INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Acoustic and articulatory feature based speech rate estimation using a convolutional dense neural network

Renuka Mannem[1], *Jhansi Mallela*[2], *Aravind Illa*[1], *Prasanta Kumar Ghosh*[1]

[1]Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, India
[2]Rajiv Gandhi University of Knowledge Technologies (RGUKT), Kadapa, 516330, India

mannemrenuka@iisc.ac.in, jhansimallela5@gmail.com, aravindi@iisc.ac.in, prasantg@iisc.ac.in

Abstract

In this paper, we propose a speech rate estimation approach using a convolutional dense neural network (CDNN). The CDNN based approach uses the acoustic and articulatory features for speech rate estimation. The Mel Frequency Cepstral Coefficients (MFCCs) are used as acoustic features and the articulograms representing time-varying vocal tract profile are used as articulatory features. The articulogram is computed from a real-time magnetic resonance imaging (rtMRI) video in the midsagittal plane of a subject while speaking. However, in practice, the articulogram features are not directly available, unlike acoustic features from speech recording. Thus, we use an Acoustic-to-Articulatory Inversion method using a bidirectional long-short-term memory network which estimates the articulogram features from the acoustics. The proposed CDNN based approach using estimated articulatory features requires both acoustic and articulatory features during training but it requires only acoustic data during testing. Experiments are conducted using rtMRI videos from four subjects each speaking 460 sentences. The Pearson correlation coefficient is used to evaluate the speech rate estimation. It is found that the CDNN based approach gives a better correlation coefficient than the temporal and selected sub-band correlation (TCSSBC) based baseline scheme by 81.58% and 73.68% (relative) in seen and unseen subject conditions respectively.

Index Terms: speech rate estimation, convolutional dense neural network, bidirectional long-short-term memory, acoustic-to-articulatory inversion, articulogram

1. Introduction

Speech rate estimation is very important for speech understanding and speech recognition. For example, 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 [8, 9]. Thus, the speech rate estimation is crucial in many speech related applications.

Speech rate is defined as the number of speech units per second in a given speech recording. In this paper, we have considered syllables as speech units similar to most of the research works [1, 10, 11]. 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. For example, the approaches presented in [2, 3, 12, 13] use HMMs for accurate speech rate estimation. 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 [11]. Thus, in such cases, the speech rate is estimated using only the acoustic features of speech data.

The approaches presented in [14–19] use the acoustic features for accurate speech rate estimation. Zhang et al. [14] detected the syllable nuclei in a Hilbert envelope based contour using a rhythm guided peak counting method. In [15], syllable nuclei are located by identifying the prominent peaks in a smoothed loudness contour. In a similar way, an intensity-based envelope with simple peak counting using voicing decisions was proposed to estimate the speech rate [16]. A Gaussian mixture model-based method was proposed for classification of speech rate into slow, medium and fast classes and these class probabilities were used to determine the speech rate [17]. Jiao et al. [18] proposed a convex weighting criterion for spontaneous speech rate estimation avoiding heuristic peak detection strategy. Wang et al. proposed approaches [11, 19] using temporal and selected sub-band correlation-based feature contour (TCSSBC) which also includes a peak detection strategy involving smoothing and thresholding operations. In [19], the robustness of the TCSSBC approach was improved using the techniques such as false envelope peak filtering, neighbouring syllable smearing prevention, pseudo peak rejection, and parameter sensitivity analysis. Dekens et al. [20] presented a comparative study of the different speech rate estimation methods. This study found that the TCSSBC method performs better than other methods. Hence, the TCSSBC method [19] is used as the baseline approach in this work. The TCSSBC approach predicts the speech rate based on the number of peaks in the sub-band correlation contour. However, for noisy speech data, the peaks in sub-band correlation contour may not correspond to the syllable nuclei. Hence, the TCSSBC approach fails to give a robust speech rate estimation for noisy speech data. In the case of noisy speech data, we hypothesize that the articulatory features, either original or estimated from speech acoustics, could help in better estimation of the speech rate. This is because the characteristics of motion of the articulators such as tongue, jaw, velum, upper lip, and lower lip significantly vary with the

