ONLINE SPEAKING RATE ESTIMATION USING RECURRENT NEURAL NETWORKS

Yishan Jiao[1]*, Ming Tu*[1]*, Visar Berisha*[1,2] *and Julie Liss*[1]

[1]Department of Speech and Hearing Science
[2] School of Electrical, Computer, and Energy Engineering
Arizona State University

ABSTRACT

A reliable online speaking rate estimation tool is useful in many domains, including speech recognition, speech therapy intervention, speaker identification, etc. This paper proposes an online speaking rate estimation model based on recurrent neural networks (RNNs). Speaking rate is a long-term feature of speech, which depends on how many syllables were spoken over an extended time window (seconds). We posit that since RNNs can capture long-term dependencies through the memory of previous hidden states, they are a good match for the speaking rate estimation task. Here we train a long short-term memory (LSTM) RNN on a set of speech features that are known to correlate with speech rhythm. An evaluation on spontaneous speech shows that the method yields a higher correlation between the estimated rate and the ground-truth rate when compared to the state-of-the-art alternatives. The evaluation on longitudinal pathological speech shows that the proposed method can capture long-term and short-term changes in speaking rate.

Index Terms— recurrent neural networks, speaking rate estimation, clinical tool

1. INTRODUCTION

Speaking rate is an important quantity in automatic speech recognition [1], speaker identification [2], speech modification [3], emotion recognition [4], etc. It is also considered as an index of the efficiency of articulatory movements over time in applications involving dysarthric speech [5, 6]. As a result, many intervention strategies in speech therapy involve exercises related to speaking rate. However, in current clinical practice, most speech language pathologists (SLPs) still use stop watches to manually calculate rate. In addition to the inefficiency of this practice, this method does not allow for continuous estimation of speaking rate - especially for long speech samples. This is problematic since some patients with neurological conditions (e.g. Huntington's disease) exhibit irregular short-term changes in speaking rate. Moreover, some patients, such as those with Parkinson's disease (PD) tend to have atypical rate and rhythm. As a result, in clinical practice, there exist a number of pacing strategies or rate control methods (RCM), such as hand tapping, pacing boards, and delayed auditory feedback [7]. However, patients require extensive training on how to use these strategies. Having a reliable online speaking rate estimation tool would allow patients to easily monitor their speaking rate either during clinical intervention or at home.

Some of the early methods for automatic speaking rate estimation [8] [9] [10] aimed to automatically detect syllables in speech by detecting maxima in a loudness function. More recently, Wang and Narayanan proposed to use subband spectral and temporal correlations with the aid of voicing information to detect syllables [11]. Jong and Wempe showed a simple syllable nuclei detector based on intensity and voicing [12]. However, these methods all involve a peak detection strategy. As a result, they may become less robust to new data since heuristically defined thresholds are introduced to select the peaks. Automatic speech recognition (ASR) has also been used to estimate speaking rate. However, the performance of ASR degrades for dysarthric speech with increased deletion and insertion errors. Thus we attempt to avoid the more difficult task of either detecting individual syllables or using ASR. We posit that a statistical learning approach to this problem is more reliable and robust. In [13], we proposed a convex optimization based linear model for speaking rate estimation. Moreover, we found that long-term statistical features were more robust on spontaneous speech. However, in [13] we used a combination of multiple acoustic features which was not applicable in real time. In [14], Faltlhauser et al. proposed an online speaking rate estimation model based on neural networks. They used GMMs to first separate data into three rate groups (fast, moderate, slow) and built a neural network with the input of the likelihood values generated by GMMs. However, like we have shown in [13], the performance of this method is less satisfactory.

In this paper, we propose to use recurrent neural network (RNN) for online speaking rate estimation. Recurrent neural networks have recently achieved great success in acoustic models [15] and language models [16] for automatic speech recognition (ASR), feature enhancement for robust ASR [17],

This research was supported in part by National Institute of Health, National Institute on Deafness and Other Communicative Disorders grants 2R01DC006859 and 1R21DC012558.

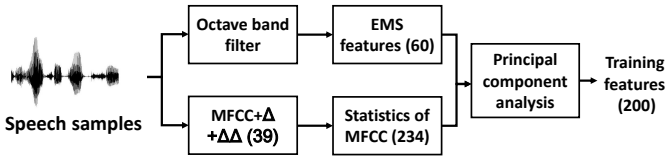

Fig. 1. Flowchart of feature extraction.

