ԵՐԵՎԱՆԻ ՊԵՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

ՄԱԹԵՄԱՏԻԿԱՅԻ ԵՎ ՄԵԽԱՆԻԿԱՅԻ ՖԱԿՈՒԼՏԵՏ

Հավանականության տեսության և վիճակագրության ամբիոն

ԿԻՐԱՌԱԿԱՆ ՎԻՃԱԿԱԳՐՈՒԹՅՈՒՆ ԵՎ ՏՎՅԱԼՆԵՐԻ ԳԻՏՈՒԹՅՈՒՆ ԿՐԹԱԿԱՆ ԾՐԱԳԻՐ

ՍԻՄՈՆՅԱՆ ԴԻԱՆԱ ԼԵՎՈՆԻ

ՄԱԳԻՍՏՐՈՍԱԿԱՆ ԹԵԶ

ԽՈՍՔԻ ԱՐԱԳՈՒԹՅԱՆ ԳՆԱՀԱՏՈՒՄ

«Վիճակագրություն» մասնագիտությամբ վիճակագրության մագիստրոսի որակավորման աստիճանի հայցման համար

ԵՐԵՎԱՆ 2024

Ուսանող՝ _____

ստորագրություն

_____

ազգանուն , անուն

Գիտական ղեկավար՝ _____

ստորագրություն

_____

գիտ. աստիճան, կոչում, ազգանուն, անուն

«Թույլատրել պաշտպանության»

Ամբիոնի վարիչ՝ _____

ստորագրություն

_____

գիտ. աստիճան, կոչում, ազգանուն, անուն

«_____»_____*2024*թ

Թեզի վերնագիրը

Հայերենով՝ Խոսքի արագության գնահատում

Ռուսերենով՝ Оценка скорости речи

Անգլերենով՝ Speaking Rate Estimation

Նպատակն ու ստացված արդյունքները

Sizes conv: chapter - 16
chapter title - 18

Contents

Abstract

Speaking rate is an important attribute of the speech signal which plays a crucial role in the performance of automatic speech processing systems. In this paper we address the problem of speaking rate estimation through two approaches: classification and regression. To tackle the problem, we use two types of architectures: LSTM-RNN and MatchBoxNet, a convolution-based architecture. We give raw waveform to the LSTM-RNN as input and use the audio feature Mel-frequency cepstral coefficients (MFCC) as input to the MatchBoxNet. Our experimental findings indicate that MatchBoxNet significantly outperforms the traditional LSTM-RNN solution. For our training and experiments we use a dataset which we constructed using subsets of LibriSpeech ASR corpus. Additionally, we evaluate the generalizability of our model by testing its performance on the five languages of Common Voice corpora: Armenian, English, Italian, Spanish and Russian. Our findings demonstrate the robustness and adaptability of our model across diverse linguistic contexts, showcasing its effectiveness in estimating speaking rate in different languages. Finally, we conduct a comparative analysis of our proposed model's performance against the ASR-based approach using small pre-trained models from the NVIDIA NeMo framework.

Index Terms: *syllable count estimation, speaking rate estimation, deep neural networks, convolutional neural network, depth-wise separable convolution, recurrent neural network, long short-term memory network*

Finally, we also
We compared our results with
Software Setup

Chapter 1

Introduction

A. Significance

Speaking rate estimation is very important for speech understanding and speech recognition. Speaking rate has been shown to be useful in several applications including pronunciation assessment [2, 8], automatic speech recognition (ASR) [1], detection of dysarthria[11, 46], computer-assisted language learning (CALL) [2] and emotion recognition [7]. Apart from these, it has also played a role in problems including perception studies, age estimation etc.
Speaking rate is an important quantity in automatic speech recognition. Morgan et al. used the speech rate to improve the robustness of the Automatic Speech Recognition (ASR) system as it gets adversely affected by the variations in speech rate [1]. Speech rate has also been used in the analysis of second language learners' fluency [2]. The speech rate variation helps in speech understanding by providing context information. Honig et al. used speech rate estimation for the appraisal of non-nativeness [3]. In [4], the authors studied the impact of the speech rate on the acoustic correlation of speech rhythm. For speaker recognition, Joseph et al. used speech rate as a distinctive characteristic between speakers [5]. Yannis et al. used the speech rate as one of the suprasegmental properties for speech modification [6]. In the emotion recognition system, speech rate variability is observed as one of the acoustic properties to distinguish between different emotions [7]. In speech therapy applications, speech rate was used to analyse the efficiency of the articulatory movements over time in dysarthric patients [9, 10].

