Attention and Encoder-Decoder based models for transforming articulatory movements at different speaking rates

Abhayjeet Singh, Aravind Illa, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India

spirelab.ee@iisc.ac.in

Abstract

While speaking at different rates, articulators (like tongue, lips) tend to move differently and the enunciations are also of different durations. In the past, affine transformation and DNN have been used to transform articulatory movements from neutral to fast(N2F) and neutral to slow(N2S) speaking rates [1]. In this work, we improve over the existing transformation techniques by modeling rate specific durations and their transformation using AstNet, an encoder-decoder framework with attention. In the current work, we propose an encoder-decoder architecture using LSTMs which generates smoother predicted articulatory trajectories. For modeling duration variations across speaking rates, we deploy attention network, which eliminates the need to align trajectories in different rates using DTW. We perform a phoneme specific duration analysis to examine how well duration is transformed using the proposed AstNet. As the range of articulatory motions is correlated with speaking rate, we also analyze amplitude of the transformed articulatory movements at different rates compared to their original counterparts, to examine how well the proposed AstNet predicts the extent of articulatory movements in N2F and N2S. We observe that AstNet could model both duration and extent of articulatory movements better than the existing transformation techniques resulting in more accurate transformed articulatory trajectories.

Index Terms: Encoder-Decoder, Attention, LSTM, Speaking rate, Electromagnetic Articulography

1. Introduction

Speech production involves planning and coordination of different articulators including, lips, jaw, tongue and velum [2]. Apart form phonetic information, various para-linguistic factors could influence the articulatory movements, and in turn, acoustics. One such factor is speaking rate, which is defined as the number of phonemes spoken per second [3]. Speaking rate affects acoustic properties such as vowel duration [4], vowel formant frequencies [5, 6, 7], consonant-vowel co-articulation [8], average syllable duration [9] and pronunciation [10] due to changes in the articulation [11, 12]. Variation in acoustic features because of changing speaking rate directly impacts the performance of several speech applications including automatic speech recognition (ASR) [13, 14, 15] that are typically designed for speech characterized by the average speaking rate. Previous studies also reported speaking rate specific changes in articulation such as the rate of articulation, range of articulatory movements and the degree of co-articulation [16]. Understanding the nature of articulatory dynamics at different speaking rates and their inter relationships could help in developing speech based systems that are robust to variations in speaking rate.

In this work, we learn a transformation function which maps the articulatory movements from neutral to fast speaking rate (N2F), and also for neutral to slow speaking rate (N2S) using an encoder-decoder framework with attention network. The transformation should be able to learn the variations in duration and range of movements of articulators from source speaking rate to target speaking rate. Preliminary work has been done on this topic [1], where variations in the range of articulatory movements are modeled using various transformation methods like an affine transformation with both diagonal and full matrix and a non-linear transformation modeled by a deep neural network (DNN). As all these transformation methods learn work at the frame level, which require dynamic time warping (DTW) to time-align articulatory movement trajectories at different rates. To learn the optimal transformation function (TF), an iterative approach was followed to optimize TF and the optimal warping path till convergence is achieved. The limitations of this approach lies in modelling the duration variations across speaking rates. To overcome these shortcomings in [1], the proposed approach utilizes attention mechanism to capture the duration variations to improve the performance. The proposed encoder-decoder based attention network used to perform Articulatory trajectories Speaking rate Transformation is denoted as AstNet. AstNet is an end-to-end model which takes articulatory movements at one rate as input and generates at a different rate. This is in contrast with the iterative approach followed in [1]. In addition, we don't need to time-align prior to training the model as the attention network learns the duration modelling and by using "stop token" method, end of sequence is predicted at the time of inference. We use long-short-term memory (LSTM) units in both encoder and decoder, and this results in preserving the smoothness characteristics in the predicted articulatory movements.

