Interspeech 2022
18-22 September 2022, Incheon, Korea

Acoustic-to-articulatory Speech Inversion with Multi-task Learning

Yashish M. Siriwardena[1], *Ganesh Sivaraman*[2], *Carol Espy-Wilson*[1]

[1]University of Maryland College park, MD, USA
[2]Pindrop, GA, USA

yashish@terpmail.umd.edu, ganesa90@gmail.com, espy@umd.edu

Abstract

Multi-task learning (MTL) frameworks have proven to be effective in diverse speech related tasks like automatic speech recognition (ASR) and speech emotion recognition. This paper proposes a MTL framework to perform acoustic-to-articulatory speech inversion by simultaneously learning an acoustic to phoneme mapping as a shared task. We use the Haskins Production Rate Comparison (HPRC) database which has both the electromagnetic articulography (EMA) data and the corresponding phonetic transcriptions. Performance of the system was measured by computing the correlation between estimated and actual tract variables (TVs) from the acoustic to articulatory speech inversion task. The proposed MTL based Bidirectional Gated Recurrent Neural Network (RNN) model learns to map the input acoustic features to nine TVs while outperforming the baseline model trained to perform only acoustic to articulatory inversion.

Index Terms: acoustic-to-articulatory speech inversion, multi-task learning, acoustic-to-phoneme mapping, biGRNNs

1. Introduction

Human speech production is a highly complex task which involves synchronized motor control of speech articulators. The inverse problem of determining the trajectories of the movement of speech articulators from the speech signal is referred to as acoustic-to-articulatory speech inversion [1, 2]. This mapping from acoustics to articulation is an ill-posed problem which is known to be highly non-linear and non-unique [3]. However, developing Speech Inversion (SI) systems have gained attention over the recent years mainly due to its potential in a wide range of speech applications like Automatic Speech Recognition (ASR) [4, 5, 6], speech synthesis [7, 8], speech therapy [9] and most recently with detecting mental health disorders like Major Depressive Disorder and Schizophrenia [10, 11]. Real articulatory data are collected by techniques like X-ray microbeam [12], Electromagnetic Articulometry (EMA) [13] and real-time Magnetic Resonance Imaging (rt-MRI) [14]. All these techniques are expensive, time consuming and need specialized equipment for observing articulatory movements directly [1]. This explains why developing a speaker-independent SI system that can accurately estimate articulatory features for any unseen speaker is of greater need.

Over the past few years, deep neural network (DNN) based models have propelled the development of SI systems to new heights. Bidirectional LSTMs (BiLSTMS) [15, 16], CNN-BiLSTMs [17, 18], Temporal Convolutional Networks (TCN) [19] and transformer models [20] have gained state-of-the-art results with multiple articulatory datasets [21]. To further improve the speech inversion task, people have tried incorporating phonetic transcriptions as an input along with acoustic features [22, 17]. One of the limitations of these models is that you need phonetic transcriptions of the speech audio file at the time of inference. To address this issue while also leveraging on the additional information that phonetic transcriptions offer, we propose a Bidirectional Gated Recurrent Neural Network (BiGRNN) model, implemented with a multi-task learning framework, to perform acoustic-to-articulatory speech inversion. The MTL based model does not need phonetic transcriptions at the time of inference, but benefits from learning the mapping from acoustics-to-phonetics to improve generalizability of SI systems.

The key contributions of the paper can be listed as follows :

- We propose a MTL based BiGRNN model to perform acoustic-to articulatory speech inversion by also learning a shared task of acoustic-to-phoneme inversion.
- We compare and contrast two training algorithms to optimize the proposed MTL model
- By conducting an ablation study we assert the importance of doing multi-task learning for speech inversion

2. Speech Inversion System

2.1. Multi-task Learning : Related Work

The idea of Multi-task learning (MTL) was formally presented by Caruana et al. [23] as an inductive transfer mechanism with the principle goal of improving generalization capability of Machine Learning (ML) models. MTL helps improve generalizability of ML models by leveraging domain-specific information of training data which can be used in related tasks. Effectively, what happens is that the training data for the parallel task serve as an inductive bias [23]. MTL has also been utilized as a solution for the data sparsity problem where one task has a limited number of labeled data and training individual models for each task is difficult. From this perspective, MTL is a useful tool which can reuse the existing knowledge and reduce the cost of collecting challenging datasets (e.g. articulatory datasets). The secret behind the success of MTL lies with the use of more data from different learning tasks compared to learning a single task, hence learning better representations and reducing the risk for overfitting [24].

