The impact of speaking rate on acoustic-to-articulatory inversion

Aravind Illa*, Prasanta Kumar Ghosh

Department of Electrical Engineering, Indian Institute of Science, Bangalore, Karnataka 560012, India

Received 16 February 2019; received in revised form 25 April 2019; accepted 25 May 2019
Available online 15 June 2019

Abstract

Acoustic characteristics and articulatory movements are known to vary with speaking rates. This study investigates the role of speaking rate on acoustic-to-articulatory inversion (AAI) performance using deep neural networks (DNNs). Since fast speaking rate causes fast articulatory motion as well as changes in spectro-temporal characteristics of the speech signal, the articulatory-acoustic map in a fast speaking rate could be different from that in a slow speaking rate. We examine how these differences alter the accuracy with which different articulatory positions could be recovered from the acoustics. AAI experiments are performed in both matched and mismatched train-test conditions using data of five subjects, in three different rates — normal, fast and slow (fast and slow rates are at least 1.3 times faster and slower than the normal rate). Experiments in matched cases reveal that, the errors in estimating vertical motion of sensors on the tongue articulators from acoustics with fast speaking rate, is significantly higher than those with slow speaking rate. Experiments in mis-matched conditions reveal that there is consistent drop in AAI performance compared to the matched condition. Further experiments performed by training AAI with acoustic-articulatory data pooled from different speaking rates reveal that a single DNN based AAI model is capable of learning multiple rate-specific mapping.

© 2019 Published by Elsevier Ltd.

Keywords: Acoustic-to-articulatory inversion; Speaking rate; Electromagnetic articulograph

1. Introduction

Estimating articulatory movement from acoustic representation is known as acoustic-to-articulatory inversion (AAI). Several techniques have been proposed in literature for AAI including codebook based procedures (Ouni and Laprie, 2005), statistical modeling of the acoustic-articulatory map such as Gaussian mixture model (GMM) (Toda et al., 2008), a trajectory hidden-Markov model (HMM) (Zhang and Renals, 2008), and neural network (Kirchhoff, 1999) based such as, mixture density network (MDN) (Richmond, 2006), and Deep Neural Networks (DNN) (Wu et al., 2015). Among all these methods, DNNs have been shown to perform the best in learning the non-linear acoustics to articulatory mapping (Uria et al., 2012; Wu et al., 2015). Articulatory information and AAI have been shown to be useful for a number of applications including human computer interaction (Ding et al., 2015; Jia et al., 2014; 2011), automatic speech recognition (ASR) (Sun and Deng, 2002; Zlokarnik, 1995; Wrench and Richmond, 2000; Frankel et al., 2000; Kirchhoff, 1999) and pathological conditions (Turner and Weismer, 1993; Turner et al., 1995;
and speech synthesis (Ling et al., 2009). The AAI performance depends on variabilities in both acoustics and articulation. Mismatch in acoustic and articulatory space degrades the performance of AAI. Major factors could be due to inter-speaker variability that leads to the mismatch of acoustic and articulatory characteristics between train and test subjects (Afshan and Ghosh, 2015; Ji et al., 2016; Sivaranman et al., 2016). On the other hand, the mismatch could occur even within the subject as well, for example, acoustic-to-articulatory map significantly changes between neutral and whispered speech (Ilia et al., 2017). One more factor within speaker variability that has not been investigated in the context of AAI is the speaking rate, which varies within the subject due to various factors like emotional state (Vroomen et al., 1993), lexical stress, etc. It is known that variation of speaking rate results in significant degradation of performance of ASR (Martinez et al., 1997; Siegler and Stern, 1995) when there is a mismatch between the train and test speaking rates. So, it is important to study how the speaking rate impacts the AAI performance in matched and mismatched train-test conditions.

Variations in speaking rate change articulatory movements as well as several acoustic characteristics (Berry, 2011). Several studies have been conducted to reveal the effect of speaking rate on acoustics and articulatory movements. In articulatory space, speaking rate affects the velocity profiles of lips (Shaiman et al., 1997), jaw (Shaiman, 2002) and tongue (Flege, 1988). It also affects the coordination among these articulators (De Nil and Abbs, 1991; Hertrich and Ackermann, 2000; Adams et al., 1993; Gay, 1981) and their gestural overlap (Hardcastle, 1985; Munhall and Löfqvist, 1988; Saltzman and Munhall, 1989). This is commonly observed in spontaneous speech (Miller et al., 1984) where substantial variation in articulation rate has been reported. Influences of speaking rate on velum movement patterns were also investigated (Krakow, 1993; Kent et al., 1974). On the other hand, in the acoustic space, variations in speaking rate result in changes in several acoustic characteristics, including consonant-vowel co-articulation and duration (Agwuele et al., 2009), spectral characteristics, formants of vowels (Gay, 1978), F2 trajectories (Weismer and Berry, 2003) and voice-onset time (Kessinger and Blumstein, 1998). Vowel space reduction (in F1-F2 plane) from normal rate to the fast rate was studied by Agwuele et al. (2009), and Fourakis (1991), Max and Caruso (1997) examined the variability of syllable and phrase-level relative timing, and showed that the relative duration of temporal intervals at the syllable-level and/or phrase-level does not remain invariant across slow, normal, and fast speech rates. Also, perceptual and acoustic properties of individual phonemes (Kuwabara, 1998; 1996) in continuous speech at different speaking rates for Japanese were investigated by Kuwabara. In fact, it has been reported that the change in speaking rate results in independent “signature” in production (Tuller et al., 1982). For example, in fast speaking rate, gestural overlap in articulation could result in segmental deletion (Davidson, 2006). Such changes in acoustic and articulatory characteristics due to change in speaking rate could alter acoustic-to-articulatory relations (Mefferd and Green, 2010). Due to the effect of speaking rate, it is unclear how and to what degree these changes in acoustics, articulation and their relation could affect the accuracy of AAI algorithms.