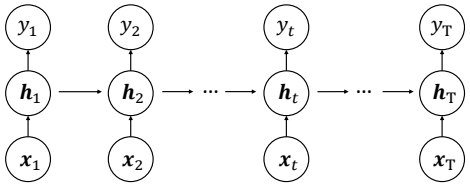

Fig. 2. Architecture of bi-directional recurrent neural network.

voice activity detection [18], etc. The main advantage of RNN over a regular neural network is that the current output not only depends on the current input but also depends on previous inputs due to the recurrent connections within hidden layers. In this way, the neural network has memory of the hidden representation of early inputs and is able to learn long dependencies among time series data. For the speaking rate estimation task, the motivation to use the RNN is that speaking rate will have longer time dependencies than other speech representations (e.g. phoneme transition, voice activity). In addition, our empirical studies show that the RNN representation for speaking rate allows for a smaller and simpler feature set when compared to our previous work in [13].

Relation to previous work. The use of RNNs for speaking rate has not been explored in the literature to the best of our knowledge. Although neural networks (NN) have been used to estimate speaking rate in [14], this model does not exploit the longer-term dependencies that RNNs exploit. Moreover, our algorithm requires training a single RNN, whereas the work in [14] uses a sequential procedure that requires training independent models for slow, moderate, and fast speech. This work is also related to other automatic speaking rate estimation methods [8] [12] [11]. However, all of the previous methods only showed statistical results at the sentence level, such as correlation, error rate, mean error, etc. While in this paper, our estimation of speaking rate is processed every 1 second with 0.1 second shift and our results show online changes in speaking rate. Finally, we also evaluate our approach on longitudinal pathological speech - this has not been done in previous work.

2. METHOD

2.1. Feature extraction

The speaking rate estimation system in Fig. 1 works as follows. The analysis window of the rate estimation algorithm is 1 second, with a 0.1 second shift. Features that are strongly related to speech rhythm are calculated for every 1 second frame. These features are described below.

Envelope modulation spectrum (EMS): The EMS is a representation of the slow amplitude modulations in a signal. It also reflects the distribution of energy in the amplitude fluctuations across designated frequencies. The 1 second speech signal is passed through a range of octave band filters, after which the envelope is extracted from each individual octave. From the envelope of each octave we extract the following features 1) Peak Frequency 2) Peak Amplitude 3) Energy from 3-6 Hz 4) Energy from 0-4 Hz 5) Energy from 4-10 Hz 6) Energy ratio between 0-4 Hz band and 4-10 Hz band. These features are primarily designed to capture the rhythmic information from the speech signal [5]. The dimension of the final EMS feature set is 60.

Mel-frequency cepstral coefficients (MFCC): The second set of features are extracted from 13$^{\text{th}}$ order MFCCs (including 0th order) and their first and second order derivatives - the total number of features is 39. Then, from each row of MFCCs, we calculate the mean, standard deviation, maximum value, skewness, kurtosis and mean absolute deviation over the 1 second interval. The total dimension of MFCC based feature is 39*6=234.

The feature sets are combined and normalized using the mean and standard deviation of the training data; finally the data are whitened and the dimension is reduced to 200 using principal component analysis (PCA).

2.2. RNN training

We use the RNN as a regression model to predict speaking rate. Different from standard neural network, RNNs use feedback connections within hidden layers to store previous events in the form of hidden activations, which builds memory into the network. We use a unidirectional RNN, which only makes use of past samples, so as to ensure a reasonable time delay (1 sec). In Fig. 2, we show an unfolded unidirectional RNN for a sequence $\mathbf{X}$. Here the input is a time series of acoustic features $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_t, ..., \mathbf{x}_\text{T}]$ with length T. After training, the RNN computes the hidden sequence $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_t, ..., \mathbf{h}_\text{T}]$ and outputs the speaking rate sequence $\mathbf{y} = [y_1, ..., y_y, ..., y_\text{T}]$ by iterating from $t = 1$ to T as follows [15]:

$$\begin{aligned} \overrightarrow{\boldsymbol{h}}_t &= f_{\boldsymbol{\theta}}(\mathbf{W}_{x\overrightarrow{h}}\mathbf{x}_t + \mathbf{W}_{\overrightarrow{h}\overrightarrow{h}}\mathbf{h}_{t-1} + \mathbf{b}_{\overrightarrow{h}}) \\ y_t &= \mathbf{w}_{\overrightarrow{h}y}\overrightarrow{\mathbf{h}}_t + b_y \end{aligned} \tag{1}$$

where $f_{\boldsymbol{\theta}}$ is the hidden layer activation function, $\mathbf{W}_{x\overrightarrow{h}}$ is the weight matrices from input to forward hidden layers, $\mathbf{W}_{\overrightarrow{h}\overrightarrow{h}}$ is the forward recurrent weight matrices, $\mathbf{w}_{\overrightarrow{h}y}$ is the weight

