Interspeech 2018
2-6 September 2018, Hyderabad

# Relating articulatory motions in different speaking rates

*Astha Singh*[1], *G. Nisha Meenakshi*[2], *Prasanta Kumar Ghosh*[2]

[1]Electrical Communication Engineering, [2]Electrical Engineering, Indian Institute of Science, Bangalore-560012, India

astha@iisc.ac.in, nishag@iisc.ac.in, prasantg@iisc.ac.in

## Abstract

Movements of articulators (e.g., tongue, lips and jaw) in different speaking rates are related in a complex manner. In this work, we examine the underlying function to transform articulatory movements involved in producing speech at a neutral speaking rate into those at fast and slow speaking rates (N2F and N2S). For this we use articulatory movement data collected from five subjects using an Electromagnetic articulograph at neutral, fast and slow speaking rates. As candidate transformation functions (TF), we use affine transformations with a diagonal matrix and a full matrix and a nonlinear function modeled by a deep neural network (DNN). Since the duration of an utterance in different speaking rates would typically be unequal, it is required to time align the articulatory movement trajectories, which, in turn, affects the TF learnt. Therefore, we propose an iterative algorithm to alternately optimize for the TF and the time alignments. Subject specific experiments reveal that while N2F transformation can be well described by an affine transformation with a full matrix, N2S transformation is better represented by a more complex nonlinear function modeled by a DNN. This could be because subjects exhibit gross articulatory movements during fast speech and hyper-articulate while producing slow speech.
**Index Terms**: Electromagnetic Articulography, Speaking rate, Deep Neural Network

## 1. Introduction

Speaking rate is one among the many important variations in the speaking style exhibited by humans. It is typically expressed as the number of phonemes per second [1]. Speaking rate depends on several speaker specific factors, including, gender and regional dialect [2, 3], age and mood of the speaker, noise in surrounding environment [4] and length of the phrase being uttered [5]. It is known that speaking rate changes several acoustic properties such as vowel duration [6], vowel formant frequencies [7], [8], [9], consonant-vowel co-articulation [10], average syllable duration [11] and pronunciation [12]. Such speaking rate specific acoustic changes pose challenges to several speech applications including automatic speech recognition (ASR) [13], [14], [15] that are typically designed for speech characterized by the average speaking rate.

In addition to changing acoustic properties, speaking rate has a direct impact on articulation. Studies report that there exists a variation in the rate of articulation, range of articulatory movements and the degree of co-articulation [16]. For example, in slow speaking rate, speakers tend to articulate slowly with an increased number of pauses and hyper-articulation in order to make speech more intelligible [17]. It is also known that the positions of the articulators vary across fast, slow and neutral speaking rates [18]. These results suggest that the motor planning is different for different speaking rates such as fast and slow [19], [20]. Although such differences in articulation strategies have been reported, the relationship between the articulation during neutral speech and speech in different speaking rates remains to be examined. Understanding the nature of articulatory differences among speech of 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 consider three speaking rates– neutral (N), slow (S) and fast (F) and examine the transformation between the articulation 1) in N and that in F and 2) in N and that in S. For this, we use the articulatory data collected using Electromagnetic Articulograph (EMA) [21]. We hypothesize that examining how neutral articulatory trajectories are transformed into those of other speaking rates could throw light in understanding the variations in speech production introduced by differences in speaking rates. We use three candidate transformations, an affine transformation with a diagonal matrix, an affine transformation with a full matrix and a nonlinear transformation modeled by a deep neural network (DNN). In order to learn these transformations, we need to first perform a time alignment between the articulatory trajectories of N and S; and N and F. For this, we use dynamic time warping (DTW) [22]. For the given time alignment via the DTW path, we find the TF. We then refine the time alignment employing the original F (or S) trajectories and the transformed N trajectories. In the next iteration, we compute the TF once again. In this iterative fashion, we optimize for the optimal TF and the optimal warping path till convergence is achieved. Our experiments reveal that while the transformation between N and F articulatory trajectories is well represented by an affine transformation with a full matrix, a nonlinear function modeled by a DNN better describes the transformation between the N and S trajectories. This could indicate that a slow rate of speech production, typically characterized by longer durations, could provide the speaker with several degrees of freedom making the articulatory movements more complex than those corresponding to neutral speaking rate.