In this work, we carried out an experimental study to systematically investigate how the AAI performance changes in different speaking rates. For this purpose, by placing sensors of electromagnetic articulograph (EMA) on different articulators, we have recorded speech and articulation data from five subjects speaking the same set of sentences in normal ($N_r$), followed by slow ($S_r$), and fast ($F_r$) speaking rates (satisfying $S_r \leq \frac{1}{3} N_r$) & ($F_r \geq 1.3N_r$). The complete details of the dataset used for this study is described in Section 2. The analysis on articulatory and acoustic data obtained at three different speaking rates and the details of observations are described in Section 3. Before performing the AAI experiments, we also present an information theoretic analysis in Section 4 to quantify the variation in statistical dependencies between acoustic and articulatory features due to changes in speaking rate. It turns out that as the speaking rate increases, the conditional dependency of articulatory features given the acoustic features decreases. This suggests that the AAI performance might decrease as the rate increases. To validate the hypothesis, we carry out AAI experiments using the DNN. In Section 5.1 we describe about DNN for AAI. In Section 6, experimental results are discussed with different train-test conditions: (1) Matched - train an AAI model specific to a speaking rate and test with the same rate for every subject, (2) Mismatched - train an AAI model with a speaking rate and test with different rates, (3) Generic AAI - here an AAI model is trained on all three speaking rates together. Conclusions are drawn in Section 7.

2. Data collection and preprocessing

To study the effect of speaking rate in AAI, we collected simultaneous speech and articulation data at slow, normal and fast speaking rates. 460 phonetically balanced English sentences from the MOCHA-TIMIT corpus (Wrench,
1999) were chosen as the stimuli for data collection. For this study, we collected data from five subjects — three males (M1, M2, M3) and two females (F1, F2) of age 19, 22, 24 and 28, 22 years respectively. All subjects reported to have no speech disorders in the past. As recommended by the institute ethics committee, all the subjects signed a consent form prior to the data collection. To avoid any pronunciation error during recording, all subjects were familiarized with the entire set of 460 sentences. For each sentence, we recorded simultaneous audio and articulatory movement data. To identify errors due to word deletions and insertions we carefully listened to each utterance. In case of any mistakes, the subject was asked to repeat the sentence.

For this study, EMA AG501 AG5 was used to record the articulatory movement data. A t.bone EM9600 shotgun, unidirectional electret condenser microphone EM9 was placed near the subject to record the audio data synchronously with the articulatory data. The articulatory movement was collected using EMA at a sampling rate of 250 Hz and with simultaneous audio data at 48 kHz. We used six sensors placed on different articulators, namely Upper Lip (UL), Lower Lip (LL), Jaw, Tongue Tip (TT), Tongue Body (TB), and Tongue Dorsum (TD). For head movement correction, two additional sensors were placed behind the two ears (Kroos, 2009). A schematic diagram of the sensor placement is shown in Fig. 1. Each of these eight sensors captures the movements of articulators in 3D space. In this study we consider the movements only in the midsagittal plane (Mermelstein, 1973), indicated by X and Y directions, in Fig. 1, resulting in twelve articulatory features denoted by \( UL_x, UL_y, LL_x, LL_y, Jaw_x, Jaw_y, TT_x, TT_y, TB_x, TB_y, TD_x, TD_y \). Before placing the sensors, the subject was made to seat comfortably in the EMA recording setup. The sensors coated with “plasty-late” latex material were glued using “Epiglu” on all the articulators Glu. The subjects were given sufficient time to get used to speaking naturally with the sensors attached to different articulators.

During recording, subjects were given breaks whenever they felt tired of speaking, and recording was resumed only when the subject felt comfortable to continue. During recording, the sentences were projected onto a screen placed in front of the subject who was allowed to navigate through the sentences by himself/herself and each utterance was listened to identify word & phone insertion & deletion. The recordings of three speaking rates of all 460 sentences were done in three different sessions. In the first session, the subject was asked to speak each sentence at his/her normal speaking rate. We observed no pause within an utterance. Hence, after the first session, the normal speaking (phone) rate for the subject was computed for every sentence by removing the silence before and after each utterance. In the second session, the subject was asked to speak at a rate twice as fast as his/her normal speaking rate for every sentence. If the subject’s speaking rate was found to be less than \( r \) times the normal speaking rate, the subject was asked to repeat the sentence. Similarly, in the third session, the subject was asked to speak at half of the normal speaking rate and to repeat if the subject failed to lower the speaking rate by at least a factor \( r \) (phones/s). For monitoring the speaking rate during recording, we developed a GUI to mark the beginning and end of each utterance, using which, the speaking rate was computed and compared with the corresponding normal speaking rate. We observed that, for sentences with 4 – 5 words, it became difficult for subjects to speak at twice their normal speaking rate. Hence, we empirically chose a factor of \( r = 1.3 \) (phones/s), a necessary criterion to be satisfied by all subjects. Since the recordings of the three speaking rates were done in three sessions, proper care was taken to keep the sensors at same locations in all the sessions. The sensors were placed on the articulators such that the inter-sensor distance closely match with the recommendation provided in Pattem et al. (2018).

