2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871909

Mapping Acoustics to Articulatory Gestures in Dutch: Relating Speech Gestures, Acoustics and Neural Data

Paolo Favero[1], Julia Berezutskaya[2], Nick F. Ramsey[1], Aleksei Nazarov[3] and Zachary V. Freudenburg[1]

Abstract— **Completely locked-in patients suffer from paralysis affecting every muscle in their body, reducing their communication means to brain-computer interfaces (BCIs). State-of-the-art BCIs have a slow spelling rate, which inevitably places a burden on patients' quality of life. Novel techniques address this problem by following a bio-mimetic approach, which consists of decoding sensory-motor cortex (SMC) activity that underlies the movements of the vocal tract's articulators. As recording articulatory data in combination with neural recordings is often unfeasible, the goal of this study was to develop an acoustic-to-articulatory inversion (AAI) model, i.e. an algorithm that generates articulatory data (speech gestures) from acoustics. A fully convolutional neural network was trained to solve the AAI mapping, and was tested on an unseen acoustic set, recorded simultaneously with neural data. Representational similarity analysis was then used to assess the relationship between predicted gestures and neural responses. The network's predictions and targets were significantly correlated. Moreover, SMC neural activity was correlated to the vocal tract gestural dynamics. The present AAI model has the potential to further our understanding of the relationship between neural, gestural and acoustic signals and lay the foundations for the development of a bio-mimetic speech BCI.**

Clinical relevance— **This study investigates the relationship between articulatory gestures during speech and the underlying neural activity. The topic is central for development of brain-computer interfaces for severely paralysed individuals.**

I. INTRODUCTION

Recent studies on bio-mimetic brain-computer interfaces (BCIs) for restoring communication have been mainly focused on three aspects of the speech production process: the vocal tract articulatory movements (kinematic data), their neural correlates (neural data) and the acoustic properties of the sounds generated from such movements (acoustic data). The simultaneous collection of neural and kinematic data is extremely challenging, especially when neural activity is recorded directly from the cortex via electrocorticography (ECoG). Acoustic-to-articulatory inversion (AAI) models are aimed at inferring vocal tract movements from acoustic data, and have previously been used in intracranial BCI studies to cope with such experimental limitations [1], [2], [3]. These studies have used AAI to predict articulatory movements, which have a non-unique mapping to acoustic data (multiple movements can lead to the same sounds) [4]. Tract variables (TVs; see Fig. 1) [5], on the other hand, measure the distance and position of constrictions in the vocal tract during speech production, capturing dynamic patterns of coarticulatory movements (articulatory gestures) and thus having a lower degree of non-uniqueness [4]. Therefore, the main goal of this study was to develop an AAI model that capitalises on TVs, rather than articulatory movements, and could lead to novel approaches in bio-mimetic BCIs. To this end, we implemented four AAI model variants (see Fig. 2), all based on a fully-convolutional neural network architecture, but varying in the type of input. After training and evaluating the performance of the AAI model variants, the best performing model was applied to a dataset collected at the University Medical Centre Utrecht from Dutch patients. The dataset contained high-density (HD) ECoG recordings acquired during a syllable production task (neural data) and microphone recordings synchronised with the ECoG data (acoustic data). We used the AAI model to predict TVs from the acoustic data. Finally, to investigate the relationship between neural data, TVs and syllable labels, a representational similarity analysis (RSA) [6] was performed. The study was approved by the Medical Ethical Committee of the University Medical Center Utrecht in accordance with the Declaration of Helsinki (2013).

II. METHODS

A. Acoustic-to-Articulatory Inversion

1) The AAI Model Variants: The decision of using TVs as output of the AAI model, was inspired by previous studies from Mitra and colleagues [7], [4]. In [7], as for this

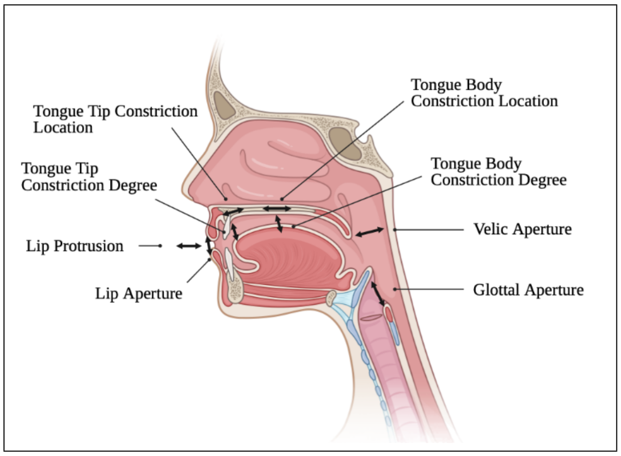
Tongue Tip Constriction Location
Tongue Body Constriction Location
Tongue Tip Constriction Degree
Tongue Body Constriction Degree
Velic Aperture
Lip Protrusion
Lip Aperture
Glottal Aperture

