Diagnosing Dysarthria with Long Short-Term Memory Networks

Alex Mayle[1], *Zhiwei Mou*[2], *Razvan Bunescu*[1], *Sadegh Mirshekarian*[1], *Li Xu*[3], *and Chang Liu*[1]

[1]School of Electrical Engineering and Computer Science, Ohio University, Athens, OH, USA
[2]Department of Rehabilitation, First Affiliated Hospital of Jinan University, Guangzhou, China
[3]School of Rehabilitation and Communication Sciences, Ohio University, Athens, OH, USA

am218112@ohio.edu, mouzhiwei@jnu.edu.cn, bunescu,sm774113,xul,liuc@ohio.edu

Abstract

This paper proposes the use of Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units for determining whether Mandarin-speaking individuals are afflicted with a form of Dysarthria based on samples of syllable pronunciations. Several LSTM network architectures are evaluated on this binary classification task, using accuracy and Receiver Operating Characteristic (ROC) curves as metrics. The LSTM models are shown to significantly improve upon a baseline fully connected network, reaching over 90% area under the ROC curve on the task of classifying new speakers, when a sufficient number of cepstrum coefficients are used. The results show that the LSTM's ability to leverage temporal information within its input makes for an effective step in the pursuit of accessible Dysarthria diagnoses.

Index Terms: Dysarthria, RNN, LSTM, speech processing

1. 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, due to neurological injuries that cause impaired or uncoordinated movement of the muscles, including the lips, tongue, lower jaw, velum, vocal folds, and diaphragm during speech production. The impact of Dysarthria is exacerbated in Mandarin-speaking individuals because Mandarin Chinese is a tone language in which variations in tone at syllable level carry lexical meaning.

With the aging population increase, the number of people with Dysarthria will continue to grow. Given the challenges it poses to effective communication, accessible means to diagnosis is paramount. Currently, there are two main categories of Dysarthria assessment: *subjective* approaches and *objective* approaches. The most common used assessments in recent rehabilitation practice and speech rehabilitation institutions are still those based on subjective auditory perception and/or subjective scales, with poor objectivity and stability. Objective assessment methods include oro-pharyngeal physical examination and electroglottography examination. These and other types of examinations however have unsatisfactory compliance of patients. Patients with Dysarthria may also turn to neurology departments and speech rehabilitation institutions, however, the lack of interdisciplinary coordination results in incomplete and subjective examinations, causing low consistency among hospitals and institutions. In China, this problem is compounded by the insufficient number of professional speech therapists. The current estimated number of speech therapists in China is less than 10,000 [1] whereas the demand for such professionals in a country of a population of 1.38 billion is 368,350, according to Enderby and Davies [2]. To release the pressure caused by the increase in the number of patients with speech disorders, and the paucity of professional speech therapists, an objective and accurate method for identifying individuals with dysarthria is deemed timely and necessary. To this end, we present a Long Short-Term Memory (LSTM) network architecture trained to identify those who suffer from Dysarthria, given Mandarin syllable pronunciations as input. Most established medical practices regarding the diagnosis of Dysarthria, such as the Frenchay Dysarthria Assessment (FDA) [3], 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 it is doubtful that such a system can completely replace diagnosis by a medical practitioner, it has the potential to provide a more accessible, less invasive, initial step in seeking care.

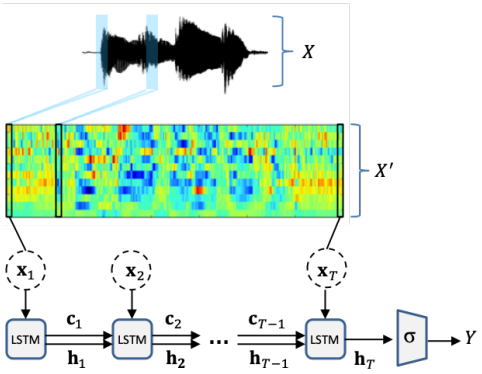

Figure 1: *From a waveform example X to its classification Y in the proposed model architecture.*

2. Model Architecture

Given an audio clip X, containing the pronunciation of a Mandarin syllable, the model is to produce a label Y, which is positive if and only if the speaker suffers from Dysarthria. Figure 1 illustrates a single training example's path through the proposed processing architecture.

