INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Speech rate task-specific representation learning from acoustic-articulatory data

Renuka Mannem[1], *Hima Jyothi R*[2], *Aravind Illa*[1], *Prasanta Kumar Ghosh*[1]

[1]Electrical Engineering, Indian Institute of Science, Bangalore-560012, India.
[2]Electronics and communications Engineering, Rajiv Gandhi University of Knowledge Technologies, Kadapa-516330, India.

mannemrenuka@iisc.ac.in, himajyothi802@gmail.com, aravindece77@gmail.com, prasantg@iisc.ac.in.

Abstract

In this work, speech rate is estimated using the task-specific representations which are learned from the acoustic-articulatory data, in contrast to generic representations which may not be optimal for the speech rate estimation. 1-D convolutional filters are used to learn speech rate specific acoustic representations from the raw speech. A convolutional dense neural network (CDNN) is used to estimate the speech rate from the learned representations. In practice, articulatory data is not directly available; thus, we use Acoustic-to-Articulatory Inversion (AAI) to derive the articulatory representations from acoustics. However, these pseudo-articulatory representations are also generic and not optimized for any task. To learn the speech-rate specific pseudo-articulatory representations, we propose a joint training of BLSTM-based AAI and CDNN using a weighted loss function that considers the losses corresponding to speech rate estimation and articulatory prediction. The proposed model yields an improvement in speech rate estimation by $\sim$18.5% in terms of pearson correlation coefficient (CC) compared to the baseline CDNN model with generic articulatory representations as inputs. To utilize complementary information from articulatory features, we further perform experiments by concatenating task-specific acoustic and pseudo-articulatory representations, which yield an improvement in CC by $\sim$2.5% compared to the baseline CDNN model.

Index Terms: speech rate estimation, task-specific representation learning, acoustic-to-articulatory inversion.

1. Introduction

Speech rate is defined as the number of speech units per second in a given speech recording. In our work, syllables are considered as speech units similar to the prior research works [1, 2]. Speech rate estimation is very important as it is used in many speech related applications [1, 3, 4, 5, 6, 7, 8, 9, 10]. Various techniques have been proposed in the literature to estimate the speech rate. For example, several approaches [6, 7, 11, 12] used hidden Markov model (HMM) to estimate the speech rate. The HMM-based methods are not robust to noise and they require a reference transcription which may not be available always [2]. Thus, typically, the speech rate is estimated using only acoustic features without using reference transcription. For example, the approaches presented in [13, 14, 15, 16, 17, 18] used the acoustic features for speech rate estimation. These approaches use Gaussian mixture model [16], intensity-based envelope [15], rhythm guided peak counting method [13], smoothed loudness contour [14] and convex weighting criterion [17] for accurate speech rate estimation. Another set of approaches [2, 18] used a temporal correlation and selected sub-band correlation (TCSSBC) based feature contour which involves peak detection with smoothing and thresholding operations. The TCSSBC method is found to be better than the above mentioned approaches. Likewise, many works have been presented in the literature for accurate speech rate estimation using acoustic representations alone. However, the prior works on speech rate estimation did not utilize the articulatory representations, although the motion of the speech articulators such as upper lip, lower lip, tongue, jaw, velum directly encodes the speech rate [19, 20, 21, 22].

In [23], the authors proposed a convolutional dense neural network (CDNN)-based speech rate estimation technique using acoustic-articulatory data. However, direct articulatory measurements may not be available in the test case unlike acoustic signal. Thus, an acoustic-to-articulatory inversion (AAI) model [24] is typically learned for this purpose. In [23], a Bidirectional Long Short Term Memory (BLSTM) network-based AAI model is trained using the input acoustic features and output articulatory features. The predicted articulatory movements from AAI are considered as pesudo ariculatory representations which are used as input to CDNN to estimate the speech rate. The CDNN-based approach has been shown to perform better than the TCSSBC approach. However, both TCSSBC and CDNN-based approaches use generic representations such as sub-band energies and Mel Frequency Cepstral Coefficients (MFCCs) respectively. The pseudo-articulatory representations are also generic as they are derived independent of the speech rate estimation task. The generic acoustic and pseudo-articulatory representations may not be optimal for all the speech tasks. Thus, both acoustic and pseudo-articulatory representations need to be learned in a task-specific manner. Unlike the generic representations, the task-specific representations are learned during the optimization of the models that are used to perform the considered speech task. Thus, using task-specific acoustic representations may help in achieving better performance in the respective speech task. In [25, 26, 27], the task-specific acoustic representations are learned from raw speech waveform using one-dimensional convolutional and max-pooling layers (CONV1D). In [25], the cascaded CONV1D and CDNN are jointly optimized to learn the task-specific representations from raw speech waveform for accurate speech rate estimation. However, these approaches [25, 26, 27] have been proposed only for task-specific acoustic representation learning and does not involve articulatory representations. In this work, we propose a joint training approach to learn the task-specific pseudo-articulatory representations. We also learn the task-specific acoustic representations using the CONV1D-based approach proposed in