MTL has widely been used in computer vision and a recent work [25] has implemented a MTL model to work on 12 different datasets while achieving the state-of-the-art with 11 of them. MTL has also been explored in Automatic Speech Recognition (ASR) tasks [26, 27], text-to-speech (TTS) [28] and in speech emotion recognition (SER) [29, 30]. Cai et al. [30] recently presented the state-of-the-art results for the SER task with IEMOCAP dataset using their model based on a MTL framework. Xie et al. [31] in their work present a MTL based SI system which has been compared against a hierarchical RNN model.

10.21437/Interspeech.2022-11164

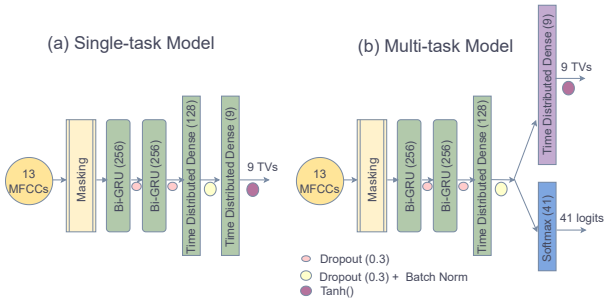

Figure 1: *Single-task and Multi-task model architectures*

2.2. Dataset Description

We used the Haskins Production Rate Comparison (HPRC) database which contains recordings from 4 female and 4 male subjects reciting 720 phonetically balanced IEEE sentences [32] at normal and fast production rates [21]. The recordings were done using a 5-D electromagnetic articulometry (EMA) system (WAVE; Northern Digital). First, every sentence was produced at speaker's preferred 'normal' speaking rate and then a 'fast' repetition of the same, without making errors. Sensors were placed on the tongue (tip (TT), body (TB), root (TR)), lips (upper (UL) and lower (LL)) and mandible, together with reference sensors on the left and right mastoids, and upper and lower incisors (UI, LI). These EMA trajectories were obtained at 100 Hz and then were low-pass filtered at 5 Hz for references and 20 Hz for articulator sensors. Synchronized audio was recorded at 22050 Hz. The following geometric transformations were used to obtain 9 TVs (namely Lip Aperture (LA), Lip Protrusion (LP), Tongue Body Constriction Location (TBCL), Tongue Body Constriction Degree (TBCD), Tongue Tip Constriction Location (TTCL), Tongue Tip Constriction Degree (TTCD), Jaw Angle (JA), Tongue Middle Constriction Location (TMCL) and Tongue Middle Constriction Degree (TMCD)). The equations to compute the geometric transformations are presented in our previous work in [1, 33].

2.3. Audio Features

All the audio files from the HPRC dataset (both 'normal' and 'fast' rate) are first segmented into 2 second long segments and the shorter audios are zero padded at the end. Previous studies with developing SI systems have shown MFCCs to be superior over conventional Melspectrograms and Perceptual Linear Predictions (PLPs) as acoustic features [1]. Based on that, in this study, we use Mel-Frequency Cepstral Coefficients (MFCCs) as the input audio feature for the proposed SI systems. MFCCs are extracted using a 20ms Hamming analysis window with a 10ms frame shift. 13 cepstral coefficients were extracted for each frame while 40 Mel frequencies were used. Each MFCC was utterance wise normalized (z-normalized) prior to model training.

