Sequence-to-sequence articulatory inversion through time convolution of sub-band frequency signals

Abdolreza Sabzi Shahrebabaki1, Sabato Marco Siniscalchi2, Giampiero Salvi1,3, Torbjørn Svendsen1

1Department of Electronic Systems, NTNU
2Department of Computer Engineering, Kore University of Enna
3KTH Royal Institute of Technology, Dept. of Electrical Engineering and Computer Science

{abdolreza.sabzi, giampiero.salvi, torbjorn.svendsen}@ntnu.no, marco.siniscalchi@unikore.it

Abstract

We propose a new acoustic-to-articulatory inversion (AAI) sequence-to-sequence neural architecture, where spectral sub-bands are independently processed in time by 1-dimensional (1-D) convolutional filters of different sizes. The learned features maps are then combined and processed by a recurrent block with bi-directional long short-term memory (BLSTM) gates for preserving the smoothly varying nature of the articulatory trajectories. Our experimental evidence shows that, on a speaker dependent AAI task, in spite of the reduced number of parameters, our model demonstrates better root mean squared error (RMSE) and Pearson’s correlation coefficient (PCC) than a both a BLSTM model and an FC-BLSTM model where the first stages are fully connected layers. In particular, the average RMSE goes from 1.401 when feeding the filterbank features directly into the BLSTM, to 1.328 with the FC-BLSTM model, and to 1.216 with the proposed method. Similarly, the average PCC increases from 0.859 to 0.877, and 0.895, respectively. On a speaker independent AAI task, we show that our convolutional features outperform the original filterbank features, and can be combined with phonetic features bringing independent information to the solution of the problem. To the best of the authors’ knowledge, we report the best results on the given task and data.

Index Terms: Acoustic-to-articulatory inversion, deep learning, sequence-to-sequence neural models, 1-D convolution.

1. Introduction

The acoustic to articulatory inversion (AAI) problem is about estimating the vocal tract shape in the form of articulator positions based on the uttered speech. The actual articulatory positions can be obtained from speakers through different techniques, such as MRI [1], X-ray microbeam [2], and electromagnetic articulography (EMA) [3]. In recent years, AAI has attracted increasing attention because of its suitability in different applications, namely speech synthesis, second language learning [6][7], and automatic speech recognition (ASR) [8]. In a companion paper submitted to this conference [9], we show that AAI is beneficial for the continuous phone recognition task. Unfortunately, this inversion problem is highly non-linear and non-unique [10][11], which means that different articulator configurations can produce the same sound. In addition coarticulation [12], i.e. the impact of adjacent phonemes on the articulators’ movement, makes the AAI problem harder.

Different machine learning techniques and various input representations have been proposed to address the AAI task. For example, search of joint acoustic and articulatory space codebooks [13], Gaussian mixture models (GMMs) [14], hidden Markov models (HMMs) [7], mixture density networks (MDNs) [15], deep neural networks (DNNs) [16], and recurrent neural networks (RNNs) [17][18]. It is reported to obtain better accuracy than the DNN-based solution proposed in [16] exploiting an RNN-based AAI approach [19]. This result was mainly due to the better capability at capturing temporal dynamics that the RNN has through its memory elements. Different acoustic representations, such as line spectral frequencies (LSF), Mel-frequency cepstral coefficients (MFCC) and filterbank energies (FBE) have also been employed as input of the AAI system [17][18].

Linguistic features have also been proven useful when used as stand-alone input features [20], or together with acoustic features [21][22]. Such linguistic features are for example: phonemic (PHN) and attribute (AF) features [19]. Those features can be estimated by using a phone recognizer or a forced phone aligner [18] whenever we have access to the transcription of the uttered speech, e.g. in language learning or speech synthesis applications.