B. General Measurement Methods

There have been two major trends in measuring speech rate. Each has its advantages and limitations.

- The first represents the use of discrete categorization— "fast," "normal" and "slow"—to describe speech rate [14]. Even though it matches human intuition, the boundaries between these three categories are fuzzy. Most of the time, human knowledge is required

to set the boundaries, and hence it is difficult to devise a completely automated engineering solution.

- In the second approach, speech rate is measured in a quantitative way by counting the number of phonetic elements per second. Words, syllables [15], stressed syllables, and phonemes [16] are all possible candidates. In this paper, we have considered syllables as speech units similar to most of the research works [1, 12, 13].

C. The Two Solution Approaches to Speaking Rate Estimation

Speaking rate is typically estimated in two ways:

1. ASR-based approach [17]

   This approach suffers from several limitations:

   - the speaking rate can be estimated only when speech with reference transcription is available, which is not typically available for spontaneous speech
   - not robust to noise
   - computationally expensive

2. Direct acoustic feature-based approach [18, 19]

   In the direct acoustic feature-based approach, the speaking rate is estimated using the features derived based on the acoustic properties of the vowels, which, in general, correspond to the syllable nuclei. This approach is computationally less expensive as compared to the ASR-based approach

In this paper, we propose two distinct solutions to address the problem: classification and regression. For our experiments, we use two architectures: LSTM-RNN, which takes raw waveforms as input, and MatchBoxNet, a convolution-based architecture that receives MFCC features as input. In our proposed approaches, the models are designed to estimate the number of syllables in speech. In the regression task to compute speaking rate we divide the estimated syllable count by the length of the audio at inference, thus measuring speaking rate by syllable count per second. To compare the models and two approaches we use Mean Absolute Error(MAE) as the main metric as it shows how many syllables we get wrong on average. After identifying the best-performing model, we assess its generalizability across five languages and compare its results with those of an ASR-based approach using NVIDIA NeMo's small models

The rest of the paper is organized as follows: the speech corpora details are described

in Sect. 2, and the proposed approach is discussed in Sect. 3. The experimental results are analyzed and elaborated in Sect. 4. Finally, conclusions are discussed in Sect. 5.

Chapter 2

Existing Datasets and Methods

2.1 Datasets

For the task of speaking rate estimation, five corpora are available: Switchboard [24], TIMIT [25], CTIMIT [26] and ISLE [27]. We will not give the detailed description of these datasets as for our training and experiments we use a dataset constructed from the subsets of LibriSpeech[28] corpus. The detailed description of the corpus and the way we constructed the dataset is given in section 3.1.

2.2. Methods

Several works in the past have dealt with the problem of speech rate estimation. Most of these are Hidden Markov Model (HMM)-based and acoustic feature-based methods. The HMM-based methods use an ASR system to obtain the syllable boundaries which are used to compute the speech rate. But the HMM-ASR based methods are not robust to noise and they need a reference transcription which is not typically available for spontaneous speech [13]. Several methods have been proposed which do not require transcriptions but only speech acoustics [18 , 20, 21]. In this regard, there are both unsupervised and supervised approaches. Among unsupervised approaches, a peak detection strategy [22] using a convex weighting criterion was used for speech rate estimation. Temporal correlation and selected subband correlation (TCSSBC) based feature contour was utilised in [13], [26] to estimate speech rate, in which peak detection was performed with smoothing and thresholding operations. On the other hand, among supervised approaches, a Gaussian mixture model (GMM)[20] based method was proposed to classify speech into slow, medium and fast rate classes and these class probabilities were used to estimate the speech rate. Recently, using neural networks, syllable rate estimation is formulated as a regression problem, and mean squared error (MSE) loss between the estimated and original speech rate is optimised to train the convolutional dense neural networks (CDNN).

Chapter 3

Proposed Method

We solve the problem of speaking rate estimation in two ways: formulating it as a regression and classification task.