2.4. Phoneme Features

The HPRC dataset contains phonetic alignment for the recorded utterances. The phone alignment is extracted using the Penn Phonetics Lab Forced Aligner(P2FA)[1]. We remove the allophonic variations of the monophones and retain only 40 monophone units. Using the forced alignment we created frame wise monophone labels for all of the HPRC dataset. The one-hot encoded frame-wise monophone labels are the phonetic features used in this study.

[1]https://web.sas.upenn.edu/phonetics-lab/facilities/

2.5. Model Architecture

In this paper, we use a novel Bidirectional Gated Recurrent Neural Network (BiGRNN) model to implement both the single-task and multi-task SI systems. Both the single-task and multi-task models have the same backbone which includes 3 bidirectional layers of Gated Recurrent Units (GRUs) followed by a time distributed fully connected layer. Single-task model which predicts TVs has an additional time distributed fully connected layer to predict the TVs (output layer). On the other hand, the multi-task model has two output layers, one a time distributed fully connected layer to predict the TVs and the other a softmax layer to predict phoneme labels. Figure 1 shows the architecture of the single task model on the left and the multi-task model on the right.

2.6. Performance Measurements

All the models are evaluated with the Pearson Product Moment Correlation (PPMC) scores computed between the estimated TVs and the corresponding ground-truth TVs. Equation 1 is used to computed the PPMC score, where X represents the estimated TVs, $\overline{X}$ the mean of the estimated TVs, Y the ground-truth TVs, $\overline{Y}$ the mean of the ground-truth TVs and N the number of TVs.

$$PPMC = \frac{\sum_i^N (X[i] - \overline{X})(Y[i] - \overline{Y})}{\sqrt{\sum_i^N (X[i] - \overline{X})^2 (Y[i] - \overline{Y})^2}} \quad (1)$$

2.7. Model training

The HPRC dataset was divided into training, development, and testing sets, so that the training set has utterances from 6 speakers (3 Males, 3 Females) and the development and testing sets have utterances of 2 speakers (1 male,1 female) equally split between them. None of the subjects in training are present in the development and testing sets and hence all the models are trained in a 'speaker-independent' fashion. The split also ensured that around 80% of the total number of utterances were present in the training, and the development and testing sets have a nearly equal number of utterances. This allocation was done in a completely random manner.

All the models were implemented with Tensorflow-Keras machine learning framework and trained with NVIDIA TITAN X GPUs. For all single-task and multi-task models, ADAM optimizer with a starting learning rate of 1e-3 and an exponential learning rate scheduler was used. The starting learning rate was maintained up to 10 epochs and then decayed exponentially after each subsequent 5 epochs. To choose the best starting 'learning rate' (LR), we did a grid search on [1e-3, 3e-4, 1e-4] and to choose the training batch size we did a similar grid search on [16,32,64,128]. The best PPMC scores were obtained for 1e-3 as the LR and 128 as the batch size for training.

2.8. Training Paradigms for multi-task SI systems

We experimented with two distinct training algorithms to optimize the MTL model. We denote the input MFCC features to the model as $x \epsilon R^{L \times d}$ where L (=200) is the number of samples in each utterance and d (=13) is the number of MFCCs. Let f_ϕ be the mapping from MFCCs to TVs from the multi-task model where ϕ defines the shared model parameters to be learned. Similarly, let g_ϕ be the mapping from MFCCs to phoneme logits. Then the output TV prediction from the TV output layer

Table 1: *Single-task vs Multi-task learning for TV predictions*

Model	LA	LP	JA	TTCL	TTCD	TMCL	TMCD	TBCL	TBCD	Average
Single-task	0.764	0.661	0.790	0.706	0.778	0.741	0.801	0.725	0.742	0.745
Multi-task (Algo 1)	0.792	0.681	0.796	0.747	0.793	0.775	0.799	0.760	0.764	0.767
Multi-task (Algo 2)	0.794	0.680	0.806	0.741	0.797	0.775	0.806	0.762	0.766	**0.770**