Although LSTM-based RNNs are promising for tackling the AAI task, the AAI accuracy could be further improved by exploiting ad-hoc connectionist components that can help remove redundant information in the speech signal. In fact, there exist many sources of information in the speech acoustic signal, which are not all relevant for the target task. Deep learning methods can reduce the effects of that irrelevant information leveraging upon large amounts of training material and parameters; however, lack of ad-hoc corpora providing an appropriate amount of data is a peculiar curse of the AAI problem. Therefore, the use of connectionist blocks that can better exploit the intrinsic characteristic of the speech signal could be beneficial to improve AAI results. We know that the vocal tract movements encode the linguistic message, and the speech signal reflects these movements. Non-linguistic components in the speech signal have a rate of change that lies outside the typical rate of the change of the vocal tract. 1-D convolutional connectionist components can intrinsically be more robust to the speech variability by suppressing spectral components that change more slowly or quickly than the typical range of change of the speech signal. Furthermore, convolutional components offer the advantage to reduce the amount of connectionist parameters with respect to fully connected components, which implies that a smaller amount of data can be sufficient to learn the 1-D convolutional filters. Bi-directional recurrent components with LSTM gates can instead be used to capture temporal relationships and better estimate the articulatory parameters.

In this work, we thus propose 1-D convolutional layers prior to the BLSTM-based recurrent blocks to project FBE features to a new space to deal with lack of data and temporal variability. Moreover, the scarcity of relevant speaker specific data makes build-
ing speaker dependent (SD) systems challenging, and the performance typically drops significantly when moving to speaker independent (SI) systems, where data from the test speaker is not used in the training stage of the neural architecture. To overcome the drop in performance caused by data scarcity in the SI configuration, we proposed to combine the feature maps from 1-D convolutions and phonetic features.

The rest of the paper is organized as follows. We describe the proposed AAI approach in Section 2. The experimental setup is given in Section 3 where the “Haskins Production Rate Comparison database” (HPRC) [23], input features and output parameters, and network parameters are presented. The experimental results are discussed in Section 4. Section 5 concludes our work.

2. Proposed method

In this work, we propose a new AAI approach, where spectral sub-bands are independently processed in time by 1-D convolutional filters of different sizes. The learned features maps are then combined and processed by an RNN with BLSTM gates for preserving the smoothly varying nature of the articulatory trajectories. We use mel filterbank energies as features in the present work to have a higher resolution for low frequency bands.

1-D convolutional layers are mostly known as the feature extraction layers from sequences and widely used in many speech applications, e.g. ASR [24, 25], speech synthesis [26], and machine translation [27]. This is the first time, to the best of the authors’ knowledge, that 1-D convolutional layers on the input sequence (acoustic space) and output sequence (articulatory space) are combined and processed by an RNN with BLSTM gates for preserving the smoothly varying nature of the articulatory trajectories. Further analysis with regards to the extracted feature maps and their representation is presented in Section 5.

For the feature extraction layers from sequences and widely used in many speech applications, we use mel filterbank energies as features in the present work to have a higher resolution for low frequency bands.

1-D convolutional layers are mostly known as the feature extraction layers from sequences and widely used in many speech applications, e.g. ASR [24, 25], speech synthesis [26], and machine translation [27]. This is the first time, to the best of the authors’ knowledge, that 1-D convolutional layers on the input sequence (acoustic space) and output sequence (articulatory space) are combined and processed by an RNN with BLSTM gates for preserving the smoothly varying nature of the articulatory trajectories. We use mel filterbank energies as features in the present work to have a higher resolution for low frequency bands.

The next convolutional layers tries to capture more temporal information and filter out undesired temporal variabilities. After those convolutional layers, two BLSTM layers are employed to capture dynamical information and estimate smoothly varying articulator trajectories. Further analysis with regards to the extracted feature maps and their representation is presented in Section 5.

\[ y_{i,j}^{m} = b_j + \sum_{k=1}^{L_{i-1}} F_i \ast y_{i-1,k}^{m}, \]

where, \( \ast \) shows the convolution operation of weights \( F_i \) in convolutional layer \( i \) with the feature maps \( y_{i-1,k}^{m} \) from the previous layer \( i-1 \). A bias \( b_j \) is added to the result of the convolution, to calculate the new feature map \( y_{i,j}^{m} \) for the \( j \)th channel feature map. Zero padding is used to guarantee that the input sequence (acoustic space) and output sequence (articulatory space) have the same length. The 1D-CNN layers are used and concatenated along the channel axis as depicted in Fig. 1.

The filter length is different in each of the CNN layers which provides more information about adjacent frames with different resolutions along the time axis. The first convolutional layer plays an important role by high-passing or low-passing different frequency bands. In our architecture, this layer has the goal of detecting the presence of specific features which are then used to derive phonetic features.