Fig. 1. Tract Variables

[1] Nick F. Ramsey, Zachary V. Freudenburg are and Paolo Favero was with Brain Center, Department of Neurology and Neurosurgery, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, the Netherlands

[2] Julia Berezutskaya is with Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Thomas van Aquinostraat 4, 6525 GD Nijmegen, the Netherlands

[3] Aleksei Nazarov is with Department of Languages, Literature and Communication, Utrecht University, Trans 10, 3512 JK Utrecht, the Netherlands

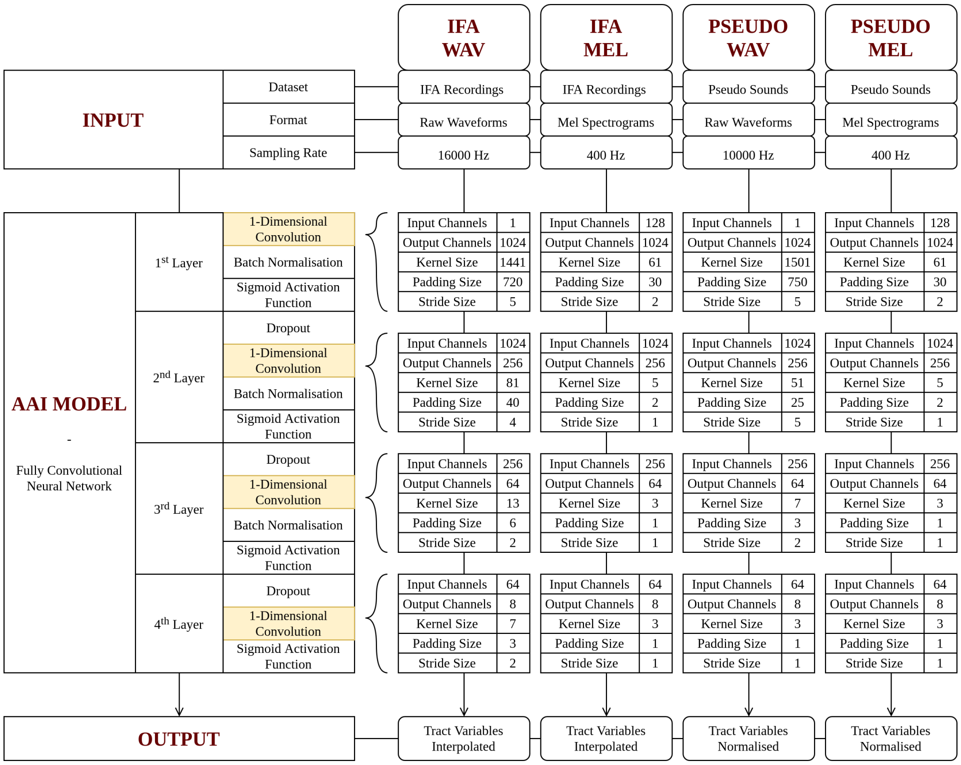

Fig. 2. Acoustic-to-articulatory inversion (AAI) variants with inputs and outputs (on the left), and model parameters (on the right).