The experimental setup and data-set for this work are similar to the preliminary work [1]. The proposed AstNet model outperforms the technique proposed in [1] in both cases (N2F and N2S) for all subjects. DTW distance is used as an evaluation metric, with lower distances indicating better performance. As it was reported that different sound units are affected differently in different speaking rates [4], we perform analysis on the duration of phonemes at the source (neutral) rate, ground-truth duration at the target rate and the same at target rate following articulatory movement prediction. We also analyze the range of articulatory movements by computing standard deviation of articulatory trajectory (SDAT) [11] at source (neutral) rate as well as original and predicted trajectories at target rate. These analysis on duration and range of articulatory movements could help to understand the modeling capabilities of AstNet.

2. Dataset

In this study, we collected articulatory data for 460 sentences spoken at three different rates: neutral, fast and slow. This data was collected from 5 subjects out of which three were male

(M1, M2, M3) and two were female (F1, F2) of age 19, 22, 24, 28 and 22 respectively. All subjects are native Indians reported to have no speech disorders. We took 460 phonetically balanced sentences from MOCHA-TIMIT corpus [17]. The phonetic transcriptions for these sentences were obtained using SONIC speech recognizer [18], which uses 55 phoneme set of American English in its lexicon [1]. Among these, five phonemes (ls, ga, SIL, br and lg) did not appear in the force-alignment output. In this study, we ignore the speaker specific phoneme variations. Figure 1 shows the placement [19] of nine articulators among which seven are along the mid-sagittal plane (indicated by X and Z directions) and the remaining two (RC and LC) are in frontal plane, to record articulatory movements using electromagnetic articulograph AG501 [20]. We obtain 18 articulatory trajectories corresponding to Upper Lip (ULx, ULz), Lower Lip (LLx, LLz), Right Commissure of Lip (RCx, RCz), Left Commissure of Lip (LCx, LCz), Jaw (JAWx, JAWz), Throat (THx, THz), Tongue Tip (TTx, TTz), Tongue Body (TBx, TBz), Tongue Dorsum (TDx, TDz), for each utterance. Articulatory movements are recorded at sampling rate of 250 Hz. It is known that the articulatory movements are low pass in nature [21], therefore we first low pass filtered with a cutoff frequency 25 Hz to avoid high frequency noise incurred in EMA data acquisition. These are further down-sampled from 250 Hz to 100 Hz. We also recorded synchronous acoustic of speech using A t.bone EM9600 shotgun [22], unidirectional electret condenser microphone.
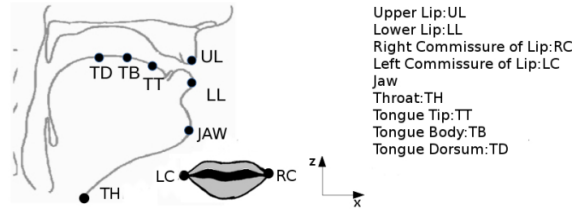

Figure 1: *Schematic diagram indicating the placement of EMA sensors [23].*

The recording of the sentences in the three speaking rates for each subject was held in three different sessions. In the first session, the subject was asked to speak in his/her normal speaking rate, from which, after silence removal, their neutral speaking (phone) rate was computed. In the subsequent sessions the subjects were required to speak at 2 times their neutral speaking rate to record fast speech and reduce their speaking rate by half to record slow speech. For this work, we utilized the same dataset used in our previous work [1], where more details regarding the dataset can be found.

3. AstNet: Transforming articulatory movements to slow/fast speaking rate

Mapping between articulatory trajectories spoken at different rates is challenging, as the duration of an articulatory trajectory is inversely proportional to its speaking rate and different phoneme durations vary differently with speaking rates [24]. To model duration variations while transforming speaking rate, we propose AstNet, an encoder-decoder architecture with a location sensitive attention to learn alignments between sequences of varying durations.