Regression Task: In the regression task, the models are not directly estimating speaking rate. Instead, they are designed to estimate the number of syllables in the input speech. This estimation is then used to calculate the speaking rate by dividing the estimated number of syllables by the length of the audio during inference.

Classification Task: In the classification task the class number indicates the number of syllables in the speech. The number of classes are determined by the maximum syllable count in the training data.

In the following subsections we present the detailed description of the architectures we used and give a comprehensive description of the data construction process.

3.1 Dataset

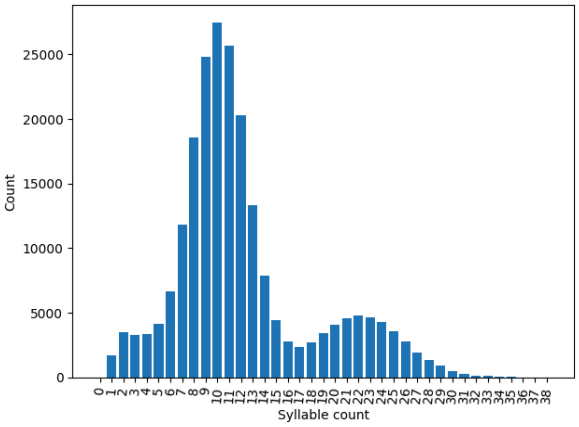

Figure 3.3 The distribution of syllable count in training data

In this work we use a dataset which we constructed using LibriSpeech ASR corpus [28], which is a corpus of approximately 1000 hours of 16kHz read English speech. The general information regarding the dataset is summarised in Table 3.1. Due to the size of the corpus, distributing it as a single large archive would be impractical or inconvenient for some users. Therefore, the training portion of the corpus was divided into three subsets, each approximately containing 100, 360, and 500 hours of data, respectively. The speakers in the corpus were ranked according to the

WER of the WSJ model's transcripts, and were divided roughly in the middle, with the lower-WER speakers designated as "clean" and the higher WER speakers designated as "other". The division into subsets was done so that the speakers in different subsets do not intersect. The detailed description of these subsets are represented in Table 3.1. We use the train-clean-100, dev-clean, and test-clean subsets to construct our training, validation, and test sets, respectively. Each subset was then divided into 2-second length chunks. Splitting is done so that it preserves the integrity of words. This preservation of word integrity is crucial for accurate label computation. For that purpose we use LibriSpeech word-based alignments. As a result, the lengths of the training data chunks are not exactly but close to 2 seconds. We then pad the audios with 0s from the right to achieve equal-length inputs for model training. The resulting number of 2 second length audios is shown in Table 3.2. Additionally, silence segments are removed from the speech chunks throughout the splitting procedure.

subset	hours	per-spk minutes	female spkrs	male spkrs	total spkrs
dev-clean	5.4	8	20	20	40
test-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-other	5.1	10	17	16	33
train-clean-100	100.6	25	125	126	251
train-clean-360	363.6	25	439	482	921
train-other-500	496.7	30	564	602	1166

Table 3.1 Data subsets in LibriSpeech

Furthermore, to enlarge our dataset and make our models robust to fast speech we implemented fast data augmentation by increasing the playback speed (speeding up) of the audios in train and validation sets by factors of 1.5 and 2. The statistics of labels in the resulting training data is illustrated in Figure 3.3.

2-second length splits	chunks count
train-clean-100	222453
val-clean	10044
test-clean	10102

Table 3.2 The count of 2-second length chunks in data splits

Labels, which in our case represent the syllable number in speech, are computed from audio transcripts by counting the number of vowels following the English convention, where each syllable generally contains one vowel.

For noise augmentation we used the ESC-50 dataset, which is a labelled collection of 2000 environmental audio recordings. The dataset consists of 5-second-long recordings organised into 50 semantical classes (with 40 examples per class) loosely arranged into 5 major categories. The specific class names can be found in Table 3.4.

Animals	Natural soundscapes & water sounds	Human, non-speech sounds	Interior/domestic sounds	Exterior/urban noises
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw

Table 3.4 The classes of ESC-50 dataset and their categorization into 5 categories

3.2 Models

In this work we use two types of architectures to solve the problem: RNN-LSTM and MatchBoxNet.

RNN-LSTM

RNN-LSTM takes as input raw waveforms. To generate the input sequence for the RNN, waveforms are segmented using a window length of 35 ms and a window shift of 10ms. The LSTM layer's number of hidden units is set to 128.