In our experiments, acoustic features are extracted from a downsampled waveform at 16 KHz using an analysis window of length 25ms with frame shift of 10ms, yielding a frame rate that matches the EMA recordings. Acoustic features are calculated from 40 filters, which are linearly spaced on the Mel-scale frequency axis. Energies in the overlapping frequency bands are called filterbank energy (FBE) features. Phonetic (PHN) features are extracted by the Penn phonetics lab forced aligner [29]. Each PHN feature is represented as one-hot 39 dimensional vector [18], and the attribute features (AF) are directly mapped from PHN features as in [18].

3.3. Neural parameters & settings

We compare three different neural architectures. In the first and most simple configuration, referred to as BLSTM, the unprocessed filterbank energies are directly fed at the input of the neural architecture, which is BLSTM-based RNN. Two fully connected layers are introduced between the FBEs and the BLSTM-based RNN in the second configuration, referred to as FC-BLSTM. The third configuration, 1D-CNN-BLSTM, is our proposal, and 1-D convolutional filters are employed between the feature maps and the input sequence.

Figure 1: Architecture of our proposed AAI method.
the FBEs and the BLSTM-based RNN. In all cases 2 BLSTM layers with 128 cells for each of the forward and backward layers are used. Sigmoid and tanh activation functions are used for the recurrent layers. The output layer has 12 nodes, corresponding to the EMA dimension with linear activation function. In FC-BLSTM the first two layers are fully connected with 512 nodes with ReLU activation functions. In the 1D-CNN-BLSTM, 5 convolutional layers are used for feature extraction with the filter size of [1, 3, 5, 7, 9], respectively for each layer with ReLU non-linearity. The channel number for each of the convolutional layers are kept the same as \( L_i = 128 \). A batch size of 5 is used.

The experimental material is chosen from the subsets “N1” and “N2”, which have the normal speaking rate. The training data consist of 576 sentences, validation and test data each contain 72 sentences. The data splitting for the HPRC database is as in [30]. Experiments were performed in an utterance by utterance fashion, which requires that all of the utterances are zero padded to 4 sec in the feature domain for ease of training implementation. The same strategy was applied to mean normalized EMA utterances in order to obtain 4 sec duration. The Adam optimizer [31] is chosen for training the network. Keras with TensorFlow backend [33] were used to train all of the neural networks. An early stopping patience of 10 iterations has been employed by checking the validation loss function to prevent over-fitting to the training data.

3.4. Performance measures

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

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y(i) - \hat{y}(i))^2},
\]

\[
PCC = \frac{\sum_{i=1}^{N} (y(i) - \bar{y})(\hat{y}(i) - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N} (y(i) - \bar{y})^2 \sum_{i=1}^{N} (\hat{y}(i) - \bar{\hat{y}})^2}},
\]

where \( y(i) \) and \( \hat{y}(i) \) are the ground-truth and estimated EMA values of the \( i^{th} \) frame, respectively; \( \bar{y} \) and \( \bar{\hat{y}} \) are mean values of \( y(i) \) and \( \hat{y}(i) \).

4. Experimental Results

In this section, we compare and contrast the three architectures described in Section 3.3. We also present additional analysis to gain a better understanding of the proposed approach.

4.1. Performance evaluation for acoustic features

For each of the three methods, we conduct 20 simulations in order to eliminate the effect of random initialization of the network parameters. Table 1 shows RMSE values, and we can observe that the proposed method outperforms both baseline approaches by almost 0.1 and 0.2 mm RMSE. A t-test shows that the reduction in RMSE with respect to both baselines is significant with p-values less than 0.05. The proposed method outperforms FC-BLSTM with lower number of parameters, as it can be observed by comparing the number of parameters reported in the table. Finally, PCC scores, related to the similarity between trajectories, are given in Table 2 and show a similar trend to that observed for RMSE.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Neural Architecture &amp; No.Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>BLSTM 1D-CNN-BLSTM</td>
</tr>
<tr>
<td>F2</td>
<td>BLSTM 1D-CNN-BLSTM</td>
</tr>
<tr>
<td>F3</td>
<td>BLSTM 1D-CNN-BLSTM</td>
</tr>
<tr>
<td>F4</td>
<td>BLSTM 1D-CNN-BLSTM</td>
</tr>
<tr>
<td>M1</td>
<td>BLSTM 1D-CNN-BLSTM</td>
</tr>
<tr>
<td>M2</td>
<td>BLSTM 1D-CNN-BLSTM</td>
</tr>
<tr>
<td>M3</td>
<td>BLSTM 1D-CNN-BLSTM</td>
</tr>
<tr>
<td>M4</td>
<td>BLSTM 1D-CNN-BLSTM</td>
</tr>
<tr>
<td>Avg.</td>
<td>BLSTM 1D-CNN-BLSTM</td>
</tr>
</tbody>
</table>

Table 1: RMSE for various baselines and proposed method for different speakers for AAI system.

