Interspeech 2018
2-6 September 2018, Hyderabad

Towards Automatic Speech Identification from Vocal Tract Shape Dynamics in Real-time MRI

Pramit Saha[1*], *Praneeth Srungarapu*[1*], *Sidney Fels*[1]

[1] Department of Electrical and Computer Engineering, University of British Columbia

pramit@ece.ubc.ca, praneethsv@ece.ubc.ca, ssfels@ece.ubc.ca

Abstract

Vocal tract configurations play a vital role in generating distinguishable speech sounds, by modulating the airflow and creating different resonant cavities in speech production. They contain abundant information that can be utilized to better understand the underlying speech production mechanism. As a step towards automatic mapping of vocal tract shape geometry to acoustics, this paper employs effective video action recognition techniques, like Long-term Recurrent Convolutional Networks (LRCN) models, to identify different vowel-consonant-vowel (VCV) sequences from dynamic shaping of the vocal tract. Such a model typically combines a CNN based deep hierarchical visual feature extractor with Recurrent Networks, that ideally makes the network spatio-temporally deep enough to learn the sequential dynamics of a short video clip for video classification tasks. We use a database consisting of 2D real-time MRI of vocal tract shaping during VCV utterances by 17 speakers. The comparative performances of this class of algorithms under various parameter settings and for various classification tasks are discussed. Interestingly, the results show a marked difference in the model performance in the context of speech classification with respect to generic sequence or video classification tasks.

Index Terms: recurrent neural network, convolutional neural network, articulatory-to-acoustics mapping, vowel-consonant-vowel, vocal tract geometry.

1. Introduction

The vocal tract [1] is the most important component of speech production that can be divided into two distinctive parts - the *oral tract* (ranging from larynx to lips) and the *nasal passage*, coupled to the oral cavity through velum. During air expulsion, this tubular passageway modifies its position and shape (length and cross sections) in such a way that results in production of various sounds and their acoustic representations.

A long-standing issue in articulatory speech research concerns the estimation of vocal tract configuration and its mapping with the acoustic parameters [2]. This is extremely challenging, not only because the vocal tract is difficult to access, but also due to its complex biological structure and the rapid movement of its articulators. The speech production process is essentially non-stationary and generally, it is the rapid transition between different articulatory states that generates the speech sounds. Hence, extraction of the dynamic information of vocal tract geometry is crucial for identifying and synthesizing speech.

The relationship between the articulatory gestures and acoustic speech characteristic has been well explored in terms of either *articulatory-to-acoustic mapping* which deals with producing audio speech signal from vocal tract configurations or the more common *acoustic-to-articulatory inversion*, which aims at recovering the vocal tract shapes from the speech [3].

In this paper, our primary goal is to investigate the end-to-end mapping of the forward pathway, *i.e* learning to identify speech sounds corresponding to subject-invariant vocal tract geometry configurations. There are two main ways of achieving this: the *modeling* approach or the *corpus-based* approach [3]. Research in the former approach mostly focus on three basic steps - *articulatory model representation*, generally achieved by speaker-specific *geometric modeling* of articulator shapes and positions; speaker-invariant *statistical modeling* of standard articulator representations; or neuro-muscular controlled *biomechanical modeling* of articulator motions, followed by *acoustic transfer function computation* step, and time or frequency based *acoustic characterization*.

Conversely, the *corpus-based* methodology utilizes parallel articulatory-acoustic database consisting of speech sounds recorded with simultaneous articulator movements, through *code-book* or *statistical* (discriminative or generative) techniques. While the *code-book* approach is a direct way to retrieve the nearest articulatory configurations, corresponding to given *'query'* configuration from the database and produce an audio spectrum interpolated from available acoustic representations, the *statistical* model utilizes supervised machine learning techniques to learn the forward connection.