In the *regression task*, a single linear layer with 1 unit is added after the final hidden state of the LSTM layer. This layer predicts the syllable count in the input speech. (See Figure 3.5).

Additionally, the output of the last layer is constrained from above by 38, representing the maximum number of syllables in our training data, and from below by 0 to prevent the prediction of negative syllable counts.
In the classification task LSTM layer follows a linear layer with 38 units and softmax activation. Class number indicates the number of syllables in the input speech. The number of classes is determined by the maximum syllable count observed in our training data, which is 38. (See Figure 3.3)

MatchBoxNet

As a second architecture we used MatchBoxNet, which was previously used for Speech Commands Recognition. MatchboxNet is a deep residual network specifically designed for devices with low computational and memory resources. It uses 1D time-channel separable convolutions to reduce model size versus regular 1D convolutions. MatchboxNet consists of a stack of blocks with residual connections. A MatchboxNet-BxRxC model has B residual blocks. Each block has R sub-blocks. All sub-blocks in a block have the same number of output channels C (see Fig. 3.6). A basic sub-block consists of a 1D-time-channel separable convolution, 1x1 pointwise convolutions, batch normalisation, ReLU and dropout. The 1D-time-channel separable convolution has C filters with a kernel of the size k. All models have four additional subblocks: one prologue layer – 'Conv1' before the first block, and three epilogue sub-blocks ('Conv2', 'Conv3', and 'Conv4') before the average global pooling layer. The whole architecture with details you can find in Figure 3.6.

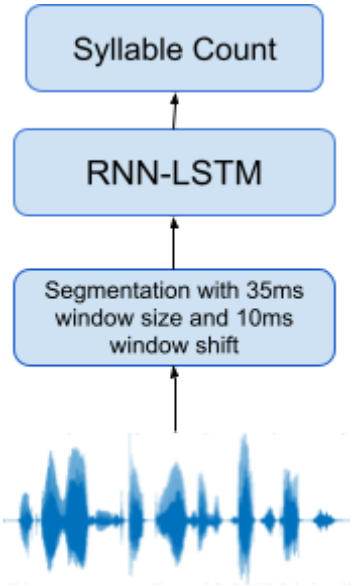

Figure 3.5 (Add Linear Layer and may be change graphic of RNN-LSTM)