The recorded audio data were first down-sampled to 16 kHz from 48 kHz. Then the silence segments before and after every sentence were removed manually in both audio and articulatory movement data. After removing silence, the duration of recordings for M1, M2, M3, F1 and F2 was found to be 18.90, 24.38, 21.83, 20.23 and 24.54 min for
normal; 37.4, 54.01, 38.15, 38.16 and 42.71 min for slow and 13.8, 15.94, 15.7, 14.27 and 15.55 min for fast speaking rates, respectively. The average speaking rates of all subjects are shown in the Table 1 in terms of phones/s and words per minute (WPM), the values in brackets indicate corresponding standard deviation (SD).

From the audio data, as acoustic features we computed Mel Frequency Cepstral Coefficient (MFCC), which was shown to be optimal for AAI using maximal mutual information criterion (Ghosh and Narayanan, 2010). We extracted 39-dim MFCC (Young and Young, 1993) feature vector with frame length and shift being 20 ms and 10 ms respectively. The EMA data had high frequency noise while most of the energy of the articulatory movement was below 25 Hz (Ghosh and Narayanan, 2010) for normal speaking rate. We found this to be true even for articulatory movement data at fast as well as slow speaking rates. To avoid high frequency noise resulting from EMA measurement error, the recorded articulatory data for all speaking rates were low-pass filtered with a cut-off frequency of 40 Hz. Then, to synchronize with the MFCC feature vectors, we down-sampled the articulatory data to 100 Hz from 250 Hz.

3. Analysis of impact of rate on articulation and acoustics

An analysis on impact of speaking rate on both articulation and acoustics is presented in this section in two parts: (1) rate of change and range of displacement in physical movements of articulators; (2) change in spectral characteristics of acoustics. Motivated by the relative spectral (RASTA) processing of speech (Hermansky and Morgan, 1994), which suppresses the spectral components that change at a rate different from the typical range of speech, we consider bandpass filter based approach for the analysis of acoustic (spectral changes) and articulatory (rate of movements) data. We also perform an analysis to characterize the change in displacement of articulatory movements across speaking rates.

Analysis of articulatory movements: Ghosh and Narayanan (2010) showed that, at normal speaking rate articulatory trajectories are smooth and low-pass in nature. In order to examine the variation in the rate of change in articulatory movements with respect to speaking rate, we perform frequency domain analysis on each articulatory trajectory at different speaking rates. In contrast to low-pass analysis by Ghosh and Narayanan (2010), here we perform a band-pass analysis on articulatory trajectories, where both lower and higher cutoff frequencies are optimized. Let \( z^p(n), (0 \leq n \leq N-1) \) be the zero-mean \( p^{th} \) articulatory trajectory for a sentence obtained at a sampling frequency \( F_s = 100 \) Hz, then the Discrete Fourier Transform (DFT) is computed as

\[
Z^p(k) = \sum_{n=0}^{N-1} z^p(n) \exp(-j(2\pi/N_F)nk) \tag{1}
\]

where, \( k = 0, 1, 2, \ldots N_f - 1 \) with a DFT order \( N_F = 2^{14} = 16384 \). From the range of \( k \), we choose an optimal passband region (with minimum bandwidth) which preserves \( \sim 95\% \) of the signal energy. The percentage of energy preserved in a given passband region (from lower cut-off \( FL \) to higher cut-off \( FH \) frequency) is computed as follows:

\[
EP = \frac{E_{\text{retain}}}{E_{\text{org}}} \times 100, \quad E_{\text{org}} = |Z^p(0)|^2 + 2 \sum_{k=1}^{N_F} |Z^p(k)|^2, \quad E_{\text{retain}} = 2 \sum_{k=k_L} |Z^p(k)|^2.
\]

For a given \( EP \), the optimal cut-off frequencies \( F_L = k_L \times \frac{F_s}{N_F} \) and \( F_H = k_H \times \frac{F_s}{N_F} \) are computed by minimizing the bandwidth \( BW = F_H - F_L \). This is done by a brute force search for \( k_L \) and \( k_H \) with a spectral resolution of \( F_s/N_F = 0.0016 \) Hz. This analysis is carried out for all 460 sentences. Fig. 2 illustrates the mean (bottom and top edge of bar...
represents $F_L$ and $F_H$, respectively) and SD (error bar) of the optimal cut-off frequencies across all 460 sentences for different articulators at three different speaking rates on a logarithmic scale. The average of $F_L$ and $F_H$ separately across all the articulators are reported in Table 2. The optimal cut-off frequencies vary across articulators and subjects. Fig. 2 and Table 2 suggest that the articulatory movements have the lowest values of both $F_L$ and $F_H$ in slow rate followed by those in normal and fast speaking rate. This is true for all subjects and all articulators. This could be due to the fact that the velocity of articulators tends to increase as the rate increases (Flege, 1988; Shaiman et al., 1997; De Nil and Abbs, 1991). Among all the articulators, the maximum increase in average bandwidth $\left(\frac{BW_{Fast} - BW_{Slow}}{BW_{Slow}} \times 100\right)$ is observed in tongue articulators, and minimum increase in average bandwidth is observed for lower lip and jaw articulators, and the respective values are reported in Table 3. Thus, different speaking rates could affect various articulators to different degrees. For example, the above analyses show that the change in speaking rate affects the tongue to a large extent while the lower lips and jaw at a lower degree.

