JOINT MODELING OF ARTICULATORY AND ACOUSTIC SPACES FOR CONTINUOUS SPEECH RECOGNITION TASKS

Vikramjit Mitra, Ganesh Sivaraman, Chris Bartels, [*]Hosung Nam, Wen Wang, [‡]Carol Espy-Wilson, Dimitra Vergyri, Horacio Franco

Speech Technology and Research Lab., SRI International, Menlo Park, CA, USA
[*]Korea University, Seoul, South Korea
[‡]University of Maryland, College Park, MD
vikramjit.mitra@sri.com

ABSTRACT

Articulatory information can effectively model variability in speech and can improve speech recognition performance under varying acoustic conditions. Learning speaker-independent articulatory models has always been challenging, as speaker-specific information in the articulatory and acoustic spaces increases the complexity of the speech-to-articulatory space inverse modeling, which is already an ill-posed problem due to its inherent nonlinearity and non-uniqueness. This paper investigates using deep neural networks (DNN) and convolutional neural networks (CNNs) for mapping speech data into its corresponding articulatory space. Our results indicate that the CNN models perform better than their DNN counterparts for speech inversion. In addition, we used the inverse models to generate articulatory trajectories from speech for three different standard speech recognition tasks. To effectively model the articulatory features' temporal modulations while retaining the acoustic features' spatiotemporal signatures, we explored a joint modeling strategy to simultaneously learn both the acoustic and articulatory spaces. The results from multiple speech recognition tasks indicate that articulatory features can improve recognition performance when the acoustic and articulatory spaces are jointly learned with one common objective function.

Index Terms— automatic speech recognition, articulatory trajectories, vocal tract variables, hybrid convolutional neural networks, time-frequency convolution, convolutional neural networks.

1. INTRODUCTION

Spontaneous speech typically includes significant variability, which is often difficult to model by automatic speech recognition (ASR) systems. Coarticulation and lenition are two sources of such variability, and speech-articulation modeling can help to account for this. Several studies [1][2][3][4] have demonstrated that speech-production knowledge (in the form of speech articulatory representations) improves ASR system performance by systematically accounting for variability such as coarticulation. Further studies [5][6][7] have demonstrated that articulatory representations provide ASR systems with some degree of noise robustness.

The mapping from acoustics to articulations is known to be highly non-linear and non-unique [8]. Studies have explored using DNNs [9][10][11] for learning the nonlinear inverse transform of acoustic waveforms to articulatory trajectories. Speaker variation adds complexity to the problem and makes speech-inversion even harder [12][13]. In this paper, we explore using CNNs and DNNs for acoustic to articulatory speech inversion. The articulatory-speech dataset used in this work is synthetically generated by employing the Task Dynamics model [14]. More details about the synthetic articulatory-speech dataset are given in sections 2 and 3.

Deep-learning techniques have become integral to current ASR systems. Convolutional neural networks (CNNs), are often found to outperform fully connected DNN architectures in ASR [15]. CNNs are noise robust [16][17] and are found to learn speaker-invariant data representations. In this paper, we present a parallel CNN architecture, where time-frequency convolution is performed on traditional gammatone-filterbank-energy-based acoustic features, and time-convolution is performed on the articulatory trajectories. Each of the parallel convolution layers is followed by a fully connected DNN, whose outputs are combined at the context-dependent (CD) state level, producing senone posteriors. The proposed hybrid CNN (HCNN) architecture learns an acoustic space and an articulatory space, and uses both to predict the CD states.

The word-recognition results from both the *Wall Street Journal* (WSJ1) data, noisy WSJ data (Aurora-4), and Switchboard (300 hour) speech recognition tasks indicate that using articulatory information is beneficial when used in addition to standard filterbank features, providing complementary information that in turn reduces the word error rates (WER) in different evaluation conditions. A detailed description of the proposed hybrid CNN architecture is given in section 4.

2. DATASETS

The articulatory dataset used to train the speech-inversion systems consists of synthetic speech with simultaneous tract variable (TV) trajectories. TVs (refer to [6] & [18] for more details) are continuous time functions that specify the shape of the vocal tract in terms of constriction degree and location of the constrictors. We used the Haskins Laboratories' Task Dynamic model (TADA) [14] along with HLsyn [19] to generate a synthetic English isolated word speech corpus along with TVs. TADA defines eight TVs altogether whose positional information is pictorially represented in Figure 1.