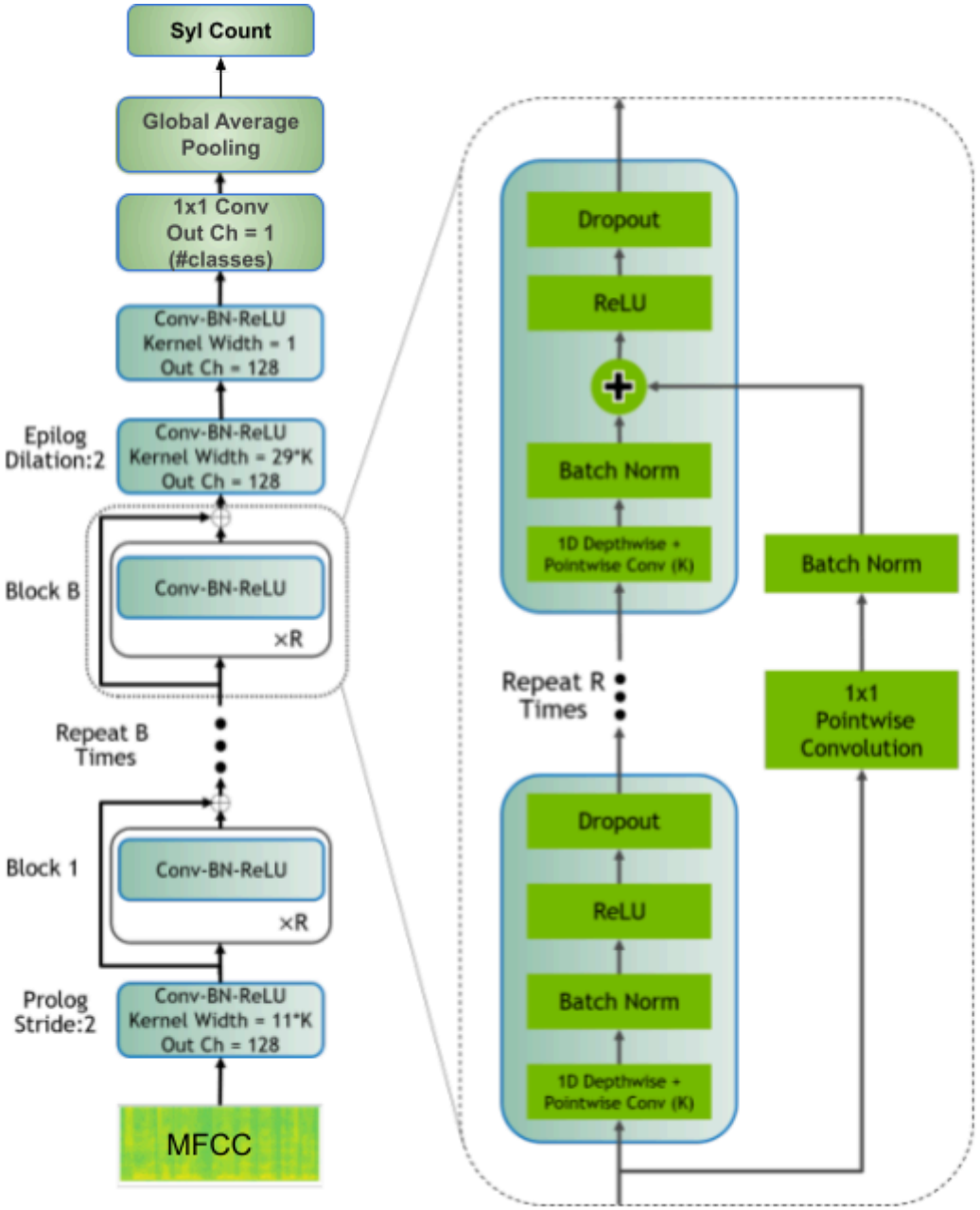

Figure 3.6

Note: Add wav pic under MFCC and (25ms window, 15ms step)(paper SylNet)

Chapter 4

Experiments and Results

4.1 Training and Evaluation

We trained our regression and classification models on the dataset described in section 3.1 with the ADAM optimizer and learning rate set to 10^{-3}. All models were trained for 200 epochs with a batch size of 200. For the regression task MSE loss was used as optimization objective and cross entropy loss was optimised in the classification task. To assess the performance of our models, we used accuracy and top-3 accuracy metrics for the classification task, and Mean Absolute Error (MAE) and the Pearson correlation coefficient (PCC) for the regression task. Additionally, to compare the best classification and regression models and determine the most effective method for solving the problem, we calculated the Mean Squared Error (MSE) between the estimated class and the actual class. This comparison is reasonable and fair, as in our scenario, the class number directly represents the syllable count in the input speech.

4.2 Experiments and Results

Bibliography

[1] N. Morgan, E. Fosler-Lussier, and N. Mirghafori, “Speech recognition using on-line estimation of speaking rate,” in *EUROSPEECH*, vol. 4, 1997, pp. 2079–2082.

[2] C. Cucchiarini, H. Strik, and L. Boves, “Quantitative assessment of second language learners’ fluency by means of automatic speech recognition technology.” *The Journal of the Acoustical Society of America*, vol. 107 2, pp. 989–99, 2000.

[3] F. Honig, A. Batliner, and E. N ¨oth, “Automatic assessment of non-native prosody annotation, modelling and evaluation,” in *International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, 2012, pp. 21–30.

[4] V. Dellwo, “Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence,” *PhD*

Dissertation, Universitat Bonn (electronic publication: http://hss. ulb. uni-bonn. de: 90/2010/2003/2003. htm), 2010.

[5] J. P. Campbell, *Speaker Recognition*. Boston, MA: Springer US, 1996, pp. 165–189.

[6] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 131–142,, 1998.