At fast speaking rate, the articulators might undergo lingual undershoot (Flege, 1988), which, in turn, may impact their displacement (Gay, 1981). The reduction in the displacement of articulatory movement may affect either one of the extremes (maximum of lower/upper displacement) of the normal range of movements, or both (Krakow, 1993; Gay, 1981). One way to quantify the influence of the speaking rate on displacement would be to compute either the amplitude range (difference between the maximum and minimum amplitudes) or the standard deviation of the articulatory movement trajectory, as the range and standard deviation typically are linearly correlated.

![Fig. 2](image.png)

**Fig. 2.** Lower and higher cut-off frequencies (Hz) optimized for each articulator at three speaking rates and for all five subjects. Top (bottom) edge of bar shows the higher (lower) cut-off frequency averaged over 460 sentences while error-bar indicates SD on logarithmic scale.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Slow</th>
<th>Normal</th>
<th>Fast</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>0.19–3.33</td>
<td>0.24–5.60</td>
<td>0.33–7.43</td>
</tr>
<tr>
<td>M2</td>
<td>0.09–2.91</td>
<td>0.20–5.72</td>
<td>0.23–7.37</td>
</tr>
<tr>
<td>M3</td>
<td>0.08–4.09</td>
<td>0.15–6.50</td>
<td>0.22–8.95</td>
</tr>
<tr>
<td>F1</td>
<td>0.10–3.36</td>
<td>0.24–5.81</td>
<td>0.32–7.94</td>
</tr>
<tr>
<td>F2</td>
<td>0.12–3.23</td>
<td>0.17–5.16</td>
<td>0.22–7.60</td>
</tr>
</tbody>
</table>

Table 2
Mean of lower and higher cut-off frequencies ($F_L–F_H$, Hz) across all the articulators for three speaking rates and for all five subjects.
We choose the standard deviation of articulatory trajectories (SDAT) for each sentence as a measure of the extent of the range of articulatory movements. This could help to characterization of the impact of speaking rate on the extent of articulatory movements. The average SDAT (bar height) across all the sentences are illustrated in Fig. 3. We observe that the sensors on the tongue, on an average, have a greater amount of displacement in slow speaking rate compared to their fast counterparts. For fast speaking rate, the maximum drop in SDAT with respect to slow rate for M1, M2, M3, F1 and F2 is observed for TBX (34.44 %), TDY (32.55 %), TDY (21.15 %), TDY (23.55 %) and TTY (35.52 %), respectively. However, the drop in SDAT is not consistent for all the articulators. Unlike tongue, we observe an increase in SDAT from slow to fast speaking rate in upper lip for M2 and F1. The increase in bandwidth and decrease in SDAT of most of the articulatory trajectories suggest that an increase in speaking rate results in an increase in the velocity and/or drop in the extent of displacement of articulatory movements.

**Analysis of acoustic data:**

For acoustics feature extraction, speech signal is passed through an 24 channel triangular filter-bank (equally spaced along the Mel-scale and spread over the frequency range 0 to 8000 Hz) which produces Mel-FILTER Bank Energies (MFBE) (Young and Young, 1993). To study the rate of change in spectral sub-band trajectories of acoustics at all three speaking rates, we perform the following analysis: For each sentence, the logarithm of MFBE over time is considered as a sub-band trajectory. We perform the band-pass filter analysis on sub-band trajectories similar to the one done for articulatory trajectories.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Max increase in Avg. BW</th>
<th>Min increase in Avg. BW</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>TBX (155.84%)</td>
<td>Jawx (107.44 %)</td>
</tr>
<tr>
<td>M2</td>
<td>TTx (159.95 %)</td>
<td>Jawx (122.60 %)</td>
</tr>
<tr>
<td>M3</td>
<td>TBX (149.56%)</td>
<td>LLx (77.29 %)</td>
</tr>
<tr>
<td>F1</td>
<td>TDY (191.35%)</td>
<td>LLx (104.25 %)</td>
</tr>
<tr>
<td>F2</td>
<td>TTx (141.57%)</td>
<td>Jawx (122.85 %)</td>
</tr>
</tbody>
</table>

Fig. 3. SDAT for each articulatory trajectory for three speaking rates and for all five subjects. Bar height shows the SDAT averaged over 460 sentences and error-bar shows the SD of SDAT.
The bandpass region of sub-band trajectories, which preserves 95% of the total energy, is considered, and corresponding lower (\(FL\)) and higher (\(FH\)) cutoff frequencies are computed. Fig. 4 illustrates the lower and higher cutoff frequencies, averaged across all sentences, for different speaking rates. We observe that as the speaking rate increases the corresponding lower cutoff-frequencies increase from \(\approx 0.05\) to \(\approx 0.1\) Hz. Also, for all subjects, the cutoff frequencies \(FL\) and \(FH\), at slow speaking rate, are significantly \((p < 0.01)\) lower than those at the fast rate in the frequency bands below 3kHz, which correspond to the range of first three formant frequencies. This could be due the fact that the spectral amplitudes change slowly in slow speaking rate compared to the fast speaking rate. We also observe that the bandwidth \((FH - FL)\) and the center of gravity (CG) for the spectrum of sub-band trajectories increase from slow to fast speaking rate. Let \(Z_{sb}(k)\) denote the spectrum for the \(sb\)th sub-band trajectory.