## 2. Data Set

To perform this study, we collected articulatory movement data for 460 English sentences spoken in three different rates : neutral, fast and slow, by five subjects. The 460 phonetically balanced sentences were chosen from the MOCHA-TIMIT corpus [23]. Fig. 1 shows the placement [24] of nine articulators along the mid-sagittal plane (indicated by X and Z directions) to record articulatory movements using electromagnetic articulograph AG501 [25]. We obtain 18 articulatory trajectories corresponding to Upper Lip ($UL_x$, $UL_z$), Lower Lip ($LL_x$, $LL_z$), Right Commissure of Lip ($RC_x$, $RC_z$), Left Commissure of Lip ($LC_x$, $LC_z$), Jaw ($JAW_x$, $JAW_z$), Throat ($TH_x$, $TH_z$) (with the sensor placed near the laryngeal prominence), Tongue Tip ($TT_x$, $TT_z$), Tongue Body ($TB_x$, $TB_z$), Tongue Dorsum ($TD_x$, $TD_z$), for each utterance. Recorded at a sampling rate


10.21437/Interspeech.2018-1862

of 250 Hz, the articulatory movements known to be low pass in nature [26], are first low pass filtered with a cutoff frequency 25 Hz and then down-sampled to 100 Hz. The five subjects under study comprised two females (F1 and F2) and three males (M1, M2 and M3) of age 28, 22, 19, 22 and 24 years respectively. All subjects reported to have no speech disorders.

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. Since for utterances with a few words (typically less than five), it would be difficult to speak at twice the neutral speaking rate, a threshold factor of 1.3 (instead of 2) was chosen empirically. The average (standard deviation) speaking rates, in phones/sec, for S, N and F for each subject turned out to be, 1) F1 : $6.22(\pm 0.75)$, $11.79(\pm 1.31)$ and $16.63(\pm 1.57)$, 2) F2 : $5.60(\pm 0.84)$, $9.78(\pm 1.34)$ and $15.32(\pm 1.86)$, 3) M1 : $6.23(\pm 1.08)$, $12.53(\pm 1.41)$ and $17.12(\pm 1.79)$, 4) M2 : $4.43(\pm 0.67)$, $9.83(\pm 1.23)$ and $14.92(\pm 1.75)$ 5) M3 : $6.21(\pm 0.85)$, $10.86(\pm 1.34)$ and $15.07(\pm 1.66)$.

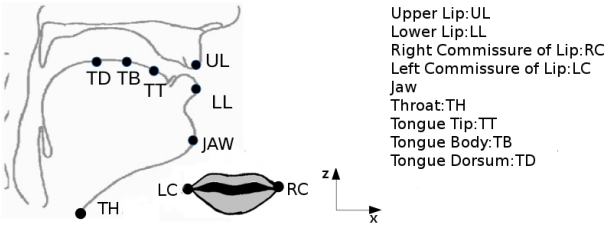


Figure 1: *Schematic diagram demonstrating the placement of the nine sensors for the study*

Figure 2 provides the trajectories of movement of sensors on the jaw and tongue in the three speaking rates for a single utterance, after mean removal. From the figure, we observe that the time duration of the utterance spoken with a slow speaking rate is much higher than that of neutral or fast rates. Correspondingly, the duration of the recordings for the subjects F1, F2, M1, M2 and M3 are 20.23, 24.54, 18.90, 24.38 and 21.83 minutes for neutral; 38.16, 42.71, 37.4, 54.01 and 38.15 minutes for slow and 14.27, 15.55, 13.8, 15.94 and 15.7 minutes for fast speaking rates, respectively. We also observe from the figure that the articulatory trajectories of S show more prominent peaks and valleys, implying hyper articulation compared to the articulatory trajectories of F [27, 28]. We find that articulatory trajectories corresponding to fast speaking rate are characterized by gross movements with less variations compared to those of N and S counterparts.

## 3. Learning the transformation functions

Motivated by the methodology proposed in [29], we consider a set of $M$ training utterances for N, F and S speaking rates, each comprising articulatory movement data from $K$ articulators. Let R stand for either F or S rate movements. Consider the N and R articulatory trajectory corresponding to $i^{th}$ training utterance to be represented by $\mathcal{N}_i \in \mathbb{R}^{K \times L_{\mathcal{N}_i}}$ and $\mathcal{R}_i \in \mathbb{R}^{K \times L_{\mathcal{R}_i}}$ , respectively. Let $\mathcal{N}_i = \left[\mathbf{n}_{i,1}, \mathbf{n}_{i,2}, \ldots, \mathbf{n}_{i,L_{\mathcal{N}_i}}\right]$ and $\mathcal{R}_i = \left[\mathbf{r}_{i,1}, \mathbf{r}_{i,2}, \ldots, \mathbf{r}_{i,L_{\mathcal{R}_i}}\right]$, such that $\mathbf{n}_{i,j}$ and $\mathbf{r}_{i,j}$ represent the $K$ dimensional articulatory movement vectors at the $j^{th}$ time instant of the $i^{th}$ training utterance, corresponding to N and R speaking rates. Since the lengths $L_{\mathcal{N}_i}$ and $L_{\mathcal{R}_i}$ need not be equal, we time align the sequences using DTW for $i = 1, \ldots, M$, in order to construct parallel training data to learn the transformation of N to R. Therefore, the TF represented by $\mathcal{G}(\cdot)$ depends on the DTW paths, $\{\boldsymbol{w}_i, i = 1, 2, \ldots, M\}$, used for the time alignment. Hence, there is a need to optimize both the warping paths and the TF by minimizing the total cost $\mathcal{D}$ as follows,