Figure 2, illustrates the AstNet architecture, which consists of encoder, attention and decoder. The encoder com-

---

[1]https://www.yumpu.com/en/document/read/9555742/sonic-the-university-of-colorado-continuous-speech-recognizer
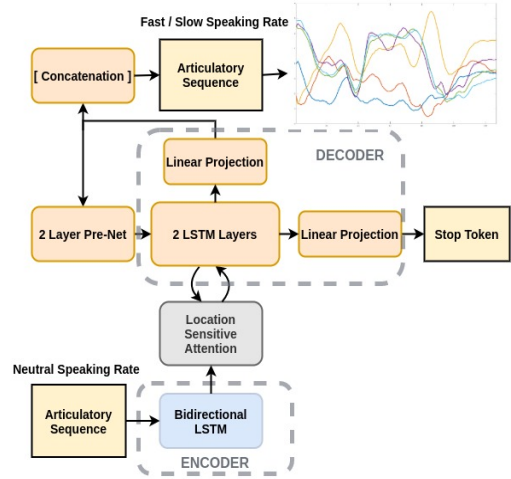

Figure 2: *Block diagram representing the AstNet architecture* prises a single BLSTM layer with 512 hidden units, which takes an articulatory sequence at a specific rate ($rate_{inp}$) as input $S_1 = \{s_1, s_2, ..s_N\}$ and maps it to the hidden states $E = \{e_1, e_2, ..e_N\}$, which acts as an input to the attention. The attention network models the time alignment between the encoder and decoder hidden states. The decoder hidden states, $D = \{d_1, d_2, ..d_M\}$, are utilized to generate the articulatory movement trajectories at a specific target speaking rate ($rate_{out}$). Attention network is a location-sensitive attention network which iterates over previous decoder output (d_{t-1}) and attention weights ($\alpha_{n,t-1}$) and all the encoder hidden states (E). The attention mechanism is governed by the equations below [25]:

$$\alpha_{n,t} = \sigma(\text{score}(d_{t-1}, \alpha_{n,t-1}, E)) \tag{1}$$

$$\text{score: } s_{n,t} = w^T \tanh(W d_{t-1} + V e_n + U f_{j,t} + b) \tag{2}$$

$$\text{Context vector: } g_t = \sum_{j=1}^{N} \alpha_{j,t} e_j \tag{3}$$

where $\alpha_{n,t}$ are *attention weights* and $n\epsilon\{1, 2..N\}$ and parameters of the attention network are denoted by W, V, U weight matrices and w, b denote the projection and bias vectors, respectively. In Eq.(2), $f_{j,t}$ is computed by $f_t = F * \alpha_{t-1}$, which incorporates location awareness to the attention network [25], where F is a weight matrix and α_{t-1} is the set of previous time-step alignments. In Eq.(2), attention scores are computed as a function of encoder outputs (e_n) and previous attention weights ($\alpha_{n,t-1}$) and decoder output (d_{t-1}), which are further normalized using *softmax* to obtain attention weights. These obtained attention weights are utilized to compute fixed context vector as described by Eq.(3).

The decoder consists of two uni-directional LSTM layers with 1024 units, followed by a linear projection layer to predict the articulatory movements at $rate_{out}$. The decoder computes the final output ($d_t \sim Decoder(d_{t-1}, g_t)$) from the previous state output (d_{t-1}) and attention context vector (g_t). To compute d_{t-1}, decoder's previous output is transformed by a two layered fully-connected network with 256 units (Pre-Net). The decoder hidden state outputs are further projected using two linear layers, one for articulatory sequence prediction and other to predict end of the sequence. For end sequence prediction, the decoder LSTM output and attention context are concatenated and projected down to a scalar and then passed through sigmoid activation to predict the probability that the output sequence has completed. This "Stop Token" prediction is used

during inference to allow the model to dynamically determine when to terminate generation instead of always generating for a fixed duration. We vary the number of maximum steps taken by decoder according to $rate_{out}$ articulatory sequences' maximum duration. This hyper parameter comes into play only at the time of inference, to bound the upper limit of the predicted sequence's duration.

