ABSTRACT

Title of Dissertation: TOWARDS EXTENDING ACOUSTIC-TO-ARTICULATORY SPEECH INVERSION AND LEARNING ARTICULATORY REPRESENTATIONS

Yashish M. Siriwardena
Doctor of Philosophy, 2023

Dissertation directed by: Prof. Carol Espy-Wilson and Prof. Shihab Shamma
Department of Electrical and Computer Engineering

Acoustic-to-articulatory speech inversion involves the challenging task of deducing the kinematic state of various constriction synergies, including the lips, tongue tip, tongue body, velum, and glottis, based on their respective constriction degree and location coordinates. These coordinates are referred to as vocal tract variables (TVs). 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), speech synthesis, speech therapy, and mental health assessments.

Over the past few years, deep neural network (DNN) based models have propelled the development of SI systems to new heights. However, the current SI systems still struggle with the lack of sufficiently larger articulatory datasets, speaker dependence, poor performance with noisy speech, and the lack of generalizability across different articulatory datasets. Moreover, one of the major drawbacks of the existing articulatory datasets is the lack of ground-truth data capturing

velar and glottal activity of speech. With this work, we try to address some of the aforementioned challenges pertaining to the development of effective SI systems. Our experiments are based on two publicly available articulatory datasets; the University of Wisconsin X-ray microbeam (XRMB) dataset, and the HPRC dataset. We show that the use of appropriate audio augmentation techniques to synthetically create data can further improve the performance of SI systems both on clean and noisy speech data. We also show that the use of multi-task learning frameworks to carry out an auxiliary, but a related task can also improve the TV prediction. A key improvement came about when the SI systems were forced to learn source features (aperiodicity, periodicity, and pitch) as additional targets. Moreover, the use of self-supervised speech representations (HuBERT) and fine tuning them to the downstream task of speech inversion resulted in improved performance.

With the aim of extending the current SI systems to estimate velar and glottal activity, data from an ongoing data collection was used to derive and validate two parameters; nasalance to capture velar constriction degree and electroglottography (EGG) envelope to capture voicing. A separate speaker-independent SI system was subsequently trained to estimate the derived parameters and is one of the first systems to achieve the feat. This SI system along with the conventional SI systems (trained to estimate lip and tongue TVs), provide a framework to estimate a complete articulatory representation of speech in speaker-interdependent fashion.

While improving and extending the current SI frameworks, we also explored an unsupervised learning algorithm inspired by sensorimotor interactions in the human brain to perform audio and speech inversion. The proposed "MirrorNet", a constrained autoencoder architecture is first used to learn, in an unsupervised manner, the controls of an off-the-shelf audio synthesizer (DIVA) to produce melodies only from their auditory spectrograms. The results demonstrate how the MirrorNet discovers the synthesizer parameters to generate the melodies that closely resemble

the original and those of unseen melodies, and even determine the best set of parameters to approximate renditions of complex piano melodies generated by a different synthesizer. To extend the same idea of learning to vocal tract controls for speech, we developed a DNN based articulatory synthesizer (articulatory-to-acoustic forward mapping) to be incorporated as the motor plant of the MirrorNet. The MirrorNet with this motor plant, once initialized with a minimal amount of ground-truth data ($\sim$ 30 mins of speech), can learn the articulatory representations (6 TVs + source features) with significantly better accuracy. Overall, this highlights the effectiveness and power of the MirrorNet's learning algorithm in enabling to solve the conventional acoustic-to-articulatory speech inversion problem with minimal use of ground-truth articulatory data.

In order to assess the practical utility of articulatory representations in real-world scenarios, we employed articulatory coordination features derived from TVs to detect and analyze articulatory-level alterations in the speech of individuals with schizophrenia. We show that the schizophrenia subjects with strong positive symptoms (e.g. hallucinations and delusions), and who are markedly ill, pose a more complex articulatory coordination pattern in facial and speech gestures compared to healthy controls. This distinction in speech coordination pattern is used to train a multimodal convolutional neural network (CNN) which uses video and audio data to distinguish schizophrenia subjects from healthy controls. Furthermore, we used TVs estimated by the best performing SI system to detect mispronunciation of /ɹ/, a common speech sound disorder in children. The classification model trained with TVs performed better compared to the state-of-the-art hand-crafted age-and-sex normalized formants.

In essence, the work in this dissertation presents steps taken towards developing effective acoustic-to-articulatory speech inversion frameworks, and highlights the importance of utilizing articulatory representations in real-world applications.

TOWARDS EXTENDING ACOUSTIC-TO-ARTICULATORY SPEECH INVERSION AND LEARNING ARTICULATORY REPRESENTATIONS

by

Yashish M. Siriwardena

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Carol Espy-Wilson, Chair/Co-Advisor
Professor Shihab Shamma, Co-Advisor
Professor Behtash Babadi
Professor Jonathan Z. Simon
Professor William J. Idsardi (Dean's Representative)

Dedication

To the kind-hearted

Teachers

who illuminate the path of knowledge,

ignite the flames of curiosity,

and nurture the seeds of potential

in every student

Acknowledgments

The successful completion of my PhD journey was made possible by the invaluable support I received from numerous individuals. I wish to express my heartfelt gratitude to each and every one of them.

To begin with, I want to express my deep appreciation to my primary research advisor, Prof. Carol Espy-Wilson. Over the years, her unwavering guidance, encouragement, and motivational support have been instrumental in my academic journey. She is not only a dedicated educator but also a fervent researcher. I cannot adequately convey my gratitude for the wealth of knowledge I've gained from her, both in the area of speech signal processing, and research in general. Our discussions were consistently enlightening and enjoyable, and she played a pivotal role in cultivating my skills as a proficient researcher, always encouraging me to surpass my limits.

I would then like to thank my co-adviser, Prof. Shihab Shamma, for giving me this great opportunity to work along side him in an interesting research project that opened my eyes into the whole area of speech perception and learning. I am forever grateful for him for always having time for me and motivating and believing me to pursue a challenging, but an extremely rewarding project.

Then, I am indebted to my PhD defense committee, Prof. Behtash Babadi, Prof. Jonathan Simon and Prof. William Idsardi for accepting my request to serve in my committee and providing valuable comments and suggestions to improve this dissertation.

I am also deeply thankful to Dr. Suzanne Boyce, Dr. Mark Tiede, Dr. Philip Resnik, Dr. Ganesh Sivaraman, Dr. Deanna L. Kelly, Dr. Liran Oren, Dr. Nina Benway and Guilhem Marion for collaborating with me in a multitude of research projects. The enriching discussions I had with them immensely helped shape my dissertation to be a success.

My colleagues at the speech communication lab have enriched my graduate life in many ways and deserve a special mention. Nadee, Rahil, Ahmed, Cherif, Gowtham and Thanushi have been my close friends over the years and had always been there when I needed them the most.

I should also make this a chance to thank University of Maryland, the National Science Foundation and UMCP & UMB Artificial Intelligence + Medicine for High Impact Challenge Award for funding my graduate studies over the last few years, which have been instrumental in enabling all my research work. I am also grateful to the free education system in Sri Lanka for funding my education from kindergarten to the undergradaute studies, without which I have not been here at first place.

Being and international student away from home, I was lucky enough to have a lovely group of friends who made me feel like home. I couldn't thank enough Dilhara, Sajani, Lasitha, Anuththara and Umesha for all the fun times we had with our road trips and board game nights. I am also grateful for Samiru and Hasini for regularly inviting us over for dinners and chit-chats. I specially thank uncle Saman and his family for constantly checking on us and for their kind advises. Not only them, the entire Sri Lankan student community at UMD played a huge role in making me feel home with all the fun events we organized and the get-togethers we had.

Most importantly, I should make this a chance to pay my sincere gratitude to my late friend Vinoj Jayasundara, one of the smartest and humble human beings I have ever met in my life. He always had time for me to discuss any machine learning related problem and was my go to person

whenever I had doubts. I will always remember the fruitful discussions I had with him during our last camping trip and that indeed motivated some of the research work in this dissertation.

Finally, I am deeply grateful to my loving wife, Dushyanthi, without whom I'd never have reached this far in my life. She has been my guiding light for the past eight years, motivating me through my undergraduate studies and then with my graduate studies. I am also forever indebted to my parents, my father Premasiri and my mother Wimala, for being the pillars of my success. They will always be the most happiest people to see where I am now with my life, and my academic achievements. I also need to mention my in-laws here, for their motivation and support given all these years. Last but not least, I want to express my profound gratitude to my three brothers, Vidusha, Charuka, and Sachintha, who have consistently offered their heartfelt support and encouragement in pursuit of my dreams.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

TV	Tract Variable
SI	Speech Inversion
ASR	Automatic Speech Recognition
DNN	Deep Neural Network
EGG	Electroglottography
CNN	Convolutional Neural Network
MFCC	Mel-frequency Cepstral Coefficient
FAU	Facial Action Unit
ACF	Articulatory Coordination Feature
EMA	Electromagnetic Articulometry
rt-MRI	real-time Magnetic Resonance Imaging
PLP	Perceptual Linear Prediction
AP	Articulatory Phonology
TADA	Task Dynamic model of speech production
XRMB	X-ray Microbeam
LA	Lip Aperture
LP	Lip Protrusion
TBCL	Tongue Body Constriction Location
TBCD	Tongue Body Constriction Degree
TTCL	Tongue Tip Constriction Location
TTCD	Tongue Tip Constriction Degree
HPRC	Haskins Production Rate Comparison
JA	Jaw Angle
TMCL	Tongue Middle Constriction Location
TMCD	Tongue Middle Constriction Degree
TCN	Temporal Convolutional Networks
VRM	Vicinal Risk Minimization
MTL	Multi-task learning
TTS	Text-to-Speech
SER	Speech Emotion Recognition
DDSP	Differentiable Digital Signal Processing
GMM	Gaussian Mixture Models
EM	Expectation Maximization
PMR	Psychomotor retardation

MDD	Major Depressive Disorder
TDEC	Time-Delay Embedded Correlation
FVTC	Full Vocal Tract Coordination
HAN	Hierarchical Attention Networks
BERT	Bidirectional Encoder Representations from Transformers
KD	Knowledge Distillation
GRU	Gated Recurrent Unit
MSE	Mean Square Error
MAE	Mean Absolute Error
PPMC	Pearson Product Moment Correlation
SSL	Self-supervised Learning
HuBERT	Hidden Unit Bidirectional Encoder Representations from Transformers
KP	Knowledge of Performance

Chapter 1: Introduction

Acoustic-to-articulatory speech inversion (SI) is a highly non-linear and a non-unique problem where the input acoustic signal is usually mapped to a set of articulatory parameters. Speaker-wise dependencies in speech make it more challenging and if not impossible to make complete speaker-independent SI systems. However, accurate estimation of articulatory parameters can hugely benefit speech applications like Automatic Speech Recognition (ASR) (Frankel and King, 2001; Mitra et al., 2010, 2011), speech synthesis (Ling et al., 2013; Richmond and King, 2016), speech therapy (Fagel and Madany, 2008) and speech in mental health assessments (Espy-Wilson et al., 2019; Siriwardena et al., 2021a). With the advent of deep neural networks (DNNs), the SI systems have reached new heights which the early parametric models could not achieve for decades. However, more work needs to be done to improve the generalizability of these DNN based SI systems while also addressing the key issues in robustness, susceptibility to domain differences, and over dependence on large volumes of ground-truth articulatory data.

One of the other key limitations of the current SI systems is their inability to estimate velar and glottal activity of speech. A prime reason for that is the lack of publicly available articulatory speech datasets that have direct articulatory level data capturing the velar and glottal constrictions (Tiede et al., 2017; Westbury, 1994a). Therefore, most of the available SI systems (trained on the available articulatory datasets) are limited to estimating the articulatory level information pertaining

to lip and tongue constrictions (Illa and Ghosh, 2018; Shahrebabaki et al., 2020; Siriwardena and Espy-Wilson, 2023; Siriwardena et al., 2023a; Udupa et al., 2021). Exploring alternative techniques (eg. nasalance for velar constriction degree and electroglottography for glottal activity) to capture the velar and glottal activity could provide ground-truth data to train independent SI systems to estimate the velar and glottal activity. Given the importance of voicing and velar constriction information, formulating methods to derive glottal and velar TVs and incorporating them to the existing SI framework can further improve the applications of acoustic-to-articulatory speech inversion.

1.1 Acoustic-to-Articulatory Speech Inversion

Speech production in humans is a coordinated process of generating sounds which involves the speech articulators like glottis, velum, tongue, lips, teeth and the jaw. The position, shapes and sizes of these articulators change slowly over time to produce desired speech sounds. The vibration of vocal folds in the glottis or the lack of it determines the periodicity of the the generated sound. The vocal tract which consists of the velum, tongue, lips and teeth acts like an acoustic tube which modulates the glottal source waveform.

Acoustic-to-articulatory speech inversion inspired by Articulatory Phonology (Browman and Goldstein, 1992) maps the acoustic speech signal to the kinematic state of each constriction synergy (lips, tongue tip, tongue body, velum, and glottis) by its corresponding constriction degree and location coordinates, which are called vocal tract variables (TVs). This mapping from acoustics to articulation is an ill-posed problem which is known to be highly non-linear and non-unique (Qin and Carreira-Perpiñán, 2007). Ground-truth articulatory data are collected

by techniques like X-ray microbeam (Westbury, 1994a), Electromagnetic Articulometry (EMA) (Schönle et al., 1987) and real-time Magnetic Resonance Imaging (rt-MRI) (Narayanan et al., 2004). All these methods are expensive, time consuming and need specialized equipment for observing articulatory movements directly (Sivaraman et al., 2019). This explains why developing a speaker-independent SI system that can accurately estimate articulatory features for any unseen speaker is of greater need.

1.2 Sensorimotor Learning inspired Audio and Speech Inversion

Sensorimotor interactions play a fundamental role in performing complex tasks engaging visual, auditory or somatosensory perceptual systems and is vital in executing motor actions (reaching, speaking and lifting etc.) (Keller et al., 2012; Wolpert and Ghahramani, 2000). Previous studies on the human cortical speech system with ECoG recordings have shown that learning complex sensorimotor mappings proceeds simultaneously and in an unsupervised fashion by listening and speaking all at once (Shamma et al., 2020). Based on these studies on the human cortical speech system, a novel autoencoder architecture was proposed by Shamma et.al (Shamma et al., 2020) which was called the 'Mirror Network' (MirrorNet). The essence of this biologically motivated algorithm is the bidirectional flow of interactions ('forward' and 'inverse' mappings) between the auditory and motor responsive regions, coupled to the constraints imposed simultaneously by the actual motor plant to be controlled.

The idea of the MirrorNet can be generalized beyond speech production and applies to any motor plant for which the control parameters need to be learned. In other words, MirrorNet can learn the parameters to drive a given motor plant (audio synthesizer, vocal tract model etc.)

to best estimate a given input in a completely unsupervised fashion. Given that there is limited ground-truth articulatory data to train SI systems, harnessing the potential of such algorithms to learn articulatory representations in a completely unsupervised or semi-supervised fashion can be extremely helpful.

1.3 Articulatory representations for assessing mental health and detecting child speech sound disorders

Apart from pushing the boundaries of the current SI systems, it is extremely important to explore the avenues where articulatory representations can be useful. To that end, the estimated articulatory representations from the best performing SI systems were utilized as acoustic features.