http://dx.doi.org/10.21437/Interspeech.2019-2295

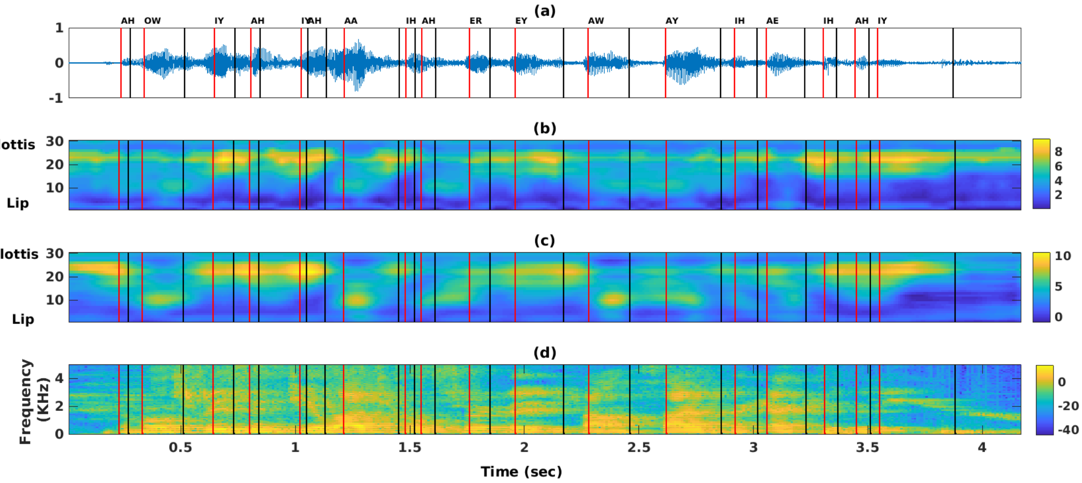

Figure 1: *Syllable nuclei representation in (a) speech waveform (b) interpolated articulogram (c) estimated articulogram (d) spectrogram (red and black lines represent the start and end boundaries of the syllable nuclei and the phonetic transcription for vowels is shown on speech waveform). The waveform is for a sentence "The most recent geological survey found seismic activity" spoken by M2 speaker.*

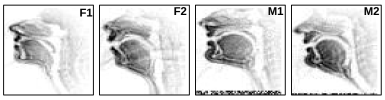

Figure 2: *Illustation of sample rtMRI video frames of F1, F2, M1, and M2 subjects*

changes in the speech rate [21–24]. In this paper, we propose a robust approach for the speech rate estimation using a Convolutional Dense Neural Network (CDNN) which uses the acoustic and articulatory features. In addition to this, the proposed approach predicts the speech rate using the articulatory features estimated from speech acoustics. To investigate the proposed CDNN based method, we need a database which contains both audio and articulatory motion data of a subject while speaking. Hence, in our work, we consider the USC-TIMIT corpus [25] which consists of audio and real-time magnetic resonance imaging (rtMRI) video in the midsagittal plane of multiple speakers while uttering a sentence. The speech data from this database is very noisy (due to the scanner noise of rtMRI) which is challenging for most of the speech rate estimation approaches.

In our work, articulograms [26] are used as articulatory features. The articulogram is defined as the sequence of vocal tract tube profile (VTTP) estimated from the rtMRI video. And from the corresponding speech recording, Mel Frequency Cepstral Coefficients (MFCCs) are obtained. In addition to the articulogram features from rtMRI video, we also experiment with estimated articulatory features as, in practice, articulatory features may not be directly available, unlike speech recording. We use an Acoustic-to-Articulatory Inversion (AAI) method [27, 28] which estimates the articulogram features from the speech data. The CDNN based approaches using MFCC ($\mathrm{CDNN_{mfcc}}$) and articulogram features ($\mathrm{CDNN_{arti}}$) are found to perform better than the baseline TCSSBC method. The CDNN based approach using the estimated articulogram features ($\mathrm{CDNN_{earti}}$) from AAI performs better than the $\mathrm{CDNN_{mfcc}}$, $\mathrm{CDNN_{arti}}$ and TCSSBC approaches. The $\mathrm{CDNN_{earti}}$ based approach requires both the articulatory and acoustic features while training but it requires only the acoustic data for testing.

