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

A phonetic-level analysis of different input features for articulatory inversion

Abdolreza Sabzi Shahrebabaki[1], *Negar Olfati*[1], *Ali Shariq Imran*[1], *Sabato Marco Siniscalchi*[2], *Torbjørn Svendsen*[1]

[1]Department of Electronic Systems, NTNU
[2]Department of Telematics, Kore University of Enna

{abdolreza.sabzi, negar.olfati, ali.imran, torbjorn.svendsen}@ntnu.no,
marco.siniscalchi@unikore.it

Abstract

The challenge of articulatory inversion is to determine the temporal movement of the articulators from the speech waveform, or from acoustic-phonetic knowledge, e.g. derived from information about the linguistic content of the utterance. The actual position of the articulators is typically obtained from measured data, in our case position measurements obtained using EMA (Electromagnetic articulography). In this paper, we investigate the impact on articulatory inversion problem by using features derived from the acoustic waveform relative to using linguistic features related to the time aligned phone sequence of the utterance. Filterbank energies (FBE) are used as acoustic features, while phoneme identities and (binary) phonetic attributes are used as linguistic features. Experiments are performed on a speech corpus with synchronously recorded EMA measurements and employing a bidirectional long short-term memory (BLSTM) that estimates the articulators' position. Acoustic FBE features performed better for vowel sounds. Phonetic features attained better results for nasal and fricative sounds except for /h/. Further improvements were obtained by combining FBE and linguistic features, which led to an average relative RMSE reduction of 9.8%, and a 3% relative improvement of the Pearson correlation coefficient.

Index Terms: Articulatory inversion, language learning, bidirectional long short term memory, Attributes, HPRC database

1. Introduction

Acoustic to articulatory inversion (AAI) is a challenging problem due to the many-to-one mapping in which different articulator positions may produce a similar sound. This many-to-one mapping makes AAI a highly non-linear problem. In AAI, the objective is to estimate the vocal tract shape, which is estimated by the articulator positions based on the uttered speech. AAI can be useful in many speech-based applications, in particular, speech synthesis [1], automatic speech recognition (ASR) [2, 3, 4] and second language learning [5, 6]. Over the years, researchers have addressed this problem employing various machine learning techniques including codebooks [7], Gaussian mixture models (GMM) [8], hidden Markov models (HMM) [9], mixture density networks [10], deep neural networks (DNNs) [11, 12, 13], and deep recurrent neural networks (RNNs) [14, 15, 16].

Exploiting RNNs for the AAI task has demonstrated better results compared to DNNs [14, 16] because the temporal dynamic behavior is better captured through the memory elements of those recurrent architectures. Acoustic features are commonly employed at the input of the AAI system [7, 8, 9, 10] , but linguistic features have been successfully used in recent years either as stand-along features [17], or together with acoustic features [15]. Moreover, representing the linguistic features in a bottleneck form extracted from a phone classifier has been used in [16]. Although leveraging knowledge from linguistic content together with acoustic features has proven to improve AAI systems, a deeper analysis explaining why redundant information makes the system perform better is missing. We think that gaining a better understanding about such a performance improvement would be helpful for some specific tasks, where the linguistic features are available from the text, e.g. language learning. This motivates us to compare state-of-the-art methods in [16, 17] and carry out additional analyses on the acoustic and linguistic features within phoneme boundaries which later can be employed in pronunciation scoring. That is, we focus on the evaluation in time intervals concerning a single phoneme instead of analyzing the whole EMA trajectory for the uttered speech. The rest of the paper is structured as follows. Section 2 presents Deep BLSTM recurrent neural networks. Section 3 describes the "Haskins Production Rate Comparison database"(HPRC) [18]. The database, feature representation, and the performance measurements undertaken in this study, followed by results in Section 4. Finally, Section 5 concludes the paper.