http://dx.doi.org/10.21437/Interspeech.2020-2259

[25]. In [23], two CDNN models are trained separately using the generic acoustic and articulatory representations. We hypothesize that using the concatenated acoustic and articulatory representations helps in better speech rate estimation as the articulatory representations contain information complementary to the acoustics [28, 29, 30]. Thus, in this work, we use the concatenated task-specific acoustic and articulatory representations as input to the CDNN for accurate speech rate estimation. We further provide an analysis comparing the learned task-specific acoustic and articulatory representations with generic acoustic and articulatory representations.

2. Dataset

In this paper, IEEE-EMA [31] and TIMIT [32] corpora are used for experiments. IEEE-EMA corpus is used to learn the articulatory representations from acoustics. TIMIT corpus is used to estimate speech rate using generic and task-specific representations. IEEE-EMA corpus contains simultaneously recorded speech and electro-magnetic articulometry (EMA) data for 720 phonetically balanced sentences from 8 speakers (4 male and 4 female) at multiple speaking rates [31]. The speech and EMA data are acquired at 44.1 kHz and 100 Hz sampling frequencies respectively. The EMA readings are obtained from 8 sensors placed on different articulators, namely, tongue rear (TR), tongue blade (TB), tongue tip (TT), upper lip (UL), lower lip (LL), mouth left (ML), lower jaw (JAW), and jaw left (JAWL). Each EMA reading has X, Y and Z coordinates which measure the movements in horizontal, lateral and vertical directions respectively in three dimensional space. In this work, we consider the horizontal (X) and vertical (Z) movements in the midsaggital plane forming a 24 dimensional articulatory feature vector comprising six EMA points' (TR, TB, TT, UL, LL, JAW) X and Z coordinates ($6\times2=12$) and their velocity ($6\times1=6$) and acceleration ($6\times1=6$) components [24]. From acoustics, the MFCC features are obtained using HTK toolkit [33]. For each sentence, the 39-dimensional MFCC features are computed using a window length of 20 msec with a shift of 10 msec. Since, the MFCCs are obtained at a frame rate of 100 Hz and the articulatory features are extracted at a sampling rate of 100 Hz, there is one-to-one correspondence between them. Thus, for a given sentence, the MFCC features have a dimension of $M \times 39$ and the corresponding articulatory features have a dimension of $M \times 24$ where M is the number of frames. Each subject on an average provides $\sim$1607 synchronous acoustic and articulatory movement recordings. In our work, we consider the entire IEEE-EMA corpus. TIMIT corpus [32] contains 630 subjects with 8 major dialects. For each subject, 10 sentences' recordings are available. Each sentence is recorded at 16 kHz sampling frequency. For our work, we consider 8 sentences for each subject, excluding 'sa1' and 'sa2' sentences resulting in a total of 5040 recordings. In IEEE-EMA and TIMIT corpora, the phonetic transcriptions for all the speech recordings are available, these are used to obtain the ground truth speech rate. The ground truth speech rate is calculated as the number of vowels (as we consider the vowels as syllables) divided by the total duration of the speech chunk. In this work, the speech rate is estimated for a fixed length of one-second duration chunks which ranges from 2 to 8 vowels per second.