study, TVs were not directly available for the datasets in use, thus, the external software TADA (Haskins Laboratories [8]) was used to generate them. This software requires a text input in either American English orthographic notation or ARPABET notation [9]. TADA was optimised for the English language, and as we wanted to use it with a Dutch corpus, we transcribed all the words included in our corpus to ARPABET notation by approximating them to English sounds, i.e. using the phonemic inventory of Standard American English. Together with TVs, TADA creates aligned pseudo-sounds in wav format [10]. The pseudo-sounds contain only the resonance pattern of a word (except nasalisation) presented over a fixed continuous voice signal [10]. In their research, Mitra and colleagues used spectral features extracted from pseudo-sounds as the acoustic input for their AAI model and the corresponding TVs as the network targets [7]. In addition to replicating their approach, we trained a variant of the model on acoustic data from a Dutch corpus. The TVs generated with TADA, as their corresponding pseudo-sounds, do not reflect any variance in pronunciation (duration, intonation, speaker differences, etc.), so we tried an alternative approach by using actual speech recordings as input. Furthermore, while multiple research groups addressing the AAI problem have used spectral features as input for their models ([2], [7], [3]), recent trends in speech recognition have proved the advantage of using raw waveforms as input for acoustic modelling [11], [12], [13]. Thus, we investigated costs and benefits of using raw waveforms compared to spectral features for AAI mapping, trying to draw from the success in speech recognition applications. To summarise, we trained four model variants: one trained on pseudo-sounds spectral features (as in [7]), one on pseudo-sounds raw waveforms, one trained on spectral features from a Dutch corpus and one from raw waveforms from the same corpus (see Fig. 2).

2) IFA Dutch Spoken Language Corpus: We used the IFA Dutch Spoken Language Corpus [14] for training and validating two of the AAI model variants. The IFA Corpus contains phonemically segmented and labelled speech from eight Dutch native speakers (4 females, age 15-66). Participants took part in three main tasks: telling about a vacation trip to an interviewer in an informal face-to-face setting, reading aloud text from a computer screen in a variety of speaking styles (e.g. sentences, words, syllables, etc.) and re-telling a previously read text. The corpus was recorded at a frequency of 44100 Hz in a sound treated and quiet room. For further information regarding the labelling protocol or the recording equipment and methodologies refer to [14], [15], [16].

3) Tract Variables Preprocessing: After running TADA and extracting the TVs, two preprocessing steps were performed to optimise the model training: normalisation (rescaling of the data to a range of $[0, 1]$) and interpolation. The latter was necessary as TADA text-based TVs did not account for speech duration variability. Therefore, when training the model on the IFA original recordings, TVs were linearly interpolated to match the duration of their corresponding acoustic trials.

4) Acoustic Preprocessing: Acoustic data (IFA corpus) were first cut into single words recordings based on the phonemic transcription provided with the corpus. The resulting dataset consists of approximately 3.5 hours of speech (44000 trials). The data were downsampled to 16000 Hz, and then we computed the mel spectrogram. We used mel-spectrogram as a spectral representation of sound, since the mel scale better approximates the human auditory system response [17].

5) Fully Convolutional Neural Network: The architecture of the AAI model developed for this study was inspired by the work of Dai et al. [11], on the benefit of fully convolutional neural networks (FCNs) for raw acoustic data modelling. Our model includes four main convolutional layers clusters as shown in the left side of Fig. 2, and was implemented in Pytorch [18]. The number of layer clusters was selected based on preliminary testing of the network, which showed performance improvement on the validation set up until the addition of the 4^{th} cluster.

A large receptive field (kernel size) in the first convolutional layer was used to compensate for the absence of recurrent layers capturing long-term relationships in the data. The convolutional layers were selected for their inherent properties of finding hierarchical patterns and structures in data, which apply to image processing applications [19] as well as audio processing ones [20], [11], [12], [21]. Multiple regularisation techniques were used to optimise the FCN model: Xavier initialisation [22], batch normalisation [23], and dropout [24]. All of them resulted in a performance improvement and a faster convergence. The training and validation sets were subdivided by randomly selecting approximately 10 percent of the trials from each participant. The Adam optimiser [25] (learning rate of 0.001) was used to minimise the mean squared error loss. Early stopping patience parameter was set to 100 epochs.

B. Neural Analyses

1) 9 Syllables Set: Six patients (2 females, age 22-59, chronically or acutely implanted with ECoG electrodes) took part in the syllable production task while their brain data were recorded with high-density ECoG grids. All patients gave written informed consent to participate in ECoG recordings and gave permission to use their data for scientific research. Patients read syllables on the screen (nine syllable in total + rest, presented one at a time, 10 repetitions per syllable, 2 runs for 2 of the participants). The syllables were obtained by combining three vowels (/a/, /i/ and /u/) and three consonants (/k/, /z/ and /m/). The vowels were selected based on their maximal variation in terms of frontness-backness, highness-lowness and roundness, whereas the consonants were selected for their variations in terms of place and manner of articulation and voicing. Synchronously with microphone recordings, brain activity was recorded from the left SMC of the patients via HD (128 electrodes, 3-4 mm centre-to-centre) ECoG.