The corresponding CG is computed by \(CG_{sb} = \frac{\sum_{k=0}^{N_F - 1} Z_{sb}(k)}{N_F - 1}\), and the average increase in CG from slow to fast is computed as \(CG = \left[\sum_{sb=1}^{24} (CG_{fast} - CG_{slow})\right]\), where \(CG_{slow}\) and \(CG_{fast}\) are the CG corresponding to slow and fast speaking rates, respectively.

Table 4 reports the maximum increase in bandwidth among all the sub-band trajectories and average (SD) increase in CG for the spectrum of sub-band trajectories for all the subjects. The increase in CG and bandwidth from slow to fast rate could suggest that there is not only an increase in spread of the spectrum but also a change in the spectral shape of sub-band trajectories.

Clearly, the speaking rate impacts both the articulatory and acoustic characteristics significantly, and this, in turn, would impact acoustic to articulatory mapping. Note that although the acoustic analysis is done on MFBE, to remove

Table 4
Maximum increase in bandwidth (sub-band) among all the sub-band trajectories and average (SD) increase in CG for the spectrum of sub-band trajectories.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Max Increase in BW</th>
<th>Average increase in CG (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>59.83 % (2nd SB)</td>
<td>1.76 (0.16)</td>
</tr>
<tr>
<td>M2</td>
<td>116.41 % (1st SB)</td>
<td>1.58 (0.57)</td>
</tr>
<tr>
<td>M3</td>
<td>89.55 % (1st SB)</td>
<td>1.97 (0.22)</td>
</tr>
<tr>
<td>F1</td>
<td>117.90 % (1st SB)</td>
<td>2.04 (0.74)</td>
</tr>
<tr>
<td>F2</td>
<td>103.07 % (1st SB)</td>
<td>2.25 (0.29)</td>
</tr>
</tbody>
</table>
high correlations among filter bank amplitudes, discrete cosine transform on logarithm of MFBE is performed to obtain MFCC (Young and Young, 1993), which is considered as a feature for performing AAI (since uncorrelated inputs perform better for neural networks (Haykin, 2009)) as described in Sections 5.1 and 6.

4. Information theoretic analysis of acoustic articulatory map

AAI is a regression problem, where the articulatory feature $Z$ is estimated from the acoustic feature $X$ by typically minimizing a mean-square error (MSE) function, $E[\hat{f}(X) - Z]^2$, where $\hat{f}(X)$ is an estimate of $Z$, which approximates the complex mapping function between $X$ and $Z$. A criterion for the choice of $X$ could be to maximize the information about $Z$, which is characterized by the mutual information (MI). Given two random variables $X$ and $Z$, the MI in terms of the entropy is defined as Cover and Thomas (2012):

$$I(X; Z) = H(Z) - H(Z|X)$$

where, $H(Z)$ is the entropy of $Z$ and $H(Z|X)$ is the conditional entropy (CE) which characterizes the uncertainty of $Z$ after observing $X$. MFCC has been shown to be the best choice of acoustic feature in an information theoretic sense for the AAI problem (Ghosh and Narayanan, 2010). Hence, we consider MFCC as the acoustic feature in this work.

Ideally, if $X$ is a deterministic function of $Z$, then $H(Z|X) = 0$. In practice, we expect to estimate $Z$ with low estimation error only if $H(Z|X)$ is small, which is governed by Fano’s inequality (Cover and Thomas, 2012). So, in order to quantify the performance of the estimator $\hat{Z} = \hat{f}(X)$, we use CE as a metric. CE is represented as follows:

$$H(Z|X) = -\mathbb{E}_{p(x,z)}[\log p(z|x)]$$

where $p(z|x)$ is the conditional distribution of the random variable $Z$ given $X$. It is known that when the distribution of estimation error is Gaussian, Laplacian or uniform, then minimizing the MSE is equivalent to minimizing the CE (Frény et al., 2013), i.e., MSE $\geq \frac{1}{2\pi^2\sigma^2}$ exp(2$H(Z|X)$), with equality for Gaussian distribution. Hence, the estimate of CE will give an insight on the minimum MSE that can be achieved. We perform experiments to observe how CE varies across different speaking rates.

Due to the unavailability of probability densities of $X$ and $Z$ directly, we quantize the $X$ and $Z$ into equal number of bins $K$ in both acoustic and articulatory space by K-means vector quantization, denoted by $Q()$. Let $x_i \in Q(X)$ and $z_j \in Q(Z)$, then discrete version of CE between two random variables $X$ and $Z$ is obtained by

$$H(Z|X) = \sum_{i=1}^{K} \sum_{j=1}^{K} p(x_i, z_j) \log \frac{p(x_i)}{p(x_i, z_j)}$$

where $p(x_i)$ is the probability of the occurrence of measurement $x_i$ for variable $X$ and $p(x_i, z_j)$ is the measured joint distribution of random variables $X$ and $Z$; $K$, the number of quantization bins, is varied from $K = 16$ to $K = 512$ and the joint density is estimated by frequency counts. The computation of CE is repeated 10 times with different random initialization in K-means clustering each time. As the number of quantization bins $K$ increases the CE increases monotonically (Cover and Thomas, 2012). Table 5 reports the mean of CE (SD) at three different rates for each subject, for $K = 256$. We observe that, consistently across all the subjects and for all the values of $K$, CE in slow rate is less than that of the fast speaking rate.