But, all the available methods rely solely on the articulatory time-series positional data from a few specified points on tongue, lip and jaw acquired using electromagnetic articulography, or sometimes combined with electropalatography and laryngograph. While it is considerably easier and computationally less expensive to work with a set of tracked input points at a handful of sensor locations within the anterior part of the tract, it misses significant articulatory information like velar and pharyngeal constrictions, and others, thereby neglecting the spatial information of the complex vocal tract geometry. Dynamic vocal tract-imaging data, on the other hand, effectively records the labial, lingual and jaw motion along with the articulation of the velum, pharynx and larynx, and structures like the upper palate, pharyngeal wall *etc* which cannot be captured by the above-mentioned techniques. Besides, such imaging data helps to comprehend the generation of coronal, pharyngeal and nasal segments and to visualize the inter-articulator coordination for generation of multi-gestural segments in speech. Hence, there is an inherent urge to explore the informations available from medical imaging modalities on the articulatory-acoustic forward pathway utilizing dynamic information of the entire upper airway.

However, it is extremely challenging and computationally expensive to achieve a direct end-to-end mapping of the entire vocal tract geometry acquired by real time imaging modalities

* indicates equal contribution

10.21437/Interspeech.2018-2537

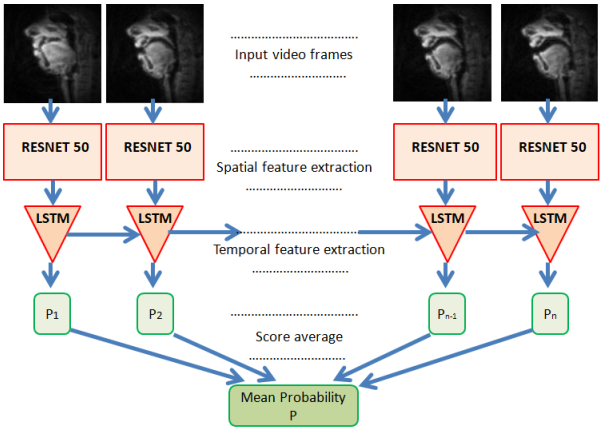

Figure 1: *Overview of the LRCN model*

to the corresponding speech sound output. Recently, real-time magnetic resonance imaging (rtMRI) has emerged as a powerful tool to acquire the complex spatio-temporal visual information about the dynamic shaping of the vocal tract during speech production [4]. This has opened a new pathway to investigate the underlying articulatory geometric representation of the speech signals in real-time, thereby developing a better understanding of the geometry-to-acoustic mapping. However, the poor spatio-temporal resolution as well as the presence of a substantial number of artifacts and noise and limited data make automatic extraction and interpretation of features an extremely difficult problem [5].

To the best of our knowledge, this work is the first discriminative approach that attempts to achieve a forward mapping of vocal tract shape dynamics from rtMRI to the speech output. In order to exploit the dynamics of open and closed vocal tract configurations along with the rapid articulator movements, we have selected 51 VCV transition tokens as the target classes. We employ a deep learning strategy to extract the per-frame spatial features and simultaneous inter-frame temporal features. The primary goal of this paper is to investigate the performance of a promising visual action detection and classification algorithm named Long term Recurrent Convolutional Neural Networks (LRCN) [6], when applied for speech identification tasks from rtMRI videos. Though the technique fails to provide satisfactory accuracy for the VCV sequence identification task, it shows promising results for classifying the individual vowels and acceptable performance for consonants. We discuss the issues that make the targeted speech identification task different from conventional video-classification [7] or action recognition tasks [8] and provide insights regarding possible ways of improvement.

2. Illustration of the Proposed Methodology

2.1. Long-term Recurrent Convolutional Networks model

2.1.1. Background

Among the most widely used algorithms for video based action recognition, 3D Convolutional Networks [9], Two stream Convolutional Networks [8] and Long-term Recurrent Convolutional Networks (LRCN) [6] deserve special mention in our context. This class of algorithms incorporate spatial and temporal feature extraction steps to capture complementary information from the individual still frames as well as between the frames, which is associated to our primary goal in this work. However, the two stream architecture involves CNN trained on multi-frame dense optical flow images for the temporal recognition stream. Since the dataset we are using [10], has a significant amount of non-stationary noise and artifacts, such algorithm gives more emphasis on the noise instead of the vocal tract shape change, which in turn leads to erroneous sound classification. On the other hand, though 3D ConvNets show faster performance on our dataset, the temporal pooling layers are too weak to extract long term temporal features spanning over a duration of 3 seconds. The end-to-end trainable LRCN networks combine CNN based deep visual feature extractor with LSTM based long-term temporal dynamics extractor and are shown to have sufficient recurrence of the latent variables over temporal domain. Thus, we utilize the LRCN model to investigate whether the speech recognition problem can be viewed similar to a video action classification/recognition task, with the vocal tract movement resembling the action and the VCV sequences as the final mapped output interpreted from the action. However, the performance of the other two is not significant enough in the current context and hence we decide to confine our discussion with LRCN.