3. Methodology

Figure 1 illustrates the steps followed in the proposed approach, which involves three models to learn acoustic representation (CONV1D), pseudo-articulatory respresentations (AAI) and to estimate speech rate (CDNN). We first present a brief review of each model followed by the proposed approach. For speech rate estimation, we use the CDNN model which has similar architecture as presented in [23]. Speech rate estimation is formulated as a regression problem; hence, mean squared error (MSE) loss is optimized to train the CDNN. The MSE loss between the estimated and ground truth speech rates for a batch of speech chunks is denoted as $\mathcal{L}_{SR}$. The generic pseudo-articulatory representations are learned using a BLSTM-based AAI which is proposed by Aravind et. al [24]. In [24], the BLSTM network has been shown to overcome the problems of capturing context and smoothing techniques and achieves the state-of-the-art AAI performance. As explained in section 2, we use IEEE-EMA corpus to train the BLSTM-based AAI model with three layers (each layer has 256 nodes) to estimate the 24 dimensional EMA points from 39 MFCCs. As the speech rate is estimated for one-second duration chunks, the articulatory points are also estimated for one-second duration chunks. Thus, the input and output of AAI have dimensions of 100×39 and 100×24, respectively. The articulatory representation estimation is formulated as a regression problem; hence, MSE loss is optimized to train BLSTM-based AAI model. The MSE loss between the estimated and ground truth articulatory representations for a batch of speech chunks is denoted as $\mathcal{L}_{EMA}$.

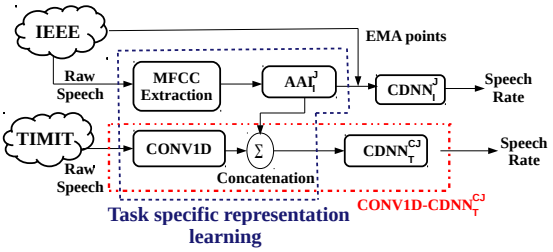

Figure 1: *Illustration of steps in the proposed speech rate estimation approach using task-specific representation learning.*

In this work, we propose a joint training approach using a weighted loss function to derive the speech rate-specific articulatory representations from acoustics. The weighted loss function to jointly train the cascaded BLSTM-based AAI and CDNN is defined as: $\mathcal{L}_{total}=w\times\mathcal{L}_{EMA} + (1-w)\times\mathcal{L}_{SR}$, where $w \in \{0.1, 0.2..., 0.9\}$. As shown in Figure 1, we use IEEE-EMA corpus which consists of parallel acoustic and articulatory data to jointly train the BLSTM-based AAI and CDNN to optimize $\mathcal{L}_{total}$. The cascaded trained model is denoted as $\text{AAI}_\text{I}^\text{J}$-$\text{CDNN}_\text{I}^\text{J}$ ('J ' and 'I ' indicate joint training and IEEE-EMA corpus respectively). Since, the $\text{AAI}_\text{I}^\text{J}$-$\text{CDNN}_\text{I}^\text{J}$ model is trained for both articulatory prediction and speech rate estimation, the $\text{AAI}_\text{I}^\text{J}$ model predicts the articulatory representations which are optimal for speech rate estimation. The $\text{AAI}_\text{I}^\text{J}$-$\text{CDNN}_\text{I}^\text{J}$ takes MFCCs as inputs using which the $\text{AAI}_\text{I}^\text{J}$ model predicts the speech rate-specific articulatory representations which are used as inputs to $\text{CDNN}_\text{I}^\text{J}$ for speech rate estimation. Likewise, for each w value, the $\text{AAI}_\text{I}^\text{J}$-$\text{CDNN}_\text{I}^\text{J}$ model is trained separately using IEEE-EMA training data. Among all these models, to obtain the pseudo-articulatory representations for TIMIT corpus, we select the $\text{AAI}_\text{I}^\text{J}$ model which provides the best performance (using $\text{AAI}_\text{I}^\text{J}$-$\text{CDNN}_\text{I}^\text{J}$) in speech rate estimation for IEEE-EMA test data.