The raw waveform X is first pre-processed into a mel-frequency cepstral coefficients (MFCC) feature vector $X' = \{\mathbf{x}_1, ..., \mathbf{x}_t, ..., \mathbf{x}_T\}$, where the number of MFCC frames T can be different from one raw input X to another. The MFCC vectors $\mathbf{x}_t$ were created using a sliding window of 25 milliseconds with a 10 millisecond stride. Each MFCC vector consists of N coefficients $\mathbf{x}_t = \{\theta_1, ..., \theta_n, ..., \theta_N\}$, where $N = 13$ unless explicitly stated. These were collected and normalized such that each coefficient θ_n had zero mean and unit variance across all training examples. An LSTM is run over the MFCC sequence

Table 1: *The distribution of positive (with Dysarthria) and negative (no Dysarthria) individuals and syllables (shown within brackets) in the dataset.*

	Ratio	Female	Male
Positive	46.6%	12 [*1001*]	19 [*1792*]
Negative	53.4%	19 [*1600*]	19 [*1605*]
Total	100%	31 [*2601*]	38 [*3397*]

X' and the last output $\mathbf{h}_T$ is provided to the logistic regression layer.

We experimented with several variants of the LSTM model, including adding layers and using a bidirectional LSTMs. For the models with one layer, L_2 regularization was used. The two-layer model employed dropout [4, 5] between the LSTM layers, as well as between the last LSTM layer and logistic regression. Bidirectional LSTM networks perform two concurrent passes on the data, left to right and right to left. The output vectors produced by the two passes are concatenated and fed to the logistic regression layer.

3. Evaluation Methodology

The evaluation data consists of samples of syllables recorded from 69 Mandarin speaking adults, 38 male and 31 female. The number of individuals in each class is presented in Table 1, together with the corresponding number of recorded syllables. The participants were from Jinan University School of Medicine, who included 31 native Mandarin-speaking patients (19 males and 12 females) with post-stroke Dysarthria. The age of the dysarthric speakers ranged from 25 to 83 years old [mean $\pm$ SD: 56.74 $\pm$ 16.40 years]. All participants went through physical examination, Frenchay Dysarthria Assessment, and other auxiliary examinations (such as brain CT, MRI). Before the stroke occurred, all patients had no speech-related impairments and were able to communicate fluently in Mandarin. They had no alexia, visual, or severe auditory comprehension impairments, and had pure-tone thresholds at 500, 1000, and 2000 Hz of $\leq$ 25 dB HL in at least one ear. The control group included 38 healthy adults (HA) (19 males and 19 females) in a similar age range (21 to 76 years old; mean $\pm$ SD: 45.89 $\pm$ 13.02 years). Some of the family members of the Dysarthria groups were recruited into the HA group. They all had pure-tone thresholds at 500, 1000, and 2000 Hz of $\leq$ 25 dB HL in at least one ear with no reported hearing or speech disorders. More details about the demographical information of the participants and the acoustic properties of the speech samples can be found in [6]. Informed consent was obtained from all participants. All research was performed in accordance with relevant guidelines and regulations.

Note that although the positive or negative labels are originally assigned only to speakers, the models are trained and tested using syllables as input. To assign labels to syllables during training, we propagate the speaker label to all the syllables recorded from that speaker. This is bound to introduce label noise in the positive labeled syllables, as not all syllables from a speaker afflicted with Dysarthria exhibit abnormal speech production. If a speaker is seen to correspond to a bag of syllables, the problem corresponds to a multiple instance learning (MIL) setting [7] [8]. In this paper, we use the simple MIL approach of projecting bag labels to all syllable instances in the bag, leaving the use of more sophisticated MIL methods for future work.

The dataset was used for the training and evaluation of four models, as follows:

1. **Baseline**: A fully connected feedforward neural network with one hidden layer.
2. **LSTM-1**: Single layer, unidirectional LSTM.
3. **LSTM-2**: Double layer, unidirectional LSTM.
4. **BiLSTM-1**: Single layer, bidirectional LSTM.