Table 1. Comparative results

	TF-Corr	GMM-NN	Praat	RNN
Correlation coefficient	0.57	0.32	0.59	**0.73**
Mean error	1.01	2.18	1.67	**0.71**
Stddev error	0.83	1.29	1.06	**0.55**
Error rate	0.34	0.51	0.40	**0.21**

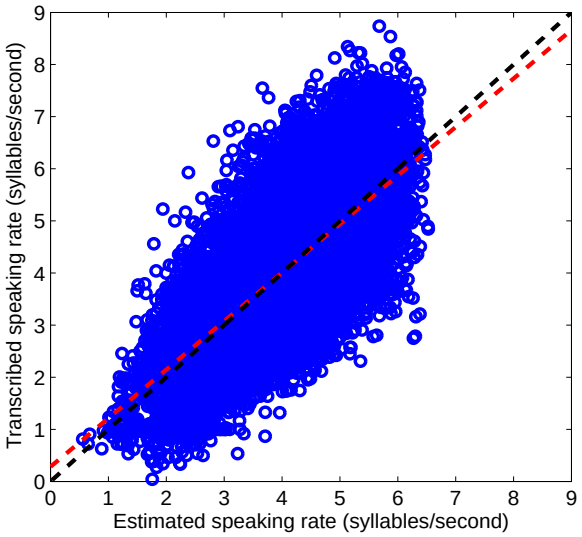

Fig. 3. Scatter plot of the estimated rates against the transcribed rates on all 1-sec test samples. The red is a fitting line of all samples. The black is a diagonal line.

vectors from forward hidden layers to output, $\mathbf{b}_{\overrightarrow{h}}$ and b_y are bias terms (since output y_t is a scalar $\mathbf{b}_{\overrightarrow{h}}$ is vector and b_y is a scalar). These weights are learned during training. We use long short-term memory (LSTM) nodes in the hidden layer, which uses an input gate, output gate and forget gate to scale the input to hidden nodes, output activation and recurrent cell states respectively in order to avoid severe gradient vanishing or exploding along time - this allows the network to learn long-term dependencies during training [19]. It is reported in [20] that LSTM can learn time dependencies as long as 1000 time steps. This is sufficient for our speaking rate estimation task. Given sequences of training samples (with ground truth speaking rate), we use the CURRENNT toolbox [21] to train our RNN model since it provides a parallel training paradigm that can speed up the training process.

3. EXPERIMENT

3.1. Evaluation on Switchboard spontaneous speech

The Switchboard corpus is a speech database that includes several hundreds informal conversations recorded over the

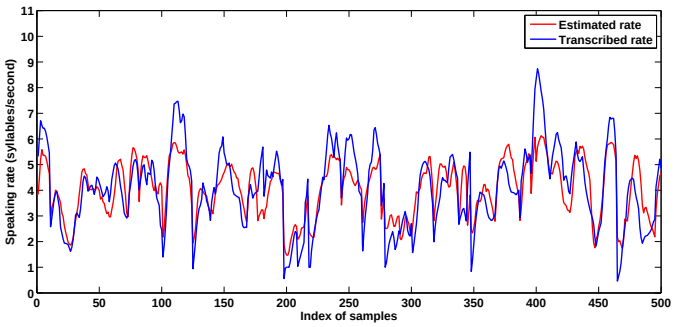

Fig. 4. Line plot of the estimated rates against the transcribed rates of 500 test samples.

telephone [22]. The International Computer Science Institute (ICSI) Switchboard corpus is a subset of the original database with phonetic transcriptions. There are 5564 speech spurts from multiple speakers sampled at 8kHz, 16-bit. Each speech sample includes a hand-corrected transcription with syllable boundary information. We used 64% of the samples to train the RNN, 16% for validation set, and 20% for testing.

The evaluation was based on a 1 second estimation window with a 0.1 second shift. The speech samples in each estimation window were analyzed using a 20ms Hamming window with a 10ms frame shift. Thirteen-order MFCC features, along with delta and delta-delta features, were extracted from each frame. Six statistical features as mentioned in Section 2 were calculated from MFCC for each 1 second window. Moreover, a 60-dimensional EMS feature vector was extracted from the 1-sec speech segment. To generate the ground-truth speaking rate for each estimation window, we calculated the number of syllables per second by using the syllable boundary information from the labels. Silence and non-speech segments were skipped.