TADA, along with HLsyn, is an articulatory-model-based text-to-speech (TTS) synthesizer that, given text as input generates vocal tract constriction variables and corresponding synthetic speech. In this work, we used the CMU dictionary and selected 111,929 words, whose Arpabet pronunciations we then fed to TADA, which generated their corresponding TVs and synthetic speech. Each word from the CMU dictionary was separately fed to TADA four or five times. For each iteration, TADA randomly selected (a) between a male and a female speaker, whose mean pitch was randomly picked from a uniform distribution; (b) a different speaking rate (fast, normal, or slow); and (c) a different set of articulatory weights to introduce speaker-specific traits. This process enabled simulating a diverse set of speakers. Altogether 534,322 audio samples were generated (approximately 450 hours of speech), of which 88% of the data was used as the training set, 2% was used as the cross-validation set, and the remaining 10% was used as the test set. Note that TADA generated speech signals at a sampling rate of 8 kHz and TVs at a sampling rate of 200 Hz.

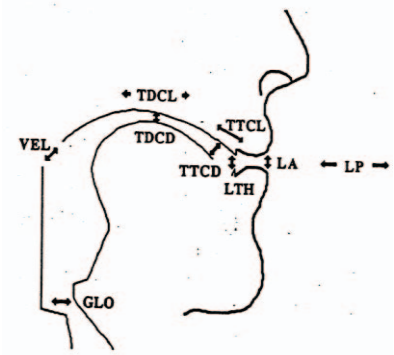

Figure 1. Eight tract variables from five constriction locations

For the speech recognition tasks presented in this paper, the DARPA WSJ1 CSR dataset was used. For training, a set of 35,990 speech utterances (77.8hrs) from the WSJ1 collection, having 284 speakers was used. For testing, the WSJ-eval94 dataset composed of 424 waveforms (0.8hrs) from 20 speakers was used. Note that for all the experiments reported here, speaker-level vocal-tract length normalization (VTLN) was not performed. We denote this dataset as WSJ1 in our experiments described in this paper.

For ASR under noisy and channel-degraded conditions, we used the Aurora-4 (noisy Wall Street Journal (WSJ0)) dataset [20]. Aurora-4 contains six additive noise versions with channel matched and mismatched conditions. It was created from the standard 5K WSJ0 database and has 7180 training utterances of approximately 15 hours duration and 330 test utterances. In all experiments, we have used 16 kHz sampled data for training and testing our speech recognition systems. Note that TADA along with HLsyn generates synthetic speech data sampled at 8 kHz, hence our speech inversion system can use a bandwidth of 0 to 4 kHz (corresponding to 8 kHz sampled data) to extract the TVs for speech recognition experiments. In Aurora-4, two training conditions were specified: (1) clean training, which is the full SI-84 WSJ training set without added noise, and (2) multi-condition training, with approximately half of the training data recorded by using one microphone, and the other half recorded by using a different microphone, with different types of added noise at different signal-to-noise ratios (SNRs). The Aurora-4 test data includes 14 test sets from two different channel conditions and six different added noises in addition to the clean condition. The SNR was randomly selected between 0 and 15 dB for different utterances. The six noise types used were: car, babble, restaurant, street, airport, and train station. The evaluation set consists of 5K words under two different channel conditions. The original audio data for test conditions 1–7 was recorded with a Sennheiser microphone, while test conditions 8–14 were recorded by using a second microphone randomly selected from a set of 18 different microphones (more details in [20]).

For the Switchboard (SWB-300) ASR task, the training data consisted of 262 hours of Switchboard data, which contained telephone-conversation speech between two strangers on a preassigned topic. The Hub5 2000 evaluation set was used to evaluate model performance, where 2.1 hours (21.4K words, 40 speakers) of Switchboard data and 1.6 hours (21.6K words, 40 speakers) of CallHome audio. The SWB-300 acoustic models were decoded with a 4-gram language model.

3. SPEECH INVERSION

The task of estimating the articulatory trajectories (in this case, the TVs) from the speech signal is commonly known as speech-to-articulatory inversion or simply speech-inversion. During speech-inversion, the acoustic features extracted from the speech signal are used to predict the articulatory trajectories, where the inverse mapping is learned by using a parallel corpus containing acoustic and articulatory pairs. The task of speech-inversion is a well known, ill-posed inverse transform problem, which suffers from both the non-linearity and non-unique nature of the inverse transform [21][7]. However, because tract variables are a relative measure (e.g., LA is a measure of the distance between the upper and lower lip, instead of an absolute flesh point location defined in Cartesian coordinates as in pellet data), they are found to

suffer less from non-linearity and non-uniqueness compared to traditional flesh-point measures such as pellet trajectories.