[7] S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, Z. Deng, S. Lee, S. Narayanan, C. Busso, "An acoustic study of emotions expressed in speech." in *Eighth International Conference on Spoken Language Processing* (2004), pp. 2193–2196

[8] M.P. Black, D. Bone, Z.I. Skordilis, R. Gupta, W. Xia, P. Papadopoulos, S.N. Chakravarthula, B. Xiao, M.V. Segbroeck, J. Kim, et al, "Automated evaluation of non-native English pronunciation quality: Combining knowledge-and data-driven features at multiple time scales." in *Sixteenth Annual Conference of the International Speech Communication Association* (2015), pp. 493–497

[9] J. Liss, L. White, S. L Mattys, K. Lansford, A. Lotto, S. M Spitzer, and J. Caviness, "Quantifying speech rhythm abnormalities in the dysarthrias," *Journal of speech, language, and hearing research : JSLHR*, vol. 52, pp. 1334–52, 09 2009.

[10] Y.-T. Wang, R. Kent, J. Duffy, and J. E Thomas, "Dysarthria associated with traumatic brain injury: Speaking rate and emphatic stress", *Journal of communication disorders*, vol. 38, pp. 231–60, 05 2005.

[11] M.P. Caligiuri, "The influence of speaking rate on articulatory hypokinesia in Parkinsonian dysarthria." Brain Lang. 36(3), 493–502 (1989)

[12] C. Heinrich and F. Schiel, "Estimating speaking rate by means of rhythmicity parameters." January 2011, pp. 1873–1876.

[13] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15,

no. 8, pp. 2190–2201, Nov 2007.

[14] B. Zellner, “Fast and slow speech rate: A characterisation for French,” *in Proc. Int. Conf. Spoken Lang. Process.*, Sydney, Australia, Dec. 1998, vol. 7, pp. 3159–3163.

[15] N. Morgan and E. Fosler-Lussier, “Combining multiple estimators of speaking rate,” in *Proc. ICASSP*, 1998, vol. 2, pp. 729–732.

[16] H. Nanjo and T. Kawahara, “Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition,” in *Proc. ICASSP*, 2002, pp. 725–728

[17] D. Wang, S.S. Narayanan, Robust speech rate estimation for spontaneous speech. IEEE Trans. Audio Speech Lang. Process. 15(8), 2190–2201 (2007)

[18] T. Pfau, G. Ruske, Estimating the speaking rate by vowel detection, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1998), pp. 945–948

[19] J. Yuan, M. Liberman, Robust speaking rate estimation using broad phonetic class recognition, in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* (2010), pp. 4222–4225

[20] R. Faltlhauser, T. Pfau, and G. Ruske, “Online speaking rate estimation using Gaussian mixture models,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP)*, vol. 3, June 2000, pp. 1355–1358 vol.3.

[21] N. H. de Jong and T. Wempe, “Praat script to detect syllable nuclei and measure speech rate automatically,” *Behavior Research Methods,* vol. 41, no. 2, pp. 385–390, May 2009.

[22] Y. Jiao, V. Berisha, M. Tu, and J. Liss, “Convex weighting criteria for speaking rate estimation,” I*EEE/ACM Transactions on Audio, Speech, and Language Processing,*

vol. 23, no. 9, pp. 1421–1430, Sep. 2015.

[23] S. Narayanan and Dagen Wang, “Speech rate estimation via temporal correlation and selected subband correlation,” in I*EEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP),* March 2005, pp. 413–416 Vol. 1.

[24] J.J. Godfrey, E.C. Holliman, J. McDaniel, SWITCHBOARD: telephone speech corpus for research and development, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1992), pp. 517–520

[25] V. Zue, S. Seneff, J. Glass, Speech database development at MIT: TIMIT and beyond. Speech Commun. 9(4), 351–356 (1990)

[26] K.L. Brown, E.B. George, CTIMIT: a speech corpus for the cellular environment with applications to automatic speech recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1995), pp. 105–108

[27] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, C. Souter, The ISLE corpus of non-native spoken English, in *2000 Language Resources and Evaluation Conference* (European Language Resources Association, 2000), pp. 957–964

[28] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, Librispeech: An ASR corpus based on public domain audio books,in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* South Brisbane, QLD, Australia, 2015, pp. 5206-5210

[46] H.Martens, G. Van Nuffelen,M. De Bodt, T. Dekens, L. Latacz,W. Verhelst, “Automated assessment and treatment of speech rate and intonation in dysarthria”,

in *Seventh International Conference on Pervasive Computing Technologies for Healthcare* (ICST, Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2013), pp. 382–384