4. Experimental Setup

In this work, we perform two different rate transformations using the AstNet, the first one is the neutral rate to fast rate transformation of articulatory movements (N2F) and the other one is the neutral rate to slow rate articulatory movement transformation (N2S). In both the cases N2F and N2S, we use a four fold setup. We divide the data into training and test set in the ratio of 3:1. Before training, articulatory trajectories are first made zero-mean and then transformations are learnt separately for N2F and N2S cases in each of the four folds from five subjects.

Training Approach: Typically neural network models demand large amount of training data to achieve better performance. To overcome the scarcity of articulatory data to train a subject specific model, we deploy the training approach proposed in [23]. At the first step, we pool the training data from all the subjects and train a generic model to learn the mapping from an articulatory trajectory at $rate_{inp}$ to an articulatory trajectory at $rate_{out}$. This is named as Generalized or Generic training. In the second stage, we fine tune the generic model weights with respect to the target speaker, to learn speaker specific models. We also train subject-dependent models for each subject.

Evaluation procedure: For a given scheme, we indicate $D_T = \{d_1, d_2, d_3, d_4\}$, where $d_i \in \mathbb{R}^{1 \times 115}$ consists of the DTW distances between the articulatory trajectory at $rate_{out}$ and the corresponding ground-truth trajectory for the 115 test utterances in the ith fold. In order to evaluate the performances of different schemes, we report the average and standard deviation of D_T for every subject. Lower the average D_T for a given scheme, better is the performance of the model. Therefore, the best scheme is the one which results in the least average D_T.

Table 1: *Average (standard deviation) of D_T (in mm) for N2F transformation.*

Model	F1	F2	M1	M2	M3
IT-DTW	6.51	6.60	6.71	6.00	5.64
	(0.85)	(1.09)	(1.18)	(0.93)	(1.04)
Baseline [1]	4.79	5.04	4.72	4.86	4.58
	(0.63)	(0.90)	(0.90)	(0.76)	(0.90)
AstNet (Subj Dep)	5.21	7.65	5.57	5.83	4.74
	(1.17)	(1.49)	(1.46)	(1.57)	(1.22)
AstNet (Generic)	4.92	5.18	4.82	4.83	4.63
	(0.60)	(0.78)	(0.90)	(0.73)	(0.77)
AstNet (finetuned)	**4.69**	**4.95**	**4.59**	**4.65**	**4.45**
	(0.57)	(0.76)	(0.84)	(0.74)	(0.78)

5. Results and Discussion

First we compare the performance of AstNet with baseline approach, then we perform analysis on the transformed articulatory trajectories to verify speaking rate specific characteristics.

Table 2: *Average (standard deviation) of D_T (in mm) for N2S transformation*

Model	F1	F2	M1	M2	M3
IT-DTW	6.46	8.27	7.32	6.03	6.30
	(0.90)	(1.27)	(1.07)	(0.77)	(0.89)
Baseline [1]	5.05	7.37	6.30	5.35	5.52
	(0.89)	(1.11)	(1.03)	(0.75)	(0.80)
AstNet (Subj Dep)	7.86	12.70	9.67	9.18	8.29
	(1.10)	(1.53)	(1.36)	(1.02)	(1.74)
AstNet (Generic)	5.80	8.33	7.44	5.89	5.98
	(0.96)	(1.31)	(1.31)	(0.82)	(0.95)
AstNet (finetuned)	**4.76**	**6.83**	**6.10**	**5.01**	**4.95**
	(0.83)	(1.14)	(1.27)	(0.89)	(0.92)