2. Deep BLSTM recurrent neural network

Recurrent neural networks (RNN) have been utilized in many speech technology areas including speech recognition [19], language modeling [20], and articulatory inversion [14, 15, 16]. They are able to estimate any output samples from dynamical systems [21], conditioned on their previous samples. Having a non-causal condition by access to both past and future input samples, we can employ a bidirectional RNN to use the past samples within the forward layer and the future samples within the backward layer as it is shown in Fig. 1. Diamonds show the merge strategy of forward and backward layers output which can be summation, concatenation, and etc. LSTM is a variant of RNN with a specific memory cell architecture for updating the hidden layers. This memory cell is formulated as follows

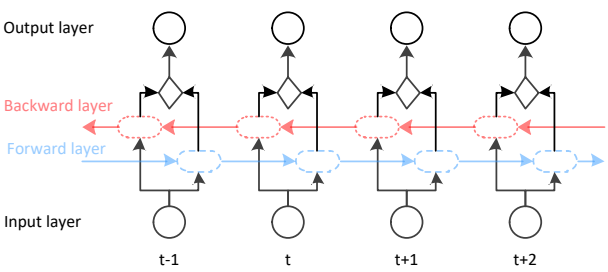

Figure 1: *A bidirectional RNN.*

http://dx.doi.org/10.21437/Interspeech.2019-2526

$$i_t = \sigma\big(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i\big) \quad (1)$$

$$f_t = \sigma\big(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f\big) \quad (2)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ g_c\big(W_{cx}x_t + W_{ch}h_{t-1} + b_c\big) \quad (3)$$

$$o_t = \sigma\big(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}c_{t-1} + b_o\big) \quad (4)$$

$$h_t = o_t \circ g_o(c_t) \quad (5)$$

where x_t and h_t are input and hidden vector, i_t, f_t, c_t and o_t are the input gate, forget gate, cell vector and output gate, respectively. The σ is sigmoid function, g_c and g_o are the activation function which is usually chosen as $tanh$, b shows the bias vector for each gate (b_f is the forget gate bias vector) and weight matrices W with different subscripts which show the connection between input/output with gates, for example, W_{ix} is the input gate-input weight matrix. The operator $\circ$ shows the element-wise multiplication. A bidirectional LSTM (BLSTM) is realizable by using the LSTM memory cells (dotted ovals) in the forward and backward layer as shown in Fig. 1.

3. Experimental Setup

3.1. EMA database

There exists several techniques to measure the articulatory movements, e.g. MRI, microbeam x-ray, and electromagnetic articulography (EMA). Among them, EMA is the most widely used technique to simultaneously capture the speech and articulatory data. MOCHA-TIMIT [22], MNGU0 [23], and USC-TIMIT [24] are speech corpora which contain EMA data. Another such database is "Haskins Production Rate Comparison"(HPRC) [18]. This database contains EMA readings from eight native American English speakers, four male and four female. This database has 720 recorded sentences at normal and fast Speaking Rate (SR), respectively. Some of the sentences are uttered two times in the normal SR by each speaker. Table 2 shows the amount of data for different SR, where "N1", "N2", and "F1"represent the normal SR, repetition of some of the sentences with the normal SR, and fast SR respectively. The sampling rate of the recorded audio files is 44.1 KHz and the EMA recordings are sampled at 100 Hz. The EMA readings are obtained from the sensors placed in different locations of the mouth, tongue, and jaw. Precisely, eight sensors are used in this case which are placed on tongue rear/dorsum (TR), tongue blade (TB), tongue tip (TT), upper lip (UL), lower lip (LL), mouth left (ML), lower incisors/jaw (JAW), and jaw left (JAWL). The EMA readings of the articulatory movements from these carefully placed sensors is measured in the midsagittal plane in X, Y, and Z directions. The X-direction denotes the movement of the articulators from posterior to anterior, the Y denotes the right to left movement, while the Z denotes the inferior to superior articulatory movements. In this paper, we used six reading locations of X and Z direction, i.e., TR, TB, TT, UL, LL, and JAW. In other databases mentioned above these six locations are mostly used, while the Y direction is omitted as the contribution of right to left movement does not contribute much under normal continuous speech.