On the other hand, the task-specific acoustic representations are learned from raw speech waveform using CONV1D filters following the approach presented in [25]. To perform the representation learning from raw speech waveform, we first downsample the speech from 16kHz to 8kHz. Then, the speech signal is converted into speech frames of short segments using a

Hamming window of length w_l=280 samples (35 msec) and shift w_s=80 samples (10 msec). The CONV1D block uses 1-D convolutional layer with n_f number of filters with a filter length of f_l=240 and a max-pooling layer with kernel size as $w_l - f_l + 1$=41. Thus, the input to the CONV1D has a dimension of $100 \times w_l$ and the corresponding output has a dimension of $100 \times n_f$ for a one-second duration chunk. The optimum value of n_f is decided based on the performance on the validation data. As shown in Figure 1, for TIMIT corpus, for a given raw speech input ($100 \times w_l$), the corresponding CONV1D block's output ($100 \times n_f$) and $\text{AAI}_\text{I}^\text{J}$ model's output discarding the velocity and acceleration components (100×12) are concatenated and fed to CDNN for speech rate estimation. During training, the CONV1D and CDNN are cascaded and jointly optimized for accurate speech rate estimation and the model is denoted as CONV1D-$\text{CDNN}_\text{T}^\text{CJ}$ ('T ' and 'C ' indicate TIMIT corpus and concatenation respectively). However, we do not update the $\text{AAI}_\text{I}^\text{J}$ model weights to preserve the articulatory information in its output. The CONV1D-$\text{CDNN}_\text{T}^\text{CJ}$ model uses concatenated task-specific acoustic and pseudo-articulatory representations as input and estimates the speech rate.

For TIMIT corpus, we also train the CDNN model using speech rate-specific pseudo-articulatory representations (obtained from $\text{AAI}_\text{I}^\text{J}$) as input for speech rate estimation which is denoted as $\text{AAI}_\text{I}^\text{J}$-$\text{CDNN}_\text{T}$. In addition to this, the CONV1D and CDNN models are cascaded and trained for accurate speech rate estimation as explained in [25] which is denoted as CONV1D-CDNN_T. Thus, in this case, the CDNN model uses task-specific acoustic representations. For baseline comparison, we also train the CDNN models using MFCCs (denoted as CDNN_T) and using pseudo-articulatory representations obtained from AAI model which is trained using IEEE-EMA data without having the knowledge of speech rate estimation (denoted as AAI_I-CDNN_T) as explained in [23]. Thus, CDNN_T and AAI_I-CDNN_T (which is nothing but $\text{AAI}_\text{I}^\text{J}$-$\text{CDNN}_\text{T}$ with w=1) models use generic acoustic and pseudo-articulatory representations, respectively, for speech rate estimation.

4. Experimental Setup

We estimate the speech rate for one-second duration speech chunks. Thus, each speech recording is divided into one-second duration chunks with an overlap of 0.5 seconds. IEEE-EMA corpus is used to train and validate AAI_I and $\text{AAI}_\text{I}^\text{J}$-$\text{CDNN}_\text{I}^\text{J}$ models. The IEEE-EMA corpus consists of 8 subjects. From each subject, we consider 80%, 10% and 10% of the data for train, validation and test sets respectively. The AAI_I and $\text{AAI}_\text{I}^\text{J}$-$\text{CDNN}_\text{I}^\text{J}$ models are trained for a maximum of 40 epochs with early stopping criterion based on the validation loss. TIMIT corpus is used to train and evaluate the CDNN models with different input representations. We consider two experimental conditions to evaluate the CDNN-based speech rate estimation: 1) Seen subject condition - train and test on same subjects and 2) Unseen subject condition - test subjects are different from those used in training. In both seen and unseen conditions, the CDNN is trained for 40 epochs with early stopping criterion based on the validation loss. The seen and unseen subject experiments are explained below:

Seen subject condition: In this case, the training and evaluation of all the models (CDNN_T,CONV1D-CDNN_T, AAI_I-CDNN_T, $\text{AAI}_\text{I}^\text{J}$-$\text{CDNN}_\text{T}$, CONV1D-$\text{CDNN}_\text{T}^\text{CJ}$) are done in a four-fold cross-validation setup using the TIMIT corpus. In TIMIT, 8 sentences are available for each speaker which are divided into four sets. We assign two sets for training, one set for validation and remaining one set for testing. Likewise, the sets are chosen in a round robin fashion forming a four-fold cross-validation setup. Each fold, on an average, consists of ∼11575, ∼5787 and ∼5787 one-second duration chunks in train, validation, and test sets respectively.

Unseen subject condition: In this case, similar to the seen subject condition, a four-fold cross-validation setup is used. We divide the TIMIT corpus into 4 sets with 157, 157, 157, 159 number of subjects. We consider two sets for training, one set for validation and remaining one set for testing. Likewise, the sets are chosen in a round robin fashion to form a four-fold cross-validation setup. Each fold, on an average, consists of ∼11575, ∼5787 and ∼5787 one-second duration chunks in train, validation, and test sets respectively.