4.2. Feature extraction layers analysis

As we discussed in Section 2, 1D-CNN extract new features from FBEs. These feature maps are weighted sums of sub-band signals which have been processed by filters with different frequency responses. Fig. 3 shows an example of FBEs, and network activations through the 1D-CNN model. We can see some channel activations match phonemic segments in the first layer. Going to the next layers, the filter outputs become sparser and activations become more intense within the phoneme boundaries. For justifying our claim about channel output activations during the phonemic segments, we picked some channels output from the first layer by using correlation analysis with PHN and AF features as the reference patterns. This analysis provided a better insight for choosing the corresponding filter outputs with regards to PHN and AF features with higher correlation. As an example, we have chosen attribute fricative and phoneme /s/ which are depicted in Fig. 4. The corresponding filters’ output which are chosen after doing correlation analysis are depicted in Fig. 4. We can see that these filters outputs have high energies when the chosen attribute and phoneme are active. Therefore, we can say these 1D-CNN layers are extracting the linguistic information from FBEs. This is inline with our expectation of sensing the significant energy changes at the phone transition. Furthermore, we can see for the second CNN layer compared to the first CNN layer, we have less activation outside the ground truth activation times of the chosen attribute and phoneme. By comparing the results for the SD experiment using i) the proposed architecture, and ii) the BLSTM model that uses PHN features along with FBEs, from Fig. 5 it can be observed, the proposed architecture’s better performance could be explained by its inherent capability of 1D-CNN layers at ex-
trating speaker dependent information not available in one-hot encoded PHN features.

Figure 3: AF feature for fricative and PHN features for phoneme /ʃ/ ( ) and channel output from the 1st 1D-CNN layer ( ) and 2nd 1D-CNN layer ( ).

4.3. Speaker independent analysis

For evaluating the proposed method for SI training, we adopt a leave-one-out strategy, where each speaker is in turn considered as the testing speaker, and the rest of speakers are used in training. For articulatory data, we use TV trajectories as targets, and FBE and PHN features are fed at the input of the neural architectures. We used PCC as the performance measure, because of its intrinsic normalized nature that makes it less dependent on the differences between speakers’ anatomy, and range of movements. We can observe from Fig. 4 that 1D-CNN improves the performance of both SD and SI configurations. Moreover, by comparing the performances of 1D-CNN with FBE, PHN and their combination, we can observe 1D-CNN has extracted more speaker dependent information while it is less speaker independent compared to FBEs. Using processed FBE features with 1D-CNN filters together with PHN features enhances the system performance in both SD and SI.

5. Conclusion

In this paper, we address the problem of articulatory inversion by employing 1D-CNNs as preprocessing layers to BLSTM layers. We show that this architecture improves the performance for the SD AAI task compared both to a BLSTM network alone, but also to BLSTM whose input is obtained with fully connected layers with a larger number of parameters. We also show that the representations obtained by the 1D-CNNs can be combined with phonetic features to improve performance both for SD and SI systems. The best result from only acoustic features for SD AAI of TV trajectories is PCC=0.895 and by considering phonetic features is PCC=0.901. As a comparison the SD results obtained by [30] on the same data set but with another architecture, is PCC=0.826. Our best results from only acoustic features for SI AAI of TV trajectories is PCC=0.755, by considering phonetic features with the proposed architecture, we reached averaged PCC equals to 0.810 for SI system. For the future works, we will focus on language learning and miss pronunciation detection by employing AAI systems while we have the transcription in this application.

6. Acknowledgements

This work has been supported by the Research Council of Norway through the project AULUS, and by NTNU through the project ArtFutt. The second author is supported by the PRIN 2007 project nr. JNKCYZ_002.
7. References