All models use a hidden layer size of 200. They are trained using Adam [9] for 40 epochs on mini-batches of size 64. The training objective is formulated as the syllable-level cross-entropy loss between the predictions and the ground truth provided by the medical practitioners who collected the data. While only the individuals were manually labeled as having Dysarthria (positive) or not (negative), the label for each individual was also assigned to all the syllables coming from that individual, and the cross-entropy loss was formulated using syllables as examples.

To prevent overfitting, a form of early stopping was employed, where training is stopped when the ratio of the current validation error ϵ_{curr} to the lowest error seen thus far ϵ_{min} exceeds a threshold $1 + \alpha$, i.e. $\epsilon_{curr}/\epsilon_{min} > 1 + \alpha$, where $\alpha = 0.075$. A grace period is used, such that training is only stopped if the threshold is met for 5 epochs in a row.

Table 2: *The syllable-level performance of baseline and LSTM models in experiment I (known speakers).*

	Accuracy	Precision	Recall	F-measure
Baseline	79.0	79.5	82.9	81.2
LSTM-1	**88.7**	88.5	81.2	**84.7**
LSTM-2	**88.7**	88.0	81.7	**84.7**
Bi-LSTM-1	87.8	86.6	81.8	84.1

4. Experiment I: Known Speakers

We first compared the LSTM models against the baseline fully connected network on the relatively easier task of non-novel speakers. Here, the entire set of syllables from the dataset is shuffled and partitioned into the training, testing, and validation sets using a number of syllables ratio of 2:1:1, respectively. Since the dataset is partitioned at syllable-level, it is possible for a patient to have their syllables partitioned among the training, validation, and test sets. Thus, syllables that appear at test time may come from patients that have been observed at training time.

We employed the syllable-level *accuracy*, *precision*, and *recall* as metrics to judge the performance of each model [10]. Accuracy is the percentage of correct syllable labels predicted by the system. Precision and recall were also considered due to the medical nature of our 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. Precision is the percentage of correct positive predictions (true positives) out of all the positive predictions (true positives + false positives). Recall is the percentage of correct positive predictions out of all the positive examples (true positives + false negatives). Because individuals who receive a negative prediction (i.e., who do not suffer from Dysarthria) are less likely to seek a second opinion, we are especially interested in a higher recall.

The baseline fully connected model obtains 79.0% accuracy, which is a significant improvement over the 53.4% of the majority classifier. Table 2 also shows the results for each

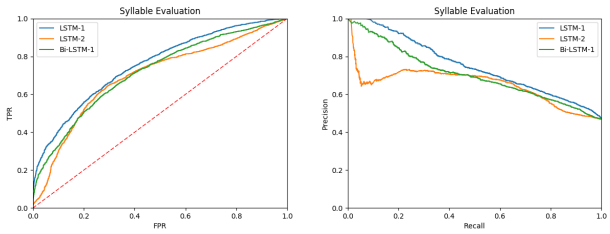

Figure 2: *Receiver-operating characteristic (ROC) and precision-recall (PR) curves for syllable-level classification.*

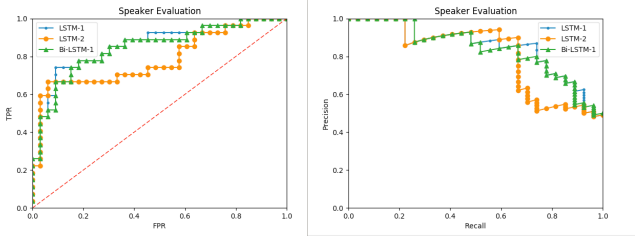

Figure 3: *Receiver-operating characteristic (ROC) and precision-recall (PR) curves for speaker-level evaluation.*

LSTM model. LSTM-1 and LSTM-2 achieve similar performance, outperforming the baseline and making a marginal improvement upon the Bi-LSTM-1 model.

5. Experiment II: Novel Speakers

While LSTM models were shown to outperform the baseline on the task of classifying syllables from known speakers, we are also interested in their performace on novel speakers. In the experiment from Section 4, the training set and the test set may contain syllables from the same speaker. To more accurately match the application to novel speakers, in the experiment from this section the training and test sets were created by partitioning the set of speakers. As such, an individual's syllables appear either in the test or in the training set, but not in both.