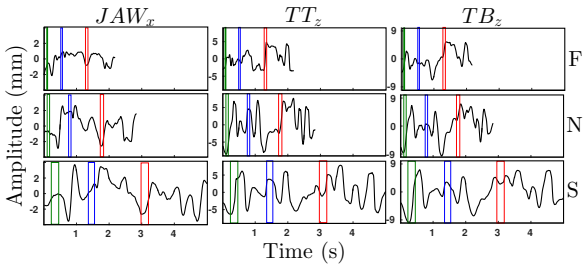


Figure 2: *Trajectories of $JAW_x$, $TT_z$ and $TB_z$ for the utterance 'Her classical repertoire gained critical acclaim', by subject F1, from speech corresponding to fast (row 1), neutral (row 2) and slow (row 3) speaking rates. The highlighted boxes in the plots indicate three exemplar phonetic units (/ɚ/, /l/, /eɪ/).*

$$[\mathcal{G}^*, \{\boldsymbol{w}_i^*\}] = \arg\min_{g, \{\boldsymbol{w}_i\}} \mathcal{D}\big(g, \{\boldsymbol{w}_i\}\big), \tag{1}$$

where, $\mathcal{G}^*(\cdot)$ and $\{\boldsymbol{w}_i^*, i = 1, 2, \ldots, M\}$ indicate the optimal TF and the optimal set of warping paths. We consider the $\mathcal{D}$, as the sum of DTW distances (using the set of warping paths $\{\boldsymbol{w}_i\}$) between the transformed N (using the TF $g(\cdot)$) and original R articulatory trajectories, across all $M$ training utterances as follows,

$$\mathcal{D}(g, \{\boldsymbol{w}_i\}) = \sum_{i=1}^{M} \mathcal{D}_{w_i}\big(g(\mathcal{N}_i), \mathcal{R}_i\big). \tag{2}$$

$\mathcal{D}_{w_i}$ represents the squared Euclidean distance between the DTW mapped trajectories along the warping path $\boldsymbol{w}_i$ of length $L_{w_i}$, between $\mathcal{R}_i$ and transformed $\mathcal{N}_i$. $\boldsymbol{w}_i$ consists of paired indices $\{w_{n,i}(p), w_{r,i}(p)\}, p = 1, \ldots, L_{w_i}$, such that $w_{n,i}(p) \in [1, L_{\mathcal{N}_i}]$ and $w_{r,i}(p) \in [1, L_{\mathcal{R}_i}]$. Along the warping path $\boldsymbol{w}_i$, we compute $\mathcal{D}_{w_i}$ as follows,

$$\mathcal{D}_{w_i}(g(\mathcal{N}_i), \mathcal{R}_i) = \frac{1}{L_{w_i}} \sum_{p=1}^{L_{w_i}} \|g(\mathcal{N}_{w,i}(p)) - \mathcal{R}_{w,i}(p)\|_2^2 \tag{3}$$

where, $||\cdot||_2$ denotes the L2 norm, $\mathcal{N}_{w,i}(p) = \mathbf{n}_{i,w_{n,i}(p)}$ and $\mathcal{R}_{w,i}(p) = \mathbf{r}_{i,w_{r,i}(p)}$, $p = 1, \ldots, L_{w_i}$ represent the time aligned N and R trajectories via the DTW warping path $\boldsymbol{w}_i$.