2. Dataset

In this paper, USC-TIMIT [25] corpus is used. The USC-TIMIT database contains rtMRI videos of the upper airway in the midsagittal plane. The database consists of the rtMRI videos of five female and five male subjects speaking 460 sentences from MOCHA-TIMIT database [29]. The rtMRI video is recorded at 23.18 frames/sec and the corresponding audio is recorded at 20kHz sampling frequency. Each rtMRI video frame has a spatial resolution of 68 × 68 pixels (with a pixel dimension of $2.9\mathrm{mm} \times 2.9\mathrm{mm}$). For our work, a total of 4 subjects, comprising two female (F1, F2) and two male (M1, M2) are used for experimentation. The rtMRI videos for a total of 456, 460, 460, 460 sentences are available for F1, F2, M1, M2 subjects respectively. Figure 2 illustrates sample rtMRI frames of F1, F2, M1, and M2 subjects. It is clear that the vocal tract morphology changes across different subjects.

3. Methodology

The proposed approach of speech rate estimation consists of two primary steps: 1) Feature extraction and 2) CDNN-based speech rate estimation. These two steps are explained below.

3.1. Feature Extraction

Feature extraction plays an important role in speech rate estimation. The input features affect the overall performance of the model significantly. In our work, we use both acoustic and articulatory features which are explained below.

3.1.1. MFCC feature extraction

The MFCC features are obtained using the procedure described in [30]. For each sentence, the MFCC features are generated using a window length of 20 msec with a shift of 10 msec. For a given sentence, the MFCC feature representation has a dimension of $N \times 13$ where N is the number of windows.

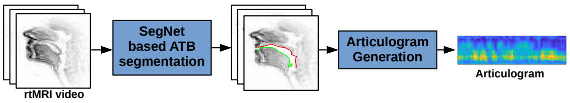

Figure 3: *Illustation of steps in the articulogram feature extraction method*

3.1.2. Articulogram feature extraction

The articulogram is a time-varying sequence of VTTP estimated from the rtMRI video. VTTP captures the vocal tract shape while a subject speaks. Thus, the articulogram represents the time-varying vocal tract shape during an utterance. The articulogram for a given rtMRI video is obtained following the steps described in [26]. Figure 3 illustrates the steps in the articulogram feature extraction method. Given an rtMRI video (corresponds to a sentence), consisting M number of frames, the obtained articulogram has a dimension of $M \times 30$. For each frame, 30 values are considered, each of which is the Euclidean distance between a point on the upper vocal tract boundary and a point on the lower vocal tract boundary. The 30 distances are considered starting from the lips to the glottis. The upper and lower vocal tract boundary shapes (red and green curves in Figure 3) are obtained using the air-tissue boundary (ATB) segmentation technique proposed in [31]. The ATB segmentation approach uses a convolutional encoder-decoder network (SegNet).

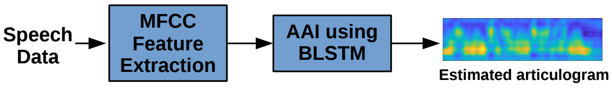

Figure 4: *Illustation of steps in the estimation of articulogram features using AAI*

3.1.3. Estimated articulogram feature generation

An acoustic-to-articulatory inversion (AAI) method [27] is used to estimate the articulograms from the acoustic features. Figure 4 illustrates the steps followed in the AAI based articulogram feature extraction method. The AAI method proposed in [27] uses a BLSTM network to estimate different articulatory points from the MFCCs of a speech recording. In a similar way, the bidirectional long-short-term memory (BLSTM) network with 128 nodes is used to generate the articulogram from the MFCCs of a given speech data. To train the BLSTM network, the articulogram features are interpolated from 23.14 frame rate to 100 Hz sampling rate using linear interpolation to have one-to-one correspondence with the input MFCCs. The dimension of the input to the BLSTM network is $K \times 13$ where K is the number of short-time windows. The output of the BLSTM network is a $K \times 30$ dimensional articulogram. Figure 1(b) and (c) illustrate the interpolated and estimated articulogram feature representations respectively with $K = 484$ for the given speech waveform.