The RNN has an input layer with 200 nodes; two hidden layers, each with 64 bidirectional LSTM nodes; the output was a scalar value approximating the speaking rate. The training objective was to minimize the root mean square error (RMSE) between the estimated speaking rate at the output of the RNN and the ground-truth speaking rate. The weights were randomly initialized using a uniform distribution between -0.1 and 0.1, which was empirically better than normal distribution initialization. The learning rate was set to 3e-6 and the maximum number of epochs was 100. Every 1 epoch the validation error was checked to see if better performance was achieved. To avoid overfitting, the training process stopped if there was no reduction in validation error for more than 5 epochs.

We compared our results with three existing methods. The first was described in [11]; we denoted it by 'TF-Corr'. We used the code that the authors made publicly available for this method. The second method was described in [14]. We denoted it by 'GMM-NN' due to its structure. The

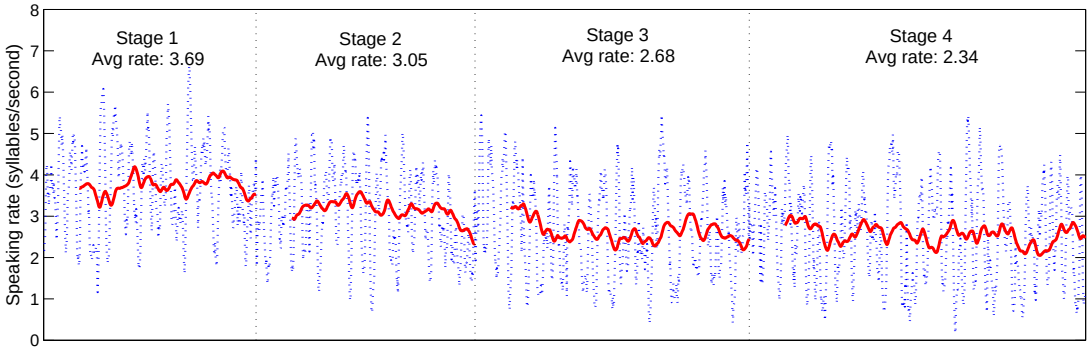

Fig. 5. Results of speaking rate estimated on longitudinal speech. Blue line is the estimated rate on each 1 second speech with a 0.1 second shift. The red line is the average speaking rate of each 10 seconds.

third method was a Praat script [12] for estimating rate - we denoted this algorithm by 'Praat'. More information about the implementation of these methods can be found in our previous paper [13]. For consistency, all algorithms were evaluated on the same test set on 1 second windows with a 0.1 second shift. Samples with a transcribed rate of zero were discarded in our evaluation. In a real online system, these can be discarded using a voice activity detection (VAD) algorithm. The results are shown in Table 1. There were four metrics evaluated: the correlation coefficient between the estimated speaking rate and the transcribed speaking rate; the mean and standard deviation of the absolute speaking rate error; the error rate, computed as the average of $\frac{|\text{estimated rate} - \text{transcribed rate}|}{\text{transcribed rate}}$ across all samples in the test set. T-test was conducted between our proposed method and the other 3 methods. All of the resulting p-values were less than 0.0001, which indicated that the improvement was statistically significant. In Fig. 3 and Fig. 4, we show the scatter plot as well as the fitting lines of the estimated speaking rate by RNN against the transcribed speaking rate. From the figures, we can see a strong correlation between the estimated rates and the transcribed rates.

3.2. Evaluation on longitudinal pathological speech

We also evaluate the speaking rate estimation algorithm on pathological speech. Our speech sample is from a female with a neurodegenerative syndrome called pallido-ponto-nigral degeneration (PPND) [23]. She was followed longitudinally at 6-month intervals over several years. In each recording session, the speaker was asked to read a standard passage, the grandfather passage, which was commonly used in clinical practice. In the first stage, she exhibited mild signs of dysarthria, including mild vocal instability, mild vocal tremor and vocal flutter, slightly slow speaking rate, etc. In the second and third stage, her speech symptoms became more pronounced, including greatly reduced loudness, voice tremor, more frequent vocal flutter, slower speaking rate, mild imprecision of consonants, etc. In the last stage, she had more difficulty in speaking and showed mixed dysarthrias, with minimally hypokinetic and flaccid features, including decreased intelligibility, very slow speaking rate, monopitch, imprecise consonant production, general voice tremor and voice flutter, breathy speaking voice with substantially reduced speech loudness. We estimated the speaking rate for these passages individually and connected the results together. The result is shown in Fig. 5. The estimation was also based on 1 second interval with 0.1 second shift. The red line represents a smooth rate based on a 10 second moving-average filter. From the result, we can see that the average speaking rate goes down from Stage 1 to Stage 4, which corresponds to the description of speech characteristics for each stage. We can also see that within each stage, the speaking rate gradually decreases over time, which was likely due to speaker fatigue.