2) Acoustic Preprocessing: Parts of the recordings included silence either at the beginning or at the end of the file. After removing rest trials, the recordings were cut from 0.1 seconds before voice onset time to a variable time after it (ranging from 0.4 to 0.6 seconds) based on the average pronunciation length of individual subjects. The waveforms were then down-sampled to 16000 Hz as done for the IFA set. Voice onset was calculated with Praat [26].

3) Neural Data Preprocessing: Data were preprocessed following a standard ECoG processing pipeline [27] that consisted of removal of noisy channels, notch-filtering of the line noise (50 Hz and harmonics), common average referencing and high-frequency component extraction (65-125 Hz) using Gabor wavelet decomposition in 1 Hz frequency bins. The final signal was averaged over frequencies, smoothed (100 ms kernel) and cut into chunks of 1.5 seconds (1 second before voice onset time and 0.5 seconds after it).

4) Representational Similarity Analysis: A common approach, Representational Similarity Analysis (RSA), was used to investigate the relationship between neural signals and TVs. The RSA consisted of three steps: 1) computation of the representational dissimilarity matrices (RDMs) that capture the difference in data points across stimulus trials per modality of the syllable dataset (such as neural data, TVs, etc.); 2) computation of the similarity of RDMs across modalities and 3) statistical inference based on permutation tests that assessed the significance of similarities computed in step 2.

Four data modalities of the syllable task were used: neural data, syllables labels, ground-truth TADA-generated TVs and TVs predicted with the best performing AAI model (variant trained on pseudo-sounds raw waveforms). One-Pearson correlation $(1 - r)$ was used as a dissimilarity measure for neural data, whereas Euclidean distance was used for TADA-generated and predicted TVs. RSA was performed separately per each participant and task run, comparing neural data RDMs to all other modalities RDMs. Pearson correlation was the measure of similarity between modality-specific RDMs, and its significance was assessed using permutation tests (5000 trial shuffles per RDM).

III. RESULTS

A. Acoustic-to-Articulatory Inversion

1) Training and validation loss: The training and validation losses for the four AAI variants are shown in Fig. 3. Comparing the loss values would not be fair given the structural differences in the four sets (IFA Wav, IFA Mel, Pseudo Wav and Pseudo Mel). Nevertheless, some common trends were observed among these AAI variants. In all of them the training loss showed a steep decrease during the first 50 epochs (steeper for the models trained on pseudo-sounds), and then slowly stabilised. This was reflected in

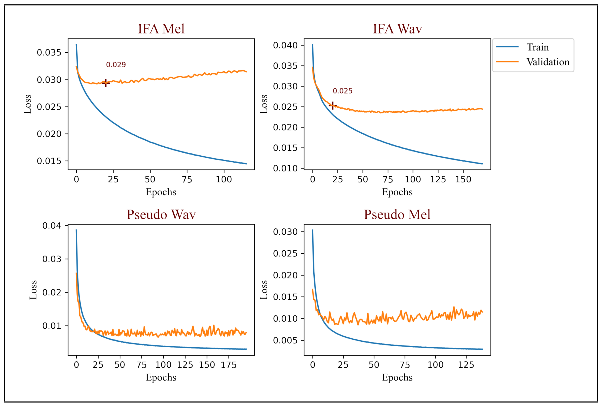

Fig. 3. Training and Validation Loss for the four AAI variants. In the two upper plots, the validation loss at epoch 20 is indicated, showing that a lower loss is reached by the model trained on raw waveforms.

the validation losses progressions, indeed all models reached their lowest validation loss between the 10^{th} and the 20^{th} epoch and increased afterwards, showing the effects of overfitting to the training set.

2) Model evaluation: Performance of the best AAI model variants (based on the early stopping criterion) was evaluated in terms of Pearson correlation between predicted and target TVs. The average correlations (per TV and overall) between the validation targets and the model predictions for the four AAI variants are shown in Table II. For all four AAI variants and for all eight TVs the correlations were significant at p-value $<$ 0.001 (as assessed with independent one-tailed t-tests between correlation of each prediction-target pair and correlations with shuffled predictions and targets).

Interestingly, when models were trained on pseudo sounds, correlations were always higher than 0.5 (apart from Lip Protrusion), whereas when trained on the IFA data the correlation values were always below 0.5. Such a large discrepancy in performance may be attributed to the temporal misalignment between the input and the target TVs. We attempted to mitigate it using temporal interpolation during preprocessing steps, although interpolation only guaranteed that input and output had the same duration, not that they were aligned correctly.