5.1. Performance of AstNet:

Estimation of articulatory movement at different speaking rates has been done in [1], where they consider learning one-to-one mapping between articulatory movements at different rates using three schemes: 1. Full affine transformation matrix, 2. Diagonal affine transformation matrix and 3. Non-Linear transformation function. To learn one-to-one mapping they align the articulatory movements at different rates using Dynamic Time Warping (DTW), so that all corresponding articulatory trajectories at $rate_{inp}$ and $rate_{out}$ are of equal durations. We consider the best results from [1] as our baseline i.e., the lowest possible average DTW values for all five subjects, and compare the results of AstNet. Table 1 reports the DTW values of different approaches for N2F transformation. In first row we report DTW distance between original neutral and fast speaking rate, as a naive baseline with identity transformation (IT-DTW), followed by DTW values reported in [1] as a baseline. The rest of the values in Table 1 from third to the last rows correspond to AstNet with different training schemes. Similarly, for N2S transformation we report DTW values of respective models in Table 2.
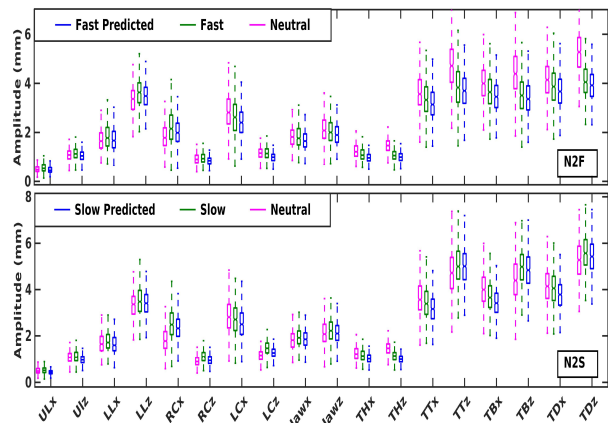

Figure 3: *Standard deviation of articulatory trajectories (SDAT) over 18 articulators (N2F and N2S)*

In both the tables subject-dependent models have the lowest performance (highest DTW distance) among different AstNet models for all subjects. It is also to be noted that the DTW distances are poor when compared with IT-DTW. This is due to the lack of training data for each subject's model. In case of generic models in which we train a single model for all subjects, the performance is better than subject-dependent models and IT-DTW. Generic models learn the transformations for N2F and N2S cases which are common for all subjects. We further
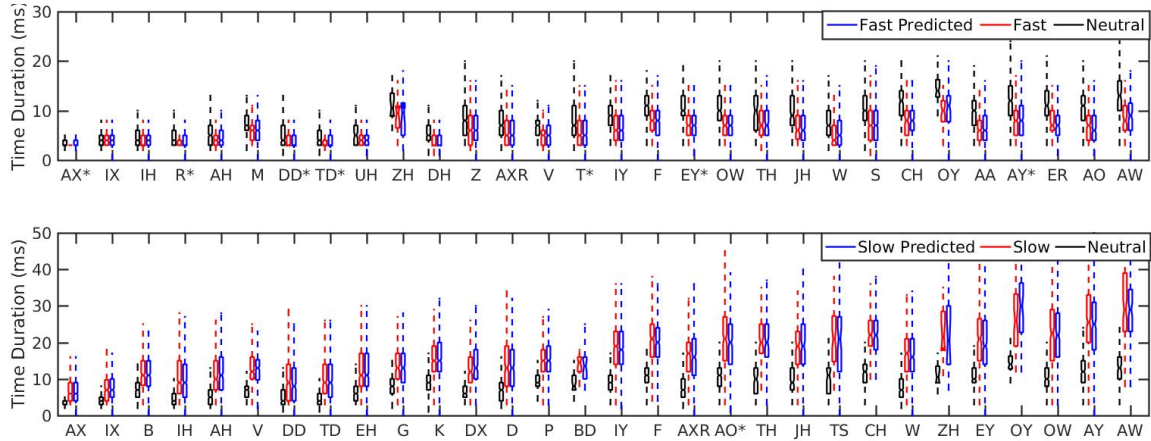
Figure 4: *Duration analysis on a 50 phoneme set for N2F and N2S transformation (only 30 phonemes are shown).*

train this generic model over each subject separately, giving us fine-tuned models, one for each subject. The finetuned AstNet performs better among all models including the baseline [1], as fine-tuned models have learnt all the generic transformations from multiple speakers' data and also the subject specific transformations in both N2F and N2S cases. In addition to reporting improvements using AstNet over the baseline [1], we predict the duration of the $rate_{out}$ articulatory trajectory. The relative percentage improvements of AstNet finetune over baseline [1] for all subjects (F1, F2, F3, M1, M2) are 2.09, 1.79, 2.75, 4.32 and 2.84% respectively and for N2F transformation 5.74, 7.33, 3.17, 6.36 and 10.33% respectively for N2S transformation. Despite the fact that the architecture complexity of AstNet is much more than the models used in [1] and the corresponding relative percentage improvements are all under 11%, it is still justifiable that AstNet is more suitable for rate transformation of articulatory movements as it also predicts the duration of the target articulatory movements at $rate_{out}$.