Evaluation Metric: The performance of the proposed approach for speech rate estimation is evaluated based on the Pearson correlation coefficient (CC) between the ground truth and the estimated speech rates (denoted as CC_SR) [2, 18]. The AAI model performance is also evaluated using CC [24] (denoted as CC_EMA).

Table 1: CC_SR *and* CC_EMA *values using* $\text{AAI}_\text{I}^\text{J}$-$\text{CDNN}_\text{I}^\text{J}$ *on the IEEE-EMA test data.*

w	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
CC_SR	0.541	0.645	0.661	0.671	0.660	0.670	0.677	0.665	0.667
CC_EMA	0.007	0.564	0.656	0.674	0.684	0.689	0.689	0.691	0.692

Table 2: *Average (± standard deviation) of* CC_SR *value for TIMIT test data across the four folds in seen and unseen subject conditions.*

Method	Seen Subject Condition	Unseen Subject Condition	Input representations to CDNN
TCSSBC	0.57 ± 0.026	0.56 ± 0.019	Sub-band energies
CDNN_T	0.79 ± 0.016	0.80 ± 0.004	MFCCs
AAI_I-CDNN_T	0.65 ± 0.039	0.60 ± 0.110	Generic pseudo EMA points
$\text{AAI}_\text{I}^\text{J}$-$\text{CDNN}_\text{T}$	0.74 ± 0.047	0.74 ± 0.020	Task-specific pseudo EMA points
CONV1D-CDNN_T	0.80 ± 0.012	0.80 ± 0.015	Task-specific acoustic representations
CONV1D-$\text{CDNN}_\text{T}^\text{CJ}$	0.82 ± 0.007	0.82 ± 0.009	Concatenated representations

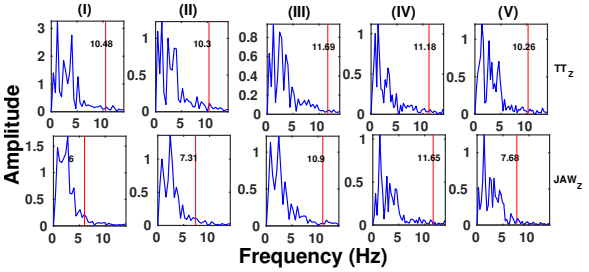

Figure 2: *Spectra of trajectories of JAW and TT in Z-direction (blue curve) (denoted as* JAW_Z, TT_Z *respectively) (I) Directly measured, taken from IEEE-EMA (II), (III) Estimated using* AAI_I *and* $\text{AAI}_\text{I}^\text{J}$ *for a sentence from IEEE-EMA data respectively. (IV), (V) Estimated using* AAI_I, $\text{AAI}_\text{I}^\text{J}$ *for a sentence from TIMIT data respectively. The vertical red line indicates the frequency (f_c) corresponding to the 99% of the energy of the entire trajectory.*

5. Results and Discussions

Table 1 shows the performance of $\text{AAI}_\text{I}^\text{J}$-$\text{CDNN}_\text{I}^\text{J}$ in terms of CC_SR and CC_EMA evaluated on IEEE-EMA test data for $w \in \{0.1, 0.2..., 0.9\}$ in the joint loss function ($\mathcal{L}_{total}$). It is observed that CC_EMA increases as the value of w increases since the contribution of $\mathcal{L}_{EMA}$ increases in $\mathcal{L}_{total}$. Thus, the model $\text{AAI}_\text{I}^\text{J}$-$\text{CDNN}_\text{T}^\text{J}$ training focuses more on accurate EMA points estimation than the accurate speech rate estimation. As w increases, CC_SR does not show consistent improvement. However, the highest CC_SR value is observed for w=0.7 and we select the corresponding trained AAI model ($\text{AAI}_\text{I}^\text{J}$) to estimate the speech rate-specific pseudo-articulatory representations for