2.1.2. Convolutional Neural Network (CNN)

CNN architectures [11] are composed of sequence of layers - Convolutional Layers (Learnable spatial filters), Pooling Layers (Downsamplers) and Fully Connected Layers (Dense layers with full connections), along with linear/non-linear activations. We use substantially deep residual learning framework, also known as ResNet [12], composed of series of Identity and Convolutional blocks in this work.

2.1.3. Recurrent Neural Network (RNN)

Recurrent neural Networks are capable of exhibiting dynamic temporal behavior [13]. LRCN network [6] uses RNN composed of LSTM units [14] which are capable of learning complex temporal dynamics by mapping the input vectors to the output vectors through a sequence of hidden steps defined by recurrence relations. They are effective in solving the exploding and vanishing gradient problems associated with conventional RNN networks [15]. The LSTM module utilized here [16], consists of a memory cell, an input gate, an output gate, a forget gate and an input modulation gate. Such networks are able to explore temporal behavior due to inter-layer directed units, as shown in Fig 1.

2.2. Vowel-Consonant-Vowel Identification through LRCN

We consider the VCV identification from the rtMRI as the '*sequential input to fixed output*' problem with videos of arbitrary length T as input and prediction of single labels corresponding to each video, as the output. The video is split into T individual frames or image sequences and fed to T convolutional network layer, each network layer implying a complete ResNet architecture and then connected through two fully connected layers

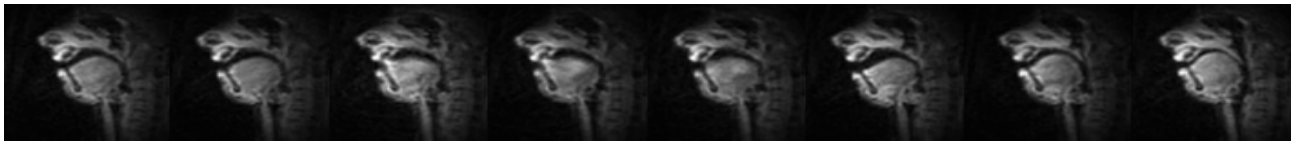

Figure 2: *Frames of rtMRI videos for speaker F1 producing [asa]. Time progresses from left to right.*

with 2048 and 1024 neurons with 0.5 drop-outs respectively to an N layered LSTM with M hidden nodes. Next, the LSTM predicts the probabilities of speech output class for each of the time frames and are averaged to yield the final class probability score across the entire sequence demonstrated in Fig 1.

3. Experiments and Results

3.1. Dataset preparation

We evaluate our architecture on the USC Speech and Vocal Tract Morphology MRI Database [10] which includes $2D$ real-time MRI of vocal tract shaping of 17 speakers (9 female and 8 male) along with simultaneous denoised audio recording. This database provides the resource to relate rtMRI data capturing dynamic vocal tract shapes and speech variability, thereby helping our attempt to explore the forward mapping pathway from imaging data. The imaging sequence has a *frame rate* of 23.18 frames/second, *slice thickness* of $5mm$, *spatial resolution* of $2.9mm^2$/pixel and *field of view* of $200mm \times 200mm$. Further details regarding the database are available in [17].