Articulatory coordination features (ACFs) derived from vocal tract variables have achieved state of the art results in depression detection and severity prediction for publicly available depression datasets (Seneviratne et al., 2020). Schizophrenia is another mental health disorder which affects around 60 million (1%) of the world's adult population (Kuperberg, 2010) and TV based ACFs have shown to be robust, as well as effective in detecting subjects with strong positive symptoms (eg. hallucinations, delusions) in schizophrenia (Siriwardena et al., 2021b). The changes in neuromotor coordination induced by these mental health disorders can be accurately captured using TV based ACFs with respect to other acoustic features (MFCCs, formants etc.) which have previously been used as proxies for measuring articulatory coordination. Hence, any improvement to the acoustic-to-articulatory SI system predicting TVs can in-turn improve the existing models (unimodal and multimodal systems) for depression and schizophrenia detection.

It is theorized that the relatively high prevalence of chronic speech sound disorders (Benway

and Preston, 2023) to be due in part to a treatment intensity gap in which common clinical practice does not typically offer best-practice treatment at the intensity specified in the evidence base. Recent research has developed computerized intervention that automates clinician-led best practices, and may be able to address this intensity gap (Maas et al., 2008); however, automated systems are limited in the clinical-grade feedback they can provide learners. Within a motor-learning framework, feedback can refer to a binary perceptual judgment (i.e., knowledge of results) or detailed reference to the auditory and somatosensory targets as well as corrective instructions to improve the next attempt (i.e., knowledge of performance; KP) (McKechnie et al., 2020). When delivered by a clinician, somatosensory KP may be based on what the clinician sees, or, in some cases based on a known set of clinical cues. Current computerized treatment systems either use clinician-deliver KP (Shadle et al., 2016) or approximate KP through random selection from a set of plausible clinical cues for the target phone. However, a mispronunciation detection system that leverages acoustic-to-articulatory speech inversion (SI) may be able to provide KP that is currently unattainable with current state-of-the-art formant-based systems, as well as circumvent known issues with child formant estimation.

1.4 Objectives of this Study

1. Improving the performance of existing state-of-the-art speech inversion systems (chapter 3) : To achieve this objective we explore multiple DNN model architectures with different input acoustic features. We also incorporate data augmentation techniques, multi-task learning frameworks and source features to experiment different means of improving the performance of SI systems

2. Extending the speech inversion systems to estimate velar and glottal activity (chapter 4) : Towards this objective we first derive parameters which can ideally capture the velar and glottal activity and validate that through gold standard measures. Data from an ongoing data collection is used to design and develop a novel SI system to estimate the derived parameters in speaker-independent fashion.

3. Exploring sensorimotor learning inspired algorithms to improve speech inversion (chapter 5) : This chapter discusses a sensroimotor learning inspired algorithm and a model architecture which can be used for audio and speech inversion with minimal exposure to ground-truth data. The algorithm is first explored with an audio/music synthesizer to synthesize unseen melodies of notes and then extended to estimate articulatory representations with a custom-trained DNN based articulatory synthesizer.

4. Application of articulatory representations estimated from speech inversion systems (chapter 6) : It is of paramount importance to explore where the articulatory representations can be used as a speech feature to enhance the performance of real-world speech applications. To this end we explore two applications; (1) using articulatory representations to detect speech changes manifested with schizophrenia, (2) detecting child speech sound disorders

Chapter 2: Background

2.1 Acoustic-to-Articulatory Speech Inversion

Acoustic-to-articulatory speech inversion is the inverse problem of retrieving articulatory dynamics for a given speech signal (Sivaraman et al., 2019). Attempts at recovering articulatory movements from the continuous speech signal have a long history (Papcun et al., 1992b), but have generally limited to identifying the movement of a specific set of articulators like upper and lower lip, tongue tip, velum closure, etc. However, it is important to derive not specifically from individual articulator movements, but from synergies among articulators;lips and jaw as individual articulators share a goal, and work together to achieve a desired vocal tract shape (Saltzman and Munhall, 1989). Hence, general speech inversion (SI) systems focus on recovering the vocal tract constriction, rather than the movement of individual articulators.

Following the initial work of Saltzman and Munhall (1989) and Browman and Goldstein (1992), Sivaraman et al. (2019) developed a SI system that estimates the constriction degree and location of functional tract variables (TVs), that is, the context-dependent synergies between articulators. These reflect the joint influence of articulators on the vocal tract shape. With this approach, constriction formation or release by five distinct constrictors (lips, tongue tip, tongue body, velum, and glottis) along the vocal tract are represented by the estimated articulatory 'gestures'. These time dependent functions or tract variables (TVs) are sensitive to changes in

constriction degree, constriction location, and also for changes in timing, so that the gestures can move apart or overlap in time.

By empirical studies of (Qin and Carreira-Perpiñán, 2007; Stevens, 1989), it has been found that the mapping from acoustics to articulatory space is highly non-linear and non-unique. The non-linearity of the acoustics-to-articulatory mapping is evident from the quantal nature (Stevens, 1989) of the articultory-to-acoustics relation. For certain ranges of articulatory parameter change, there can be very little change in the acoustic parameter, whereas for some other range, the acoustic parameter is more sensitive to changes in articulation (Stevens, 1989). Moreover, it has also been observed that similar acoustic consequences can result in from completely different articulatory configurations, which makes the SI task a one-to-many mapping, and hence a non-unique solution. Maeda (1990) and Gunther et al. (Guenther et al., 1999) have explained the reasons for this non-uniqueness based on the coordinated compensatory movements of the articulators to achieve acoustic targets in different contexts. One perfect example for this non-uniqueness mapping is the existence of two distinct vocal tract configurations (bunched and retroflexed) for the American English /r/ sound (Zhou et al., 2008) when we consider the first 3 formants of speech.

General speech inversion pipeline includes a step for extracting acoustic features from the speech signal. Section 2.1.2 discusses different types of acoustic features widely used in the SI task. Then, an inverse mapping is learned by associating these acoustic features through training on a corpus of matched acoustic and observed articulatory data. These DNN based SI systems are capable of real-time inversion and the models are typically trained with the use of Graphical Processing Units (GPUs) for faster machine learning. However, most research in SI has been focused on developing speaker-dependent systems. Previous work with codebook search (Ghosh and Narayanan, 2010), feed-forward neural networks and Mixture Density Networks (Richmond,

2006) have all looked into developing speaker-dependent SI systems. The attempts of developing speaker-independent systems (Afshan and Ghosh, 2015; Girin et al., 2017; Ji, 2014; Ji et al., 2016) have mostly limited to two speakers from the MOCHA-TIMIT dataset (Wrench, 2000). Recent work with BiLSTM models (Illa and Ghosh, 2018, 2019b) and transformer models (Udupa et al., 2021) for SI task have tried pooling a percentage of all the subject's data for training. Given that they use a portion of target subject's data at training, the system can not be considered a completely speaker-independent system.

2.1.1 Speaker-independent acoustic-to-articulatory speech inversion

Sivaraman et al. (2019) developed a speaker-independent SI system using a simple feed-forward neural network architecture. The block diagram in figure 2.1 shows the steps involved in the speech inversion system. Mel Frequency Cepstral Coefficients (MFCCs) are used as the acoustic features and the 13 cepstral coefficients are extracted from a 20ms Hamming analysis windows with a 10ms frame shift. To counter the inter-speaker variations in measurements of acoustics and articulatory data, speaker-wise mean and variance normalization is done on both input MFCCs and the artiuclatory data. As articulatory data, the system uses vocal tract variables (TVs) which is also a relative measure computed using the absolute articulatory measurements (X-Y positions of the pallet data). As shown in figure 2.1, the MFCCs are contextualized by concatenating every other feature frame in a 340ms window. This results in 8 frames of MFCCs on either side of each frame being concatenated to form the contextualized MFCC features (total of 17 frames including the current frame). By skipping two frames when splicing the frames, the current analysis frame is centered when concatenating every other frame. The optimal frame size

of 17 frames was empirically found by Mitra et al. (2010). The SI model is trained to minimize the mean squared error between the actual TVs and the estimated TVs. The estimated TVs from the DNN model are finally low-pass filtered using a Kalman filter since the feed-forward only layers are unable to produce smoothed TVs. The best performing SI system consists of 5 hidden layers with 1024 neurons in each layer.

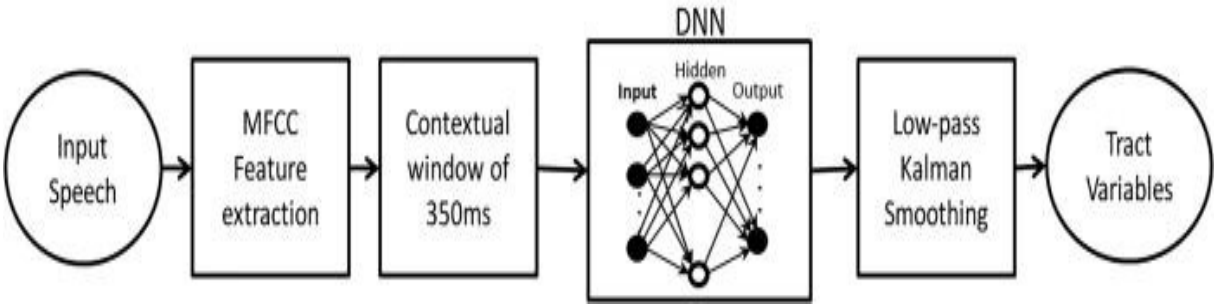

Figure 2.1: Block diagram of the speech inversion system (Sivaraman et al., 2019)

2.1.2 Acoustic features for speech inversion

In building a better SI system, it is imperative that we use the most effective representations for the acoustic signal. Mitra et al. (2010) and Sivaraman et al. (2019) explored different acoustic features like Mel-Frequency Cepstral Coefficients(MFCCs), Perceptual Linear Prediction (PLP) and mel-spectrum (MELSPECT) as inputs for the speech inversion task. The results suggest that the 13-MFCCs perform better over the other acoustic features in estimating the TVs. With the advent of deep neural networks (DNNs), people have also tried using convolutional neural network (CNN) layers to process the waveform itself to learn the best representation for the downstream SI task (Illa and Ghosh, 2019b).

Self-Supervised Learning (SSL) has proven to be a highly effective approach for enhancing the performance of DNNs by leveraging unlabeled data to acquire speech representations in

learning (Hsu et al., 2021; Schneider et al., 2019). These representations have demonstrated their effectiveness across various applications, including Automatic Speech Recognition (ASR) systems (Conneau et al., 2020), speech separation, and enhancement (Wang et al., 2023). Recent research has also shown that SSL speech representations can significantly boost the performance of SI models when dealing with Electromagnetic Articulography (EMA) data (Cho et al., 2023), surpassing the traditional acoustic features such as Mel-frequency Cepstral Coefficients (MFCCs). In their extensive evaluation of existing SSL speech representations for the SI task, Cho et al. (2023) found that HuBERT-based SSL speech representations (Hsu et al., 2021) outperform other SSL features like wav2vec2 (Baevski et al., 2020) and tera (Liu et al., 2021), as well as conventional acoustic features such as MFCCs.

2.1.3 Vocal tract variables (TVs)

According to Articulatory Phonology (AP), lexical units in speech are described as events that unfold during the process of speech production (Browman and Goldstein, 1992). These events are commonly known as articulatory 'gestures' which can be observed in the movement of speech articulators. Speech in AP is discussed as a constellation of overlapping gestures, which is different from the phonetic view of speech which describes speech as beads on a string with phones as individual static units of speech.

The notion of vocal tract variables or Tract Variables (TVs) was defined to characterize the kinematic state of articulatory constrictors by its corresponding degree and location coordinates. The Task Dynamic model of speech production (TADA) uses ARPABET transcription of an English word to compute the gestural scores and the inter-articulatory gestural coordination

(Sivaraman et al., 2017). The time functions of the TVs (degree and location variables of the constrictors) are then produced as outputs. Table 2.1 lists all the 8 TVs defined by the TADA model along with their associated constrictors. Even though the TADA model provides a sound theoretical framework for speech production, the amount of variability observed in real speech is not accurately modeled by the synthetic speech and hence the TVs produced by the TADA model. That is where deep neural network (DNN) based acoustic-to-articulatory speech inversion systems come into play to estimate TVs directly from natural speech.

Table 2.1: Constriction organ and corresponding tract variables (from (Browman and Goldstein, 1992) and (Sivaraman et al., 2017))

Constrictors	**Vocal tract variables (TVs)**
Lip	Lip Aperture (LA) Lip Protrusion(LP)
Tongue Tip	Tongue tip constriction degree (TTCD) Tongue tip constriction location (TTCL)
Tongue Body	Tongue body constriction degree (TBCD) Tongue body constriction location (TBCL)
Velum	Velum (VEL)
Glottis	Glottis (GLO)

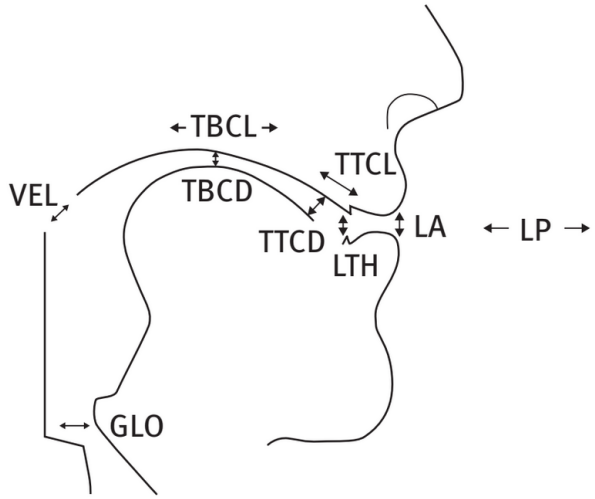

Figure 2.2: Visual representation of the vocal tract variables at five distinct constriction organs (taken from (Saltzman and Munhall, 1989)), along with a listing of constrictors and their vocal tract variables. See Table 2.1 for TV labels

2.1.4 Articulatory datasets

The X-ray microbeam (XRMB) articulatory dataset

The original University of Wisconsin XRMB database (Westbury, 1994b) comprises of naturally spoken isolated sentences and short read paragraphs collected from 32 male and 25 female subjects. These speech utterances were recorded along with trajectory data captured by X-ray microbeam cinematography of the midsagittal plane of the vocal tract using pellets placed on several articulators: upper (UL) and lower (LL) lip, tongue tip (T1), tongue blade (T2), tongue dorsum (T3), tongue root (T4), mandible incisor (MANi), and (parasagittally placed) mandible molar (MANm). However, some of the articulatory recordings were marked as mistracked in the database and eliminating these samples left us with 46 speakers (21 males and 25 females) with a total of around 4 hours of speech data.

The anatomy of the speaker's vocal tract defines the absolute positions of the articulators. Since the the X-Y positions of the pellets strongly depend on the anatomy of the speakers and variability of pellet placements, the measurements can vary significantly across speakers. Hence, to better represent vocal tract shape, relative measures were used to calculate the TVs from the X-Y positions of the pellets. TVs lead to a relatively speaker independent representation of speech articulation and characterize salient features of the vocal tract area function (McGowan, 1994). The TVs also provides a theoretical framework for speech production analysis and articulatory phonology (Browman and Goldstein, 1992). Thus using geometric transformations defined in the TADA, the XRMB trajectories were converted to TV trajectories as outlined in (Mitra et al., 2012). The transformed XRMB database comprises of six TV trajectories: Lip Aperture (LA), Lip Protrusion (LP), Tongue Body Constriction Location (TBCL), Tongue Body Constriction Degree

(TBCD), Tongue Tip Constriction Location (TTCL) and, Tongue Tip Constriction Degree (TTCD). Figure 2.3 shows a rough schematic of how the X-Y positions of the pallets in XRMB dataset is transformed to TV trajectories.