This implies that the statistical dependency of $X$ and $Z$ decreases as the rate increases. This, in turn, could imply that the MSE of AAI in slow rate would be lower than that in the fast speaking rate.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Slow</th>
<th>Normal</th>
<th>Fast</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>3.92 (.014)</td>
<td>3.93 (.031)</td>
<td>3.99 (.020)</td>
</tr>
<tr>
<td>M2</td>
<td>3.84 (.018)</td>
<td>3.95 (.018)</td>
<td>3.96 (.019)</td>
</tr>
<tr>
<td>M3</td>
<td>3.63 (.022)</td>
<td>3.74 (.018)</td>
<td>3.72 (.025)</td>
</tr>
<tr>
<td>F1</td>
<td>3.79 (.017)</td>
<td>3.73 (.021)</td>
<td>3.83 (.015)</td>
</tr>
<tr>
<td>F2</td>
<td>3.90 (.019)</td>
<td>3.93 (.021)</td>
<td>4.00 (.022)</td>
</tr>
</tbody>
</table>
5. Inversion model and experiment setup

5.1. Acoustic-to-articulatory inversion using deep neural network

The mapping from acoustic to articulatory features is known to be a nonlinear problem (Wu et al., 2015). For AAI, we choose a Deep Neural Network (DNN), since DNN can learn a nonlinear mapping efficiently (Wu et al., 2015). For a DNN with \( L \) layers, the input acoustic feature vector \( x \) is given to the first layer. The last layer of the DNN is considered to be a linear regression layer, where we obtain the output — the predicted articulatory feature vector \( v_L(x) \) for the input \( x \). Given the weight matrix \( W_l \), and hidden bias \( b_l \), the output of the \( l \)-th hidden layer \( v_l(x) \) is given by,

\[
v_l(x) = \text{sigmoid}(u_l(x)), l = 2, 3, \ldots, L-1
\]

where,

\[
u_l(x) = W_{l-1}v_{l-1}(x) + b_l,
\]

\[
\text{sigmoid}(x) = \frac{1}{1 + e^{-x}}
\]

\( \text{sigmoid}(x) \) is the activation function, and \( x \) is the input for DNN. We consider an objective of minimizing the MSE between the desired output and the predicted output, following the work by Wu et al. (2015). While the weight matrices are learnt by the back-propagation algorithm, the weights are updated using ADAM (Kingma and Ba, 2014), a first-order gradient-based optimization algorithm. ADAM is an efficient algorithm for stochastic objective functions that computes the gradients based on the adaptive estimates of lower order moments. DNN is implemented by using Keras libraries Chollet (2015).

5.2. Experimental setup

The AAI is performed separately for slow, normal and fast speaking rates for each subject. At a particular speaking rate for each subject, the AAI is performed in a 10-fold cross-validation setup, where the entire set of 460 sentences is divided into ten groups, among which eight groups are used for training, and remaining two groups for validation and testing respectively in a round robin fashion. To incorporate the contextual information in the DNN, we concatenate MFCCs of five frames before and after the current frame, resulting in a 429-dim acoustic feature vector. A cepstral mean subtracted and variance normalized (Young and Young, 1993) acoustic feature vector is then used as the input for the DNN. Since average position for each sensor could change across sentences (Kirchhoff, 1999), we remove the mean sensor position and divide by the SD for each articulatory feature trajectory for every sentence.

We choose a \( L = 5 \)-layer (input, 3-hidden, and output layers) DNN which contains 300 units in each hidden layer (Uria et al., 2011). The DNN predicts the 36-dim articulatory feature (including velocity and acceleration coefficients of 12 articulatory features) when provided with the 429-dim acoustic feature vector as the input. For learning the weights of the DNN, the parameters in ADAM optimization are chosen as follows: learning rate = 0.001, \( \beta_1 = 0.9 \), \( \beta_2 = 0.999 \), \( (\beta_1, \beta_2) \in [0, 1) \): exponential decay rates for the moment estimates), \( \epsilon = 10^{-8} \) and batch size of 128. To compute the AAI performance, two evaluation metrics, namely, the root mean squared error (RMSE) and correlation coefficient (CC) \( \psi \) (Ghosh and Narayanan, 2010) between the actual and predicted articulatory trajectories are chosen.

For the \( k \)-th articulatory feature the RMSE and \( \psi \) are given by

\[
RMSE^k = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left( z^k(n) - v^k_L(n) \right)^2}, \tag{7}
\]

\[
\psi^k = \frac{\sum_{n=1}^{N} \left( z^k(n) - \bar{z}^k \right) \left( v^k_L(n) - \bar{v}^k_L \right)}{\sqrt{\sum_{n=1}^{N} \left( z^k(n) - \bar{z}^k \right)^2 \sum_{n=1}^{N} \left( v^k_L(n) - \bar{v}^k_L \right)^2}}. \tag{8}
\]
where, $z_k(n)$ and $v_k^L(n)$ are the original and predicted articulatory features at the $n$-th frame for a test data of duration $N$ frames. $z_k = \frac{1}{N} \sum_{n=1}^{N} z_k(n)$ and $v_k^L = \frac{1}{N} \sum_{n=1}^{N} v_k^L(n)$ are the averages for original and predicted articulatory movement data.