$\hat{y}_{tv} \epsilon R^{L \times T}$ can be defined from equation 2 and similarly the output logits from the phoneme prediction, $\hat{y}_{ph} \epsilon R^{L \times V}$ can be defined from eqaution 3. Here T (=9) is the number of TVs predicted and V (=41) is the number of phonemes in the dictionary + the symbol for zeros (padded for shorter utterances). We used the Mean Absolute Error (MAE) loss between ground truth TVs y_{tv} and predicted TVs $\hat{y}_{tv}$ and cross entropy error loss between ground truth one-hot encoding labels of phonemes y_{ph} and the predicted phonemes $\hat{y}_{ph}$.

$$\hat{y}_{tv} = f_{\phi}(x) \; ; \; x \epsilon R^{L \times d} \tag{2}$$

$$\hat{y}_{ph} = g_{\phi}(x) \; ; \; x \epsilon R^{L \times d} \tag{3}$$

2.8.1. Training Algorithm 1

Here the multi-task model is optimized for each task in an alternating fashion. In each epoch, the model weights ϕ are first learned from the TV prediction task and the learned weights are then used for computing phoneme labels $\hat{y}_{ph}$. The final model weights $\phi[i]^*$ are then updated with the phoneme prediction task and the process is repeated for the given number of $Epochs$.

Algorithm 1 Iterative Loss Optimization

Require: : $x \epsilon R^{L \times d}, y_{ph}, y_{tv}, Epochs(\epsilon R)$
  while $i < Epochs$ **do**
    $\hat{y}_{tv} \leftarrow f_{\phi[i-1]}(x)$
    $L_{tv} \leftarrow MAE(\hat{y}_{tv}, y_{tv})$
    $\phi[i] \leftarrow \min_{\phi} L_{tv}$
    $\hat{y}_{ph} \leftarrow g_{\phi[i]}(x)$
    $L_{ph} \leftarrow CrossEntropy(\hat{y}_{ph}, y_{ph})$
    $\phi[i]^* \leftarrow \min_{\phi} L_{ph}$
    $i \leftarrow i + 1$
  end while

2.8.2. Training Algorithm 2

In this training algorithm we optimize a joint loss L_{joint}, where the phoneme prediction loss L_{ph} is weighted to combine with the TV prediction loss L_{tv}. The contribution of L_{ph} is controlled by the weight $\alpha \epsilon (0, 1)$, which is a hyper-parameter to be tuned. Here the model is trained with an early stopping criteria monitoring the validation loss ($ValLoss$) with a patience p (=10).

Algorithm 2 Joint Loss Optimization

Require: : $x \epsilon R^{L \times d}, ValLoss, p \epsilon R, \alpha (0 < \alpha < 1), y_{ph}, y_{tv}$
  while $ValLoss[i] < ValLoss[i-p]$ **do**
    $\hat{y}_{ph} \leftarrow g_{\phi[i-1]}(x)$
    $\hat{y}_{tv} \leftarrow f_{\phi[i-1]}(x)$
    $L_{ph} \leftarrow CrossEntropy(\hat{y}_{ph}, y_{ph})$
    $L_{tv} \leftarrow MAE(\hat{y}_{tv}, y_{tv})$
    $L_{joint} \leftarrow L_{tv} + \alpha L_{ph}$
    $\phi[i] \leftarrow \min_{\phi} L_{joint}$
    $i \leftarrow i + 1$
  end while

3. Experiments and Results

3.1. Single-task vs Multi-task Learning for TV prediction

Table 1 shows the results of the single-task model when compared to the two multi-task models trained with two training algorithms. The reported PPMC scores are from evaluations of the speaker-independent test set. Figure 2 shows ground-truth TVs and predicted TVs, LA, TBCD, TTCD and TMCD for an example utterance estimated by the multi-task and the single-task models. Figure 3 shows ground-truth TVs and predicted TVs, LP, TBCL, TTCL and TMCL for the same utterance estimated by the multi-task and the single-task models.