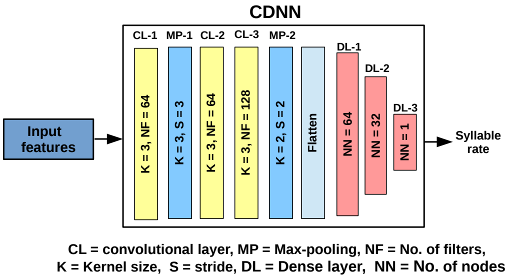

Figure 5: *Illustration of CDNN architecture*

3.2. CDNN-based speech rate estimation

The CDNN consists of two-dimensional convolutional, max-pooling, and dense layers. The two-dimensional convolutional filters extract the spatial and temporal characteristics from the given acoustic and articulatory features. Further, dense layers estimate the syllable rate using the convolutional layers' output. The number of dense layers, nodes, and the convolutional layers (with their filters' depth and width) of the CDNN architecture are optimized based on the validation loss so that the optimized network yields better performance for the given input features. At every convolutional and dense layer, relu activation function is used. To avoid overfitting, dropout of 0.20 is used at every dense layer. The syllable rate estimation is treated as a regression problem; hence, mean squared error (MSE) loss is optimized to train the CDNNs. We also experiment using the correlation coefficient based loss function to train the CDNN network. However, no improvement in the speech rate estimation accuracy is found compared to the MSE loss based optimized network. Figure 5 illustrates the CDNN architecture. The CDNN consists of four convolutional layers and three dense layers to predict the speech rate. We estimate the speech rate for a fixed length of input by dividing the feature sequence into one-second duration chunks.

The articulogram represents the vocal tract shape which varies significantly for different syllables. Thus, extracting the spatial and temporal characteristics of the articulogram features helps in better estimation of the speech rate. The CDNN based approach using the articulogram features is denoted as $\mathrm{CDNN}_{\mathtt{arti}}$. The articulogram from the rtMRI video have a dimension of 24×30 while the interpolated and estimated articulogram has a dimension of 100×30 corresponding to every one-second duration chunk. Thus, the CDNN architecture has different input layer dimensions depending on the type of the articulogram. The CDNN based approaches using the interpolated and estimated articulogram features are denoted as $\mathrm{CDNN}_{\mathtt{iarti}}$ and $\mathrm{CDNN}_{\mathtt{earti}}$ respectively. MFCCs capture the spectral energy distribution in a perceptually meaningful way. The convolutional filters extract the spectral and temporal characteristics of the MFCCs that are significant for speech rate estimation. For one-second duration of the speech recording, the input MFCC matrix has a dimension of 100×13. The CDNN based approach using the MFCC features is denoted as $\mathrm{CDNN}_{\mathtt{mfcc}}$.

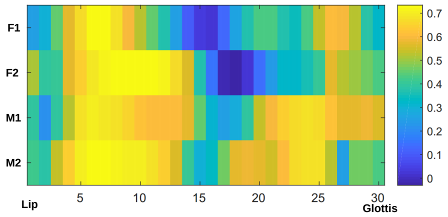

Figure 6: *Average CC computed for all the 30 distances in articulogram across all the sentences for the four subjects*

4. Experimental Setup

In our work, experiments are performed in two conditions using both acoustic and articulatory features: 1) seen subject condition and 2) unseen subject condition. In both seen and unseen subject conditions, the CDNN-based models are trained for a maximum of 60 epochs by imposing early stopping criterion based on the validation loss. The seen and unseen subject experiments are explained below.

Seen subject: In this experiment, all the CDNN based models are trained and tested using four-fold cross-validation. The total number of videos from each subject is divided into four sets where each set from F2, M1, M2 subjects contain 115 videos and each set from F1 subject contains 114 videos. Each fold comprises, a total of four sets considering one set from each subject. Likewise, four folds are created. From the four folds, two folds are used for training, one fold for validation and one fold for testing in a round robin fashion. For speech rate estimation, each video is used to create one-second long chunks with a shift of 0.5 second. Each fold, on average, consists of ∼4077, ∼2448, and ∼2272 train, validation and test chunks respectively.

Unseen subject: In this experiment, four-fold cross-validation is used to train and test the CDNN. In each fold, three subjects correspond to train and validation and one subject corresponds to the test data. Likewise, the subjects are chosen in a round robin fashion for the four folds. In the train and validation data, the total number of videos in each subject is divided into four sets where each set from F2, M1, M2 subjects contain 115 videos and each set from F1 subject contains 114 videos. For training, three sets from the three subjects (total 9 sets) are considered. For validation, remaining one set from three subjects (total 3 sets) are considered. Each video is used to create one-second long chunks and with a shift of 0.5 second. Each fold, on average, consists of ∼4893, ∼1704, and ∼2199 train, validation and test chunks respectively.