TIMIT data. On the other hand, the $\mathrm{AAI_I}$ model provides the generic pseudo-articulatory representations. The $\mathrm{AAI_I}$ and $\mathrm{AAI_I^J}$ (for w=0.7) models provide $\mathrm{CC_{EMA}}$ values of 0.7310 and 0.6893 on the IEEE-EMA test data respectively. $\mathrm{CC_{EMA}}$ from $\mathrm{AAI_I^J}$ is less than $\mathrm{CC_{EMA}}$ from $\mathrm{AAI_I}$, as the $\mathrm{AAI_I^J}$ model is trained not only for accurate articulatory representation estimation but also for accurate speech rate estimation. We further examine the extent to which the outputs of $\mathrm{AAI_I}$ and $\mathrm{AAI_I^J}$ networks have characteristics similar to those of articulatory movements which are smoothly varying and low-pass in nature [34, 35]. Figure 2 illustrates the spectra of trajectories of JAW and TT in Z-direction (denoted as $\mathrm{JAW_Z}$, $\mathrm{TT_Z}$) (a) for a sentence from the IEEE-EMA (b), (c) estimated using $\mathrm{AAI_I}$ and $\mathrm{AAI_I^J}$,respectively, a sentence from the IEEE-EMA data (d),(e) estimated using $\mathrm{AAI_I}$, $\mathrm{AAI_I^J}$, respectively, for a sentence from the TIMIT data. The vertical red line indicates the frequency (f_c) corresponding to the 99% of the energy of the entire trajectory. It is observed that, all the estimated trajectories are low-pass in nature similar to a directly measured articulatory trajectory. The f_c values corresponding to the $\mathrm{AAI_I}$ and $\mathrm{AAI_I^J}$ models do not vary much from each other. Thus, the speech-rate specific articulatory representations preserve the original spectral characteristics of the articulators although they are optimized for the speech rate task.

In this work, we use n_f=32 CONV1D filters for both $\mathrm{CONV1D\text{-}CDNN_T}$ and $\mathrm{CONV1D\text{-}CDNN_T^{CJ}}$ based on the performance on the validation data. Table 2 shows the average ($\pm$ standard deviation) of $\mathrm{CC_{SR}}$ values for TIMIT test data across the four folds in seen and unseen subject conditions respectively for the baseline and proposed approaches. It is observed that, due to the supervised nature, the CDNN-based approaches perform better than the TCSSBC approach. In seen subject case, the task-specific acoustic representations provide better performance compared to MFCCs with a percentage improvement of 1.3% [25]. However, in unseen subject case, the learned representations are on par with MFCCs. The task-specific pseudo EMA points provide better performance compared to the generic pseudo EMA points with percentage improvements of 13.85% and 23.33% in seen and unseen subject conditions respectively. Thus, the task-specific pseudo-articulatory representations help in better speech rate estimation compared to the generic pseudo-articulatory representations. The $\mathrm{CONV1D\text{-}CDNN_T^{CJ}}$, which uses concatenated speech rate-specific acoustic and pseudo-articulatory representations, performs better than $\mathrm{CDNN_T}$, $\mathrm{CONV1D\text{-}CDNN_T}$, $\mathrm{AAI_I\text{-}CDNN_T}$, and $\mathrm{AAI_I^J\text{-}CDNN_T}$ with a relative improvement of 3.79%, 2.50%, 26.15%, and 10.81% respectively. Hence, concatenated acoustic and articulatory representations help in better speech rate estimation compared to using either of them alone. In [25], an analysis on the learned representations from CONV1D is done compared to MFCCs. It is interesting to see the variation in the frequency response of the learned 1-D convolutional filters when the articulatory representations are involved. For this, we observe the center frequencies of the filters used in computation of MFCCs, and those of the learned filters from $\mathrm{CONV1D\text{-}CDNN_T^{CJ}}$ and $\mathrm{CONV1D\text{-}CDNN_T}$ which are illustrated in Figure 3. The log-magnitude responses of the learned 32 1-D convolutional filters in $\mathrm{CONV1D\text{-}CDNN_T^{CJ}}$ and $\mathrm{CONV1D\text{-}CDNN_T}$ are illustrated in Figure 4. The x-axis and y-axis indicate the frequencies and the filter index in the sorted order, respectively. The color intensity variations represents the magnitude response of the filters.