3.2. Baseline comparison on previous work with HPRC dataset

Table 2 lists the reported PPMC scores of the work in Shahrebabaki et. al. [2, 17] for speaker-independent SI task using the HPRC dataset. Both the studies and our study use same target TVs (derived from same transformations [1]) and use both 'normal' and 'fast' rate utterances without any speaker rate matching at evaluations. The comparison is still not a perfect one given that there can be differences in the test splits used for evaluation.

Table 2: *Baseline models with HPRC dataset*

Model	Average PPMC score
Feed-forward model* [2]	0.705
CNN-BiLSTM model* [17]	0.755
Single-task BiGRNN model	0.745
Proposed Multi-task BiGRNN model	0.770

3.3. Ablation Study

We changed the weight α in the MTL model trained with algorithm 2 to explore how the phoneme learning task would help the desired SI task. Recall that α controls the amount of contribution from the phoneme prediction loss L_{ph} to the joint loss L_{joint}. Here setting $\alpha = 0$ is equivalent to the single-task model.

Table 3: *Contribution of phoneme learning task for the SI task*

	Average PPMC	Phoneme Accuracy (%)
$\alpha = 0.0$	0.743	2.25
$\alpha = 0.1$	0.762	70.60
$\alpha = 0.3$	0.766	72.53
$\boldsymbol{\alpha = 0.5}$	**0.770**	72.88
$\alpha = 0.8$	0.759	72.90
$\alpha = 1.0$	0.758	73.60

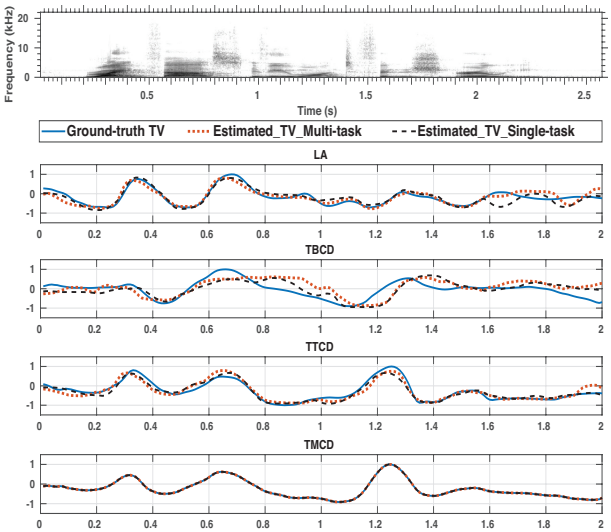

Figure 2: *LA and constriction degree TV plots for the utterance 'Write fast if you want to finish early' estimated using Multi-task model and the Single-task model. Solid blue Line - actual TV (from HPRC database), red dotted line - estimated TV from Multi-task model, black dashed Line - estimated TV from Single-task model*

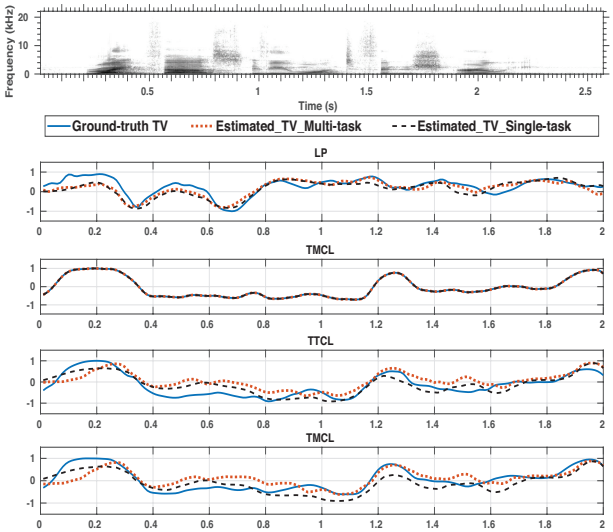

Figure 3: *LP and constriction location TV plots for the utterance 'Write fast if you want to finish early' estimated using Multi-task model and the Single-task model. Solid blue Line - actual TV (from HPRC database), red dotted line - estimated TV from Multi-task model, black dashed Line - estimated TV from Single-task model*

