ԵՐԵՎԱՆԻ ՊԵՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

ՄԱԹԵՄԱՏԻԿԱՅԻ ԵՎ ՄԵԽԱՆԻԿԱՅԻ
ՖԱԿՈՒԼՏԵՏ

ՀԱՎԱՆԱԿԱՆՈՒԹՅՈՒՆՆԵՐԻ ՏԵՍՈՒԹՅԱՆ ԵՎ
ՄԱԹԵՄԱՏԻԿԱԿԱՆ ՎԻՃԱԿԱԳՐՈՒԹՅԱՆ ԱՄԲԻՈՆ

ԱՌԿԱ ՈՒՍՈՒՑՄԱՆ ՄԱԳԻՍՏՐԱՏՈՒՐԱՅԻ
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

«_____»_____20__թ

Թեզի վերնագիրը

Հայերենով՝ Խոսքի արագության գնահատում

Ռուսերենով՝ Оценка скорости речи

Անգլերենով՝ Speaking Rate Estimation

Նպատակն ու ստացված արդյունքներ

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 this problem, we implement two distinct architectures: BLSTM and MatchBoxNet, a convolution-based architecture. The BLSTM network takes raw waveform as input, while the MatchBoxNet utilises Mel-frequency cepstral coefficients (MFCC) as audio features. Our experimental findings indicated that MatchBoxNet significantly outperforms the traditional BLSTM solution. To train our model we used a dataset which we obtained from LibriSpeech ASR corpus. Additionally, we evaluated the generalizability of our model by testing its performance on the Common Voice corpora across five languages: Armenian, English, Italian, Spanish, and Russian. Our findings demonstrated 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 an 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, bidirectional 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: BLSTM, 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 used a dataset which we got from LibriSpeech[28] corpus. The detailed description of the corpus and the way we constructed the dataset is described 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.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, our models 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 classification task class division is based on 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

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. Specifically we use

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

Aparna Srinivasan, Diviya Singh, Chiranjeevi Yarra, Aravind Illa, and Prasanta Kumar Ghosh (2021, dec.). A Robust Speaking Rate Estimator Using a CNN-BLSTM Network. *Circuits Systems and Signal Processing,* 40(1). doi: 10.1007/s00034-021-01754-1