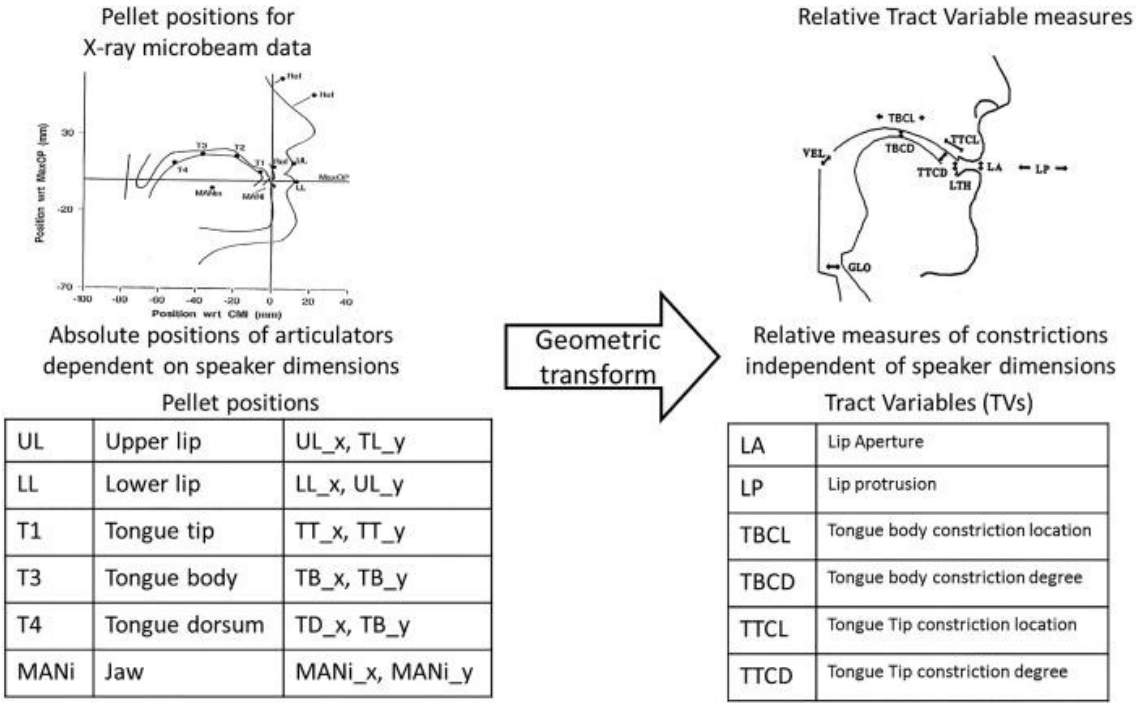

Figure 2.3: Transformation of XRMB database from pellets to TV trajectories (Sivaraman et al., 2017)

The Haskins Production Rate Comparison (HPRC) articulatory dataset

The Haskins Production Rate Comparison (HPRC) database contains recordings from 4 female and 4 male subjects reciting 720 phonetically balanced IEEE sentences (IEE, 1969) at normal and fast production rates (Tiede et al., 2017). 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 previous work (Seneviratne et al., 2019; Sivaraman et al., 2019).

2.1.5 Deep learning model architectures for speech inversion

Advancements in deep neural networks (DNNs), especially in processing time series data to capture contextual information has propelled the development of SI systems to new heights. Bidirectional LSTMs (BiLSTMS) (Aravind Illa and Prasanta Kumar Ghosh, 2020; Illa and Ghosh, 2018), CNN-BiLSTMs (Illa and Ghosh, 2019b; Shahrebabaki et al., 2020), Temporal Convolutional Networks (TCN) (Shahrebabaki et al., 2021) and transformer models (Udupa et al., 2021) have gained state-of-the-art results with multiple articulatory datasets (Tiede et al., 2017). Sivaraman et al. (2019) reported the state-of-the-art SI results with the XRMB dataset, but with a rather simple feed-forward neural network, using manually contextualized MFCCs as input features and doing speaker adaptation with vocal tract length normalization.

2.1.6 Audio data augmentation and noise robust speech inversion

Most of the available SI systems are usually trained with a single corpus of data and have shown to perform poorly in cross-corpus (Seneviratne et al., 2019) or speaker-independent experiments (Shahrebabaki et al., 2020). Difficulty collecting a larger corpus of data with many subjects and the differences in procedures of data collection (placement of sensors, EMA vs XRMB) have also made things difficult. To address the issue of data scarcity, the machine learning community has recently turned in the direction of data augmentation and generating synthetic data to train DNN models. The idea of training a DNN model on similar but different examples is known as data augmentation, which was initially proposed in (Simard et al., 2012) and later formalized by the Vicinal Risk Minimization (VRM) principle in (Chapelle et al., 2000). Data augmentation is widely utilized in automatic speech recognition (ASR) e.g. (Jaitly and Hinton, 2013; Kanda et al., 2013; Ko et al., 2015; Park et al., 2019) and it was recently explored in speech emotion recognition (Pappagari et al., 2021). In previous work with developing a noise robust articulatory SI system, Seneviratne et al. (2018) explored the idea of generating additive noise audio data. The goal of that work was to develop a SI system which is robust against noisy acoustic data which is a common issue with rt-MRI articulatory datasets that contain significant amounts of MRI machine noise.

2.1.7 Multi-task learning for speech inversion

The idea of Multi-task learning (MTL) was formally presented by Caruana (1997) 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 (Caruana, 1997). 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 (Zhang and Yang, 2021).

MTL has widely been used in computer vision, and a recent work (Lu et al., 2020) 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 (Hori et al., 2017; Kim et al., 2017), text-to-speech (TTS) (Chien et al., 2021) and in speech emotion recognition (SER) (Cai et al., 2021; Li et al., 2019). Cai et al. (2021) recently presented the state-of-the-art results for the SER task with IEMOCAP dataset, where a MTL based model was used.

2.1.8 Incorporating source features to improve speech inversion

Proxy source features which capture the glottal activity have previously been used as input features to improve the performance of speech applications (Seneviratne et al., 2020). These features primarily capture aperiodicity, periodicity and pitch in speech and hence are called source features. Section 2.2 discusses in brief the techniques used to capture the glottal activity and the tools which are built on these ground-truth data to estimate the aforementioned source features.

MTL frameworks discussed in section 2.1.7 differ from conventional supervised learning frameworks which tries to incorporate additional features, like the source features to the same input space. MTL frameworks work by incorporating additional features or tasks to the target output space, thereby solving the need for any additional parameters used at the time of inference. To the best of our knowledge, none of the previous work has experimented with using any sort of source features either in the input space or the output space to explore how it could benefit the overall SI task.

2.2 Extending speech inversion frameworks to estimate velar and glottal activity

Laryngography (Baken, 1996b) or electroglottography (EGG) (Rothenberg and Mahshie, 1988) uses two electrodes which are placed on the throat of the speaker, positioned on each side of the thyroid cartilage. A weak, constant voltage is passed from one electrode to the other allowing the current to fluctuate with the contact variations between the vocal cords. The resulting signal called the laryngographic or electroglottographic signal includes high frequencies and resembles the actual audio of the speech signal (Toutios and Margaritis, 2003). The electroglottographic signal has successfully been used for pitch tracking (Deshmukh et al., 2005; Hui et al., 2015) and to determine places of voicing in speech (Mandal et al., 2020).

The action of the velum can not be easily observed through visual means, and there is no significant proprioceptive feedback associated with velar movements. The velopharyngeal (oronasal) passageway needs to be opened when producing nasal consonants like /m/ or /n/ in English. Methods for monitoring velopharyngeal closure during speech have been studied extensively by Baken (1996a). These various methods can broadly be divided into two types based

on the aspect of velar control being measured : (i) methods that measure velar control during those consonants requiring an oral pressure buildup (e.g., stops), (ii) methods that measure velar control during vowels and sonorant consonants. Baken (1996a) has also shown that the methods of type (ii) are more difficult to be implemented successfully.

Fletcher and Frost (1974) have used the word 'nasalance' to describe different measures of the balance between the acoustic energy at the nares (A_n) and the acoustic energy at the mouth (A_o) when producing voiced speech. They used the simple ratio (A_n/A_o) referred to as 'Nasalance Ratio' (NR) or ($A_n/(A_o + A_n)$) which was referred to as 'Nasalance' as metrics. In recent practice, clinicians and speech researchers have reported nasalance values based on particular combination of microphone type, microphone placement, separator dimensions, and bandpass filter parameters with a standard Nasometer. Pneumotachography is a known technique for measuring the nasal and oral airflow velocities which has widely been used for the diagnosis of respiratory problems such as asthma. Toutios and Margaritis (2003) presents the idea of using that technique to compute an articulatory variable to capture the nasal activity of speech.

The acoustic-to-articulatory SI systems discussed so far have been trained and validated against articulatory 'ground truth' primarily through direct observation of tongue and lip movement. However, these oral and labial articulators account for only part of the articulatory contrasts used in language. Previous work with incorporating information about nasalization from acoustics have shown to improve results in applied systems for automatic speech recognition (Pruthi and Espy-Wilson, 2007, 2004). Similarly, recent work in depression detection by Seneviratne et al. (2020) has shown that the incorporation of glottal activity by aperiodicity and periodicity measures improved the absolute accuracy by around 8%.

2.3 Speech synthesis from articulatory variables

Articulatory variable based speech synthesis or the process of generating acoustic speech signals from parameters which capture the human speech production process has been a widely researched area in speech processing (Kröger and Birkholz, 2008; Scully, 1990). Previous work has also elaborated that articulatory parameters or gestures can be used to synthesize continuous, co-articulated and intelligible speech which can replicate realistic models of the vocal tract (Assaneo et al., 2013; Birkholz et al., 2006; Dang and Honda, 2004; Engwall, 2003; Kello and Plaut, 2004; Maeda, 1982, 1990; Sondhi and Schroeter, 1987; Toda et al., 2008). Articulatory based speech synthesizers can be mainly divided in to two categories: physical or geometrical approaches which model the geometry of the vocal tract along with its physical properties (Assaneo et al., 2013; Story, 2013; Toutios et al., 2011), and the machine learning based approaches (Bocquelet et al., 2014b; Toda et al., 2008) which learn the non-linear relationships between articulatory and acoustic data. Machine learning approaches typically need large articulatory-acoustic data for model training where as geometric and physical models require high computation power and are comparatively slower than machine learning models at real-time synthesis or inference (Bocquelet et al., 2016). However, the machine learning based articulatory synthesizers also have the limitation of speaker dependence where a synthesizer trained with one subject's articulatory data will not produce intelligible or as good acoustic outputs for a different speaker whose data has not been used during training (Bocquelet et al., 2016).

Most of the machine learning based articulatory synthesizers do not learn a direct mapping from the articulatory variables to the acoustic waveform. But, they actually learn an intermediate representation or an acoustic feature (MFCCs, Melspectrogram etc.) which will then be used

as input to an off-the-shelf vocoder (parametric vocoders like World (MORISE et al., 2016) or neural vocoders like LPCNet (Valin and Skoglund, 2018) for example) to synthesize the final waveform (Bocquelet et al., 2016; Georges et al., 2020; Illa and Ghosh, 2019a). One key limitation of these vocoder based approaches is that the DNN model only learns a mapping from articulatory variables to the filter level parameters like MFCCs, and the source level information like pitch and aperiodicity (voicing) are directly extracted from the original waveform itself and then fed to the vocoder at the time of synthesis (Georges et al., 2020; Illa and Ghosh, 2019a).

2.4 Sensorimotor learning inspired audio and speech inversion

With the goal of characterizing sensorimotor interactions in the human cortical speech system, Shamma et al. (2020) recorded and analyzed the sensorimotor neural interatcions with electrocorticography (ECoG) in humans while they spoke (S), listened (L), or simulated speaking by moving their articulators in the vocal tract without producing any sound (M). The experiment was designed primarily with the goal of characterising the nature of spectral and temporal representations of the auditory and motor cortical responses. The observations of this experiment resulted in a computational architecture simulating the sensorimotor interactions and clarifying their functional role in action and perception.

During speaking, motor areas control the vocal tract and its articulators to generate the desired speech signal. However, it is also proposed that certain motor cortical areas send a parallel internal neural copy of the speech signal to the auditory cortex, which is often discussed as the 'forward' prediction signal and is compared with the responses induced by the listened speech signal (Hickok and Poeppel, 2007). On the other hand during listening, and 'inverse' mapping

from the auditory to the motor areas is expected to create a motor representation of the acoustic signal (Wilson et al., 2004). It is this bidirectional flow of interactions between the auditory and motor responsive regions (L and M in figure 2.4), defined the phenomenological network as the 'Mirror Network' (MirrorNet) in (Shamma et al., 2020).

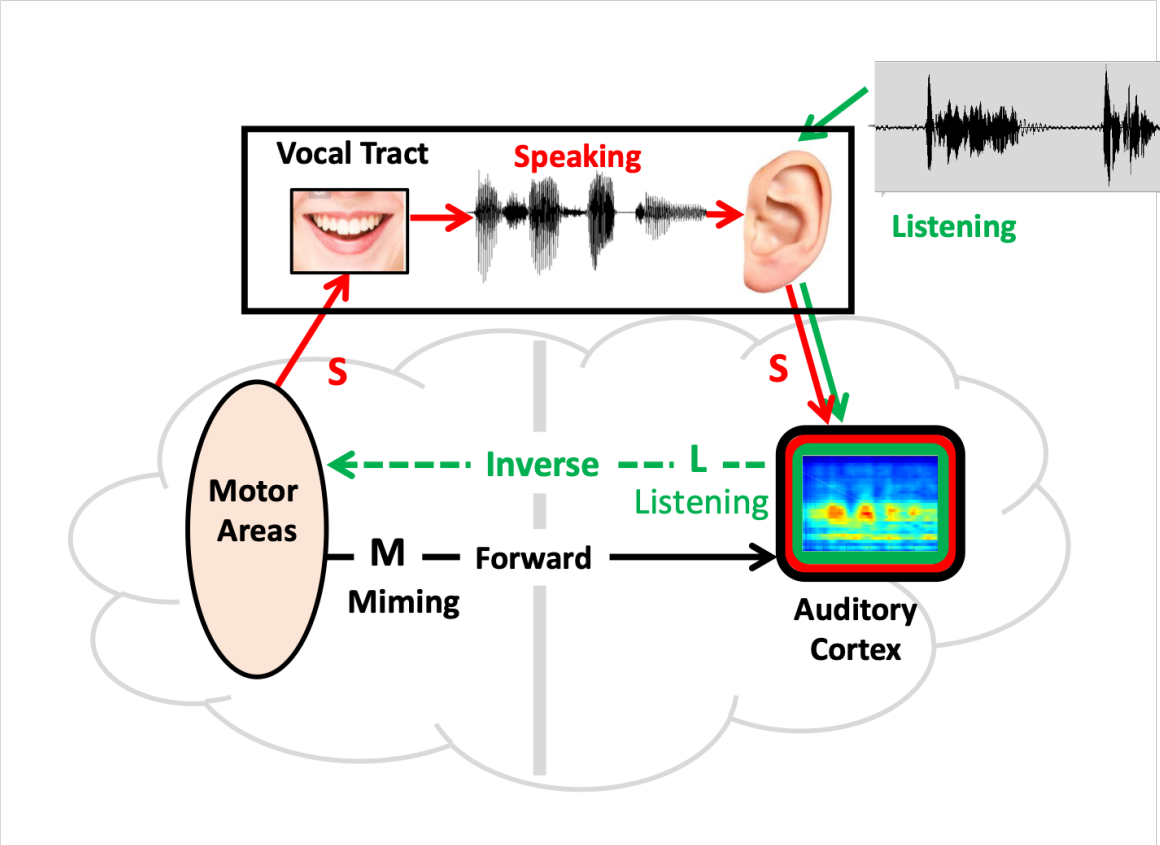

Figure 2.4: Schematic depicts the four types of recordings. Miming (M), Speaking (S) and Listening (L). Taken from 'Learning Speech Production and Perception through Sensorimotor Interaction' by (Shamma et al., 2020)

In Shamma et al. (2020), a preliminary work was also done to simulate the proposed MirrorNet model with the goal of inverting a given speech input to a set of control parameters that can be used to approximate the input speech signal. This simulation has shown its ability to learn the control parameters to drive a given speech synthesizer (World vocoder (MORISE et al., 2016)) in a completely unsupervised fashion.