The predicted articulatory trajectory from the DNN is, in general, jagged in nature (Ghosh and Narayanan, 2010), but the original articulatory trajectory is typically smooth. So the predicted articulatory data, for each articulator $v_k^L(n)$, $1 \leq k \leq 12$ is smoothened by low-pass filtering. In this regard, we varied the cutoff frequency of the low-pass filter from 1 to 25 Hz, with a step of 1Hz. We choose the cut-off frequency for each articulatory feature that minimizes the corresponding RMSE on the validation set.

### 6. Results and discussion

We present the AAI performance in three conditions — (1) Train an AAI model specific to the speaking rate and test with the same rate (matched train-test) for every subject. (2) Train an AAI model with a speaking rate and test with unseen rate (mismatched train-test). (3) Train a generic AAI model which is trained with all the three speaking rates together and test with each speaking rate separately.

#### 6.1. Speaking rate specific matched train-test condition

To observe the effect of speaking rate on the inversion performance, for all three different speaking rates separately, we trained and tested the AAI model on the same speaking rate data. Fig. 5 shows the bar plot of CC for the 12 articulatory features for all five subjects (each row in Fig. 5 corresponds to one subject) in three speaking rates. Bar height indicates the average value and the error-bar shows the SD. Last row shows the barplot of CC averaged across five subjects. Table 6 reports accuracy of AAI in terms of average RMSE across all the articulators. In general, average CC drops and average RMSE increases as the speaking rate increases for most of the articulatory features, as seen from the last row in Fig. 5. For example, by comparing the slow and fast speaking rates, the AAI performance drops for all articulatory features except $UL_x$, as seen from the last row in Fig. 5. Among these
articulatory features, average RMSE increases significantly ($p < 0.01$, all statistical tests in this work are done by t-test (Mendenhall and Sincich, 2016)) for the fast rate compared to the slow speaking rate for Jawx, Jawy, TTx, TTy, TBx, TBy, TDx, and TDy. Similarly, average CC drops significantly ($p < 0.01$) for the fast rate compared to the slow rate for all these articulators except Jawx. The articulators for which the AAI performance for the fast rate significantly drops compared to that for the slow rate, varies from one subject to another. However, it is interesting to observe that there is a significant ($p < 0.01$) drop in the CC and a significant ($p < 0.01$) increase in RMSE for TTy, TBy and TDy, consistently, for all subjects. These results suggest that, the predicted movement of TT, TB, and TD in Y (vertical) direction in case of the fast rate are less accurate compared to the slow rate. It could be because of the fact that the manner in which the tongue movement and velocity profiles are encoded in speech could be different when spoken in these two rates (Flege, 1988). This is also observed in Table 3, where tongue articulators show maximum increase in average BW and also drop in SDAT.

The AAI performance for the normal rate does not appear to be always in between that for the slow and fast rates for different articulators. This is clear from the AAI performance (AvgS, AvgN, and AvgF in Fig. 5) averaged across all articulators for all subjects as well as when averaged across subjects. Unlike, slow vs. fast speaking rate, the AAI performance in slow and fast speaking rates does not change significantly from that in normal speaking rate for most of the articulators. It could be that the differences in the acoustic due to slow and normal speaking rates may not cause significant change in the acoustic-articulatory map resulting in similar AAI performance in these two rates. This could also be the case for acoustic-articulatory map in normal and fast speaking rates.

The drop in the AAI performance using DNN in fast compared to the slow speaking rate is consistent with the information theoretic analysis of acoustic-articulatory map (Table 5). This drop in performance could be due to the fact that an increase in rate is achieved by increasing the velocity and/or decreasing the range of articulatory movement (Flege, 1988; Krakow, 1993), resulting in higher overlap among articulatory gestures (Hardcastle, 1985; Munhall and Lofqvist, 1988; Saltzman and Munhall, 1989) and vowel space reduction (Fourakis, 1991; Weismer and Berry, 2003). This may cause an increase in the variation of articulatory movements and their corresponding acoustic characteristics within the speech units, which results in a drop in the AAI performance during fast rate. On the other hand, at slow speaking rate there is a decrease in variation within the speech units. This is supported, for example, by the speed-accuracy trade-off in the neural network model of speech production (Guenther, 1994; 1995), where target size reduces for clear articulation during slow speaking rate, which could help in a better AAI performance in the slow rate.

6.2. Speaking rate mismatched train-test condition:

To verify the consistency of the AAI performance with change in speaking rate, an AAI model, trained with one particular speaking rate, is used on the test cases speaking rate different from that in training.