The BLSTM network in the AAI model is trained and tested using four-fold cross-validation. In each fold, three subjects are considered for training and one subject is considered for testing. And the network is trained for 30 epochs.

Evaluation Metric: The performance of the proposed CDNN based approach is evaluated based on the Pearson correlation coefficient (ρ) between the ground truth syllable rate and the estimated syllable rate across all the test sentences.

Table 1: *Average ρ for each subject in seen subject condition (green and blue colours indicate the first and second best ρ values for each subject respectively))*

Sub	F1	F2	M1	M2	*Avg* ± *std*
TCSSBC	0.44	0.39	0.24	0.44	0.38 ± 0.10
CDNN$_{\mathbf{arti}}$	0.65	0.58	0.43	0.64	0.57 ± 0.10
CDNN$_{\mathbf{iarti}}$	0.66	0.60	0.43	0.64	0.58 ± 0.10
CDNN$_{\mathbf{earti}}$	**0.70**	**0.71**	**0.64**	**0.69**	0.69 ± 0.03
CDNN$_{\mathbf{mfcc}}$	**0.68**	**0.65**	**0.52**	**0.64**	0.62 ± 0.07

5. Results and Discussions

Table 1 and Table 2 show the average (Avg) ρ (± standard deviation (std)) values for TCSSBC and all the CDNN based methods in seen and unseen subject conditions respectively. From the results, it is observed that all the proposed CDNN based approaches perform better than the baseline TCSSBC approach. This improvement in performance is primarily because the CDNN based approaches are supervised unlike TCSSBC approach. The acoustic feature based method $\mathrm{CDNN_{mfcc}}$ performs better than the articulatory feature based methods such as $\mathrm{CDNN_{arti}}$ and $\mathrm{CDNN_{iarti}}$. Although we use the segnet model [31] for ATB prediction due to its best performance so far in the literature, the predicted upper and lower vocal tract boundaries are not as accurate as the ground truth ATBs. Thus in VTTP generation, the distance between the upper and lower vocal tract boundaries may not always be accurate enough to follow the vowel and consonant variation which could affect the final speech rate estimation. We assess the performance of AAI using correlation coefficient (CC) for each dimension of the VTTP from the original and estimated articulograms for every sentence. Figure 6 illustrates the average CC obtained across 460 sentences for all 30 dimensions of VTTP. It is observed that the correlation coefficients at the tongue region are high for all the subjects. When averaged across all 30 dimensions, the AAI method yields the mean CC values of 0.42, 0.45, 0.54, 0.55 for F1, F2, M1, M2 subjects respectively.

Table 2: *Average ρ for each subject in unseen subject experiments (green and blue colours indicate the first and second best ρ values for each subject respectively))*

Sub	F1	F2	M1	M2	*Avg* ± *std*
TCSSBC	0.44	0.39	0.24	0.44	0.38 ± 0.10
CDNN$_{\mathbf{arti}}$	**0.53**	0.35	0.37	0.44	0.44 ± 0.08
CDNN$_{\mathbf{iarti}}$	0.48	0.33	**0.57**	0.40	0.45 ± 0.11
CDNN$_{\mathbf{earti}}$	**0.68**	**0.70**	**0.72**	**0.56**	0.66 ± 0.07
CDNN$_{\mathbf{mfcc}}$	0.55	**0.50**	0.44	**0.48**	0.49 ± 0.05

The $\mathrm{CDNN_{earti}}$ performs better than all the other CDNN based approaches. Considering the Avg ρ value across all the subjects, the $\mathrm{CDNN_{earti}}$ approach yields 81.58% and 73.68% improvement over the baseline TCSSBC approach for seen and unseen subject conditions respectively. In the AAI method, the BLSTM network is trained using MFCC features as input and corresponding articulogram features as output. Thus the estimated articulogram features can be treated as function of both acoustic and articulatory features. The $\mathrm{CDNN_{earti}}$ based approach exploits the multimodal acoustic and articulogram features. Figure 1 shows the syllable nuclei on the interpolated articulogram, estimated articulogram and spectrogram of a sample utterance. It is observed that in the estimated articulogram, the distance between the upper and lower vocal tract boundary is increased in the anterior tongue region for the syllable nuclei (in the vowel region) compared to that in the original articulogram. The tongue movement is significantly affected by the vowels. This could be a reason why, the estimated articulogram performs better than the other methods. Hence, in noisy speech data condition, considering the articulatory features helps in a better speech rate estimation.