From Figure 3 and 4, it is observed that the frequency responses of the CONV1D filters are low pass in nature and majority of the filters are centred below 1000 Hz unlike filter banks of MFCCs. For speech rate estimation, the primary focus is on identifying vowel nuclei regions [36]. The energy of a vowel typically lies in low frequency regions. This could be a reason why the CONV1D filters learn the speech rate-specific representations which lie in the low-frequency regions. Interestingly, incorporating articulatory representations further reduces the effect of high frequency components and emphasize more on low frequencies. From Figure 4, it is observed that the magnitude of the filters from $\mathrm{CONV1D\text{-}CDNN_T^{CJ}}$ and $\mathrm{CONV1D\text{-}CDNN_T}$ is high in low frequency regions. However, in the case of $\mathrm{CONV1D\text{-}CDNN_T}$, the magnitude of the side lobes is high. In contrast to this, in the case of $\mathrm{CONV1D\text{-}CDNN_T^{CJ}}$, the magnitude of the side lobes is attenuated in the high frequency regions. Thus, involving articulatory representations along with acoustic representations, helps in learning better task-specific representations.

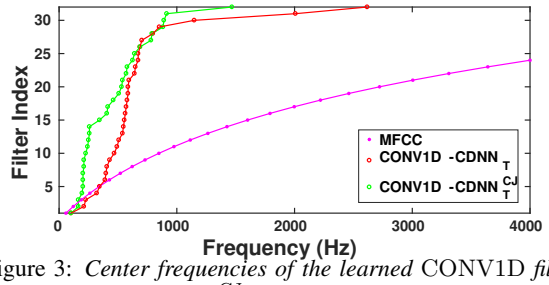

Figure 3: *Center frequencies of the learned* CONV1D *filters from* $\mathrm{CONV1D\text{-}CDNN_T^{CJ}}$ *and* $\mathrm{CONV1D\text{-}CDNN_T}$ *(with and without articulatory points) in comparison to those in MFCC.*

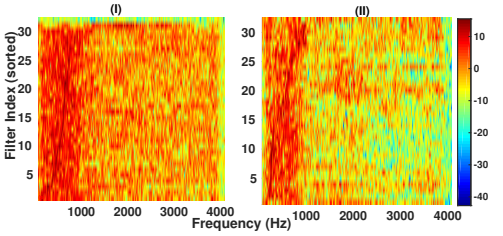

Figure 4: *Log-magnitude response of the learned filters in (I)* $\mathrm{CONV1D\text{-}CDNN_T}$ *(without articualtory representations) and (II)* $\mathrm{CONV1D\text{-}CDNN_T^{CJ}}$ *(with articulatory representations)*

6. Conclusion

In this work, we proposed a joint training approach to learn the task-specific pseudo-articulatory representations. We used the concatenated task-specific acoustic and articulatory representations to utilize the benefit from complementary information provided by articulatory representations compared to acoustics. From experiments in seen and unseen conditions, we observed that the task-specific representations provide better performance for speech rate estimation compared to the generic representations. The concatenated acoustic and articulatory representations have shown to provide better performance compared to using either of them alone. From the frequency response of the learned CONV1D filters, it is observed that the filters emphasize low-frequency regions indicating emphasis on the vowel regions. Involving articulatory representations further helps in suppressing the high frequency components which lead to even more accurate speech rate estimation. Our future work includes estimating the syllable boundaries using CONV1D output.

7. Acknowledgement

Authors thank the Department of Science and Technology (DST), Government of India for their support in this work.

8. References

[1] N. Morgan, E. Fosler-Lussier, and N. Mirghafori, "Speech recognition using on-line estimation of speaking rate," in *EUROSPEECH*, vol. 4, Jan 1997, pp. 2079–2082.

[2] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, Nov 2007.

[3] J. P. Campbell, "Speaker recognition," *Biometrics: Personal Identification in Networked Society*, vol. 479, pp. 165–189, Apr 1996.

[4] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. S. Narayanan, "An acoustic study of emotions expressed in speech," in *INTERSPEECH*, Oct 2004, pp. 2193–2196.

[5] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar 1998.

[6] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology." *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–99, Feb 2000.

[7] F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody annotation, modelling and evaluation," in *International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, Jun 2012, pp. 21–30.

[8] V. Dellwo, "Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence," *PhD Dissertation, Universität Bonn*, July 2010.

[9] J. Liss, L. White, S. L Mattys, K. Lansford, A. Lotto, S. M Spitzer, and J. Caviness, "Quantifying speech rhythm abnormalities in the dysarthrias," *Journal of speech, language, and hearing research (JSLHR)*, vol. 52, pp. 1334–52, Sept 2009.