Importantly, even low correlations remained significant (p-value $<$ 0.001) indicating the predicted TVs preserved the distinctive features of different words.

In addition, we found that the use of mel spectrograms as input has led to faster model convergence. This is likely due to the fact that the the preprocessed mel spectrogram is less complex and contains less features compared to raw waveform. However, the TV average correlation for model variants trained on mel features was lower compared to model variants trained on raw waveforms both for pseudo-sounds and IFA recordings. Thus, in the long run, the models benefited from the richness and complexity of information present in the raw acoustic data (see Fig. 3).

B. *Neural Analyses*

The results of the RSA that tested the relationship between neural data and TVs (either TADA generated or predicted with the AAI model) were significant in all subjects and runs except for S2 (Table I). However, S2 did not show a significant relationship also between neural data and syllable labels, which means that it was difficult to establish the overall relationship between S2's neural data and the task. Only one other subject lacked significance for similarity between neural and predicted TVs (S3) confirming that overall the AAI model provided high-quality predictions of TVs, that exhibited as much similarity with the brain data as the ground-truth TVs, extracted with TADA software. Altogether, the results of this analysis confirmed the previously reported high degree of similarity between TVs and SMC responses during a speech production task.

IV. DISCUSSION

The present study aimed to 1) relate articulatory gestures (captured by the tract variables) to their acoustic outputs by building and training an acoustic-to-articulatory inversion (AAI) model; and 2) investigate the relationship between articulatory gestures and their neural correlates by means of a representational similarity analysis (RSA).

The predicted and target TVs were correlated significantly across all eight TVs. Thus, the present AAI model can be used in future studies to link vocal tract configurations to produced speech. Furthermore, the RSA results indicated a consistent significant correlation across neural data and gestural data (either generated or predicted) for all subjects (apart from one). While articulatory gestures have already been shown to correlate with SMC neural activity [3], [28], this is the first study in which SMC response is related to eight TVs that were predicted from acoustic data, setting a precedent in bio-mimetic approaches for speech BCIs.

The present study has a number of limitations. Regarding the AAI mapping: the model variants overfitted the training set within 100 epochs, suggesting that the learning rate was set too high or the network architecture was too complex to model the data; second, here, in the absence of experimentally collected vocal tract movements, we made use of software that approximated them, which may have affected some

TABLE I
RSA results per subject (S) and run, comparing neural RDMs to the other modalities. Non-significant results (at $p > .005$ based on permutation tests) are highlighted in bold.

Subjects	Neural vs TADA-TVs	Neural vs Predicted-TVs	Neural vs Labels
S1 run 1	0.0002	0.0002	0.0002
S1 run 2	0.0002	0.0002	0.0002
S2 run 1	**0.3379**	**0.8076**	**0.5376**
S2 run 2	**0.9908**	**0.9988**	**0.8042**
S3	0.0030	**0.1468**	0.0048
S4	0.0002	0.0002	0.0002
S5	0.0002	0.0002	0.0002
S6	0.0002	0.0002	**0.0106**

TABLE II
INDIVIDUAL AND OVERALL CORRELATIONS OF PREDICTED AND TARGET TVS FOR THE FOUR AAI VARIANTS. ALL CORRELATIONS ARE SIGNIFICANT AT $p < .001$

Variants	Lip Protrusion	Lip Aperture	Tongue Body CL	Tongue Body CD	Velum	Glottis	Tongue Tip CL	Tongue Tip CD	Average Correlation
Pseudo wav	0.30	0.84	0.93	0.90	0.67	0.59	0.95	0.96	0.77
Pseudo mel	0.29	0.84	0.90	0.88	0.59	0.58	0.94	0.94	0.74
IFA wav	0.16	0.16	0.38	0.36	0.53	0.36	0.40	0.44	0.35
IFA mel	0.13	0.17	0.32	0.23	0.40	0.26	0.21	0.28	0.25

of the results. Furthermore, the study included a relatively small number of participants for the neural analyses. Follow-up work with a larger number of subjects could provide group statistics for these results and could shed more light on the relationship between neural and gestural (TV) data. Also, background noise in the recordings likely added variance to the predicted TVs that is not related to SMC activity; this can be addressed in future studies that employ better recording equipment (such as a directional microphone). Finally, both AAI mapping and neural analyses were restricted to isolated words. Follow-up projects using continuous speech data are needed to provide more insight for real-time decoding from SMC to restoring communication in locked-in patients. In conclusion, we feel that our approach of predicting TVs from recorded acoustics is a promising approach to examine the link between SMC activity and TVs during speech.