The database contains 3 repetitions corresponding to each speaker for each of 51 VCV utterances (*apa, upu, ipi, ata, utu, iti, aka, uku, iki, aba, ubu, ibi, ada, udu, idi, aga, ugu, igi, aθa, ithi, uthu, asa, usu, isi, aʃa, uʃu, iʃi, ama, umu, imi, ana, unu, ini, ala, ulu, ili, afa, ufu, ifi, aɹa, uɹu, iɹi, aha, uhu, ihi, awa, uwu, iwi, aja, uju, iji*). We pre-segment these into a total of 2754 videos, each containing the entire length of single VCV utterance. We further divide the total available data into training dataset having 2268 videos for 14 speakers and testing dataset with the remaining 486 videos. The testing dataset containing videos corresponding to 3 speakers with 3 repetitions were absolutely unseen during the training phase. Eight frames of a sample test data, where a female speaker F1 produces the VCV token *[asa]* has been shown in Fig 2.

3.2. Training

Sixteen image frames are extracted from each of the videos with a stride of 3, excluding the silent frames. The target is to classify the videos into any of the 51 VCV labels. For the CNN part, we utilize the 50 layered Residual Network (ResNet50) that is pre-trained on the ILSVRC-2012 dataset [18]. This speeds up the training process with the optimized weights and further restricts overfitting in the current small dataset. We vary the parameters one by one keeping all others fixed and simultaneously conduct evaluations for various combinations to select the set of optimum values. The batch size is varied from 32 to 128 and set to the optimum value of 64. Similarly, the number of steps/epoch for the training is changed from 100 to 1000 and fixed at 500, which gives better accuracy. The range of values tried for layers and nodes of LSTM network are 1 to 5 and 32 to 512 respectively. We conclude that the respective optimal values are 1 and 256. The total number of epochs is kept at 140. The drop out value for the LSTM network is varied from .5 to .97 and finally fixed at 0.9. The performance considerably decreases when any activation function except softmax is used.

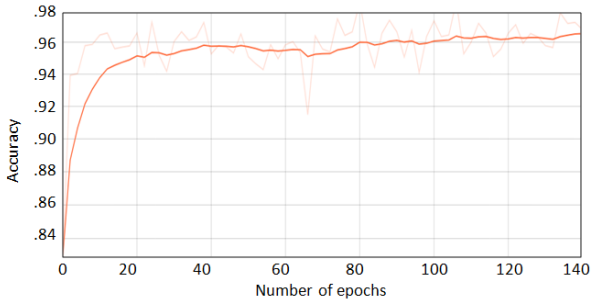

Figure 3: *The Top-1 accuracy for vowel identification*

With the optimal value of the set of parameters remaining fixed, the number of target classes were changed from 51 to 3 and 17 for individual vowels and consonants, respectively. Subsequently, the target class labels were also modified.

3.3. Performance Evaluation

We conducted evaluation for various combinations of parameters and classes (V, C and VCV). In the objective evaluation, the loss measure was given by categorical cross-entropy and various metrics such as Top-1, Top-5 and Top-10 categorical accuracy were employed to analyze its performance.

From Table 1, it can be observed that the mapping accuracy markedly improves from VCV to vowels. The algorithm achieves highest performance of 0.96 when we classify the data into 3 vowel classes, as shown in Fig 3. The performance unexpectedly drops to 0.68 in terms of Top-1 accuracy for 17 consonant identifications. A careful analysis of the predicted classes shows that, the videos are wrongly classified because of their spatial characteristics rather than the temporal features. In other words, it tends to shift towards speaker based classification, rather than the targeted consonant based identification. This is because the number of speakers is also 17 and hence, in this case, the ResNet architecture dominates the trained LSTM layer which makes the identification task much more difficult than vowel identification.

Top-1 classification accuracy for test VCV sequences is 0.42 as shown in Fig 4. Top-5 and Top-10 accuracies reach 0.76 and 0.93 respectively. Though it has demonstrated Top-1 accuracy of .8 for video action recognitions, it fails to maintain the accuracy for the current sound identification task. The degradation of accuracy may be due to excessive increase in number of controllable parameters with increase in number of classes, which would require more training data for estimation.