From Eq. (1), (2) and (3) we find that in order to optimize for the TF, we require the optimal warping paths and vice-versa. Therefore, we perform an alternate minimization of the objective function (Eq. (1)) by alternately optimizing the TF and the warping paths, in an iterative fashion. Given the TF $g^{(l)}(\cdot)$ at the $l^{th}$ iteration, we optimize for the warping paths as follows,

$$\{\boldsymbol{w}_i^{(l)}\} = \arg\min_{w_i} \mathcal{D}_{w_i}(g^{(l)}(\mathcal{N}_i), \mathcal{R}_i) \quad i = 1, 2, .., M \tag{4}$$

Given the set of warping paths $\{\boldsymbol{w}_i^{(l)}\}$, we construct $\mathbf{N}_{\{\boldsymbol{w}_i^{(l)}\}} = [\mathcal{N}_{w^{(l)},1}, \ldots, \mathcal{N}_{w^{(l)},M}] \in \mathbb{R}^{K \times \mathcal{L}^{(l)}}$ and $\mathbf{R}_{\{\boldsymbol{w}_i^{(l)}\}} = [\mathcal{R}_{w^{(l)},1}, \ldots, \mathcal{R}_{w^{(l)},M}] \in \mathbb{R}^{K \times \mathcal{L}^{(l)}}$ with $\mathcal{L}^{(l)} = \sum_{i=1}^{M} L_{w_i^{(l)}}$. Then the optimal TF for the $(l+1)^{th}$ iteration can be computed as,

$$g^{(l+1)}(\cdot) = \arg\min_{g} \left\| \mathbf{N}_{\{\boldsymbol{w}_i^{(l)}\}} - \mathbf{R}_{\{\boldsymbol{w}_i^{(l)}\}} \right\|_2^2 \quad (5)$$

At each iteration $l$, we calculate the total cost $\mathcal{D}^{(l)}$ using $g^{(l)}(\cdot)$ and $\{\boldsymbol{w}_i^{(l)}\}, i = 1, \ldots, M$ in Eq. (2).

In the first iteration, we consider the TF $g^{(0)}(\cdot)$ to be an identity transformation and initialize the total cost $\mathcal{D}^{(0)} = \infty$. Using $g^{(0)}(\cdot)$ we compute the optimal warping paths $\{\boldsymbol{w}_i^{(0)}\}, i = 1, \ldots, M$ using Eq. (3) and Eq. (4). Using Eq. (5) and $\{\boldsymbol{w}_i^{(0)}\}, i = 1, \ldots, M$, we compute $g^{(1)}(\cdot)$. In this manner, we optimize for the TF and the warping paths till the condition $\mathcal{D}^{(l)} < \mathcal{D}^{(l-1)}$ is satisfied. Upon convergence, we obtain the optimal TF, $\mathcal{G}^*(\cdot)$ and the optimal warping paths $\{\boldsymbol{w}_i^*\}, i = 1, \ldots, M$ corresponding to the least $\mathcal{D}$. In this work, we consider three candidate TFs [1].

***(1) Full Affine Transformation Matrix, $\mathcal{G}_{\mathcal{F}}$ scheme:***

We hypothesize that several articulators could be involved in transforming one N articulatory trajectory into its R counterpart. Therefore we learn an affine mapping between $\mathcal{N}_i$ and $\mathcal{R}_i$. Consider $\mathbf{A} \in \mathbb{R}^{K \times K}$, $\mathbf{b} \in \mathbb{R}^{1 \times K}$ and $\mathbf{N}_{\{\boldsymbol{w}_i\}}^T \in \mathbb{R}^{\mathcal{L} \times K}$, then the optimal affine transformation can be computed as,

$$\mathcal{G}_{\mathcal{F}}(\mathbf{N}_{\{\boldsymbol{w}_i\}}^T) = \underbrace{\begin{bmatrix} \mathbf{N}_{\{\boldsymbol{w}_i\}}^T & \mathbf{1}^{\mathcal{L}X1} \end{bmatrix}}_{\mathbf{N}_{aug}} \begin{bmatrix} \mathbf{A}^{K \times K} \\ \mathbf{b}^{1 \times K} \end{bmatrix}. \quad (6)$$

Let $\mathbf{R}_{aug}$ as $\begin{bmatrix} \mathbf{R}_{\{\boldsymbol{w}_i\}}^T & \mathbf{1}^{\mathcal{L}X1} \end{bmatrix}$, such that $\mathbf{N}_{aug}, \mathbf{R}_{aug} \in \mathbb{R}^{\mathcal{L} \times (K+1)}$. Then Eq. (6) can be rewritten as, $\mathcal{G}_{\mathcal{F}} = \left[(\mathbf{N}_{aug})^T \mathbf{N}_{aug}\right]^{-1} (\mathbf{N}_{aug})^T \mathbf{R}_{aug}$. This is similar to the approach proposed in [30].

***(2) Diagonal Affine Transformation Matrix, $\mathcal{G}_{\mathcal{D}}$ scheme:***

In order to examine the effects of the transformation if we consider only the $k^{th}$ N articulatory trajectory to obtain the $k^{th}$ R articulatory trajectory, we modify the TF described in Section 3(1) such that $\mathbf{A}$ is a diagonal matrix.