Fig. 6 shows the bar plot of CC averaged across all 12 articulatory features in three different speaking rates for five subjects. The height of the bar plot and error bar indicate the average and SD of CC across all articulators respectively. Three different shades of gray in the bar plot in Fig. 6 indicate the AAI model trained on three different speaking rates namely, SlowM, NormalM, and FastM for slow, normal and fast rates respectively. Rows in Fig. 6 illustrate the AAI model performance on test cases with fast, normal and slow speaking rates, denoted by FastT, NormalT and SlowT respectively. From Fig. 6 we observe that there is a significant drop in the AAI performance in
the mismatched cases compared to the respective matched cases. For example, the first row illustrates that the performance on a test set with slow speaking rate with model trained on slow train set (matched condition) is better than the other two mismatched cases. We examine the accuracy with which an AAI model trained with one speaking rate, performs, when the speech from other speaking rate is presented as a test case (mismatched condition). We compute the percentage drop in CC (PDCC) for mismatched condition as follows:

$$PDCC_{\alpha}^{\beta_k} = \frac{\psi_{\alpha_k}^k - \psi_{\beta_k}^k}{\psi_{\alpha_k}^k} \times 100$$

where, for the $k$-th articulator, $\psi_{\alpha_k}^k$ and $\psi_{\beta_k}^k$ are the CCs in the matched (train and test with $\alpha$ rate) and mis-matched conditions ($\beta$ rate train and test with $\alpha$ rate) respectively. The PDCC averaged across all the subjects is found to be larger (16.92%) in fast($\beta$) train and slow($\alpha$) test mismatched condition compared to slow-fast (14.39%), normal-slow (9.15%), and normal-fast (7.09%) mismatched conditions. Similarly, for normal rate the PDCC for fast-normal is found to be larger (9.85%) than the slow-normal (7.80%) mismatched condition. We also observe that the PDCC drop for each subject is related to their relative drop in speaking rate from fast to slow. From Table 1, the maximum decrease in speaking rate of 70.31% is found for M2 and the minimum of 58.79% for M3. Interestingly, the drop in PDCC averaged across all combinations of mismatched conditions is found to be the maximum (13.84%) for M2, and minimum (7.41%) for M3 among all subjects. From Table 4 and Table 2 we observe a change in acoustic features (BW and CG) and articulatory trajectories (BW and SDAT) between slow to fast speaking rates, which, in turn, alters their statistical distributions. This results in a mismatch in mapping function across fast and slow speaking rate, which could leads to a drop in performance of AAI in mismatched cases. The mismatched results with RMSE are illustrated in Fig. 7, where bar plot illustrates RMSE averaged across all the 12 articulatory features and the conclusions are found to be consistent with CC (Fig. 6).

### 6.3. Generic AAI model

The AAI performance summarized in Sections 6.1, 6.2 indicates that (1) performance of AAI is rate dependent (2) acoustic-to-articulatory mapping is different for each rate. It brings up few questions: (a) Do we need a rate dependent AAI model? (b) Can a single model learn the acoustic-articulatory mappings for different rates? To answer these questions, the following experiment is conducted.

We trained an AAI model together on slow, normal and fast speaking rates training data and tested on all three rates. The AAI performance is summarized in Fig. 8 (using CC) and Fig. 9 (using RMSE). By comparing Figs. 5 and 8, we, interestingly, find that both rate specific and generic model appear to result in a similar performance. We verify this by performing statistical test between the performance of the generic AAI model and the rate specific AAI model with matched train-test condition in all three speaking rates. For all the articulators, we found that there is no significant ($p < 0.01$) drop in performance for both RMSE and CC for every subject.

This experiment indicates that an AAI model trained together on three different speaking rates performs equally well as the speaking rate dependent AAI model. By providing training data from different rates, DNN is able to learn the rate dependent variations in the AAI mappings within a single model. Thus, when AAI needs to be performed on
test sentences at different speaking rates, it is not necessary to have a speaking rate dependent AAI model, rather a single model trained on all three speaking rates is sufficient.

We presented the AAI performance in three conditions: (1) In matched train-test condition where we observed that performance in fast rate with respect to slow rate decreases particularly for jaw and tongue articulators. (2) The evaluation of AAI performance in mismatched train-test condition revealed that due to mismatch in acoustic and articulatory feature characteristics there is a drop in performance of AAI in mis-matched condition compared to that in matched condition. (3) The Generic AAI model revealed that when a single DNN model is trained using acoustic-articulatory data pooled from different speaking rates, the AAI model is capable of learning multiple rate-specific mapping.

Fig. 7. RMSE averaged across all articulators for AAI tested in both matched (in △) and mismatched conditions with all possible combinations of slow, normal & fast speaking rates for five subjects.

Fig. 8. Correlation coefficient ψ using Generic AAI model trained together on slow, normal and fast speaking rates. ○ denotes the articulatory features for which there is a significant drop in AAI performance for fast speaking rate compared to the slow speaking rate.
7. Conclusions

Analysis on the articulatory data shows that the speaking rate affects different articulators to different degrees. The conditional entropy of articulatory features given acoustic features decreases as the rate increases, which implies that the performance of AAI could decrease for fast rate. In fact, this matches with the results from AAI experiments done using acoustic and articulation data as well. AAI experiments at different speaking rates reveal that the AAI performance at fast speaking rate drops for tongue compared to those at slow speaking rate. We also find that the performance of the AAI decreases more in slow-fast mismatched conditions compared to the normal-slow and normal-fast cases. Fast articulatory movement along with rapid acoustic changes in fast speaking rate could result in an acoustic-articulatory map quite different from those in slow speaking rate. How this map changes for different sounds require further investigations. In future we will also investigate on low resource subject and rate independent AAI models using transfer learning techniques which could benefit in applications like ASR.

Acknowledgements

We thank all subjects and G. Nisha Meenakshi for helping in EMA data collection and the Pratiksha Trust for their support.

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.csl.2019.05.004.

References