3.2. Input representation.

3.2.1. Acoustic representation

The acoustic features are extracted from audio downsampled to 16 KHz, using 25 ms frame length and 10 ms frame shift. The resulting features have a 100 Hz sampling rate, the same as the articulatory features. The acoustic features are calculated from smoothed spectrum by the STRAIGHT method [25] with 40 filters which are linearly spaced on Mel-scale frequency axis. The energies in the overlapping frequency bands are called filter bank energy (FBE) features. The extracted feature frame is concatenated with the M previous and future time for each frame as the network input.

3.2.2. Phonetic representation

Spoken utterances have been labeled with the Penn phonetics lab forced aligner [26]. There are 61 phonetic categories which are folded onto 39 categories [27] for TIMIT database [28] which are depicted in the first row of Table 1. Each phoneme is represented as a one-hot 39-dimensional vector [17].

3.2.3. Attribute representation

With the reduced phoneme set from 61 to 39, we use a mapping from phoneme to their phonological features known as attributes which are depicted in Table 1. We consider 22 attributes in this study, comprising manner and place of articulation for both vowel and consonant categories [29]. The attribute features are binary and more than one attribute feature is often active at the same time. These features are more language universal [30] compared to the phonetic representations.

3.3. Performance measurements

To measure the performance of the AAI methods, root mean squared error (RMSE) and Pearson's correlation coefficient (PCC) are chosen. The first criterion reports the deviation and the latter measures the similarity between estimated and the ground-truth trajectories. These measures are defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_i \big(y(i) - \hat{y}(i)\big)^2}, \quad (6)$$

$$\mathrm{PCC} = \frac{\sum_i (y(i) - \bar{y})(\hat{y}(i) - \bar{\hat{y}})}{\sqrt{\sum_i \big(y(i) - \bar{y}\big)^2 \sum_i \big(\hat{y}(i) - \bar{\hat{y}}\big)^2}}, \quad (7)$$

where $y(i)$ and $\hat{y}(i)$ are the ground-truth and estimated EMA value of the i^{th} frame respectively and $\bar{y}$ and $\bar{\hat{y}}$ are mean values of $y(i)$ and $\hat{y}(i)$. All results are based on training on the N1 subset and test on the N2 subset. 5% of the training data is used as the validation data which is used to stop training, to prevent the network from getting over-fitted to the training data.

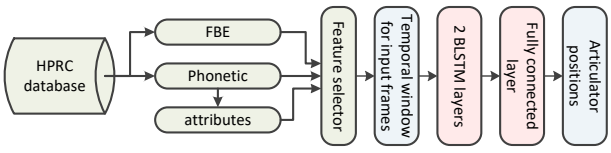

Figure 2: *Network structure for the articulatory inversion system*

3.4. Deep neural network architecture

The block diagram of the network architecture is shown in Fig. 2. It contains a feature selector module that selects among the input features to either output them individually or combine them two by two. The output of the feature selector goes to

Table 1: *American-English phonemes and associated attributes in terms of manner and place of articulations*

	ɑ	æ	ʌ	aʊ	aɪ	b	ʧ	d	ð	ɾ	ɛ	ɜ	eɪ	f	ɡ	h	ɪ	i	ʤ	k	l	m	n	ŋ	oʊ	ɔɪ	p	ɹ	s	ʃ	t	θ	ʊ	u	v	w	j	z	sil
Vowel	1	1	1	1	1	0	0	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0
Fricative	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	1	0	0	1	0
Nasal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Stop	0	0	0	0	0	1	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
Approx	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0
High	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0	0
Coronal	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0
Dental	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Glottal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Labial	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0
Low	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Mid	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Retroflex	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Velar	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Voiced	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	1	1	1	0	1	1	1	1	1	1	0	1	0	0	0	0	1	1	1	1	1	1	0
Round	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	1	1	1	1	1	0	0
Tense	1	1	0	1	1	0	1	0	0	0	0	0	1	1	0	1	0	1	0	1	0	0	0	0	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0
Anterior	0	0	0	0	0	1	0	1	1	1	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	1	0	1	0	1	1	0	0	1	1	0	1	0
Back	1	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0
Continuant	1	1	1	1	1	0	0	0	1	0	1	1	1	1	0	0	1	1	0	0	1	0	0	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0
Vocalic	1	1	1	1	1	0	0	0	0	0	1	1	1	0	0	0	1	1	0	0	1	0	0	0	1	1	0	1	0	0	0	0	1	1	0	1	1	0	0
Silence	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 2: *Available amount of data for different speaking rates for the HPRC database.*