Because the number of speakers in the dataset is relatively small, we opted to evaluate the models using 10-fold cross validation. We randomly sampled 9 of the 69 speakers to use as a validation set. The remaining 60 were partitioned into 10 groups of 6. In each of the 10 evaluation rounds, the models were trained on 54 speakers and tested on a different group of 6 novel speakers. The trained models are then evaluated in two scenarios: *syllable-level* and *speaker-level* classification.

5.1. Syllable-Level Evaluation

In the syllable-level classification, the trained models are evaluated by how good they are at classifying syllables from the speakers in the test set. This is similar to the experiment from Section 4, except that now the test syllables are now coming from novel speakers. Figure 2 shows the receiver-operating characteristic (ROC) and precision-recall (PR) curves. To gain perspective on the ROC behavior, we consider a model which produces a positive classification with probability p. The majority classifier can be seen as the extreme case, where $p = 1$. As p increases from 0 to 1, the red ROC line in Figure 2 is produced, with an area under the curve (AUC) of 0.5. The three LSTM models clearly improve upon this baseline for all three methods of inference, with LSTM-1 edging out the other two models. In terms of AUC, the 3 systems obtain the following scores: 75.4 for LSTM-1, 69.5 for LSTM-2, and 71.4 for Bi-LSTM-1.

5.2. Speaker-Level Evaluation

In the speaker-level classification, we take the logistic regression outputs for all syllables belonging to a test speaker and aggregate them into a single probability score that can be used to classify the speaker. To achieve this, we investigated two aggregation methods: *soft-majority* and a normalized version of *noisy-OR*. Given a speaker with m syllables, let σ_k be the the logistic regression output for the k-th syllable of that speaker. Then the two aggregation methods compute the speaker-level probability as follows:

$$\textit{soft-majority} = \frac{1}{m}\sum_{k=1}^{m}\sigma_k \tag{1}$$

$$\textit{noisy-OR} = 1 - \frac{1}{m}\sum_{k=1}^{m}\log\left(1-\sigma_k\right) \tag{2}$$

Because the traditional noisy-OR calculation would be affected by the number of syllables each speaker has, we computed it in log-space and normalized the probability of the negative class by the number of that speaker's syllables. This allowed us to directly compare the noisy-OR scores between speakers with a varying number of examples.

Figure 3 presents the receiver-operating characteristic (ROC) and precision-recall (PR) curves for the soft-majority method of inference. The noisy-OR method produced identical results, therefore its curves are not shown. The LSTM-1 model again obtains the best results. The first line in Table 3 shows the performance of the three LSTM models in terms of the area under the ROC curve (AUC). The speaker-level performance is higher than at syllable level, likely because not all syllables from speakers afflicted with Dysarthria exhibit abnormal speech production. Because an accurate diagnosis cannot be expected to result from a single syllable, the speaker-level method is therefore more appropriate for practical purposes.

Table 3: *AUC scores: speaker-level vs. syllable-level.*

	LSTM-1	LSTM-2	Bi-LSTM-1
Speaker-level	85.4	78.5	84.7
Syllable-level	75.4	65.9	71.4

5.3. Effects of Syllables Types on Classification Accuracy

There are three types of syllables in the dataset: (1) syllables with monophthongs, (2) syllables with compound vowels, and (3) syllables with consonant-/a/. In order to evaluate the classification accuracy based on various types of syllables, we created three combinations of the syllable dataset. The first combination (no prefix) consists of all three types of syllables. The second combination (prefixed with 'cv-') does not contain syllables with compound vowels. The third combination (prefixed with 'c/a/-') does not contain syllables with consonant-/a/. In this experiment, we test how well the models perform when they are trained on the three different combinations of syllable types.

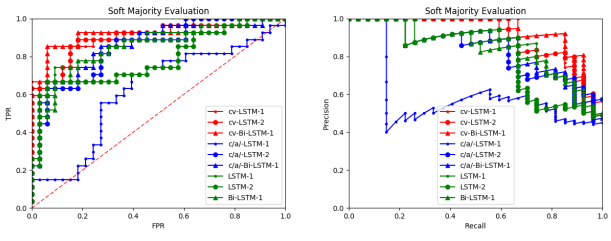

Figure 4: *Receiver-operating characteristic (ROC) and precision-recall (PR) curves using all syllables vs. without syllables with compound vowels (prefixed with 'cv-') vs. without syllables with consonant-/a/ (prefixed with 'c/a/-').*

Figure 4 shows the receiver-operating characteristic (ROC) and precision-recall (PR) curves for the three LSTM models and the three different combinations of syllable types. The corresponding AUC scores are shown in Table 4.