Comparing Table 1 and 2 results, it is observed that the CDNN based methods' results are not consistent across all the subjects in the unseen subject condition. This is due to having a limited number of subjects (four) which affects the generalizability of the network for a new test subject.

6. Conclusion

In this paper, a CDNN based approach using acoustic and articulatory features is proposed for speech rate estimation. The proposed approach gives significantly higher performance than the baseline TCSSBC approach in both seen and unseen subject conditions. In order to avoid the need for articulatory features directly from rtMRI video during test, inversion method is used to estimate articulatory features from acoustics. We found that the articulatory features estimated from speech acoustics perform better than acoustic and articulatory features separately. As a part of the future work, 1) we plan to extend our work by implementing an end-to-end model which extracts the articulatory features from the speech data and estimates the speech rate using acoustic and estimated articulatory features. 2) We use more number of speakers in training so that the model can have more generalizability to unseen test speakers. 3) We experiment with different noisy conditions to obtain a robust speech rate estimation model.

7. Acknowledgement

We thank the Department of Science & Technology, Government of India and the Pratiksha Trust for their support.

8. References

[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. Hönig, A. Batliner, and E. Nöth, "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, Universität 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. Cappé, 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. Min Lee, A. Kazemzadeh, Z. Deng, S. Lee, S. Narayanan, and C. Busso, "An acoustic study of emotions expressed in speech." 01 2004.

[8] 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.

[9] 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.

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

[11] 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.

[12] T. Cincarek, R. Gruhn, C. Hacker, E. Noeth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-natives first language," *Computer Speech & Language*, vol. 23, pp. 65–88, 01 2009.

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

[14] Y. Zhang and J. R. Glass, "Speech rhythm guided syllable nuclei detection," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3797–3800.

[15] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, vol. 2, May 1998, pp. 945–948 vol.2.

[16] 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.

[17] R. Faltlhauser, T. Pfau, and G. Ruske, "On-line 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.

[18] Y. Jiao, V. Berisha, M. Tu, and J. Liss, "Convex weighting criteria for speaking rate estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1421–1430, Sep. 2015.

[19] S. Narayanan and Dagen Wang, "Speech rate estimation via temporal correlation and selected sub-band correlation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, March 2005, pp. I/413–I/416 Vol. 1.

[20] T. Dekens, M. Demol, W. Verhelst, and P. Verhoeve, "A comparative study of speech rate estimation techniques," in *INTERSPEECH*, January 2007, pp. 510–513.

[21] S. G. Adams, G. Weismer, and R. D. Kent, "Speaking rate and speech movement velocity profiles," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 1, pp. 41–54, 1993.

[22] T. Gay, "Mechanisms in the control of speech rate," *Phonetica*, vol. 38, no. 1-3, pp. 148–158, 1981.

[23] T. Gay, T. Ushijima, H. Hirose, and F. S. Cooper, "Effect of speaking rate on labial consonant-vowel articulation," *The Journal of the Acoustical Society of America*, vol. 55, no. 2, pp. 385–385, 1974.

[24] O. Engstrand, "Articulatory correlates of stress and speaking rate in swedish VCV utterances," *The journal of the Acoustical society of America*, vol. 83, no. 5, pp. 1863–1875, 1988.

[25] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

[26] A. Prasad, V. Periyasamy, and P. K. Ghosh, "Estimation of the invariant and variant characteristics in speech articulation and its application to speaker identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4265–4269.

[27] A. Illa and P. Kumar Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory," in *INTERSPEECH*, September 2018, pp. 3122–3126.

[28] A. Illa and P. K. Ghosh, "The impact of speaking rate on acoustic-to-articulatory inversion," *Computer Speech & Language*, vol. 59, pp. 75–90, 2020.

[29] A. A. Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *5th Seminar of Speech Production*, 2000, pp. 305–308.

[30] S. Young and S. Young, "The HTK Hidden Markov Model Toolkit: Design and philosophy," *Entropic Cambridge Research Laboratory Ltd*, vol. 2, pp. 2–44, 1994.

[31] C. Valliappan, K. Avinash, R. Mannem, K. Girija Ramesan, and G. Prasanta Kumar, "An improved air tissue boundary segmentation for real-time magnetic resonance imaging video using segnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.