Based on previous observations [22], we explored using speech subband amplitude modulation features such as normalized modulation coefficients (NMCs) [23]. NMCs are noise-robust acoustic features obtained from tracking the amplitude modulations (AM) of gammatone-filtered subband speech signals in the time domain. The modulation information after root power compression is used to create a cepstral feature, where the first 13 discrete cosine transform (DCT) coefficients are retained. These cepstral NMCs are usually known as the NMC cepstral or (NMCC). In addition, we also explored using the above features without the DCT transform, which resulted in a 40-dimensional feature vector, and we denote them as NMCs. The features were Z-normalized before being used to train the DNN/CNN models. Further, the input features were contextualized by splicing multiple frames. In this work, we separately explored the optimal splicing window for the DNN and CNN models.

4. ACOUSTIC MODELS

We trained different acoustic models for the speech recognition tasks, where we explored traditional DNNs, CNNs, and time-frequency convolutional nets (TFCNNs) [24]. The acoustic models were trained with gammatone filterbank energies (GFBs). It was shown in [25] that CNNs give lower WERs compared to DNNs when using filterbank features and GFBs are better baseline features than mel-filterbank energies (MFBs). To generate the alignments necessary for training the CNN system, a GMM-HMM model was used to produce the senone labels. Altogether, the GMM-HMM system produced 3.2K context-dependent (CD) states for Aurora-4, 1.7K CD states for WSJ1, and 5.6K CD states for SWB-300. The input features to the acoustic models were formed by using a context window of 15 frames (7 frames on either side of the current frame), except for the TFCNN where 17 frames of feature information were used.

The acoustic models were trained by using cross-entropy (CE) on the alignments from the GMM-HMM system. For the CNN, 200 convolutional filters of size 8 were used in the convolutional layer, and the pooling size was set to 3 without overlap. The subsequent, fully connected network had four hidden layers, with 1024 nodes per hidden layer, and the output layer included as many nodes as the number of CD states for the given dataset. The networks were trained by using an initial four iterations with a constant learning rate of 0.008, followed by learning-rate halving based on cross-validation error decrease. Training stopped when no further significant reduction in cross-validation error was noted or when cross-validation error started to increase. Backpropagation was performed using stochastic gradient descent with a mini-batch of 256 training examples. For the DNN systems, we used five layers with 1024 neurons in each layer, with similar learning criteria as the CNNs. Note that for SWB-300, we evaluated a six hidden layer DNN acoustic model with 2048 neurons.

In this work, we investigated modified deep neural network architecture to jointly model the acoustic and the articulatory spaces, as shown in Figure 2. In this modified architecture, two parallel neural networks are trained simultaneously. These two parallel neural networks model two things: (1) learning the acoustic space from the GFB features and (2) learning the articulatory space from the TV trajectories.

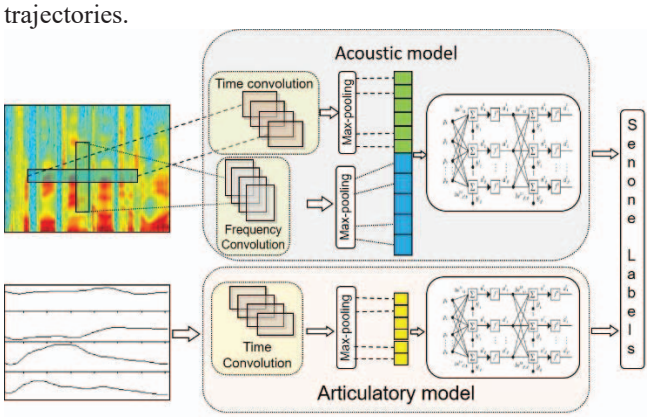

Figure 2. Hybrid convolutional neural network (HCNN). The top layer represents the acoustic model, whose input is filterbank features, and the bottom layer represents the articulatory model, whose input is TV trajectories.

The acoustic space is learned by using a time-frequency convolution layer, where two separate convolution filters operate on the input GFB features. These two convolution layers have the same parameter specification as that used in the TFCNNs. The articulatory space is learned by using a time-convolution layer that contains 75 filters, followed by max-pooling over five samples. Note that the cross-TV convolution operation may not produce any meaningful information, whereas time-convolution on the TVs can help in extract TV modulation-level information, which was the motivation behind selecting a time-convolution layer for learning the articulatory space. The fully connected DNN layers are different in size; we observed 800 neurons to be near optimal for learning the acoustic space, and 256 neurons to be near optimal for learning the articulatory space. Note that both the parallel networks are jointly trained.