2.4.1 Learning audio synthesizer controls to drive audio synthesizers

Previous work with DNNs on determining music and speech synthesizer controls are all based on at least partially supervised techniques which often involve large databases of audio and control parameter pairs (order of 1000s) (Esling et al., 2020; Georges et al., 2021; Le Vaillant et al., 2021; Yee-King et al., 2018). Furthermore, previous efforts have mostly demonstrated the ability to compute the controls for single notes or single vowels for speech (Esling et al., 2020; Saha and Fels, 2020).

Differentiable Digital Signal Processing (DDSP) based models (Engel et al., 2020a,b) have recently gained a lot of attention as a method of learning control parameters to synthesize a given acoustic signal. However, the MirrorNet is fundamentally different from DDSP based models which effectively learn a differentiable music/audio synthesizer, whereas the goal of the MirrorNet is to learn controls to drive a given non-differentiable, off-the-shelf music/audio synthesizer.

2.4.2 Learning articulatory representations in semi-supervised fashion

The idea of determining articultory parameters to control a vocal tract model was explored by Yoshikawa et al. (2003), where the idea was related to a parrot-like teaching, based on the studies of mother-infant interactions. Westerman and Miranda (2002) proposed a model which uses prototypical representations for motor commands (two motor parameters, jaw opening and the position of the styloglossus muscle) and auditory stimuli (2 formants extarcted from the audio) to learn the perception-action coupling which leads to the development of the ability to imitate sounds. The input audio used to learn the coupling between audio and motor features is a sequence of vowels generated by the 'vocal model', which is a structure of chained short pipes,

each representing a different section of the compound resonating system. The 'hearing model' which learns the inverse mapping uses a simple Hebbian-based weight update. Westerman et al. also claims that the model suggests an account of mirror neurons that have, for example been found in pre-motor cortex of the monkey experiments (Gallese et al., 1996).

Moulin-Frier et al. (2014) showed in their work, that doing an intrinsically motivated spontaneous exploration first would allow the learner to autonomously structure its own vocal experiments. They used the DIVA vocal tract model (Guenther, 2006-03-01) as the motor plant for which the controls are learned using the first two formants and an intensity parameter extracted from audio. They propose an internal sensorimotor model using Gaussian Mixture Models (GMMs) and optimize it using the Expectation Maximization(EM) algorithm. The model undergoes two stages of training; In the initial self-exploration phase the model learns from a non-speech like sounds which shares the idea of infants learning to speak. During this phase the vocal learner first discovers how to control phonation and focuses on vocal variations of unarticulated sounds. Then the model is exposed to babbling with articulated proto-syllables for which the model will automatically discover right motor parameters. Once the vocal learner becomes more proficient, the idea is that the model is able to imitate vocalizations of peers to provide high learning progress where the model shifts from self-exploration to vocal imitation.

Warlaumont et al. (2013) used neural networks with a reinforcement mechanism to simulate the vocal motor development in infants to learning to control the process of phonation. The idea of using a reinforcement mechanism aligns with the social supervision a child gets during the phase of learning to speak. The model uses the vocal tract model in (Boersma, 1999) and uses formant frequencies as auditory features. The neural network model is based on the combination of self-organizing topographic map learning and reinforcement gating. Recently, Chen et al. (2021)

explored the same idea with the Maeda vocal tract model (Maeda, 1982) and used an initialization step with ground truth articulatory parameters. During the subsequent unsupervised learning of the model, the inverse and forward mappings are learned at the same time. The model is only able to synthesize a limited set of vowels by only learning tongue and jaw parameters of the Maeda vocal tract model. Similarly, Philippsen et al. (2014) tried to learn the articulatory parameters to a commercially available vocal tract model by first learning the inverse and forward mapping with a supervised approach and then learning the model in an unsupervised approach, inspired by the sensorimotor learning paradigm. The smaller set of data used for supervised learning has been synthetically synthesized by the same vocal tract model, by manually designing the articulatory parameters to produce a sequence of phonemes.

For the first time, Georges et al. (2022) developed a model to learn vocal tract variables using the sensorimotor learning paradigm where a DNN based articulatory speech synthesizer is used as the motor plant. The inverse mapping of the model is implemented by an LSTM model where it takes in MFCCs as the input acoustic feature. The forward mapping is implemented with feed-forward layers similar to the architecture used in the articulatory speech synthesizer. The DNN based speech synthesizer is speaker-dependent (unlike the parametric model discussed in previous work) and hence the forward and inverse models converge better for only that speaker and works poorly for the two other speakers experimented with the model. The DNN based speech synthesizer and the forward model only learn a mapping from the articulatory variables to a melcepstrum, which is then passed to a vocoder (LPCNet (Valin and Skoglund, 2018)). The vocoder then synthesizes the input speech by using the estimated melcepstrum, pitch and periodicity, where pitch and periodicity are directly extracted from the input speech itself.

Similar approaches have recently been explored in (Beguš et al., 2022; Sun et al., 2022) to

do acoustic-to-articulatory speech inversion with parametric vocal tract models (Birkholz, 2013) or DNN based articulatory synthesizers. Majority of these work focus on how well the input speech is re-synthesized and/or fail to show strong quantitative results to verify the similarity between the learned articulatory representations and the ground-truth. Some of the work here are also limited to learning articulatory representations to a given set of words or phonemes (Beguš et al., 2022; Westerman and Miranda, 2002).

2.5 Application of TVs for Schizophrenia Detection

2.5.1 Articulatory coordination features (ACFs)

The motivation for ACFs comes from the observed condition of Psychomotor retardation (PMR), which is clinically referred to as a condition of slowed neuromotor output that manifests changes in speech, motility and cognition (Flint et al., 1993). PMR is regarded as a key characteristic of depression and is viewed as a necessary feature of Major Depressive Disorder (MDD) (Greden and Carroll, 1981). PMR is known to cause changes of articulatory coordination in speech. Williamson et al. (2013) showed that these articulatory coordination features can be used to characterize the level of articulatory coordination and timing. Williamson et al. (2013) in their early work used formant tracks as low level acoustic features and computed a multi-scale structure of correlations across the time series signals. Since the auto- and cross-correlations of the time acoustic features can reveal the hidden parameters in the stochastic-dynamical systems which generated these time series data, the ACFs have shown promise in MDD detection (Williamson et al., 2013, 2014, 2019).

The technique of computing correlation structure features or ACFs in Williamson et al.

(2013) is referred to as time-delay embedded correlation (TDEC) analysis. Recently, new multi-scale full vocal tract coordination (FVTC) features generated with a dilated CNN have shown further improvement in detecting MDD (Huang et al., 2020).

Time Delay Embedded Correlation (TDEC) Analysis

Williamson et al. (2013) used the first three formant tracks as the proxi acoustic features to compute the articulatory coordination. For each speech signal, a channel-delay correlation matrix is computed from the low-level multi-channel signals (first three formants in this case), using time-delay embeddings at multiple different delay scales. This correlations structure provides details about which time series signal is correlated with which, and at which time delays, and is therefore rich with information about the mechanisms underlying the coordination level. The computed correlation matrix R_j has a dimensionality of $MN \times MN$, where $M = 3$ channels and $N = 15$ time delays per channel.

From the correlation matrix R_j, an eigenspectrum is computed, which takes the form of an MN -dimensional (45-dimensional) feature vector. This eigenspectrum characterizes the within-channel and cross-channel distributional properties of the multivariate formant time series. In the work for MDD detection, four different delay scales j = 1, 2, 3, 4 which correspond to delay spacings 1, 3, 7, 15 have been experimented. Given that the sampling rate of formants is 100Hz, the four delay scales correspond to 10ms, 30ms, 70ms and 150ms increments respectively. In the same work, the authors claim that the low rank eigenvalues (ranked according to the descending order of magnitude) being larger for subject recordings with severe depression relative to healthy/mild depression. The trend is reversed towards high rank eigenvalues which taken together is a signature observation associated with depression severity (Williamson et al., 2013, 2014). This implies that the time series channels are highly correlated in the case of severe depression which in-turn imply

that there is simpler articulatory coordination or more coupled movements going on in depressed speech. On the other hand, in the case of mild depression or healthy speech, it can be thought of as more complex articulatory coordination going on in speech (Williamson et al., 2019).

Full vocal tract coordination (FVTC) analysis

In time delay embedded correlation analysis discussed above, a M-channel feature vector results in a channel delay correlation matrix which has a dimensionality of $MN \times MN$ where N is the time delays per channel. In order to incorporate multiple delay scales p, the correlation matrices computed at different delay scales are stacked yielding a $p \times MN \times MN$ dimensional matrix. To address this high dimensionality of the coordination feature and to overcome some limitations of the conventional channel delay correlation structure, Huang et al. (2020) proposed a novel method to construct the channel-delay correlation matrix which they called as the full vocal tract coordination (FVTC) method. The FVTC method addresses repetitive sampling and matrix discontinuity issues of the TDEC correlation structure.

In the FVTC formulation, for an M-channel feature vector X, the delayed correlations $r_{i,j}^{d}$ between the ith channel x_i and jth channel x_j delayed by d frames is computed by equation 2.1. Here N is the length of the channels.

$$r_{i,j}^{d} = \frac{\sum_{t=0}^{N-d-1} x_i[t]x_j[t+d]}{N - |d|} \tag{2.1}$$

Then the correlation vector for each pair of channels with delays $d\epsilon[0, D]$ frames is constructed as follows :

$$R_{i,j} = [r_{i,j}^{0}, r_{i,j}^{1}, ..., r_{i,j}^{D}]\epsilon\mathbb{R}^{1\times(D+1)} \tag{2.2}$$

The final FVTC correlation structure is then constructed by stacking the delayed auto-correlations and cross-correlations in the following manner :

$$\hat{R}_{ACF} = [R_{1,1}, ..., R_{i,j}, ..., R_{M,M}]^T \epsilon \mathbb{R}^{M^2 \times (D+1)} \tag{2.3}$$

Unlike the TDEC correlation structure, $\hat{R}_{ACF}$ contains every correlation only once. The new ACF representation also contains information pertaining to multiple delay scales and can be incorporated into a DNN based classification model by dilated CNN layers which have dilation factors matched to chosen delay scales (1, 3, 7 and 15), while also maintaining a low input dimensionality. Huang et al. (2020) also proposed a dilated convolution based CNN model where each $R_{i,j}$ is processed as a separate input channel in the CNN model, and thereby overcoming discontinuities in the input 2D representation.

2.5.2 TV based ACFs for detecting mental health disorders

Espy-Wilson et al. (2019) used TDEC based ACFs computed from TVs for MDD classification. They used six TVs estimated from the feed-forward DNN based speech inversion system in (Sivaraman et al., 2019) as the low-level acoustic features to compute the channel-delay correlation structure features. Their work, for the first time, showed that ACFs derived from TVs effectively capture the changes in neuromotor coordination in MDD. Seneviratne et al. (2020) then showed that adding aperiodicity and periodicity to the six TVs to compute the ACFs will further improve the detection of MDD speech from healthy. In the same work they showed that the TV based ACFs outperform the MFCC based ACFs computed in the same fashion.

With the development of the FVTC based ACFs and the proposed dilated CNN network

for depression classification using MFCCs in (Huang et al., 2020; Seneviratne and Espy-Wilson, 2020) developed a dilated CNN classification model inspired by the work of Huang et al. (2020) and used the new FVTC based ACFs derived from TVs for depression detection. In the same work, they showed that the new FVTC based ACFs derived from TVs outperform the MFCC based ACFs computed in the same manner.

The same FVTC based ACFs derived from TVs were used for the first time to detect positive symptoms in schizophrenia and more details on that work is discussed in section 6.2.

2.6 Application of Articulatory Representations for Mispronunciation Detection of /ɹ/in Child Speech Sound Disorders

Previous studies have identified several barriers to the uptake of effective clinical speech technologies for child use (Furlong et al., 2017; J. McKechnie and Ballard, 2018). The lack of datasets for system training, inadequate technical details of tools, low accuracy when rating sounds in error are some of them Benway et al. (2023b). A recently published open-access dataset, the PERCEPT Corpora (Benway et al., 2023a) has been instrumental in solving some of the previously identified barriers in the context of speech sound disorders, including for /ɹ/(i.e., rhotics) in American English, which is the most common such error.

Previous attempts with detecting mispronounciation of /ɹ/have focused on ultrasound image features with CNN based models (Ribeiro et al., 2021). Acoustically derived feature based models have recently gained more traction and have reported state-of-the-art results with certain datasets. Benway et al. (2023b) has contrasted the use of MFCCs with age and sex normalized formant features to show that the latter is an effective acoustic feature for detecting fully rhotic

versus derhotic /ɹ/. However, none of the previous studies have looked into using articulatory representations which can ideally capture articulatory configurations that will result in fully rhotic versus derhotic /ɹ/. In other words, TVs estimated by a SI system can be a critical acoustic feature compared to proxy acoustic features (eg. MFCCs and formants) in rhoticity detection in children.

Chapter 3: Improving Acoustic-to-Articulatory Speech Inversion

3.1 Overview

This chapter discusses one of the primary contributions of this dissertation with respect to improving the performance of acoustic-to-articulatory speech inversion systems. We discuss the main improvements made under four main aspects, (i) Exploring audio data augmentation techniques to increase the variability of training data, (ii) Learning acoustic-to-phoneme mapping with a multi-task learning framework, (iii) Incorporating source features as additional targets, and (iv) Using self-supervised speech representations with enhanced TV targets. Under each aspect, we discuss the motivation behind the work, the experiments conducted and finally the conclusions drawn from each respective experiment.