Table 4: *Speaker-level AUC scores over all syllables (All) vs. without syllables with compound vowels (No 'cv') vs. without syllables with consonant-/a/ (No 'c/a/').*

	All	No 'cv'	No 'c/a/'
LSTM-1	85.4	**92.3**	62.1
LSTM-2	78.5	**90.2**	84.5
Bi-LSTM-1	84.7	**92.0**	85.9

The models performed significantly better according to their AUC scores when syllables with compound vowels were removed. All three models scored above 90% when trained without this type of syllables. Therefore, it may be a reasonable heuristic to not include syllables with compound vowels when diagnosing a Dysarthria patient. This intuitively follows from the observation that, even for healthy speakers, these syllables are more difficult to produce and variability of their acoustic properties is greater than syllables with monophthongs.

5.4. Varying the Number of Cepstrum Coefficients

While including $N = 13$ cepstrum coefficients in each feature has produced promising results, there may still be room for improvement by adding more coefficients. To this end, three 10-fold cross-validation evaluations were conducted in the same manner as before, with $N = 13$, 19, and 25 coefficients used as input, respectively. When less than 25 cepstrum coefficients are used, they are taken starting from the cepstrum with the lowest quefrency. Table 5 shows the speaker-level AUC scores for the three LSTM models using the soft-majority inference method.

Table 5: *Speaker-level AUC scores for different numbers of cepstrum coefficients.*

	$N = 13$	$N = 19$	$N = 25$
LSTM-1	85.4	**90.1**	81.7
LSTM-2	78.5	**88.2**	87.1
Bi-LSTM-1	84.7	88.4	**90.4**

Adding more cepstrum coefficients leads to substantial improvements in the performance of BiLSTM-1, matching LSTM-1's best performance. LSTM-1 and LSTM-2 have a similar behavior in the sense that their maximum performance is achieved for 19 coefficients. When all 25 coefficients are used, their performance decreases, which could be due to a lack of capacity.

6. Related Work

Carmichael et al. [11] employed multilayer perceptrons and decision trees to classify the different forms of Dysarthria, using as input a computerised Frenchay Dysarthria Assessement (CFDA) profile, essentially a vector of articulatory dysfunction values measured using acoustic signal processing techniques. Unlike our work, however, the system is trained and tested on a distribution of English-speaking people already 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 a manual Frenchay Dysarthria Assessement of each patient as input [3][12]. The more advanced topic of recognizing speech produced from someone with Dysarthria using RNN networks has also been investigated recently for English speaking individuals, using Elman recurrent neural networks in [13] and a hybrid deep neural network – hidden Markov model (DNN-HMM) architecture in [14]. Wu et al. [15] presented a personalized model adaptation for automatic speech recognition (ASR) targeted at Mandarin-speaking individuals afflicted with articulation disorders due to mild-to-moderate hearing impairment.

7. Conclusion and Future Work

This paper investigated the effectiveness of three LSTM networks, two uni-directional and one bi-directional, for the task of Dysarthria diagnosis based on recordings of syllables from both afflicted and healthy Mandarin speakers. In the first experiment, all LSTM architectures outperformed a fully connected baseline when evaluated using syllable-level accuracy, with the bi-directional variant slightly trailing the uni-directional variants. The second experiment assumes the test syllables come from novel speakers, and evaluates the three LSTM models at both syllable-level and speaker-level. When the syllables with compound vowels are removed from the dataset, all models obtain over 90% AUC. Furthermore, we found that the LSTM models' performance could be improved by increasing the number of cepstrum coefficients. While these methods may not be yet practical as a stand-alone medical test, they do suggest that LSTM networks may provide a fruitful avenue for the realization of autonomous Dysarthria diagnosis.