[10] Y.-T. Wang, R. Kent, J. Duffy, and J. E Thomas, "Dysarthria associated with traumatic brain injury: Speaking rate and emphatic stress," *Journal of communication disorders*, vol. 38, pp. 231–60, May 2005.

[11] T. Cincarek, R. Gruhn, C. Hacker, E. Noeth, and S. Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-native's first language," *Computer Speech & Language (CSL)*, vol. 23, pp. 65–88, Jan 2009.

[12] J. Yuan and M. Liberman, "Robust speaking rate estimation using broad phonetic class recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2010, pp. 4222–4225.

[13] Y. Zhang and J. R. Glass, "Speech rhythm guided syllable nuclei detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 2009, pp. 3797–3800.

[14] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, May 1998, pp. 945–948.

[15] N. H. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, May 2009.

[16] R. Faltlhauser, T. Pfau, and G. Ruske, "On-line speaking rate estimation using Gaussian mixture models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (ICASSP)*, vol. 3, Jun 2000, pp. 1355–1358.

[17] Y. Jiao, V. Berisha, M. Tu, and J. Liss, "Convex weighting criteria for speaking rate estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1421–1430, Sep 2015.

[18] S. Narayanan and Dagen Wang, "Speech rate estimation via temporal correlation and selected sub-band correlation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Mar 2005, pp. 413–416.

[19] S. G. Adams, G. Weismer, and R. D. Kent, "Speaking rate and speech movement velocity profiles," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 1, pp. 41–54, Feb 1993.

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

[21] T. Gay, T. Ushijima, H. Hirose, and F. S. Cooper, "Effect of speaking rate on labial consonant-vowel articulation," *The Journal of the Acoustical Society of America*, vol. 55, no. 2, pp. 385–385, Jan 1974.

[22] O. Engstrand, "Articulatory correlates of stress and speaking rate in swedish VCV utterances," *The journal of the Acoustical society of America*, vol. 83, no. 5, pp. 1863–1875, May 1988.

[23] R. Mannem, J. Mallela, A. Illa, and P. K. Ghosh, "Acoustic and Articulatory Feature Based Speech Rate Estimation Using a Convolutional Dense Neural Network," in *INTERSPEECH*, Sep 2019, pp. 929–933.

[24] A. Illa and P. Kumar Ghosh, "Low resource acoustic-to-articulatory inversion using Bi-directional long short term memory," in *INTERSPEECH*, Sep 2018, pp. 3122–3126.

[25] R. Mannem, H. Jyothi, A. Illa, and P. K. Ghosh, "Speech rate estimation using representations learned from speech with convolutional neural network," in *International Conference on Signal Processing and Communication (SPCOM)*, 2020. [Online]. Available: https://tinyurl.com/ycwq3g8v

[26] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *INTERSPEECH*, Sep 2015, pp. 1–5.

[27] J. Millet and N. Zeghidour, "Learning to detect dysarthria from raw speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5831–5835.

[28] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *Fifth International Conference on Spoken Language Processing*, Oct 1998.

[29] G. Srinivasan, A. Illa, and P. K. Ghosh, "A study on robustness of articulatory features for automatic speech recognition of neutral and whispered speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5936–5940.

[30] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Conversational speech recognition using acoustic and articulatory input," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (ICASSP)*, vol. 3, Jan 2000, pp. 1435–1438.

[31] M. Tiede, C. 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, pp. 3580–3580, May 2017.

[32] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," vol. 93, p. 27403, Jan 1993.

[33] S. Young and S. Young, "The HTK Hidden Markov Model Toolkit: Design and philosophy," *Entropic Cambridge Research Laboratory Ltd*, vol. 2, pp. 2–44, Jan 1994.

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

[35] A. Illa and P. K. Ghosh, "The impact of speaking rate on acoustic-to-articulatory inversion," *Computer Speech & Language (CSL)*, vol. 59, pp. 75–90, Jan 2020.

[36] C. Yarra, O. D. Deshmukh, and P. K. Ghosh, "A mode-shape classification technique for robust speech rate estimation and syllable nuclei detection," *Speech Communication*, vol. 78, pp. 62–71, Apr 2016.