***(3) Non-linear transformation function, $\mathcal{DNN}$ scheme:***

In order to consider a more complex nonlinear TF, we use a DNN which considers $\mathbf{N}_{\{\boldsymbol{w}_i\}}^T$ and $\mathbf{R}_{\{\boldsymbol{w}_i\}}^T$ as the input and output, respectively. While the DNN at the first iteration is initialized with random weights, those corresponding to the subsequent iterations are initialized with the weights from that of the previous iteration.

# 4. Experiments

Since the articulation strategy is subject dependent [16] we learn the optimal TFs in a subject specific manner, using a four fold setup. We divide the data into training and test set in the ratio $4:1$ such that $M = 345$. Articulatory trajectories are first made zero-mean. The optimal TFs are learnt separately for N2S and N2F transformations in each of the four folds from the five subjects. Specifically we denote $\mathcal{G}_{\mathcal{F}}^{(N2F)}$ and $\mathcal{G}_{\mathcal{F}}^{(N2S)}$ for the TF described in Section 3(1) for N2F and N2S, respectively. Similarly, we have $\mathcal{G}_{\mathcal{D}}^{(N2F)}$ and $\mathcal{G}_{\mathcal{D}}^{(N2S)}$ for the TF described in Section 3(2) and notations $\mathcal{DNN}^{(N2F)}$ and $\mathcal{DNN}^{(N2S)}$ to describe the TF described in Section 3(3). We also consider the case when $g^{(0)} = \mathcal{G}^*(\cdot)$ is the identity matrix denoted by $\mathcal{G}_{\mathcal{I}}^{(N2F)}$, for N2F and $\mathcal{G}_{\mathcal{I}}^{(N2F)}$ for N2S. The input and output feature dimensions for all the schemes are $K = 18$, corresponding to the 18 articulatory trajectories described in Section 2.

For the DNN, we use 15% of the training set as the validation set and optimize for different parameters for each subject in every fold. Specifically, we optimize the number of hidden layers (candidates: 1, 2 and 3), number of neurons in each hidden layer (candidates: 32, 64, 128, 256 and 512), activation functions corresponding to each hidden layer (candidates: 'tanh' and 'relu') and different batch size (16, 32 and 64). The 'linear' activation is used in the output layer. We also use batch normalization followed by a dropout of 0.1 between the hidden layers and between the final hidden layer and the output layer. The implementation of the DNN is done using Keras [31] and Theano [32] libraries.

For a given scheme, we indicate $\mathbf{d}_T = [\mathbf{d}_1, \ldots, \mathbf{d}_4]$, where $\mathbf{d}_i \in \mathbb{R}^{1 \times 115}$ consists of the DTW distances between the transformed N and original R trajectories for the 115 test utterances in the $i^{th}$ fold. In order to evaluate the performances of different schemes, we report the average and standard deviation of $\mathbf{d}_T$ each subject. Lower the average $\mathbf{d}_T$ for a given scheme, better is the TF. Therefore, the best scheme is the one which results in the least average $\mathbf{d}_T$.

# 5. Results

Across all subjects and all folds, the average number of iterations for estimating $\mathcal{G}_{\mathcal{D}}^{(N2F)}$, $\mathcal{G}_{\mathcal{F}}^{(N2F)}$ and $\mathcal{DNN}^{(N2F)}$ is found to be $6.05(\pm 1.32)$, $6.65(\pm 1.35)$ and $3.6(\pm 1.27)$, respectively. Similarly the average number of iterations for estimating $\mathcal{G}_{\mathcal{D}}^{(N2S)}$, $\mathcal{G}_{\mathcal{F}}^{(N2S)}$ and $\mathcal{DNN}^{(N2S)}$ is found to be $6.75(\pm 1.83)$, $7.90(\pm 1.88)$ and $2.25(\pm 0.55)$, respectively. For the DNN based TF, the optimal DNN architecture was found to have one hidden layer with 'relu' activation function using a batch size of 64, for all subjects and folds. Out of the 20 folds (5 subjects $\times 4$ folds) the optimal number of neurons turned out to be 32 for 4 folds, 64 for 10 folds, 128 for 5 folds and 256 for 1 fold for N2F conversion and 64 for 12 folds, 128 for 7 folds and 512 for 1 fold for N2S conversion.