5. RESULTS

5.1 SPEECH INVERSION

We explored using DNNs and CNNs for training speech-inversion models, where contextualized (spliced) acoustic features in the form of NMCs (for CNNs) and NMCCs (cepstral version of NMCs for DNNs) were used as input, and the TV trajectories were used as the targets. The network parameters and the splicing window were optimized by using a held-out development set. Pearson's product-moment correlation coefficient (r_{PPMC}) between the actual or ground truth and the estimated articulatory trajectories (averaged

across all the TVs) was used as the performance measure. A four-hidden-layer DNN containing 2048 neurons in each layer was found to be optimal. The convolution layer of the CNN had 200 filters, where max-pooling was performed over three samples. The CNN had three fully connected hidden-layers with 2048 neurons in each layer.

Table 1 presents the r_{PPMC} values from the test set obtained from the DNN and CNN systems, showing that both the systems exhibited similar performance across all eight TVs of the test set. This similar performance can be attributed to the diversity of the training data, which contained a large number of speaker configurations, consequently making the DNN system more robust to speaker variation.

Table 1: r_{PPMC} for each TV obtained from the best DNN and CNN systems.

	GLO	VEL	LA	LP	TTCD	TTCL	TBCD	TBCL
DNN	0.97	0.95	0.91	0.97	0.95	0.94	0.94	0.96
CNN	0.97	0.96	0.91	0.97	0.95	0.94	0.94	0.97

5.2 SPEECH RECOGNITION

We selected the DNN speech-inversion model and used that to estimate the articulatory trajectories (TVs) for all datasets used in our speech recognition studies. Note that the DNN speech-inversion model was trained with neither real conversational speech nor any noise/channel-degraded data. From our past experiments we have noticed that the eight TVs are insufficient by themselves for use as ASR features [26]; hence, we combined them with standard acoustic features. Our initial ASR experiments were on Aurora-4, where the baseline system is the TFCNN system reported in [24], using GFBs as acoustic features. Given that Aurora-4 has 14 different evaluation conditions depending on the noise conditions and microphone types, we used the standard partition of the evaluation set to report our results, where A and C represent the clean matched and mismatched channel conditions, respectively, and B and D represent the noisy matched and mismatched channel conditions, respectively. We investigated using bottleneck (BN) features by introducing a BN layer at the CNN's third layer and combining the resulting BN features with GFBs to train a HCNN system; the fourth row in Table 2 corresponds to the results obtained from the BN-CNN features.

We applied the HCNN architecture to the clean WSJ1 evaluation task and Table 3 presents the WERs from the different systems. For the SWB-300 baseline model, we trained 6-hidden layer DNN having 2048 neurons, with fMLLR transformed damped oscillator cepstral coefficient (DOCC) [27] features as input. The estimated TVs from the DNN speech inversion model was appended with the DOCC features and they were fMLLR transformed to train a 6-layered DNN with 2048 neurons. The results from sequence trained MMI models are given below in Table 4.

Table 2: WERs on multi-conditioned training task of Aurora-4 (16 kHz) from the baseline system using GFB feature and the HCNN using GFB + estimated TVs.

feature	model	A	B	C	D	avg.
GFB	TFCNN	3.1	5.7	6.1	14.6	9.4
GFB+TV_{DNN}	HCNN	3.3	5.7	5.5	14.2	9.2
GFB+TV_{CNN}	HCNN	3.0	5.7	5.5	14.2	9.1
GFB+$TV_{BN\text{-}CNN}$	HCNN	2.9	5.5	5.2	14.0	**8.9**

Table 3: WER from WSJ1 ASR experiments using baseline GFB features and GFB + estimated TVs.

Features	Models	WER
GFB	TFCNN	5.7
GFB+TV_{DNN}	HCNN	5.4
GFB+TV_{CNN}	HCNN	5.6
GFB+$TV_{BN\text{-}CNN}$	HCNN	5.3

Table 4: WER from SWB-300 ASR experiments using SWB part of the Hub5 eval data.

Features	Models	WER
DOCC	DNN-MMI	12.8
DOCC+TV_{DNN}	DNN-MMI	12.1

6. CONCLUSION

We investigated deep neural network (DNN)- and convolutional neural network (CNN)- based speech-inversion systems for estimating articulatory trajectories from the speech signal. We observed that additional hidden layers consisting of a large number of neurons and longer contextual windows gave better inversion performance. We presented a hybrid convolutional neural network (HCNN), in which two parallel layers are used to jointly model the acoustic and articulatory spaces, which were trained with one objective function. Speech recognition results on Aurora-4, WSJ1, and SWB-300 speech recognition tasks showed that the proposed architecture using articulatory features demonstrated reduction in WERs in each of the clean, noisy, and channel-mismatched conditions. For the Aurora-4 and WSJ1 ASR tasks, the best WERs from the HCNN system were found to be 8.9% and 5.3%, respectively, which, to the best of our knowledge, are state-of-the-art results for these datasets. On SWB-300, the proposed architecture gave a WER of 12.1%.