3.2 Audio Data Augmentation for Acoustic-to-Articulatory Speech Inversion

Data augmentation has proven to be a promising prospect in improving the performance of deep learning models by adding variability to training data. In a previous work with developing a noise robust acoustic-to-articulatory SI system, data augmentations have been successfully used to improve the performance of speech inversion particularly on 'noisy' speech conditions (Seneviratne et al., 2018). In this work, we extend this idea of data augmentation to improve the SI

systems on both the clean speech and noisy speech data by experimenting three data augmentation techniques. We also propose a Bidirectional Gated Recurrent Neural Network as the speech inversion system instead of the previously used feed forward neural network. The inversion system uses mel-frequency cepstral coefficients (MFCCs) as the input acoustic features, and six TVs as the output articulatory targets. The work in this section has been published in (Siriwardena et al., 2023a)

3.2.1 Bidirectional Gated RNN (BiGRNN) Speech Inversion Model

Gated Recurrent Unit (GRU) was first proposed by Cho et al. (2014) and has only 2 gates compared to 3 gates in a conventional LSTM unit, hence resulting in relatively smaller models, and takes lesser time for training. In this work, we propose a Bidirectional Gated RNN model as the speech inversion system. The model has 3 bidirectional layers of Gated Recurrent Units (GRUs) followed by two time distributed fully connected layers. Figure 3.1 shows the detailed model architecture of the proposed BiGRNN model. We specifically used a Masking layer at the input to avoid the affect of padded zeros for TV predictions. Dropout layers were used after every layer to minimize over-fitting.

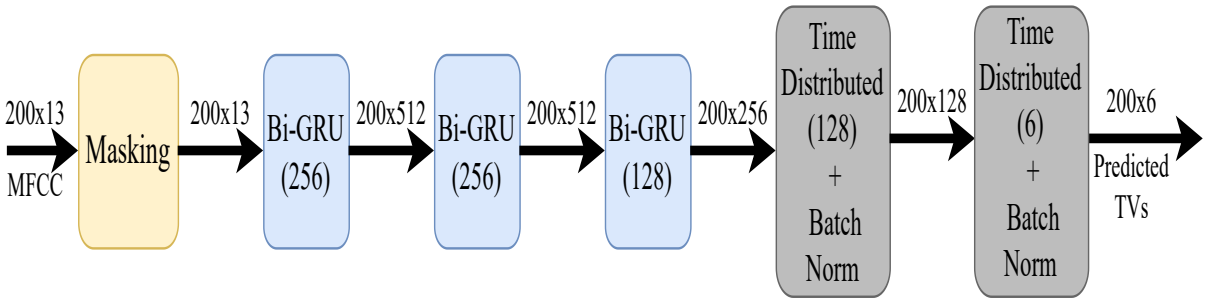

Figure 3.1: Proposed BiGRNN model architecture

Input audio features

In our experiments, all the audio files are first segmented to 2 seconds long segments and

the shorter audios are zero padded at the end. Mel-Frequency Cepstral Coefficients (MFCCs) and Melspectrogram (MSPEC) features are then extracted as the acoustic input features for the speech inversion systems. Both MFCCs and MSPECs were extracted using a 20ms Hamming analysis window with a 10ms frame shift. For MFCCs, 13 cepstral coefficients were extracted for each frame while 40 Mel frequencies were used for both MFCCs and MSPECs. Δ and $\Delta\Delta$'s for MFCCs were also computed to be used as an extended input feature set along with MFCCs. Both MFCCs (and Δs) and MSPECs are utterance wise normalized (z-normalized) prior to model training. Table 3.1 shows the results for how well each audio feature type performed with the SI systems.

Previous work with developing SI systems have shown MFCCs to be superior over MSPECs and Perceptual Linear Predictions (PLPs) as acoustic features (Sivaraman et al., 2019). Our results with different audio features as shown in Table 3.1 are consistent with the previous studies which suggest MFCCs can be effective especially with Bidirectional RNN based models. We also show that adding Δ and $\Delta\Delta$ features derived from MFCCs does not necessarily improve the performance.

Performance measurements

All the models are evaluated with Mean Squared Error (MSE), Mean Absolute Error (MAE) and Pearson Product Moment Correlation (PPMC) scores computed between the estimated TVs and the corresponding ground-truth TVs. Equation 3.1 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.

$$PPMC = \frac{\sum_i^N (X[i] - \overline{X})(Y[i] - \overline{Y})}{\sqrt{\sum_i^N (X[i] - \overline{X})^2 (Y[i] - \overline{Y})^2}} \tag{3.1}$$

Baseline BiLSTM and CNN-BiLSTM models

To compare against the proposed BiGRNN model, we used a BiLSTM model inspired by the work in Illa and Ghosh (2018) and a CNN-BiLSTM model similar to that in Shahrebabaki et al. (2020) as the baseline models. The BiLSTM model was trained with MFCCs and the CNN-BiLSTM model was trained with MSPECs.

The BiLSTM model has 3 bidirectional LSTM layers followed by two time distributed fully connected layers. The model has the same architecture as the proposed BiGRNN model except for the fact that it uses BiLSTM layers instead of the BiGRNN layers in the front-end of the model. The CNN-BiLSTM model consists of 5 CNN layers, whose outputs are then concatenated together and fed to 2 BiLSTM layers, followed by a final CNN layer.

Table 3.1: PPMC scores for models trained with clean speech when tested with clean speech test set

Model	Audio features	LA	LP	TBCL	TBCD	TTCL	TTCD	Average
BiGRNN	**MFCC**	**0.8801**	**0.6200**	**0.8580**	**0.7382**	**0.6922**	**0.9206**	**0.7848**
BiLSTM	MFCC	0.8742	0.6236	0.8535	0.7189	0.6792	0.9142	0.7773
BiGRNN	MFCC +$\Delta + \Delta\Delta$	0.8708	0.6256	0.8565	0.7167	0.7020	0.9125	0.7807
BiLSTM	MFCC +$\Delta + \Delta\Delta$	0.8480	0.6112	0.8347	0.7055	0.6667	0.8975	0.7606
CNN-BiLSTM	Melspectrogram	0.8285	0.5651	0.8028	0.6827	0.6193	0.8551	0.7256

Model training

The input XRMB dataset was divided into training, development, and testing sets, so that the training set has utterances from 36 speakers and the development and testing sets have 5 speakers each (3 males,2 females). None of the training, development and testing sets have overlapping speakers and hence all the models are trained in '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 augmented audio files were included in the same split as its original audio file to preserve the 'speaker-independence' and also not affecting the splitting ratio.

All the models were implemented with Tensorflow-Keras machine learning framework and trained with NVIDIA TITAN X GPUs. MSE, MAE and PPMC were experimented as loss functions to optimize the models. PPMC loss can be defined as $1 - PPMC$, where PPMC is defined in equation 3.1. Table 3.2 shows the average correlation on the test set for the best performing models with the 3 types of loss functions. It can be seen that the BiLSTM and BiGRNN models converged better with MAE loss where as CNN-BiLSTM model worked better with the PPMC loss. One limitation with the PPMC loss is it can predict some TVs with an offset from the ground-truth which is expected. To address that, we combined PPMC loss with MAE loss where a weight ($0 < \alpha < 1$) was assigned to PPMC loss and $1 - \alpha$ was assigned to MAE loss. All the CNN-BiLSTM models are trained with the new weighted loss function which outperformed MSE and MAE losses.

Table 3.2: Comparison of loss functions with each model type

Loss Fn. / **Model**	**MSE**	**MAE**	**PPMC**	**Weighted Loss** ($\alpha = 0.8$)	**Weighted Loss** ($\alpha = 0.5$)
BiGRNN	0.7910	**0.7959**	0.7917	0.7928	0.7940
BiLSTM	0.7850	**0.7870**	0.7858	0.7862	0.7865
CNN-BiLSTM	0.7163	0.7214	**0.7333**	0.7256	0.7203

3.2.2 Audio Data Augmentation for Speech Inversion

Seneviratne et al. (2018) developed a noise robust SI system by training the feed-forward SI system in Sivaraman et al. (2019), with noise augmented audio data. But, compared to the SI system trained entirely with clean speech data, the noise robust model did not perform well with the clean speech test set. Hence in this work, the goal is to explore data augmentation techniques that would improve the performance of the SI system with both noisy speech and clean speech data. With that goal in mind, we conducted 3 types of audio data augmentations using the Audiomentation library[1].

Background noise and music

We added the noise and music data from MUSAN corpus (Snyder et al., 2015) with the audio files in the XRMB dataset. The music files from the corpus were manually checked to remove any files with singing voices/human voices. We created two copies of the original speech data, one adding noise and the other adding music. The gain of noise and music added are set to be proportional to the Root Mean Square (RMS) value of the input sound in the audio. The added noise/music can range from 5dB to 20dB SNR and the level is randomly chosen.

Gaussian noise

Two copies of the original audio file are generated by adding Gaussian noise ranging from 5dB SNR to 20dB SNR. The noise SNR is sampled uniformly from the [5dB, 20dB] range when creating the noisy copy of the data.

Environmental impulse response (IR) functions

To add reverberation noise, we used the environmental IR functions from the MIT Acoustical

[1]https://github.com/iver56/audiomentations

Reverberation Scene Statistics Survey corpus (Traer and McDermott, 2016). Similar to the previous noise types, two noise copies of the original audio files were generated by convolving with randomly chosen IR functions from the corpus.

Effectiveness of different types of audio data augmentations

Figure 3.2 shows the PPMC scores for each TV and the resulting average score from the proposed BiGRNN model on the clean speech test set. Each model is trained with a dataset generated with one type of data augmentation in section 3.2.2.

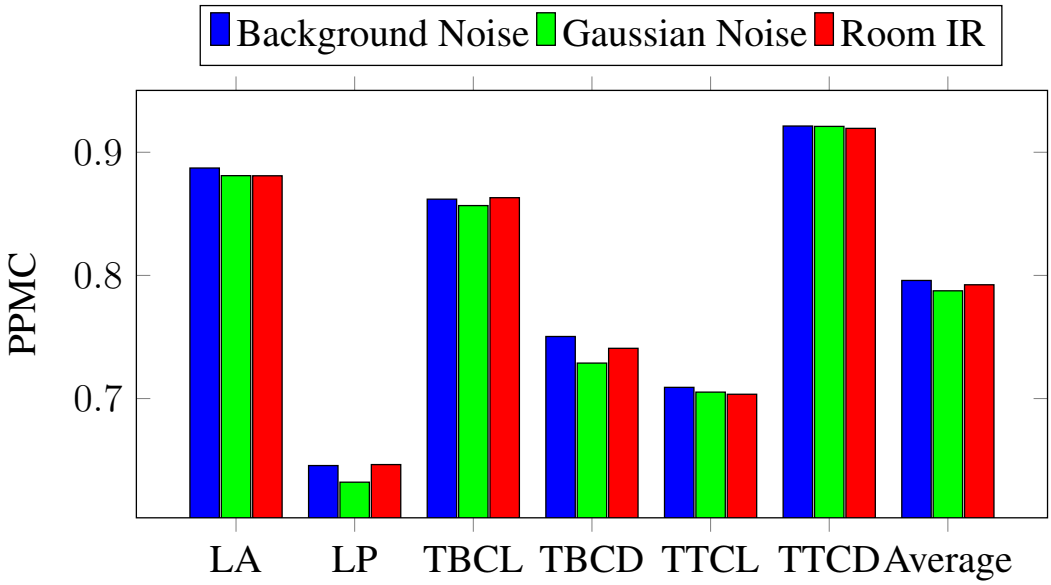

Figure 3.2: PPMC scores for each TV and the average score from the proposed BiGRNN model on clean speech test set

Figure 3.2 shows how the 3 different data augmentations compare each other when predicting each of the 6 TVs and the final average across all the 6 TVs. It can be observed that adding background noise and music from MUSAN dataset performs slightly better compared to adding reverberation noise from IR functions or adding random Gaussian noise. We also did a cross augmentation type experiment by testing the models trained with one type of augmentation with the other two types. From Table 3.4 it is clear that models trained with one data augmentation

type tends to perform well with data augmented in the same fashion.

Table 3.3: PPMC scores for models trained on one type of data augmentation when tested with other data augmentation types

Train Aug. \ Test Aug.	Background Noise & Music	Gaussian Noise	Room IR
Background Noise & Music	**0.7830**	0.7735	0.7572
Gaussian Noise	0.7726	**0.7843**	0.7395
Room IR	0.7624	0.7337	**0.7797**

3.2.3 Effect of data augmentation for speech inversion

Figure 3.3 shows the average PPMC score across the 6 TVs for the proposed BiGRNN, BiLSTM and CNN-BiLSTM baseline models when trained with clean speech only (clean-train) and when trained with clean speech + augmented data (augment-train). The two SI systems for each model type are tested with clean only, clean+augmented and augmented only test sets to evaluate the robustness of the models not only for noisy/augmented speech, but also for clean speech.

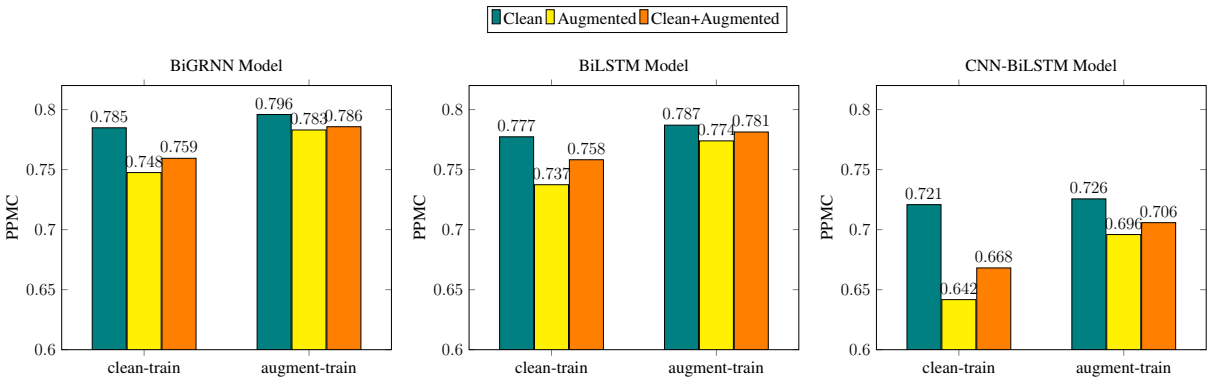

Figure 3.3: Performance of each model with and without augmented data on clean, augmented and clean+augmented test sets

Since data augmentation with background noise and music resulted in the best SI models, we

trained all the subsequent models with music and noise augmentation. Figure 3.3 shows how each model architecture (BiGRNN, BiLSTM and CNN-BiLSTM) trained with and without augmented data performed on clean, augmented only and clean+augmented test sets. The best performing SI model is reported with the BiGRNN model when trained with augmented data and it predicts TVs with an average PPMC score of 0.7959 on the clean speech data split. The previous noise-robust model (Seneviratne et al., 2018) was only able to achieve a best PPMC score of 0.741, thus the new SI system has a 5% relative improvement over the previous noise-robust SI system. This highlights one of the key take away points of this work, which suggests that data augmentation can be used as a technique not only for making DNN models more noise-robust, but also for improving the power of generalizability.