Table 1: *Average (standard deviation) of $\mathbf{d}_T$ (in mm) using different TFs for N2F transformation*

| Function | F1 | F2 | M1 | M2 | M3 |
|---|---|---|---|---|---|
| $\mathcal{G}_{\mathcal{I}}^{(N2F)}$ | 6.51 (0.85) | 6.60 (1.09) | 6.71 (1.18) | 6.00 (0.93) | 5.64 (1.04) |
| $\mathcal{G}_{\mathcal{D}}^{(N2F)}$ | 5.32 (0.66) | 5.21 (0.86) | 5.38 (0.93) | 5.24 (0.79) | 4.90 (0.94) |
| $\mathcal{G}_{\mathcal{F}}^{(N2F)}$ | 4.82 (0.62) | **5.04** (0.90) | 4.74 (0.93) | 4.88 (0.78) | 4.59 (0.92) |
| $\mathcal{DNN}^{(N2F)}$ | **4.79** (0.63) | 5.09 (0.93) | **4.72** (0.90) | **4.86** (0.76) | **4.58** (0.90) |

### 5.1. N2F transformation

Table 1 provides the average $\mathbf{d}_T$ obtained using the different TFs for N2F transformation. The scheme for which average $\mathbf{d}_T$ is the least is indicated in bold for each subject. From the table we observe that the relative decrease in average $\mathbf{d}_T$ of the

[1] Codes for $\mathcal{G}_{\mathcal{F}}$ and $\mathcal{G}_{\mathcal{D}}$ schemes are available at `https://spire.ee.iisc.ac.in/spire/software.php`

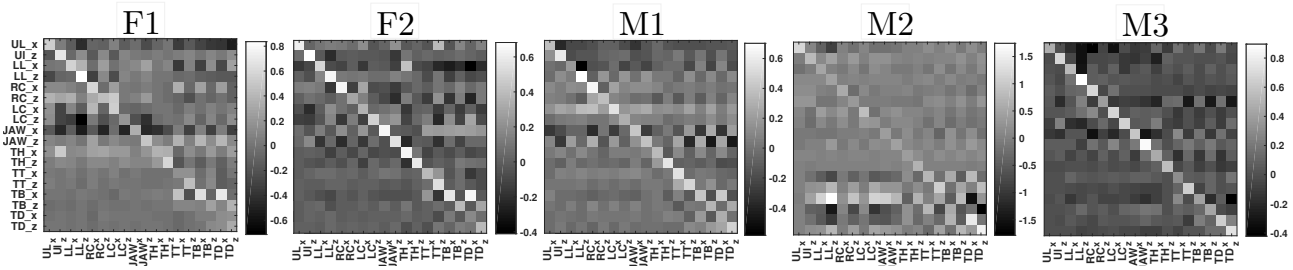


Figure 3: $\mathcal{G}_{\mathcal{F}}^{(N2F)}$ *matrices of one fold for the five subjects.*

best scheme compared to the identity TF, $\mathcal{G}_{\mathcal{I}}^{(N2F)}$, is $26.42\%$, $23.63\%$, $29.65\%$, $19\%$ and $18.79\%$ for the subjects F1, F2, M1, M2 and M3, respectively. This indicates that there indeed exist differences in the articulation strategies between the speech production during neutral and fast speaking rates. From Table 1 we find that the relative decrease in the average $\mathbf{d}_T$ using the $\mathcal{G}_{\mathcal{F}}^{(N2F)}$ scheme compared to that using $\mathcal{G}_{\mathcal{D}}^{(N2F)}$ scheme is $9.39\%$, $3.26\%$, $11.89\%$, $6.87\%$ and $6.32\%$ for the five subjects. This implies that there is a significant contribution from several N articulatory trajectories to transform one N trajectory into its F counterpart.

Although for most subjects the $\mathcal{DNN}^{(N2F)}$ scheme secures the least average $\mathbf{d}_T$, an analysis based on pairwise *t-test* reveals that in 12 out of 20 folds the performance of $\mathcal{DNN}^{(N2F)}$ scheme is *not* statistically significantly better than that of $\mathcal{G}_{\mathcal{F}}^{(N2F)}$ scheme ($0.06 \leq p$-val$\leq 0.90$). In 9 out of 20 folds, the DNN scheme outperforms the $\mathcal{G}_{\mathcal{F}}^{(N2F)}$ scheme ($p$-val $\leq 0.04$). For subject F2, the $\mathcal{G}_{\mathcal{F}}^{(N2F)}$ scheme performs better than the DNN in three out of four folds (p-val $\leq 7.18e-05$). The comparable performance of the $\mathcal{DNN}^{(N2F)}$ and $\mathcal{G}_{\mathcal{F}}^{(N2F)}$ in most folds reveals that the transformation of articulatory movements from N to F can be represented by an affine transformation with a full matrix. We hypothesize that the transformation from N to F is not a very complex nonlinear function since speech production in fast speaking rates results in gross articulatory movements (Fig. 2) compared to that in neutral speaking rate.