In the future, we will investigate merging the feature maps of the HCNN such that one DNN is trained for both the acoustic and articulatory information. We will also investigate the performance of the proposed architecture for languages other than English.

7. ACKNOWLEDGEMENT

This research was supported by NSF Grant # IIS-0964556, IIS-1162046, BCS-1435831 and IIS-1161962.

8. REFERENCES

[1] J. Frankel and S. King, "ASR - Articulatory Speech Recognition," in *Eurospeech*, 2001, pp. 599–602.

[2] L. Deng and D. X. Sun, "A statistical approach to automatic speech recognition using the atomic speech units constructed fro m overlapping articulatory features," *J. Acoust. Soc. Am.*, vol. 95, no. 5, p. 2702, 1994.

[3] V. Mitra, W. Wang, A. Stolcke, H. Nam, C. Richey, J. Yuan, and M. Liberman, "Articulatory trajectories for large-vocabulary speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7145–7149.

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

[5] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition.," *J. Acoust. Soc. Am.*, vol. 121, no. 2, pp. 723–42, Feb. 2007.

[6] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Articulatory Information for Noise Robust Speech Recognition," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 7, pp. 1913–1924, Sep. 2011.

[7] V. Mitra, "Articulatory Information For Robust Speech Recognition." Ph.D. dissertation, University of Maryland, College Park, 2010.

[8] C. Qin and M. Carreira-Perpiñán, "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping.," *INTERSPEECH*, 2007.

[9] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Retrieving Tract Variables From Acoustics: A Comparison of Different Machine Learning Strategies.," *IEEE J. Sel. Top. Signal Process.*, vol. 4, no. 6, pp. 1027–1045, Sep. 2010.

[10] B. Uria, S. Renals, and K. Richmond, "A Deep Neural Network for Acoustic-Articulatory Speech Inversion," 2011.

[11] C. Canevari, L. Badino, L. Fadiga, and G. Metta, "Relevance-weighted-reconstruction of articulatory features in deep-neural-network-based acoustic-to-articulatory mapping," in *Interspeech*, 2013.

[12] G. Sivaraman, V. Mitra, H. Nam, M. K. Tiede, and C. Y. Espy-Wilson, "Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion," in *Interspeech 2016*.

[13] P. K. Ghosh and S. S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4624–4627.

[14] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable Task Dynamics model in MATLAB," *J. Acoust. Soc. Am.*, vol. 115, no. 5, p. 2430, May 2004.

[15] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8614–8618.

[16] V. Mitra, W. Wang, H. Franco, Y. Lei, and C. Bartels, "Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions.," in *Interspeech*, 2014.

[17] V. Mitra, W. Wang, and H. Franco, "Deep convolutional nets and robust features for reverberation-robust speech recognition," in *2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings*, 2014, pp. 548–553.

[18] C. P. Browman and L. Goldstein, "Articulatory Phonology : An Overview *," *Phonetica*, vol. 49, pp. 155–180, 1992.

[19] H. M. Hanson and K. N. Stevens, "A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn.," *J. Acoust. Soc. Am.*, vol. 112, no. 3 Pt 1, pp. 1158–82, Sep. 2002.

[20] G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task," 2001.

[21] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," University of Edinburgh, UK., 2001.

[22] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, and E. Saltzman, "Articulatory features from deep neural networks and their role in speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3017–3021.

[23] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4117–4120.

[24] V. Mitra and H. Franco, "Time-frequency convolutional networks for robust speech recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 317–323.

[25] V. Mitra, W. Wang, H. Franco, Y. Lei, C. Bartels, and M. Graciarena, "Evaluating robust features on Deep Neural Networks for speech recognition in noisy and channel mismatched conditions," in *Interspeech*, 2014, pp. 895–899.

[26] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, and E. Saltzman, "Articulatory features from deep neural networks and their role in speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3017–3021.

[27] V. Mitra, H. Franco, and M. Graciarena, "Damped oscillator cepstral coefficients for robust speech recognition," in *Interspeech*, 2013, pp. 886–890.