3.2.4 Improvement with model adaptation

With the goal of creating individual speaker adapted models, we took the best performing BiGRNN model pre-trained with augmented data, and further trained the model with each speaker's data in the original test set. The pre-trained model weights are only used for initialization and the new training was carried out with a smaller batch size (=4), reduced starting LR (=1e-4) and a quick decay with LR scheduler (every 2 epochs). The early stopping patience was increased to 30 to account for slight fluctuations in the validation loss due to smaller amount of training data. Here we used 80 % of all the subject's data (average 12 mins of speech) for training and used 10 % each for validation and testing. Table 3.4 shows the PPMC results of BiGRNN models when tested with same test split of subject's data before and after model adaptation.

In previous work with SI systems, it has been shown that model adaptation with a generalized (pre-trained) model trained with a larger number of subjects can perform better than a speaker-dependent model trained only with target speaker's data (Illa and Ghosh, 2018). Based on that observation, we performed model adaptation for speakers in the test set to see how much of an improvement the models can gain when an already pre-trained model is further trained with a portion of the subject's data. Turns out for the 5 speakers in the test set, an average PPMC score of 0.8506 can be achieved when the speaker-adapted models are tested with held-out data from the same subject. The pre-trained model without any adaptation, tested on the same splits of the target subjects can only achieve a PPMC score of 0.7942.

Figure 3.4 shows ground-truth and predicted TVs, LA, TBCD, and TTCD for an example utterance of JW31 subject. The TVs are estimated by the pre-trained and the speaker-adapted SI systems. The effect of speaker adaptation is clearly evident with the estimated TV trajectories, where the predicted TVs from the speaker adapted model looks significantly better than that estimated by the generalized model. However, it should also be noted, that for certain speakers (e.g. JW61) this improvement is not clearly evident. Hence, further work needs to be done to understand what speaker-specific characteristics are captured by these speaker-adapted SI systems and to devise effective modifications to infuse speaker characteristics to improve speaker-adapted SI systems.

Table 3.4: PPMC score for TV predictions, before and after model adaptation

Subject	**Before Model Adaptation**	**After Model Adaptation**
JW31	0.8071	**0.9106**
JW39	0.7907	**0.8836**
JW18	0.7936	**0.8633**
JW33	0.8334	**0.8490**
JW61	0.7463	**0.7466**

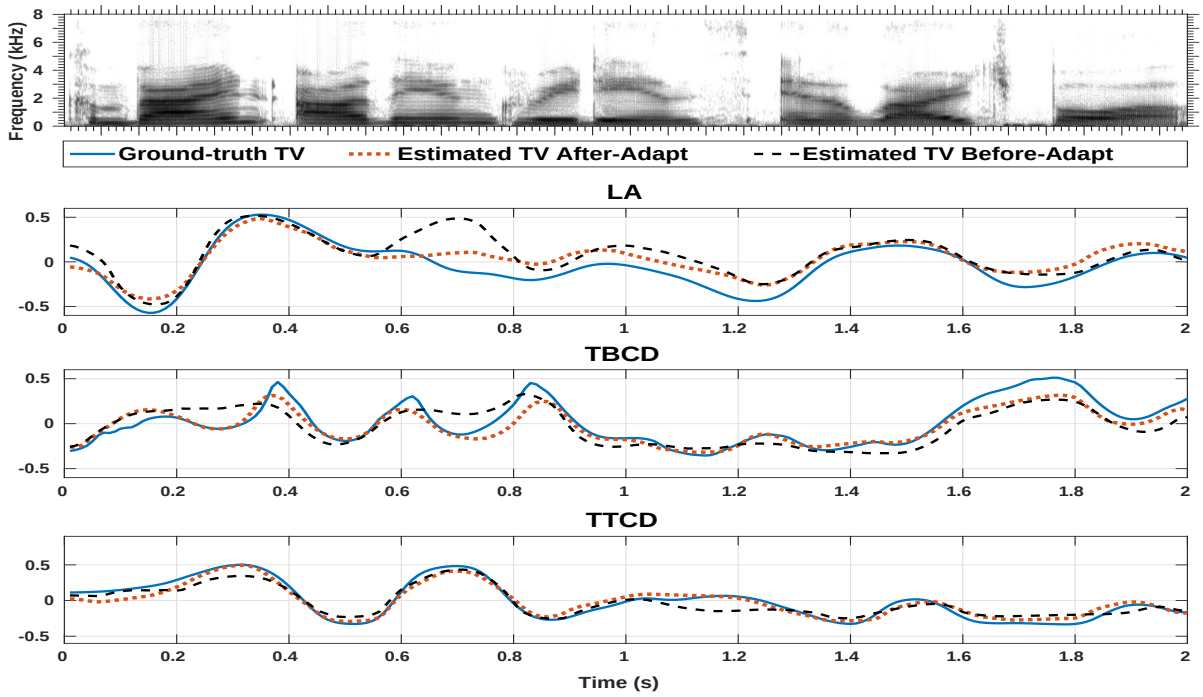

Figure 3.4: LA and constriction degree TV plots for the utterance 'Combined are the ingredients in a large bowl' estimated by the SI systems before and after adaptation on the JW31 subject. Solid blue Line - actual TV (from XRMB database), red dotted line - estimated TV after model adaptation, black dashed Line - estimated TV before model adaptation

3.2.5 Comparison between the new BiGRNN and previous feed-forward SI systems

Table 3.5 shows the results of the TV estimation from the new BiGRNN model and previously used feed-forward model in Sivaraman et al. (2019). Both the BiGRNN and feed-forward models are trained and tested in a speaker-independent fashion. The BiGRNN model is trained with the clean+augmented dataset and tested with both clean and augmented/noisy only test splits separately. The feed-forward model is evaluated in similar fashion. Results clearly show that the BiGRNN model trained with augmented data outperforms the previously used feed-forward model in both clean only and noisy only test splits. Figure 3.5 shows the predicted

TVs from the BiGRNN and feed-forward models for a sample utterance in the clean test set.

Table 3.5: TV-wise PPMC scores for feed-forward model and the BiGRNN model evaluated on noisy and clean test sets

Model	test set	LA	LP	TBCL	TBCD	TTCL	TTCD	Average
Feed-forward	clean	0.856	0.613	0.866	0.745	0.707	0.907	0.782
BiGRNN	clean	**0.887**	**0.646**	**0.862**	**0.750**	**0.709**	**0.921**	**0.796**
Feed-forward	noisy	0.750	0.520	0.795	0.670	0.643	0.830	0.701
BiGRNN	noisy	**0.873**	**0.631**	**0.855**	**0.724**	**0.702**	**0.914**	**0.783**

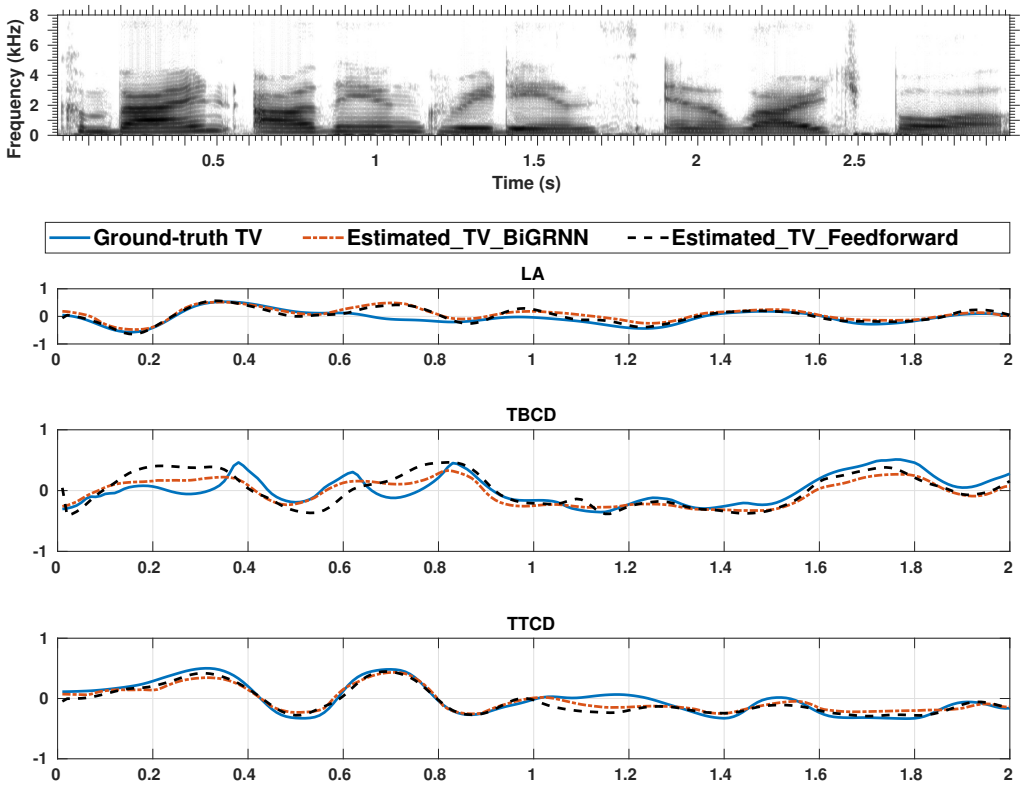

Figure 3.5: LA and constriction degree TV plots for the utterance 'Combined are the ingredients in a large bowl' estimated using the BiGRNN model and the feed-forward model. Solid blue Line - actual TV (from XRMB database), red dotted line - estimated TV from BiGRNN model, black dashed Line - estimated TV from feed-forward model

3.3 Multi-task Learning for Acoustic-to-Articulatory Speech Inversion

We explored the idea of learning the acoustic-to-phoneme mapping as a parallel task along with the SI task to investigate the effectiveness of a multi-task learning (MTL) framework for speech inversion. To generate the acoustic features for the SI task, we used all the audio files from the HPRC dataset (both 'normal' and 'fast' rate) and segmented them into 2 second long segments. Then, the shorter audios are zero padded at the end to make fixed size input embeddings. In this study, we use MFCCs as the input audio feature for the MTL based SI system. 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. The study in this section has been published in Siriwardena et al. (2022b).

3.3.1 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://github.com/jaekookang/p2fa$_{p}y3$

3.3.2 Multi-task model architecture

We use the 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 5.1 shows the architecture of the single task model on the left and the multi-task model on the right.

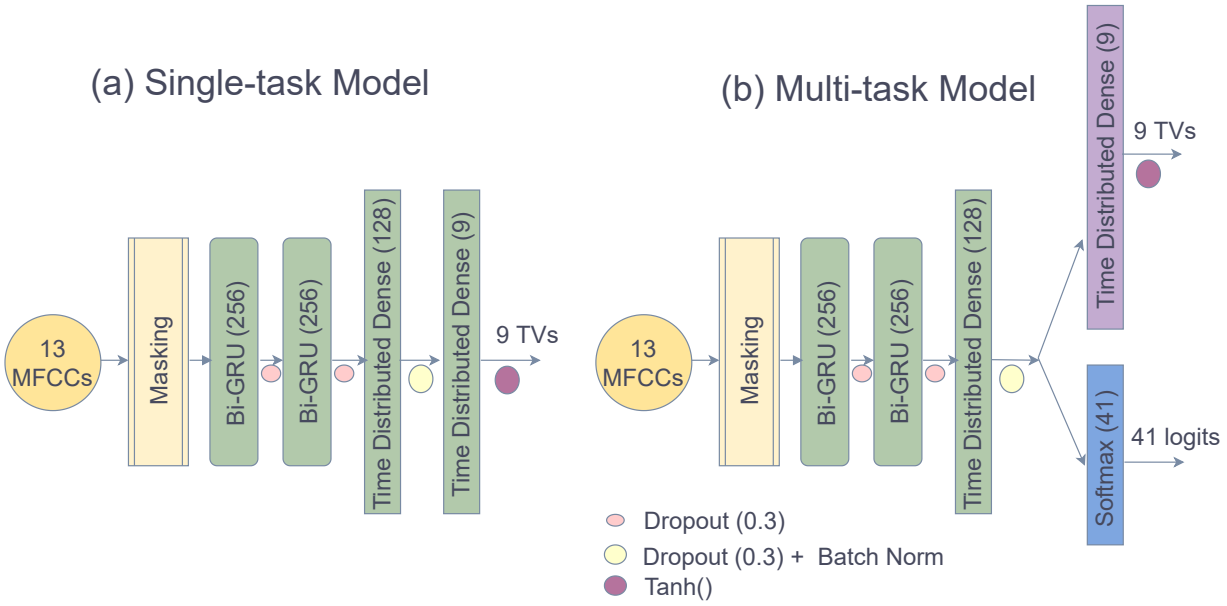

Figure 3.6: Single-task and Multi-task model architectures

The HPRC dataset was divided into training, development, and test sets, so that the training set has utterances from 6 speakers (3 Males, 3 Females) and the development and test 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 test 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 test 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.

3.3.3 Multi-task model training algorithms

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 $\hat{y}_{tv} \epsilon R^{L \times T}$ can be defined from equation 3.2 and similarly the output logits from the phoneme prediction, $\hat{y}_{ph} \epsilon R^{L \times V}$ can be defined from eqaution 3.3. 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 $\hat{y}_{ph}$.

$$\hat{y}_{tv} = f_{\phi}(x); x \epsilon R^{L \times d} \tag{3.2}$$

$$\hat{y}_{ph} = g_{\phi}(x); x \epsilon R^{L \times d} \tag{3.3}$$

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

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

3.3.4 Single-task vs multi-task learning for TV prediction

Table 3.6 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.

Table 3.6: 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**

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.

The results in Table 3.6 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,

Table 3.7: 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
$\alpha = 0.5$	**0.770**	72.88
$\alpha = 0.8$	0.759	72.90
$\alpha = 1.0$	0.758	73.60

algorithm 2 has a slight edge in TV prediction. However, when training time for the two algorithms are considered (table 3.8), 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 (Seneviratne et al., 2019), a similar training procedure to algorithm 1 was used.

Table 3.8: 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

Figure 3.11 and figure 3.8 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 3.11 shows the TVs which characterise the constriction degree of articulators, whereas figure 3.8 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 3.6. 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.