Fig. 3 shows the $\mathcal{G}_{\mathcal{F}}^{(N2F)}$ matrices from one fold of each of the five subjects. We observe that the matrices are not diagonal in nature. This indicates that several articulators contribute to transforming one N articulatory trajectory into its F counterpart. This is supported by the observation that the average $\mathbf{d}_T$ using $\mathcal{G}_{\mathcal{D}}^{(N2F)}$ scheme is higher than that using $\mathcal{G}_{\mathcal{F}}^{(N2F)}$ scheme. From the figure, it is also evident that the optimal TFs are subject dependent. It could be due to subject specific motor control plans to produce speech with different speaking rates.

### 5.2. N2S transformation

Table 2 provides the average $\mathbf{d}_T$ obtained using the different schemes for N2S transformation. Similar to the observations made in Section 5.1, we find that $\mathcal{G}_{\mathcal{I}}^{(N2S)}$ results in the highest average $\mathbf{d}_T$ for all subjects. The relative decrease in average $\mathbf{d}_T$ using $\mathcal{G}_{\mathcal{F}}^{(N2S)}$ compared to that using $\mathcal{G}_{\mathcal{D}}^{(N2S)}$ is $12.98\%$, $8.30\%$, $7.90\%$, $7.10\%$ and $7.67\%$ for the five subjects. This implies that for N2S transformation also, several articulators contribute in transforming one N articulator trajectory into its S counterpart. From the table we see that the $\mathcal{DNN}^{(N2S)}$ scheme outperforms all other schemes (bold entries). Analysis for significance test using pairwise *t-test* reveals that in most folds (15 out of 20) the $\mathcal{DNN}^{(N2S)}$ exhibits statistically significantly lower ($p$-val $\leq 0.0095$) DTW distances between the transformed N and original S articulatory trajectories compared to the $\mathcal{G}_{\mathcal{F}}^{(N2S)}$ scheme. It could be that an affine function with a full matrix cannot capture the complex transformation between N and S better than a highly nonlinear TF modeled by a DNN. Interestingly, we find that the average $\mathbf{d}_T$ values are higher for N2S transformation (Table 2) than that for N2F transformation (Table 1). This could be due to 1) the increase in DTW distances with increase in the duration of the articulatory movements in slow speech compared to that in fast speech and 2) a more complex transformation involved in N2S than in N2F, which in turn, could be due to the fact that subjects to tend to hyper-articulate while speaking slowly [17].

Table 2: *Average (standard deviation) of* $\mathbf{d}_T$ *(in mm) using different TFs for N2S transformation*

| **Function** | **F**1 | **F**2 | **M**1 | **M**2 | **M**3 |
|---|---|---|---|---|---|
| $\mathcal{G}_{\mathcal{I}}^{(N2S)}$ | 6.46 (0.90) | 8.27 (1.27) | 7.32 (1.07) | 6.03 (0.77) | 6.30 (0.89) |
| $\mathcal{G}_{\mathcal{D}}^{(N2S)}$ | 5.93 (0.91) | 8.14 (1.24) | 6.90 (1.02) | 5.76 (0.75) | 6.12 (0.87) |
| $\mathcal{G}_{\mathcal{F}}^{(N2S)}$ | 5.16 (0.87) | 7.46 (1.13) | 6.35 (1.02) | **5.35** (0.71) | 5.65 (0.81) |
| $\mathcal{DNN}^{(N2S)}$ | **5.05** (0.89) | **7.37** (1.11) | **6.30** (1.03) | **5.35** (0.75) | **5.52** (0.80) |

## 6. Conclusion

In this work, we examine the function that transforms articulatory trajectories of neutral speech into those of fast and slow speech. We find that the transformation of N to F articulatory movements can be well represented by an affine function with a full matrix, indicating that several articulators contribute in transforming one N articulatory movement into its F counterpart. For the N2S transformation, we observe that a DNN provides the best TF. This could be due to the hyper-articulation exhibited in slow speech. Our future work includes 1) performing a suitable time scaling of the transformed neutral articulatory trajectories to match the duration of those from different speaking rates in a subject independent manner, and 2) analyzing the effects of sentence stress on the articulation in different rates.

## 7. Acknowledgements

We thank the five subjects and Mr. Aravind Illa for the data collection. We also thank the Pratiksha Trust for their support.