Speaking rate	NO. utterances	Amount of data
N1	5756	$\sim$ 244 (minutes)
N2	1379	$\sim$ 55 (minutes)
F1	5735	$\sim$ 173 (minutes)

the next module, namely the temporal window for input frames. It takes M past and future frames and feeds this temporal context window to the BLSTM layers consisting of 128 cells in both forward and backward directions. $M = 15$ is used for this work which is resulted from some primary experiments. Sigmoid and tanh activation functions are used for the recurrent layers. At last, a fully connected network with the linear activation is used to map the BLSTM outputs to the articulator positions. The implementations used Keras [31] with TensorFlow backend [32].

4. Results

In this section, we evaluate the performance of different inputs to the AAI system. For having a fair comparison between different features, we used the same architecture for both state-of-the-art methods [16, 17] introduced in 3.4. However, feeding inputs directly to the BLSTM, instead of having several fully connected layers prior to BLSTM performs slightly better. We used the bottleneck features proposed in [16] but we got same performance as phonetic features. We argue that the reason is that the phonetic features are already a parsimonious representation of the input speech capturing information similar to that captured by bottlenecks. The experiments are done speaker dependently. The same context window of 15 past and future frames is used for all input features. Tables 4 and 5 show the PCC and RMSE results for each speaker, considering the acoustic (FBE), phonetic (Phn) and attribute (AF) features in the first three columns, and the combined features in the second three columns (FBE+Phn, FBE+AF, Phn+AF). Comparing the results, we observe that attribute features give worse results than the phonetic features for all speakers. The RMSE results of 4 show that combining features improve performance relative to stand-alone features. The RMSE improvement is in the range of 0.1 mm to 0.15 mm. Similarly, the PCC improvement for each speaker after combining the features is in the range of 0.01 to 0.03, as shown in table 5.