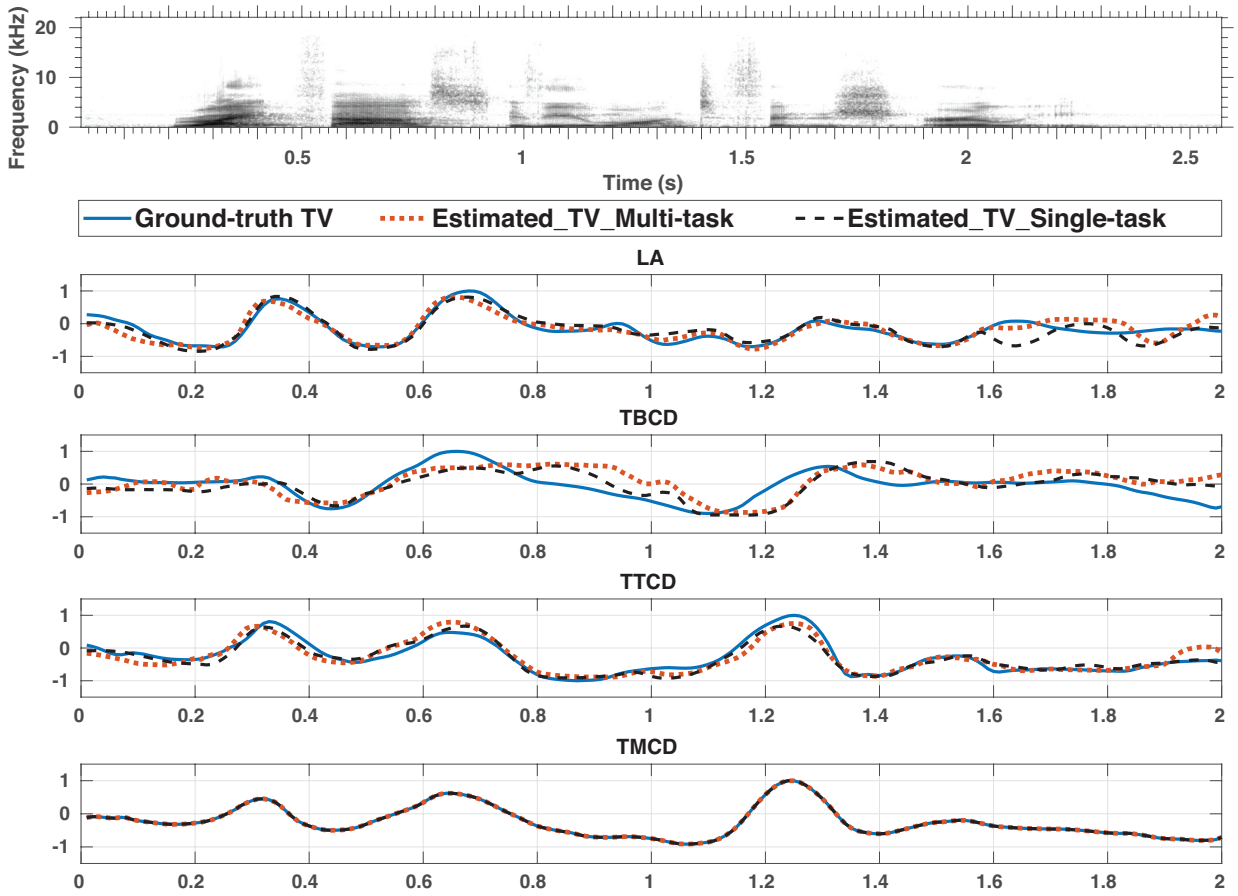

Figure 3.7: LA and constriction degree TV plots for the utterance 'Write fast if you want to finish early' estimated using Multi-task model and the Single-task model. Solid blue Line - actual TV (from HPRC database), red dotted line - estimated TV from Multi-task model, black dashed Line - estimated TV from Single-task model

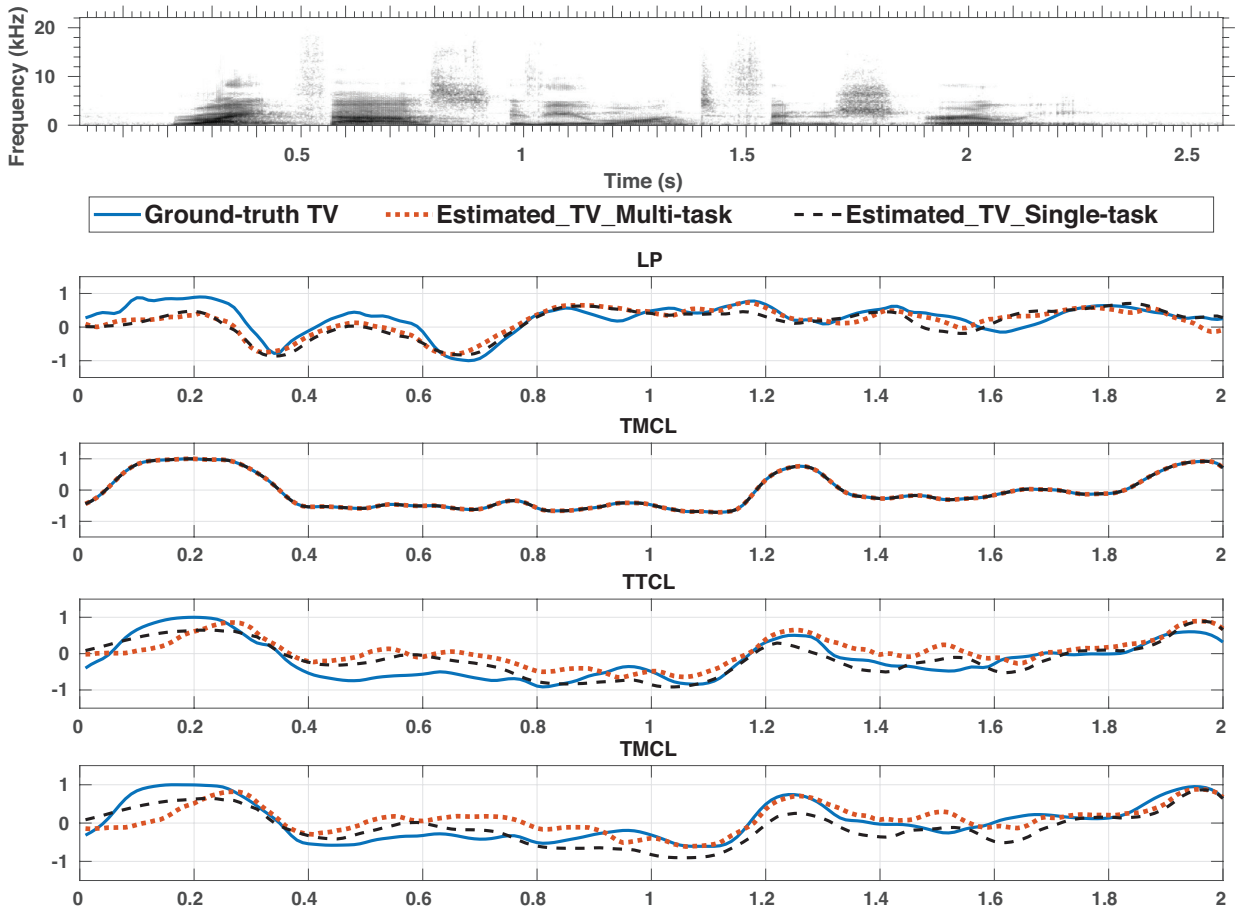

Figure 3.8: LP and constriction location TV plots for the utterance 'Write fast if you want to finish early' estimated using Multi-task model and the Single-task model. Solid blue Line - actual TV (from HPRC database), red dotted line - estimated TV from Multi-task model, black dashed Line - estimated TV from Single-task model

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.

3.4 Incorporating Source Features to Improve Acoustic-to-Articulatory Speech Inversion

The vocal tract which consists of the velum, tongue, lips and teeth, acts like an acoustic tube which modulates the source waveform. The acoustics of all vowel productions and most of the consonants have been described by a linear source-filter theory (Stevens, 2000). This theory is based on the assumption that the source of speech production is independent of the vocal tract filter. However, the actual process of speech production is nonlinear since the aerodynamics inside the glottis and vocal tract is governed by non-linear equations, and most importantly it has been shown that there exists a mutual interaction between the source and filter in certain cases of speech production (Titze et al., 2008). In this work, we take into account this source-filter interaction to improve the acoustic-to-articulatory speech inversion task.

In this work, we investigated the idea of estimating source level features along with the articulatory trajectories as targets to leverage any source-filter interactions to improve the overall SI task. Here we used source features from the Aperiodicity, Periodicity and Pitch (APP) detector (Deshmukh et al., 2005) as proxies for the source activity. We also experimented with multiple input representations of speech with different DNN model architectures and used two publicly available articulatory datasets to show the significance of learning source level information in improving the acoustic-to-articulatory speech inversion task. The work in this section has been published in Siriwardena and Espy-Wilson (2023).

3.4.1 Input Speech Representations for proposed and Baseline SI models

All the audio files are first segmented into 2 second long segments and the shorter audios are zero padded at the end. However for the HPRC dataset, the audio files are first down-sampled to 16 KHz before segmentation and padding. The following features are then extracted from the segmented audio utterances.

Auditory Spectrograms

We converted the one-dimensional pressure time waveform into a two-dimensional pattern of neural activity distributed along the tonotopic axis (roughly a logarithmic frequency). This two-dimensional representation, which is defined as an 'auditory spectrogram' (Audspec) (Wang and Shamma, 1994) is used in our proposed SI system as the input speech representation. It has been shown that this Audspec is an enhanced and a noise-robust estimate of the Fourier-based spectrogram (Wang and Shamma, 1994).

MFCCs and Mel-spectrograms

Mel-Frequency Cepstral Coefficients (MFCCs) and Mel-spectrograms (MSPECs) are extracted as the input acoustic features for baseline SI systems. Both MFCCs and MSPECs were extracted using a 20ms Hamming analysis window with a 10ms frame shift. For MFCCs, 13 cepstral coefficients were extracted for each frame while 40 Mel frequencies were used for both MFCCs and MSPECs. Both MFCCs and MSPECs are utterance wise normalized (z-normalized) prior to model training.

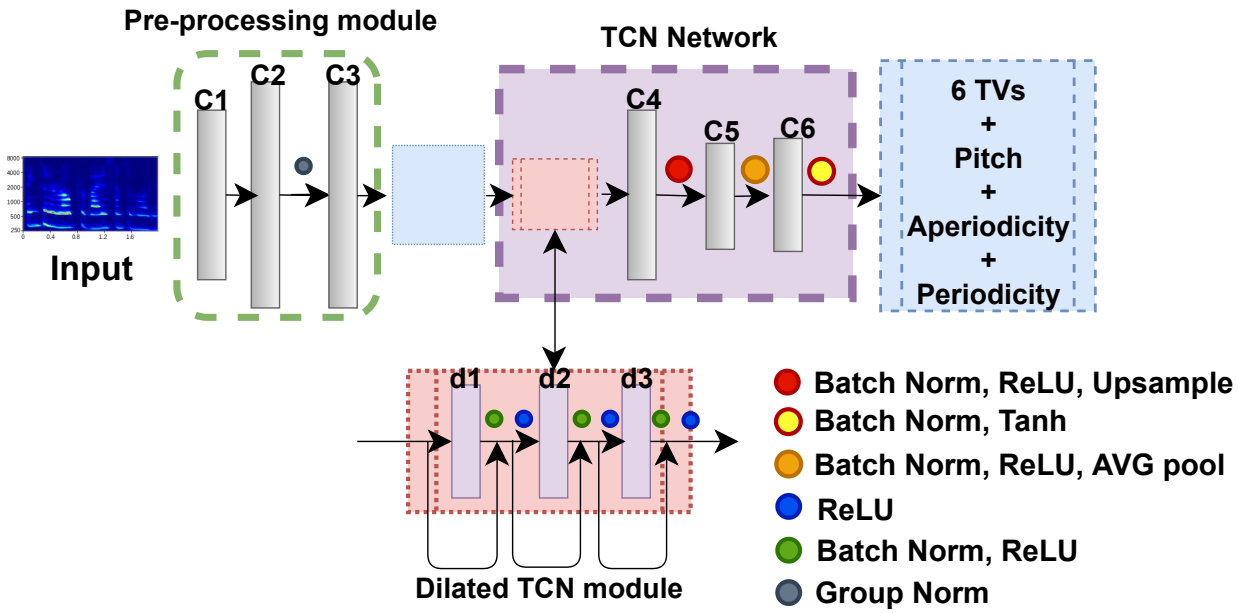

Figure 3.9: Model architecture of the SI system. Here C1-C6 represent 1D-CNN layers where as d1-d3 represent 1D dilated CNN layers

3.4.2 Proposed Temporal Convolution Based Speech Inversion System

We propose a Temporal Convolution Network (TCN) based acoustic-to-articulatory SI system which takes in the Audspecs as input. The proposed system, unlike conventional SI systems, estimates both TVs and source level features (aperiodicity, periodicity and pitch) as the output. The model is optimized using the Mean Squared Error (MSE) loss computed between the predicted articulatory variables and the ground truth (TVs from articulatory datasets and source features from APP detector (Deshmukh et al., 2005)).

The SI system is implemented in PyTorch with 1-D convolutional (CNN) layers. The complete network is inspired by the multilayered Temporal Convolution Network in Lea et al. (2017). Figure 5.1 shows the proposed model architecture with its sub-modules used for pre processing and dilated TCN. The Pre-processing module contains three 1-D CNN layers with 1×1 kernels (C1, C2 and C3), which have 128, 256 and 256 filters, respectively. The d1, d2 and d3

dilated CNN layers have a kernel size of 3 with 1,4 and 16 dilation rates respectively. Upsampling (window size 4) is done after C4 layer and average pooling (window size 5) is done after C5 layer along with BatchNorm layers after every CNN layer in the TCN network. The upsampling and average pooling operations take care of matching the time dimension of the input spectrograms to the target time dimension of TVs.

To train the SI system, learning rates were determined based on a grid search by testing all combinations from [1e-2, 1e-3, 1e-4, 3e-4] that resulted in 1e-3 as the best pick. A similar grid search was done to choose the batch size from [16, 32, 64, 128] and 64 gave the best validation MSE. The objective function was optimized using the ADAM optimizer with an 'ExponentialLR' learning rate scheduler and a decay of 0.5. All models were trained by monitoring the validation loss.

To train the models with the XRMB dataset, the dataset was divided into training, development, and testing splits, so that the training set has utterances from 36 speakers and the development and testing sets have 5 speakers each (3 males,2 females). To train the models with the HPRC dataset, similar to the XRMB dataset, the 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. We used audio samples with both the normal and fast production rates in the HPRC dataset. To create the validation and test sets, the audio samples from the two speakers were randomly split, so that both the splits have samples corresponding to normal and fast rates.

For both the datasets, none of the training splits have overlapping speakers with the development and testing sets and hence all the models are trained in a 'speaker-independent' fashion. The splits also ensured that around 80% of the total number of utterances were present

Table 3.9: PPMC scores for articulatory variable prediction on the XRMB dataset. Model names with 'SF' uses source features as additional targets. The AVG. TVs column for those models also show the percentage increase in TV prediction with respect to the same model which does not use source features

Model	**LA**	**LP**	**TBCL**	**TBCD**	**TTCL**	**TTCD**	**Ap.**	**Per.**	**Pitch**	**AVG. TVs**	**AVG. all**
TCN-Audspec	0.7977	0.7942	0.7883	0.7836	0.7743	0.7684	-	-	-	0.7844	-
TCN-SF-Audspec	0.8448	**0.8640**	0.8604	**0.8818**	**0.9029**	0.9005	**0.9082**	0.8860	**0.9021**	**0.8770** (9.3%)	**0.8834**
TCN-Mspec	0.7432	0.7427	0.7366	0.7244	0.7179	0.6993	-	-	-	0.7273	-
TCN-SF-Mspec	0.8364	0.8639	**0.8727**	0.8607	0.8807	0.8917	0.8732	0.9005	0.8638	0.8677 (14%)	0.8715
BiGRNN-MFCC	0.8801	0.6200	0.8580	0.7382	0.6922	**0.9206**	-	-	-	0.7848	-
BiGRNN-SF-MFCC	**0.8810**	0.6211	0.8628	0.7365	0.7019	0.9191	0.8693	0.9163	0.7209	0.7871 (0.2%)	0.8032
CNN-BiGRNN-Mspec	0.8801	0.6165	0.8505	0.7355	0.7146	0.9171	-	-	-	0.7858	-
CNN-BiGRNN-SF-Mspec	0.8799	0.6246	0.8566	0.7302	0.7065	0.9175	0.8794	**0.9296**	0.7441	0.7859 (0.01%)	0.8076
CNN-BLSTM-Mspec	0.8770	0.6184	0.8463	0.7200	0.6915	0.9197	-	-	-	0.7788	-
CNN-BLSTM-SF-Mspec	0.8774	0.6202	0.8525	0.7172	0.6941	0.9180	0.8734	0.9263	0.7442	0.7799 (0.1%)	0.8026

Table 3.10: PPMC scores for HPRC dataset.