Table 4: *Training Time : Single-task and Multi-task models*

Model Type	No. of Trainable Parameters	Training Time
Single-task	2.19 M	10 ($\pm$2) min
Multi-task (Algo 1)	2.20 M	61 ($\pm$5) min
Multi-task (Algo 2)	2.20 M	15 ($\pm$2) min

4. Discussion

The results in Table 1 clearly confirms the impact of multi-task learning for the SI task with a relative improvement of 2.5% over the single-task model. Over the two training algorithms, algorithm 2 has a slight edge in TV prediction. However, when training time for the two algorithms are considered (table 4), algorithm 2 has a considerable advantage by only taking nearly quarter of the time of algorithm 1. Hence for the subsequent experiments and comparisons we used the MTL model trained with algorithm 2. It should also be mentioned that in a previous work with developing a multi-corpus SI system [33], a similar training procedure to algorithm 1 was used.

Figure 2 and figure 3 shows the ground-truth TVs and the predicted TVs from the multi-task and single-task models. The key difference between the two figures is that figure 2 shows the TVs which characterise the constriction degree of articulators, whereas figure 3 shows TVs which characterizes the constriction location. It is usually observed that SI systems tend to do better with constriction degree related TVs compared to ones which capture constriction location mainly due to the fact that the same speech sound can be produced with different vocal tract configurations (speaker-dependent characteristics). The same can be observed with the PPMC scores for each TV in Table 1. But an interesting observation is that the multi-task models mostly improve in estimating location related TVs with respect to the single-task model suggesting that learning the phoneme mapping is helping the SI task with additional subject-dependent information.

Table 2 shows that the proposed MTL based SI system achieves the best PPMC scores over the existing SI systems on the HPRC dataset. The results also suggest that the BiGRNN and the CNN-BiLSTM models clearly outperform the conventional feed-forward neural network models in the SI task. Moreover, with the ablation study in section 3.3, we show the importance of multi-task learning (i.e. learning a related, shared task) on improving the SI systems for TV prediction. This also suggests that with the joint loss L_{joint} optimization, an additional hyper-parameter α needs to be fine-tuned properly to achieve the best results.

Finally, it should also be highlighted that the proposed MTL based SI system only uses phoneme transcriptions for training. At the time of inference, only the acoustic features are needed and it draws the key difference between the proposed SI system and the SI systems using both phoneme and acoustic features as inputs.

5. Future Work

The lack of larger articulatory datasets for training DNN based models is a key challenge in developing generalizable SI systems. One of the envisions of developing a MTL based SI system lies with the idea of tapping into larger, existing datasets of audio, phonetic transcriptions (e.g. Librispeech [34]). The authors wish to work on transfer-learning and model adaptation paradigms to improve the current MTL framework by pre-training the models with existing corpora of audio, phonetic transcriptions.

6. Acknowledgements

This work was supported by the National Science Foundation grant #1764010

7. References

[1] G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson, "Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 316–329, 2019. [Online]. Available: https://doi.org/10.1121/1.5116130

[2] A. S. Shahrebabaki, G. Salvi, T. Svendsen, and S. M. Siniscalchi, "Acoustic-to-articulatory mapping with joint optimization of deep speech enhancement and articulatory inversion models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 135–147, 2022.

[3] C. Qin and M. Á. Carreira-Perpiñán, "An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping." *Interspeech*, pp. 74–77, 2007.

[4] J. Frankel and S. King, "Asr - articulatory speech recognition," in *INTERSPEECH*, 2001.

[5] 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 Journal on Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1027–1045, sep 2010.

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

[7] Z.-H. Ling, K. Richmond, and J. Yamagishi, "Articulatory control of hmm-based parametric speech synthesis using feature-space-switched multiple regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 207–219, 2013.

[8] K. Richmond and S. King, "Smooth talking: Articulatory join costs for unit selection," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5150–5154.