Table 3: *Average RMSE for phoneme*

	FBE	**Phn**	**AF**	**FBE+Phn**	**FBE+AF**	**Phn+AF**
/ɑ/	**2.05**	2.16	2.31	**1.87**	1.89	2.16
/æ/	**1.96**	2.21	2.22	**1.83**	1.88	2.17
/ʌ/	**1.94**	1.99	2.01	**1.77**	1.78	1.98
/aʊ/	**2.04**	2.21	2.32	1.80	1.86	2.21
/aɪ/	**2.00**	2.15	2.32	1.83	1.83	2.15
/ɛ/	**1.88**	1.94	1.95	1.74	1.74	1.96
/ɜ/	**1.87**	1.9	1.91	**1.71**	1.72	1.88
/eɪ/	**1.71**	1.82	1.83	**1.62**	1.63	1.83
/ɪ/	1.83	**1.81**	1.85	**1.65**	1.66	1.81
/i/	1.76	**1.71**	1.72	1.59	**1.58**	1.69
/oʊ/	2.06	**2.00**	2.05	1.79	**1.77**	2.02
/ɔɪ/	**1.99**	2.15	2.7	1.82	**1.81**	2.13
/ʊ/	1.89	**1.87**	1.91	**1.64**	1.70	1.94
/u/	1.84	**1.82**	1.84	**1.66**	**1.66**	1.79
/ʧ/	1.68	**1.66**	1.67	**1.56**	**1.56**	1.66
/ð/	1.96	**1.86**	**1.86**	**1.75**	1.76	1.86
/f/	1.84	**1.80**	1.82	**1.65**	1.66	1.77
/h/	**2.05**	2.26	2.27	**1.85**	1.92	2.28
/ʤ/	1.67	**1.64**	1.68	1.57	**1.55**	1.67
/s/	1.61	**1.57**	1.59	**1.48**	1.49	1.57
/ʃ/	**1.61**	1.63	1.60	1.49	**1.48**	1.59
/θ/	1.97	**1.75**	1.78	**1.63**	1.64	1.75
/v/	2.06	**1.89**	1.90	**1.72**	1.74	1.86
/z/	1.66	**1.61**	**1.61**	**1.51**	1.52	1.58
/m/	2.06	**1.95**	1.98	**1.80**	1.82	1.94
/n/	1.97	**1.89**	1.92	**1.73**	**1.73**	1.88
/ŋ/	2.13	**1.90**	1.92	**1.75**	**1.75**	1.9
/b/	**1.89**	1.90	1.90	1.75	**1.72**	1.87
/d/	**1.91**	1.92	1.93	**1.72**	1.74	1.89
/ɾ/	–	–	–	–	–	–
/g/	2.04	**1.85**	1.88	**1.70**	1.72	1.84
/k/	**1.97**	**1.97**	1.99	1.76	**1.75**	1.94
/p/	2.01	**1.96**	2.03	**1.72**	1.74	1.94
/t/	**1.91**	1.94	1.96	**1.72**	1.73	1.91
/l/	1.87	**1.85**	1.87	1.69	**1.68**	1.82
/ɹ/	**1.95**	2.09	2.09	**1.82**	1.83	2.09
/w/	2.02	**1.98**	2.02	**1.79**	1.82	1.98
/j/	1.77	**1.74**	1.82	1.61	**1.58**	1.72

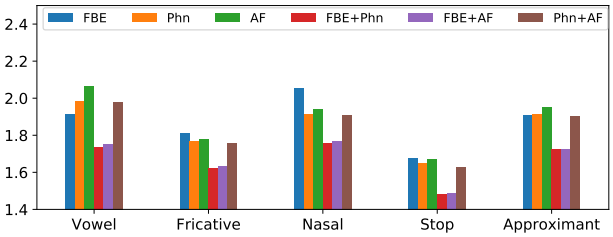

Figure 3: *Average RMSE for manner of articulation from estimated trajectory by different input features*

To gain a better understanding of the performance of each input feature, RMSE results for all phones averaged over all speakers and articulator positions is calculated for the input features investigated. This is depicted in table 3. A compact form of table 3 in terms of five phonetic classes is represented in Fig. 3. These phonetic classes are "vowel", "fricative", "nasal", "stop"and "approximant". We can conclude that FBE works better in case of vowels for stand-alone input features. By a deeper inspection on the results in Table 3, we can say FBE works better for phones where the place of articulation is low (/ɑ/, /æ/, /aʊ/, /aɪ/ and /ɔɪ/), whilst Phn works better in the case of high (/ɪ/, /i/, /ʊ/ and /u/). For all fricatives except /h/, Phn and AF perform better than FBE according to the Fig. 3 and Table 3. Phn and AF features are also better for nasals. In case of stops, all of the features are performing more or less the same except /g/ in which Phn is superior to FBE by 0.19 mm in RMSE.