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a recurrent neural network (RNN) based speaking rate estimation method. In contrast to previous work, we used fewer long-term acoustic features and provided an online estimate of the rate that depended on past estimates due to the long-term dependencies in the RNN. The evaluation on spontaneous speech revealed a high correlation between the estimated speaking rate and the ground-truth. We also evaluated the method on longitudinal pathological speech. The results showed that the proposed method can capture the decreasing trend of speaking rate not only along different disease stages but also during each individual session. Our future work includes testing the performance of the method under different conditions and implementing it on mobile devices, such as a smart phone or a tablet, so that clinicians can use it in practice.

5. REFERENCES

[1] Nelson Morgan, Eric Fosler-Lussier, and Nikki Mirghafori, "Speech recognition using on-line estimation of speaking rate.," in *Eurospeech*, 1997, vol. 97, pp. 2079–2082.

[2] Joseph P Campbell Jr, "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[3] Yannis Stylianou, Olivier Cappé, and Eric Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 131–142, 1998.

[4] Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Zhigang Deng, Sungbok Lee, Shrikanth Narayanan, and Carlos Busso, "An acoustic study of emotions expressed in speech," in *Annual Conference of the International Speech Communication Association*, 2004.

[5] Julie M Liss, Laurence White, Sven L Mattys, Kaitlin Lansford, Andrew J Lotto, Stephanie M Spitzer, and John N Caviness, "Quantifying speech rhythm abnormalities in the dysarthrias," *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 5, pp. 1334–1352, 2009.

[6] Yu-Tsai Wang, Ray D Kent, Joseph R Duffy, and Jack E Thomas, "Dysarthria associated with traumatic brain injury: speaking rate and emphatic stress," *Journal of communication disorders*, vol. 38, no. 3, pp. 231–260, 2005.

[7] Gwen Van Nuffelen, Marc De Bodt, Floris Wuyts, and Paul Van de Heyning, "The effect of rate control on speech rate and intelligibility of dysarthric speech," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 2, pp. 69–75, 2009.

[8] Paul Mermelstein, "Automatic segmentation of speech into syllabic units," *The Journal of the Acoustical Society of America*, vol. 58, no. 4, pp. 880–883, 1975.

[9] Jan P Verhasselt and J-P Martens, "A fast and reliable rate of speech detector," in *Spoken Language Processing (ICSLP), Fourth International Conference on*. IEEE, 1996, vol. 4, pp. 2258–2261.

[10] H Pfitzinger, "Local speaking rate as a combination of syllable and phone rate," *Proceeding of ICSLP 1998*, 1998.

[11] Dagen Wang and Shrikanth S Narayanan, "Robust speech rate estimation for spontaneous speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2190–2201, 2007.

[12] Nivja H de Jong and Ton Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[13] Yishan Jiao, Visar Berisha, Ming Tu, and Julie Liss, "Convex weighting criteria for speaking rate estimation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 23, no. 9, pp. 1421–1430, 2015.

[14] Robert Faltlhauser, Thilo Pfau, and Günther Ruske, "On-line speaking rate estimation using gaussian mixture models," in *Acoustics, Speech, and Signal Processing (ICASSP), 2000 IEEE International Conference on*. IEEE, 2000, vol. 3, pp. 1355–1358.

[15] Alan Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6645–6649.

[16] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Honza Černockỳ, and Sanjeev Khudanpur, "Extensions of recurrent neural network language model," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5528–5531.

[17] Felix Weninger, Shinji Watanabe, Yuuki Tachioka, and Bjorn Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4623–4627.

[18] Tim Hughes and Keir Mierle, "Recurrent neural networks for voice activity detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7378–7382.

[19] Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 5, pp. 855–868, 2009.

[20] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] Felix Weninger, Johannes Bergmann, and Björn Schuller, "Introducing currennt: The munich open-source cuda recurrent neural network toolkit," *Journal of Machine Learning Research*, vol. 16, pp. 547–551, 2015.

[22] John J Godfrey, Edward C Holliman, and Jane McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*. IEEE, 1992, vol. 1, pp. 517–520.

[23] Julie M Liss, Kari Krein-Jones, Zbigniew K Wszolek, and John N Caviness, "Speech characteristics of patients with pallido-ponto-nigral degeneration and their application to presymptomatic detection in at-risk relatives," *American Journal of Speech-Language Pathology*, vol. 15, no. 3, pp. 226–235, 2006.