Table 1: *Speech identification performance*

Identification task	Top-1 accuracy
Vowel	0.96
Consonant	0.68
VCV	0.42

To investigate this issue further, we divided the dataset into 3 parts with 17 target classes, to avoid ambiguity among tokens involving similar articulator movements. As a consequence, the results increased by a considerable margin of 0.12 and the maximum accuracy reached around 0.55 which proves that the similar articulatory movements getting mapped to different sounds is a vital issue that the LRCN algorithm is unable to detect.

4. Discussion and Limitation

Automatic speech recognition from the movements and deformations of vocal tract articulators as visualized from an imaging modality is an incredibly challenging task. The first issue associated with this, is the inter-speaker anatomical differences in vocal tract. This is not a considerable problem in EMA, as the EMA recordings are more concerned with tracking the time varying trajectories of selected points on the tongue and palate. However, while extracting features from imaging modalities, these anatomical differences restrict a particular model from generalizing the structural features like the shapes and lengths of the articulators, which play secondary role in speech identification. Hence, it is difficult to identify the articulator positions and contours due to this inter-subject variability. This makes the conventional methods like Hidden Markov Models [19], Maximum Likelihood Mapping Models [20], Gaussian Mixture Models [21] quite ineffective in extracting appropriate sequential, structural features from the imaging dataset.

Secondly, the subtle and rapid changes of the tract physiology makes the speech recognition task more difficult. This is conceptually different from the conventional action recognition tasks, where the target is to classify the video action space based on widely varying user-movements. In our case, a speaker-specific frame-by-frame analysis of the image sequences corresponding to various VCV tokens demonstrate nearly similar tongue motion and vocal tract contour variation. This makes it difficult to detect significant changes in articulator movements for the consonant part of the VCV utterances. Also, the action recognition tasks involve an underlying spatial object detection task to interpret the actions from the video-clips, through spatial networks. However, such networks [6, 8, 9] fail to extract meaningful features from rtMRI frames to map them to distinct articulatory movements towards speech identification.

Thirdly, more than two-third of the total time duration for each video clip is occupied for uttering the two 'V's and the remaining time is utilized for the intermediate 'C'. So there are more frames corresponding to the vowels than the consonants, which means more feature extraction from those frames and heavier emphasis on the vowels in the classification task. Besides, the training data corresponding to each vowel class is more than that corresponding to each consonant or each transition. The availability of training data is seen to have an enormous influence on the accuracy of the methods. Thus, while the algorithm

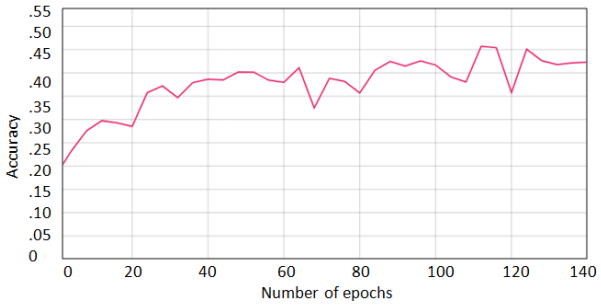

Figure 4: *The Top-1 accuracy for VCV identification*

shows great potential for vowel identification from the tokens, the accuracy reduces for consonants and gets further decreased for the entire VCV identifications.

Fourthly, rtMRI images have low spatial and temporal resolution and are infested with various noises and reconstruction artifacts [5]. This work mostly focuses on addressing the speech identification issue from the raw images and hence, we did not incorporate any preprocessing module. However, it might be interesting to investigate how noise and artifact reduction or resolution enhancement effects the performance accuracy.

Lastly, we observed that the acoustic pitch and energy level of the audio signals exhibit significant variations for several VCV tokens, which could be utilized to achieve better classification. One way to incorporate this is to associate the target class labels with such acoustic variables and take advantage of multi-variate loss function towards accurate prediction scores.

5. Conclusion and Future Directions

A promising deep learning based video action classification technique named Long Term Recurrent Convolutional Network (LRCN) has been trained and tested on 2754 videos with 51 VCV tokens. Results showed that the targeted articulatory-to-acoustic mapping was satisfactorily approximated for vowels and acceptably for consonants. However, the classification performance demonstrated a considerable decrease in accuracy, while identifying the entire VCV transitions, which implies that the model alone is not sufficient to distinguish these transitional tokens.