Table 4: *RMSE for different input features.*

Spk.	FBE	Phn	AF	FBE+Phn	FBE+AF	Phn+AF
F1	**1.356**	1.365	1.405	**1.196**	1.221	1.352
F2	**1.601**	1.631	1.663	**1.449**	1.451	1.632
F3	1.308	**1.302**	1.320	**1.201**	1.202	1.293
F4	**1.469**	1.601	1.625	**1.308**	1.329	1.585
M1	1.208	**1.158**	1.173	1.074	**1.073**	1.151
M2	**1.667**	1.715	1.745	1.536	**1.530**	1.707
M3	1.565	**1.539**	1.566	**1.426**	1.441	1.523
M4	1.259	**1.235**	1.264	**1.124**	1.128	1.231
Avg.	**1.429**	1.443	1.470	**1.289**	1.296	1.434

Table 5: *PCC for different input features.*

Spk.	FBE	Phn	AF	FBE+Phn	FBE+AF	Phn+AF
F1	0.918	**0.921**	0.915	**0.937**	0.936	0.921
F2	0.848	**0.850**	0.845	**0.879**	0.878	0.852
F3	0.821	**0.830**	0.826	**0.850**	**0.850**	0.834
F4	**0.901**	0.894	0.890	**0.922**	0.920	0.895
M1	0.869	**0.883**	0.880	**0.897**	0.896	0.886
M2	**0.860**	0.855	0.850	**0.880**	**0.880**	0.856
M3	0.821	**0.831**	0.823	**0.850**	0.847	0.834
M4	0.832	**0.846**	0.839	0.863	**0.865**	0.845
Avg.	0.859	**0.864**	0.858	**0.885**	0.884	0.865

Moreover, RMSE for each of the articulator positions is calculated by different input features and shown in Table 6. By comparing different individual features we can see there are 0.01 to 0.05 mm differences in RMSE. In the combined input features, combination of FBE and phonetic features gives a better performance in most of cases. Moreover, there is not a big difference (less than 0.01 mm RMSE) between combining FBE with phonetic and attribute features, which are more universal among languages and the network architecture would not need any changes for using it in transfer learning for new languages.

Table 6: *Performance of AAI system in terms of RMSE.*

	FBE	Phn	AF	FBE+Phn	FBE+AF	Phn+AF
TD_x	**1.539**	1.568	1.596	1.426	**1.424**	1.555
TD_z	1.904	**1.868**	1.941	**1.667**	1.680	1.861
TB_x	**1.729**	1.759	1.788	**1.563**	1.564	1.742
TB_z	**1.864**	1.928	2.005	**1.690**	1.699	1.916
TT_x	**1.851**	1.878	1.896	1.665	**1.662**	1.857
TT_z	**1.922**	1.966	1.989	**1.711**	1.715	1.959
UL_x	**0.715**	0.722	0.727	**0.665**	0.668	0.718
UL_z	**1.214**	1.288	1.297	**1.129**	1.138	1.279
LL_x	0.863	**0.822**	0.844	**0.794**	0.802	0.821
LL_z	0.817	**0.750**	0.754	**0.726**	0.727	0.747
JAW_x	1.010	**0.999**	1.016	**0.910**	0.920	0.992
JAW_z	**1.725**	1.772	1.794	**1.527**	1.552	1.765

5. Conclusions

The problem of acoustic to articulatory inversion is addressed in this paper for different input feature types for a two-hidden layer BLSTM with 128 cells in each of its forward and backward layers. FBE features are chosen as the acoustic features and phonetic and attribute features are selected as the linguistic features. The experiments are conducted on a multi-speaker database which will be useful for further investigations on the speaker independent AAI systems. RMSE and PCC are computed for both stand-alone and combined features. Phonetic features have better capability of modelling vowels where the place of articulation is high whilst the vowels with the low place of articulation are better modelled by FBE features. Attribute features combined with acoustic features improve the articulatory inversion performance and will be helpful for transfer learning in case of new languages. Future works will focus on the jointly training of speakers and try building up a speaker independent framework by using linguistic features as the initial estimates.

6. Acknowledgements

This work has been supported by the Research Council of Norway through the project AULUS, and by NTNU through the project ArtiFutt.

7. References

[1] Z.-H. Ling, K. Richmond, and J. Yamagishi, "Articulatory control of hmm-based parametric speech synthesis using feature-space-switched multiple regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 207–219, 2013.