# 8. References


[1] R. H. Stetson, "Motor phonetics: A study of speech movements in action," vol. 2nd edition, 1951.

[2] E. Jacewicza and R. A. Fox, "Between-speaker and within-speaker variation in speech tempo of american english," *Journal of Acoustical Society of America*, vol. 128(2), p. 839850, 2010.

[3] H. Quene, "Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo," *Journal of Acoustical Society of America*, vol. 123, pp. 1104–1113, 2008.

[4] E. Jacewicza, R. A. Fox, C. ONeill, and J. Salmons, "Articulation rate across dialect, age, and gender," *Language Variation and Change*, vol. 21(2), pp. 233–256, 2009.

[5] H. Quene, "Modeling of between-speaker and within-speaker variation in spontaneous speech tempo," *Interspeech*, 2005.

[6] H. Kuwabara, "Acoustic and perceptual properties of phonemes in continuous speech as a function of speaking rate," *Eurospeech*, 1997.

[7] B. Lindblom, "Spectrographic study of vowel reduction," *Journal of Acoustical Society of America*, vol. 90, 1963.

[8] T. Gay, "Effect of speaking rate on vowel formant movements," *Journal of Acoustical Society of America*, vol. 63(1), pp. 223–230, 1978.

[9] S.-J. Moon and B. Lindblom, "Interaction between duration, context, and speaking style in english stressed vowels," *Journal of Acoustical Society of America*, vol. 40, 1994.

[10] A. Agwuele, H. Sussman, and B. Lindblom., "The effect of speaking rate on consonant vowel coarticulation," *Phonetica*, vol. 65(4), pp. 194–209, 2008.

[11] J. L. Miller, F. Grosjean, and C. Lomanto, "Articulation rate and its variability in spontaneous speech: A reanalysis and some implications," *Phonetica*, 1984.

[12] E. Fosler-Lussier and N. Morgan, "Effects of speaking rate and word frequency on pronunciations in conversational speech," *Speech Communication*, vol. 29, pp. 137–158, 1999.

[13] M. Benzeghiba, O. D. Renato De Mori, S. Dupont, T. Erbes, and et al, "Automatic speech recognition and speech variability: a review," *Speech Communication*, vol. 49 (10-11), p. 763, 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, 2011.

[15] R. Stern, A. Acero, F. Liu, and Y. Ohshima, "Signal processing for robust speech recognition," *Automatic Speech and Speaker Recognition, edited by C.-H. Lee, F. K. Soong, and K. K. Paliwal (Springer, Berlin)*, pp. 357 – 384, 1996.

[16] J. Berry, "Speaking rate effects on normal aspects of articulation: Outcomes and issues," *American Speech-Language-Hearing Association - Perspectives on Speech Science and Orofacial Disorders*, vol. 21(1), pp. 15–26, 2011.

[17] R. Smiljani and A. Bradlow, "Temporal organization of english clear and conversational speech," *Journal of Acoustical Society of America*, vol. 124(5), pp. 3171–3182, 2008.

[18] J. Gooze, L. Lapointe, and B. Murdoch, "Effects of speaking rate on ema-derived lingual kinematics: a preliminary investigation," *Clinical Linguistics and Phonetics*, vol. 17(4-5), pp. 375–381, 2003.

[19] Y. Payan and P. Perrier, "Synthesis of v-v sequences with a 2d biomechanical tongue model controlled by the equilibrium point hypothesis," *Speech Communication*, vol. 22(1), 1997.

[20] M. D. Mcclean, "Patterns of orofacial movement velocity across variations in speech rate," *Journal of speech, language, and hearing research : JSLHR*, vol. 43 1, pp. 205–216, 2000.

[21] P. Schnle, K. Grbe, P. Wenig, J. Hhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic felds for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, vol. 31(1), pp. 26–35, 1987.

[22] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

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

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

[25] "3d electromagnetic articulograph," *http://www.articulograph.de*.

[26] 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 4, pp. 205–16, 2010.

[27] 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.

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

[29] G. N. Meenakshi and P. K. Ghosh, "Reconstruction of articulatory movements during neutral speech from those during whispered speech," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3352–3364, 2018. [Online]. Available: https://doi.org/10.1121/1.5039750

[30] Y. Qiao and M. Yasuhara, "Affine invariant dynamic time warping and its application to online rotated handwriting recognition." *Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh.*, 2006.

[31] F. Chollet, "keras," *https://github.com/fchollet/keras*, 2015.

[32] T. D. Team, "Theano: A python framework for fast computation of mathematical expressions," *arXiv e-prints, http://arxiv.org/abs/1605.02688*, 2016.