ZCA whitening is employed as a pre-processing step in many audio classification tasks [16][17][18], as such it is a compelling next step in an effort to improve performance. CNNs can often performed competitively on sequence processing tasks [19], therefore we plan to comparatively evaluate CNNs and long-term recurrent CNNs [20], as well as dilated RNNs [21]. The model presented in [22] takes features much closer to the raw waveform when compared to MFCCs. Applying this approach to Dysarthria classification may also prove to be effective. The type of training data available for speaker classification falls under the multiple instance learning (MIL) setting [7, 23]. Correspondingly, we plan to use LSTMs with models that are specifically designed for the MIL setting.

8. Acknowledgements

This study was supported in part by the NIH NIDCD Grant No. R15-DC014587.

9. References

[1] S. Li, *Speech Therapy*. People's Health Press, 2013.

[2] P. Enderby and P. Davies, "Communication disorders: Planning a service to meet the needs," *International Journal of Language and Communication Disorders*, vol. 4, no. 3, pp. 301–331, 1989.

[3] P. Enderby, "Frenchay Dysarthria assessment," *British Journal of Disorders of Communication*, vol. 15, no. 3, pp. 165–173, 1980.

[4] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2627435.2670313

[5] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16, 2016, pp. 1027–1035. [Online]. Available: http://dl.acm.org/citation.cfm?id=3157096.3157211

[6] Z. Mou, Z. Chen, J. Yang, and L. Xu, "Acoustic properties of vowel production in Mandarin-speaking patients with post-stroke Dysarthria," *Scientific Reports*, vol. 8, no. 14188, 2018.

[7] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, Jan. 1997.

[8] S. Ray, S. Scott, and H. Blockeel, "Multiple-instance learning," in *Encyclopedia of Machine Learning and Data Mining*, 2017, pp. 882–892.

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

[10] L. Torgo and R. Ribeiro, "Precision and recall for regression," in *International Conference on Discovery Science*. Springer, 2009, pp. 332–346.

[11] J. Carmichael, V. Wan, and P. D. Green, "Combining neural network and rule-based systems for Dysarthria diagnosis," in *Interspeech*, 2008, pp. 2226–2229.

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

[13] S. S. Nidhyananthan, R. S. S. Kumari, and V. Shenbagalakshmi, "Assessment of Dysarthric speech using Elman back propagation network (recurrent network) for speech recognition," *International Journal of Speech Technology*, vol. 19, no. 3, pp. 577–583, 2016.

[14] C. España-Bonet and J. A. 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*. Springer, 2016, pp. 97–107.

[15] C.-H. Wu, H.-Y. Su, and H.-P. Shen, "Articulation-disordered speech recognition using speaker-adaptive acoustic models and personalized articulation patterns," *ACM Transactions on Asian Language Information Processing*, vol. 10, no. 2, pp. 7:1–7:19, Jun. 2011.

[16] Y. L. Gwon, W. M. Campbell, D. Sturim, and H. Kung, "Language recognition via sparse coding," in *Interspeech*, 2016.

[17] C. Chen, R. Bunescu, L. Xu, and C. Liu, "Tone classification in Mandarin Chinese using convolutional neural networks," *Interspeech 2016*, pp. 2150–2154, 2016.

[18] O. Vinyals and L. Deng, "Are sparse representations rich enough for acoustic modeling?" in *Interspeech*, 2012, pp. 2570–2573.

[19] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of cnn and rnn for natural language processing," *CoRR*, vol. abs/1702.01923, 2017.

[20] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, Apr. 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2599174

[21] S. Chang, Y. Zhang, W. Han, M. Yu, X. Guo, W. Tan, X. Cui, M. Witbrock, M. A. Hasegawa-Johnson, and T. S. Huang, "Dilated recurrent neural networks," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 77–87. [Online]. Available: http://papers.nips.cc/paper/6613-dilated-recurrent-neural-networks.pdf

[22] J. Lee, J. Park, K. L. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," in *Proceedings of the 14th Sound and Music Computing Conference (SMC)*, Espoo, Finland, 2017. [Online]. Available: smc2017.aalto.fi/media/materials/proceedings/SMC17_p220.pdf

[23] S. Ray, S. Scott, and H. Blockeel, "Multiple-instance learning," *Encyclopedia of Machine Learning and Data Mining*, pp. 1–13, 2014.