5.2. Analysis on the range of articulatory movements:

Due to the variation in the speaking rate, the range of movement of articulators is affected. For example at fast speaking rate, the articulators might undergo lingual undershoot thus impacting there displacement [26, 27] and similarly at slower rate due to hyper-articulation articulatory movements for some phones peak. To verify to what extent AstNet has learnt amplitude scaling, we compute the standard deviation of articulatory trajectories (SDAT) for each sentence [11]. Figure 3 report the analysis on characterizing range of movements of articulators which are transformed using AstNet in comparison with input neutral and original trajectories at target speaking rate. In Figure 3 we perform amplitude analysis over all 18 articulators for both N2F (top subplot) and N2S (bottom subplot) transformations. For each articulator, we box-plot SDAT over 460 utterances original trajectories at neutral rate as well as original and predicted ones at fast/slow rates. For most of the articulators in N2F, we observe there is a reduction in SDAT in fast rate compared to neutral rate, when original articulators are considered. The same trend of reduction in SDAT is also observed for transformed articulatory trajectories. Similar trends are observed in N2S, SDAT increases from neutral to slow rate for both original and transformed trajectories.

5.3. Phone specific duration analysis:

Figure 4 illustrates phoneme duration for N2F and N2S trans-

formations in top and bottom subplots, respectively. We have performed this analysis over a set of 50 phonemes and for each phoneme we box-plot the duration of all utterances for neutral, fast and the predicted fast articulatory movements. We have rank ordered (ascending) them in terms of the absolute median different between the duration at source neutral rate and target fast/slow rates. The durations of the top 15 and bottom 15 from this ranked ordered phonemes are illustrated in Figure 4. We observe in Figure 4 (top plot) that, for all phones, neutral duration values are higher than both fast and fast-predicted, as time taken to speak any phoneme at the neutral rate will be longer than that at the fast rate. It is also observed that the durations of the original and transformed articulatory trajectories are similar. A phoneme specific paired t-test between the original durations and the ones predicted by N2F AstNet shows that the durations of the original and transformed articulatory trajectories are significantly ($p <0.01$) different only for seven phonemes, namely 'AX', 'R', 'DD', 'TD', 'T', 'EY', 'AY'. This suggests that AstNet is able to learn phone level duration transformations. This is also true with N2S AstNet, where we observe an increase in the duration from neutral to slow speaking rate and the duration of the original and transformed trajecotries are significantly ($p <0.01$) different only for one phoneme, namely 'AO'.

6. Conclusions

In this work, we have modeled transformation of articulatory movements (N2F and N2S) using encoder-decoder framework with attention network (AstNet), which outperforms the baseline [1]. Using this model not only do we estimate articulatory movement trajectory from one to another speaking rate, but also predict the duration of the resulting trajectory thus eliminating the need of pre-processing time-alignment of trajectories. We analyze phoneme specific transformations among articulatory movements corresponding to different speaking rates, as it is reported that different sound units are affected differently in different speaking rates [4]. We also perform an amplitude analysis of predicted and ground-truth trajectories using SDAT and validating that apart from learning time-scaling, model also learns amplitude-scaling.

7. Acknowledgements

8. References

[1] A. Singh, G. N. Meenakshi, and P. K. Ghosh, "Relating articulatory motions in different speaking rates," in *Proc. Interspeech 2018*, 2018, pp. 2992–2996.