[9] S. Fagel and K. Madany, "A 3-d virtual head as a tool for speech therapy for children," in *INTERSPEECH*, 2008.

[10] C. Espy-Wilson, A. C. Lammert, N. Seneviratne, and T. F. Quatieri, "Assessing Neuromotor Coordination in Depression Using Inverted Vocal Tract Variables," in *Proc. Interspeech 2019*, 2019, pp. 1448–1452.

[11] Y. M. Siriwardena, C. Espy-Wilson, C. Kitchen, and D. L. Kelly, *Multimodal Approach for Assessing Neuromotor Coordination in Schizophrenia Using Convolutional Neural Networks*. New York, NY, USA: Association for Computing Machinery, 2021, p. 768–772. [Online]. Available: https://doi.org/10.1145/3462244.3479967

[12] J. R. Westbury, "Speech Production Database User ' S Handbook," *IEEE Personal Communications - IEEE Pers. Commun.*, vol. 0, no. June, 1994.

[13] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, vol. 31, no. 1, pp. 26–35, may 1987.

[14] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, mar 2004. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.1652588

[15] A. Illa and P. K. Ghosh, "Low Resource Acoustic-to-articulatory Inversion Using Bi-directional Long Short Term Memory," in *Proc. Interspeech 2018*, 2018, pp. 3122–3126.

[16] Aravind Illa and Prasanta Kumar Ghosh, "Speaker Conditioned Acoustic-to-Articulatory Inversion Using x-Vectors," in *Proc. Interspeech 2020*, 2020, pp. 1376–1380.

[17] A. S. Shahrebabaki, S. M. Siniscalchi, G. Salvi, and T. Svendsen, "Sequence-to-Sequence Articulatory Inversion Through Time Convolution of Sub-Band Frequency Signals," in *Proc. Interspeech 2020*, 2020, pp. 2882–2886. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1140

[18] A. Illa and P. K. Ghosh, "Representation learning using convolution neural network for acoustic-to-articulatory inversion," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5931–5935.

[19] A. S. Shahrebabaki, S. M. Siniscalchi, and T. Svendsen, "Raw Speech-to-Articulatory Inversion by Temporal Filtering and Decimation," in *Proc. Interspeech 2021*, 2021, pp. 1184–1188.

[20] S. Udupa, A. Roy, A. Singh, A. Illa, and P. K. Ghosh, "Estimating Articulatory Movements in Speech Production with Transformer Networks," in *Proc. Interspeech 2021*, 2021, pp. 1154–1158.

[21] M. Tiede, C. Y. 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, no. 5, pp. 3580–3580, 2017. [Online]. Available: https://doi.org/10.1121/1.4987629

[22] A. Singh, A. Illa, and P. K. Ghosh, "A comparative study of estimating articulatory movements from phoneme sequences and acoustic features," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7334–7338, 2020.

[23] R. Caruana, "Multitask Learning," 1997. [Online]. Available: https://doi.org/10.1023/A:1007379606734

[24] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.

[25] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2020, pp. 10 434–10 443.

[26] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4835–4839, 2017.

[27] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," in *INTERSPEECH*, 2017.

[28] C.-M. Chien, J.-H. Lin, C.-y. Huang, P.-c. Hsu, and H.-y. Lee, "Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8588–8592.

[29] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *INTERSPEECH*, 2019.

[30] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech Emotion Recognition with Multi-Task Learning," in *Proc. Interspeech 2021*, 2021, pp. 4508–4512.

[31] X. Xie, X. Liu, and L. Wang, "Deep Neural Network Based Acoustic-to-Articulatory Inversion Using Phone Sequence Information," in *Proc. Interspeech 2016*, 2016, pp. 1497–1501.

[32] "Ieee recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[33] N. Seneviratne, G. Sivaraman, and C. Espy-Wilson, "Multi-Corpus Acoustic-to-Articulatory Speech Inversion," in *Proc. Interspeech 2019*, 2019, pp. 859–863.

[34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.