REFERENCES

[1] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, pp. 493–498, Apr. 2019.

[2] J. Chartier, G. K. Anumanchipalli, K. Johnson, and E. F. Chang, "Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex," *Neuron*, vol. 98, pp. 1042–1054.e4, June 2018.

[3] E. M. Mugler, M. C. Tate, K. Livescu, J. W. Templer, M. A. Goldrick, and M. W. Slutzky, "Differential Representation of Articulatory Gestures and Phonemes in Precentral and Inferior Frontal Gyri," *The Journal of Neuroscience*, vol. 38, pp. 9803–9813, Nov. 2018.

[4] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Noise robustness of tract variables and their application to speech recognition," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[5] C. P. Browman and L. Goldstein, "Articulatory Phonology: An Overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[6] N. Kriegeskorte, M. Mur, and P. A. Bandettini, "Representational similarity analysis-connecting the branches of systems neuroscience," *Frontiers in systems neuroscience*, vol. 2, p. 4, 2008.

[7] 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)*, (Florence, Italy), pp. 3017–3021, IEEE, May 2014.

[8] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable Task Dynamics model in MATLAB," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2430–2430, 2004. Publisher: Acoustical Society of America.

[9] A. Klautau, "ARPABET and the TIMIT alphabet," 2001.

[10] H. Nam, C. Browman, L. Goldstein, P. Rubin, M. Proctor, and E. Saltzman, "TADA (TAsk Dynamics Application) manual," p. 32, 2007.

[11] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very Deep Convolutional Neural Networks for Raw Waveforms," *arXiv:1610.00087 [cs]*, Oct. 2016. arXiv: 1610.00087.

[12] S. Qu, J. Li, W. Dai, and S. Das, "Understanding audio pattern using convolutional neural network from raw waveforms," *arXiv preprint arXiv:1611.09524*, p. 6, 2016.

[13] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (South Brisbane, Queensland, Australia), pp. 4580–4584, IEEE, Apr. 2015.

[14] R. van Son, D. Binnenpoorte, H. van den Heuvel, and L. C. W. Pols, "The IFA Corpus: a Phonemically Segmented Dutch "Open Source" Speech Database," p. 9, 2001.

[15] R. van Son, "IFA Label Protocol," May 2001.

[16] R. van Son and L. Pols, "Structure and access of the open source IFA Corpus," in *Proceedings of the IRCS workshop on Linguistic Databases, Philadelphia*, pp. 245–253, 2001.

[17] E. C. Zsiga, *The Sounds of Language: An introduction to phonetics and phonology*. John Wiley & Sons, 2013.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *arXiv:1912.01703 [cs, stat]*, Dec. 2019. arXiv: 1912.01703.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[20] T. Sercu, C. Puhrsch, B. Kingsbury, and Y. LeCun, "Very Deep Multilingual Convolutional Neural Networks for LVCSR," *arXiv:1509.08967 [cs]*, Jan. 2016. arXiv: 1509.08967.

[21] H. Muckenhirn, V. Abrol, M. Magimai-Doss, and S. Marcel, "Understanding and Visualizing Raw Waveform-Based CNNs," in *Interspeech 2019*, pp. 2345–2349, ISCA, Sept. 2019.

[22] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, JMLR Workshop and Conference Proceedings, 2010.

[23] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv:1502.03167 [cs]*, Mar. 2015. arXiv: 1502.03167.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[25] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017. arXiv: 1412.6980.

[26] P. Boersma and V. van Heuven, "Speak and unSpeak with PRAAT," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[27] N. Ramsey, E. Salari, E. Aarnoutse, M. Vansteensel, M. Bleichner, and Z. Freudenburg, "Decoding spoken phonemes from sensorimotor cortex with high-density ECoG grids," *NeuroImage*, vol. 180, pp. 301–311, Oct. 2018.

[28] K. E. Bouchard, D. F. Conant, G. K. Anumanchipalli, B. Dichter, K. S. Chaisanguanthum, K. Johnson, and E. F. Chang, "High-Resolution, Non-Invasive Imaging of Upper Vocal Tract Articulators Compatible with Human Brain Recordings," *PLOS ONE*, vol. 11, p. e0151327, Mar. 2016.