of machine learning research*, 3(Aug):115–143, 2002.

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[6] Luis Torgo and Rita Ribeiro. Precision and recall for regression. In *International Conference on Discovery Science*, pages 332–346. Springer, 2009.

[7] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

[8] James Carmichael, Vincent Wan, and Phil D Green. Combining neural network and rule-based systems for dysarthria diagnosis. In *INTERSPEECH*, pages 2226–2229, 2008.

[9] James N Carmichael. *Introducing objective acoustic metrics for the Frenchay Dysarthria Assessment procedure*. University of Sheffield, 2007.

[10] S Selva Nidhyananthan, V Shenbagalakshmi, et al. Assessment of dysarthric speech using elman back propagation network (recurrent network) for speech recognition. *International Journal of Speech Technology*, 19(3):577–583, 2016.

[11] Cristina España-Bonet and José AR Fonollosa. Automatic speech recognition with deep neural networks for impaired speech. In *Advances in Speech and Language Technologies for Iberian Languages: Third International Conference, IberSPEECH 2016, Lisbon, Portugal, November 23-25, 2016, Proceedings 3*, pages 97–107. Springer, 2016.

[12] Youngjune L Gwon, William M Campbell, Douglas Sturim, and HT Kung. Language recognition via sparse coding. In *INTERSPEECH*, 2016.

[13] Charles Chen, Razvan Bunescu, Li Xu, and Chang Liu. Tone classification in mandarin chinese using convolutional neural networks. *Interspeech 2016*, pages 2150–2154, 2016.

[14] Oriol Vinyals and Li Deng. Are sparse representations rich enough for acoustic modeling? In *INTERSPEECH*, pages 2570–2573, 2012.

[15] Tim Cooijmans, Nicolas Ballas, César Laurent, and Aaron C. Courville. Recurrent batch normalization. *CoRR*, abs/1603.09025, 2016.