Along with the structural vocal tract model, it might be interesting to explore whether augmentation from the vocal tract flow model [22] can assist the algorithm to differentiate between consonants that involve similar articulator movements. In future, we will investigate this issue as well as address whether its performance can be increased by coupling it with an underlying biomechanical model for speech production [23, 24, 25] or using Deep belief Networks (DBN) [26].

6. Acknowledgements

This work was funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada and Canadian Institutes for Health Research (CIHR). The authors would like to thank Amir Hossein Abdi and C. Antonio Sanchez of HCT Lab for providing their valuable insights.

7. References

[1] S. Veena, N. S. Wankhede, and M. S. Shah, "Study of vocal tract shape estimation techniques for children," *Procedia Computer Science*, vol. 79, pp. 270–277, 2016.

[2] V. Mitra, G. Sivaraman, C. Bartels, H. Nam, W. Wang, C. Espy-Wilson, D. Vergyri, and H. Franco, "Joint modeling of articulatory and acoustic spaces for continuous speech recognition tasks," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5205–5209.

[3] T. Hueber, P. Badin, G. Bailly, A. Ben-Youssef, F. Elisei, B. Denby, and G. Chollet, "Statistical mapping between articulatory and acoustic data, application to silent speech interface and visual articulatory feedback," in *Proceedings of the 1st International Workshop on Performative Speech and Singing Synthesis (p3s)*, 2011, p. 3.

[4] A. Toutios and S. S. Narayanan, "Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research," *APSIPA Transactions on Signal and Information Processing*, vol. 5, 2016.

[5] S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, "Recommendations for real-time speech mri," *Journal of Magnetic Resonance Imaging*, vol. 43, no. 1, pp. 28–44, 2016.

[6] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[8] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[9] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4489–4497.

[10] T. Sorensen, Z. Skordilis, A. Toutios, Y.-C. Kim, Y. Zhu, J. Kim, A. Lammert, V. Ramanarayanan, L. Goldstein, D. Byrd *et al.*, "Database of volumetric and real-time vocal tract mri for speech science," *Proc. Interspeech 2017*, pp. 645–649, 2017.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013, pp. 6645–6649.

[14] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth annual conference of the international speech communication association*, 2014.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] W. Zaremba and I. Sutskever, "Learning to execute," *arXiv preprint arXiv:1410.4615*, 2014.

[17] 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.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[19] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an hmm-based speech production model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 175–185, 2004.

[20] J. Hodgen and P. Valdez, "A stochastic articulatory-to-acoustic mapping as a basis for speech recognition," in *Instrumentation and Measurement Technology Conference, 2001. IMTC 2001. Proceedings of the 18th IEEE*, vol. 2. IEEE, 2001, pp. 1105–1110.

[21] 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.

[22] V. Zappi, A. Vasuvedan, A. Allen, N. Raghuvanshi, and S. Fels, "Towards real-time two-dimensional wave propagation for articulatory speech synthesis," in *Proceedings of Meetings on Acoustics 171ASA*, vol. 26, no. 1. ASA, 2016, p. 045005.

[23] F. Vogt, J. Lloyd, S. Buchaillard, P. Perrier, M. Chabanas, Y. Payan, and S. Fels, "An efficient biomechanical tongue model for speech research." in *Proceedings of the 7th International Seminar on Speech Production.* Cephala, ISBN 85-99598-02-3, 2006, pp. 51–58.

[24] B. Gick, P. Anderson, H. Chen, C. Chiu, H. B. Kwon, I. Stavness, L. Tsou, and S. Fels, "Speech function of the oropharyngeal isthmus: a modelling study," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 2, no. 4, pp. 217–222, 2014.

[25] I. Stavness, J. E. Lloyd, and S. Fels, "Automatic prediction of tongue muscle activations using a finite element model," *Journal of biomechanics*, vol. 45, no. 16, pp. 2841–2848, 2012.

[26] J. Wei, Q. Fang, X. Zheng, W. Lu, Y. He, and J. Dang, "Mapping ultrasound-based articulatory images and vowel sounds with a deep neural network framework," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5223–5245, 2016.