[2] L. Goldstein and C. A. Fowler, "Articulatory phonology: A phonology for public language use," *Phonetics and phonology in language comprehension and production: Differences and similarities*, pp. 159–207, 2003.

[3] P. Barbosa, "R.h. stetson, motor phonetics: A study of speech movements in action, 2nd ed., amsterdam, north holland publishing co., 1951," *Phonetica*, vol. 74, pp. 255–258, 11 2017.

[4] H. Kuwabara, "Acoustic and perceptual properties of phonemes in continuous speech as a function of speaking rate," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[5] B. Lindblom, "Spectrographic study of vowel reduction," *The journal of the Acoustical society of America*, vol. 35, no. 11, pp. 1773–1781, 1963.

[6] T. Gay, "Effect of speaking rate on vowel formant movements," *The journal of the Acoustical society of America*, vol. 63, no. 1, pp. 223–230, 1978.

[7] S.-J. Moon and B. Lindblom, "Interaction between duration, context, and speaking style in English stressed vowels," *The Journal of the Acoustical society of America*, vol. 96, no. 1, pp. 40–55, 1994.

[8] A. Agwuele, H. M. Sussman, and B. Lindblom, "The effect of speaking rate on consonant vowel coarticulation," *Phonetica*, vol. 65, no. 4, pp. 194–209, 2008.

[9] J. L. Miller, F. Grosjean, and C. Lomanto, "Articulation rate and its variability in spontaneous speech: A reanalysis and some implications," *Phonetica*, vol. 41, no. 4, pp. 215–225, 1984.

[10] E. Fosler-Lussier and N. Morgan, "Effects of speaking rate and word frequency on pronunciations in convertional speech," *Speech Communication*, vol. 29, no. 2-4, pp. 137–158, 1999.

[11] 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.

[12] S. R. Kuberski and A. I. Gafos, "The speed-curvature power law in tongue movements of repetitive speech," *Plos one*, vol. 14, no. 3, p. e0213851, 2019.

[13] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris *et al.*, "Automatic speech recognition and speech variability: A review," *Speech communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[14] B. T. Meyer, T. Brand, and B. Kollmeier, "Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes," *The Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 388–403, 2011.

[15] R. M. Stern, A. Acero, F.-H. Liu, and Y. Ohshima, "Signal processing for robust speech recognition," in *Automatic Speech and Speaker Recognition*. Springer, 1996, pp. 357–384.

[16] J. Berry, "Speaking rate effects on normal aspects of articulation: Outcomes and issues," *Perspectives on Speech Science and Orofacial Disorders*, vol. 21, no. 1, pp. 15–26, 2011.

[17] A. Wrench, "Mocha-TIMIT speech database," *The 18th International Conference on Pattern Recognition*, 1999.

[18] B. Pellom, "Sonic: The university of colorado continuous speech recognizer," Tech. Rep., 2001.

[19] A. K. Pattem, A. Illa, A. Afshan, and P. K. Ghosh, "Optimal sensor placement in electromagnetic articulography recording for speech production study," *Computer Speech & Language*, vol. 47, pp. 157–174, 2018.

[20] "3d electromagnetic articulograph," available online: http://www.articulograph.de/, last accessed: 4/2/2020.

[21] P. K. Ghosh and S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2162–2172, 2010.

[22] "t.bone em 9600," available online:https://www.tbone-mics.com/en/product/information/details/the-tbone-em-9600-richtrohr-mikrofon/ , last accessed: 13/5/2020.

[23] A. Illa and P. K. Ghosh, "Low resource acoustic-to-articulatory inversion using bi-directional long short term memory," in *Proc. Interspeech*, 2018, pp. 3122–3126.

[24] M. Muto, H. Kato, M. Tsuzaki, and Y. Sagisaka, "Effect of speaking rate on the acceptability of change in segment duration," *Speech Communication*, vol. 47, no. 3, pp. 277 – 289, 2005.

[25] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[26] J. E. Flege, "Effects of speaking rate on tongue position and velocity of movement in vowel production," *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 901–916, 1988.

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