[2] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[3] P. K. Ghosh and S. Narayanan, "Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. EL251–EL257, 2011.

[4] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "Integrating articulatory data in deep neural network-based acoustic modeling," *Computer Speech & Language*, vol. 36, pp. 173–195, 2016.

[5] P. Badin, A. B. Youssef, G. Bailly, F. Elisei, and T. Hueber, "Visual articulatory feedback for phonetic correction in second language learning," in *Second Language Studies: Acquisition, Learning, Education and Technology*, 2010.

[6] T. . B. P. . B. G. Youssef, Atef Ben / Hueber, "Toward a multi-speaker visual articulatory feedback system," in *Proc. Interspeech 2011*, 2011, pp. 589–592.

[7] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatorytoacoustic transformation in the vocal tract by a computersorting technique," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1535–1555, 1978.

[8] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[9] L. Zhang and S. Renals, "Acoustic-articulatory modeling with the trajectory HMM," *IEEE Signal Processing Letters*, vol. 15, pp. 245–248, 2008.

[10] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Ninth International Conference on Spoken Language Processing*, 2006.

[11] B. Uria, I. Murray, S. Renals, and K. Richmond, "Deep architectures for articulatory inversion," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[12] P. L. Tobing, H. Kameoka, and T. Toda, "Deep acoustic-to-articulatory inversion mapping with latent trajectory modeling," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec 2017, pp. 1274–1277.

[13] N. Seneviratne, G. Sivaraman, V. Mitra, and C. Espy-Wilson, "Noise robust acoustic to articulatory speech inversion," in *Proc. Interspeech 2018*, 2018, pp. 3137–3141. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1509

[14] P. Liu, Q. Yu, Z. Wu, S. Kang, H. Meng, and L. Cai, "A deep recurrent approach for acoustic-to-articulatory inversion," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4450–4454.

[15] L. . C. Y. Zhu, Pengcheng / Xie, "Articulatory movement prediction using deep bidirectional long short-term memory based recurrent neural networks and word/phone embeddings," in *Proc. Interspeech 2015*, 2015, pp. 2192–2196.

[16] X. Xie, X. Liu, and L. Wang, "Deep neural network based acoustic-to-articulatory inversion using phone sequence information," in *Interspeech 2016*, 2016, pp. 1497–1501. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-659

[17] T. Biasutto-Lervat and S. Ouni, "Phoneme-to-articulatory mapping using bidirectional gated rnn," in *Proc. Interspeech 2018*, 2018, pp. 3112–3116. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1202

[18] M. Tiede, C. Y. Espy-Wilson, D. Goldenberg, V. Mitra, H. Nam, and G. Sivaraman, "Quantifying kinematic aspects of reduction in a contrasting rate production task," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3580–3580, 2017.

[19] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[20] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh annual conference of the international speech communication association*, 2010.

[21] A. M. Schäfer and H.-G. Zimmermann, "Recurrent neural networks are universal approximators," *International journal of neural systems*, vol. 17, no. 04, pp. 253–263, 2007.

[22] A. Wrench, "The MOCHA-TIMIT articulatory database," http://www.cstr.ed.ac.uk/artic/mocha.html, Queen Margaret Univ. College, Edinburgh, U.K., 1999.

[23] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the MNGU0 articulatory corpus," in *Proc. Interspeech*, Florence, Italy, August 2011, pp. 1505–1508.

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

[25] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveign, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds1speech files available. see http://www.elsevier.nl/locate/specom1," *Speech Communication*, vol. 27, no. 3, pp. 187 – 207, 1999.

[26] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.

[27] K. . Lee and H. . Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Nov 1989.

[28] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[29] S. M. Siniscalchi and C.-H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition," vol. 51, 2009, pp. 1139–1153.

[30] C.-H. Lee and S. M. Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification, and recognition," *Proceedings of the IEEE*, vol. 101(5), pp. 1089–1115, 2013.

[31] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[32] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.