Model	**AVG. 9 TVs**	**AVG. all**
TCN-Audspec	0.4805	-
TCN-SF-Audspec	**0.7573** (27.7%)	**0.7636**
TCN-Mspec	0.4763	-
TCN-SF-Mspec	0.6503 (17.4%)	0.6621
BiGRNN-MFCC	0.7118	-
BiGRNN-SF-MFCC	0.7153 (0.3%)	0.7263
CNN-BiGRNN-Mspec	0.7277	-
CNN-BiGRNN-SF-Mspec	0.7290 (0.1%)	0.7461
CNN-BLSTM-Mspec	0.7245	-
CNN-BLSTM-SF-Mspec	0.7259 (0.1%)	0.7428

in training, and the development and testing sets have a nearly equal number of utterances. This allocation was done in a completely random manner.

3.4.3 Baseline Speech Inversion Systems

This section discusses the baseline SI systems implemented for comparison. Detailed information on the model architectures and implementation can be found in a GitHub repository. [1]

BiGRNN model

We used the BiGRNN model architecture implemented in (Siriwardena et al., 2022b, 2023a)

[1] https://github.com/Yashish92/Speech-Inversion-TCN

as one of our baseline models for comparison. The model has 2 bidirectional layers of Gated Recurrent Units (GRUs) followed by two time distributed fully connected layers. Dropout layers are also used after every layer to minimize the issue of over-fitting. 13 MFCCs are used as input to this SI system.

CNN-BiLSTM and CNN-BiGRNN model

A CNN-BiLSTM model inspired by the work in Shahrebabaki et al. (2020) was implemented as another baseline SI system. The model consists of 5 CNN (1D CNNs) layers, whose outputs are then concatenated together and fed to 2 BiLSTM layers. The output from the last BiLSTM layer is then passed through two time distributed fully connected layers, where the final fully connected layer serves as the output layer.

A similar architecture was used to implement a CNN-BiGRNN model where the only difference is that the 2 BiLSTM layers in the CNN-BiLSTM model are now replaced with bidirectional layers of Gated Recurrent Units (GRUs). CNN-BiGRNN model is comparatively light weight due to the GRUs used in the model instead of LSTM units. A BiGRNN based SI system has also shown to outperform a conventional BiLSTM based SI system in Siriwardena et al. (2023a) which motivated this new CNN-BiGRNN model as a baseline for comparison.

Similar to the BiGRNN model in section 3.4.3, dropout layers are used after every layer to minimize possible over-fitting in both the CNN-BiLSTM and CNN-BiGRNN models. Both models used MSPECs as the input speech representation.

3.4.4 Comparison with baseline SI systems

The baseline models discussed in section 3.4.3 were trained and evaluated with the same train-dev-test splits for comparison. For every model architecture, two versions of the model were implemented with one only predicting the TVs as targets (6 TVs for XRMB dataset, 9 TVs for HPRC dataset) and the other predicting both TVs and source features.

A baseline TCN based SI system (TCN-Audspec) was trained with Audspecs as input and only TVs as targets for both XRMB and HPRC datasets. A similar TCN architecture was implemented to use MSPECs as inputs and, two versions of this model (TCN-Mspec and TCN-SF-Mspec) were trained similar to the other baseline models. Table 3.9 shows the PPMC scores for TV estimation on the XRMB dataset. The PPMC scores for individual TVs and source features are listed here along with average scores across TVs and all the predicted articulatory variables (TVs + source features). Similarly, Table 3.10 lists the average PPMC scores across the 9 TVs and all the articulatory variables for the HPRC dataset.

3.4.5 Estimated TVs and source features

Figure 3.10 shows the estimated constriction degree TVs (LA, TBCD, TTCD) and source features from the proposed TCN-SF-Audspec and the TCN-Audspec models. As can be observed in the plots, the source features are predicted with a considerably better accuracy, which is an added advantage of the proposed SI system. This also gives an almost complete articulatory representation of speech (only missing velar activity) that can be useful in various speech applications (e.g articulatory speech synthesis).

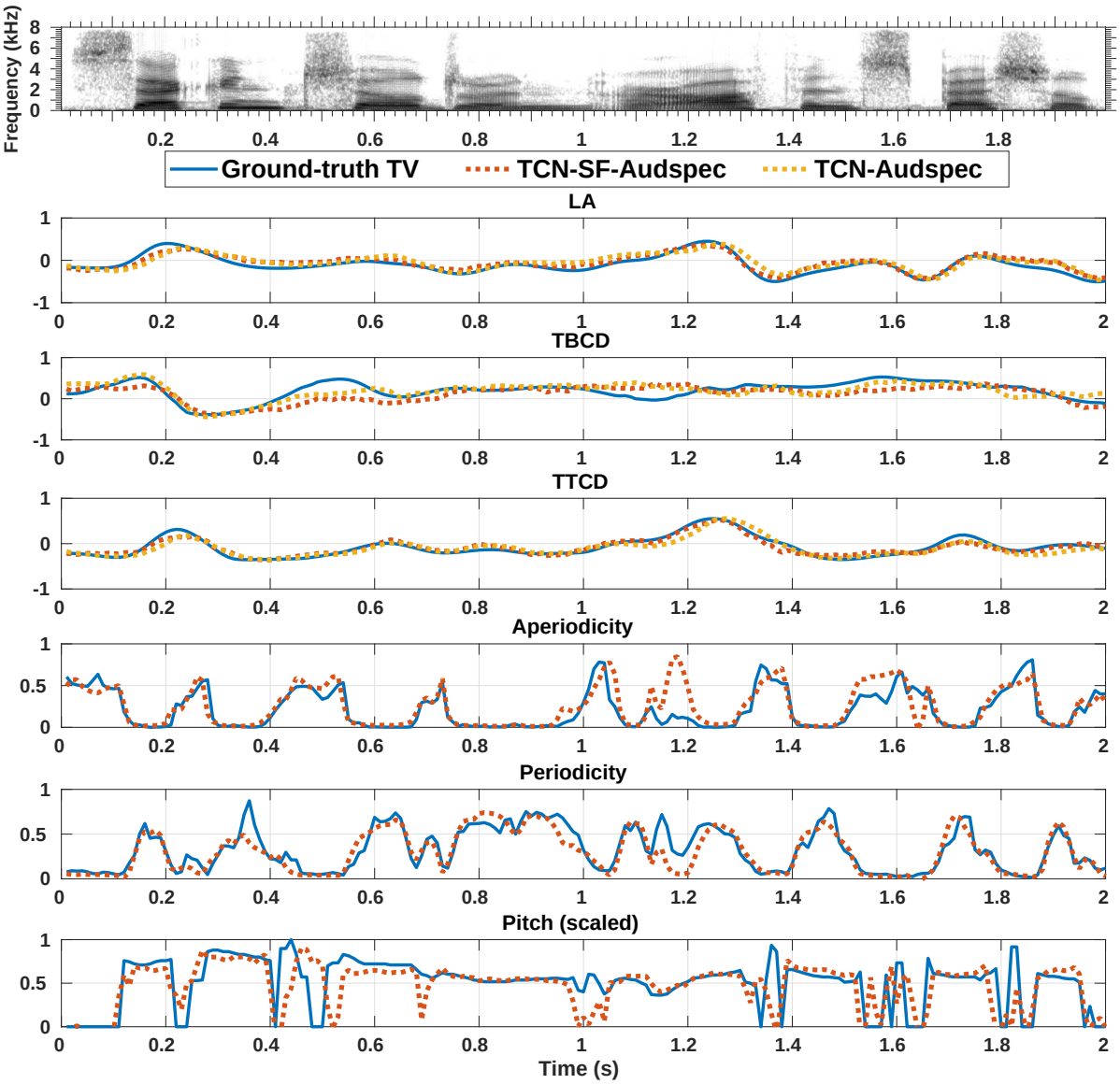

Figure 3.10: LA and constriction degree TVs + source features for the utterance 'second children are often special' estimated by the proposed TCN-SF-Audspec model compared to the TCN-Audspec. Solid blue Line - ground truth, red dotted line - predictions by the TCN-SF-Audspec, yellow dotted Line - predictions by TCN-Audspec.

3.4.6 Discussion on results with incorporating source features and the proposed TCN model

Deep Neural Networks have been effective in learning complex non-linear relationships between input speech representations and articulatory movements that has lead to the recent success in acoustic-to-articulatory SI task. However, there is still a lot of room for improvement, especially in developing speaker-independent and more generalizable SI systems that can be effectively utilized in speech applications. To this end, we explored the idea of incorporating source characteristics to exploit any source-filter interactions, and thereby mutually learn, and improve, the acoustic-to-articluatory SI task.

We used aperiodicity, periodicity and pitch as source features, which are used as additional targets to train the SI systems. For the proposed TCN based SI system and for every baseline model, two versions of the model were trained with the goal of investigating the real effect of incorporating source features. As shown in Tables 3.9 and 3.10, the results are consistent across both articulatory datasets, and support the fact that incorporating source features into the mix of TVs is definitely helping the estimation of articulatory variables. This observation is quite evident with the proposed TCN models which use Audspecs or MSPECs as inputs. For example, with the XRMB dataset, the TCN model which uses MSEPCs gain an absolute improvement of around 14% with respect to the same model which does not use source features as targets. Similarly with the HPRC dataset, the TCN model which uses Audspecs gain an absolute improvement of close to 28% when the source features are used as additional targets. Most importantly, when the best PPMC scores for average TV estimation with the proposed TCN model is considered, it is around a 9% improvement over the current best performing SI systems in (Siriwardena et al., 2023a;

Sivaraman et al., 2019), trained and evaluated on the same splits of the XRMB dataset.

However, a key observation here is that both the input speech representation and the DNN model architecture play a significant role in learning these complex relationship between the source features and TVs. For example, with the 13 MFCCs, which is the most commonly used speech representation in SI systems, adding source features as targets does not significantly improve the PPMC scores. This is consistent with both the XRMB and HPRC datasets with the BiGRNN model which uses 13 MFCCs as input. This can be mainly due to the fact that the 13 MFCCs do not contain important source information and is usually limited to capturing the filter characteristics (vocal tract) in speech production. Moreover, having richer speech representations that contain source information does not necessarily mean it will improve on the SI task. A fine observation to support that is the CNN-BiGRNN and CNN-BiLSTM models that use MSPECs as inputs, which necessarily contain valuable source information unlike the 13 MFCCs. This elucidates the fact that the DNN model architecture too plays a critical role in learning these complex dependencies between the source and articulatory targets.

TCN based models have shown to be extremely effective in speech applications (Pandey and Wang, 2019) with learning long-range temporal (and contextual) dependencies. These models have shown to outperform typical RNN and CNN based models, especially in applications where subtle and complex contextual information needs to be extracted from input representations (Lea et al., 2017). This was one of the key motivations for the proposed SI system, which ultimately outperformed all the other baseline systems on both XRMB and HPRC datasets. However, as mentioned earlier, the input representation used to train these models also plays a role which can be clearly observed with results in Tables 3.9 and 3.10. The TCN model trained with Audspecs as input is outperforming the same TCN based model architecture trained with MSPECs which

suggests that the Audspecs might be capturing important spectral and temporal information that is helpful in mutually learning both the source features and TVs.

Conventional SI systems usually predict constriction degree related TVs significantly better (eg., LA, TTCD) with respect to the constriction location related TVs (eg., LP, TTCL). The same can be more or less observed (with the exception of TBCL and TBCD) for individual TV predictions in Table 3.9 for all the 'baseline models'. Surprisingly, this is not as evident with the TCN based models which tend to predict both the location and degree TVs with close to similar accuracies. Moreover, further analysis needs to be done to investigate the ways and instances by which the source features are actually interacting with the TVs, and also to understand what the TCN models are actually capturing as source-filter interactions that is ultimately helping the overall SI task.

3.5 Self-Supervised Speech Representations with Enhanced TVs for Speech Inversion

This section highlights the results of a speech inversion system trained with self-supervised learning based acoustic representations and enhanced TV targets extracted from the XRMB dataset.

Self-Supervised Learning (SSL) has shown to be an effective method of improving DNN performance through the utilization of unlabeled data in learning speech representations (Hsu et al., 2021; Schneider et al., 2019). These representations have shown to be effective in Automatic Speech Recognition (ASR) systems (Conneau et al., 2020), speech separation and enhancement (Wang et al., 2023). Recent works have also shown that SSL speech representations have the

capacity to improve the performance of SI models for EMA data (Cho et al., 2023) outperforming the conventional acoustic features like MFCCs. Cho et al. (2023) have extensively evaluated the existing SSL speech representations for the SI task and have found that HuBERT based SSL speech representations (Hsu et al., 2021) works the best over both the other SSL features (eg. wav2vec2, tera (Liu et al., 2021)) and conventional acoustic features like MFCCs.

In a recent work, Attia et al. (2023b) proposed a novel geometric transformation which improved the performance of SI systems through better output feature space representation. In this work we explore the idea of using HuBERT (Hsu et al., 2021) SSL speech representation as the input speech representation and learn a mapping to the enhanced TVs proposed in (Attia et al., 2023a,b). We show that using better input and output feature representations lead to better SI performance and more robust TV estimates.

3.5.1 SI Architecture with HuBERT features

SSL speech representations when used in the SI task with EMA data have shown to outperform the conventional acoustic features (eg. Mel-spectrograms, MFCCs) (Cho et al., 2023). Here the SSL representations only need to be fine-tuned for the down stream task of speech inversion and can be expected to generalize better even with limited ground-truth articulatory data. Based on the previous work in Cho et al. (2023) for using SSL features for the SI task with EMA data, we explored the idea of using HuBERT SSL features (Hsu et al., 2021) as the input acoustic representation to train our best performing Bidirectional Gated Recurrent Unit (BiGRNN) SI architecture.

We used the HuBERT-large model pre-trained with the Librilight dataset (60,000h) to extract

the HuBERT speech embeddings. All the audio files (sampled with 16 KHz) are first segmented to 2 second long segments and the shorter ones are zero padded at the end. The HuBERT embeddings are then extracted from the 2 second long segments using the speechbrain open-source AI toolkit (Ravanelli et al., 2021). The HuBERT embeddings are sampled at 50 Hz and have a dimensionality of 1024.

We used the BiGRNN SI system proposed in Siriwardena et al. (2023a), and adapted the input layer to match the input dimensionality of the HuBERT embeddings.

3.5.2 SI Architecture with MFCC features

We trained the same SI system architecture used in Attia et al. (2023b) which is identical to that discussed in section 3.5.1, with the only difference being 13 MFCCs used as the input acoustic feature. The MFCCs were extracted using a 20ms Hamming analysis window with a 10ms frame shift. The MFCCs are also utterance wise normalized (z-normalized) prior to model training.

3.5.3 Model Training

Both the SI architectures described above were trained in similar fashion. The input XRMB dataset was first divided into training, development, and test sets, so that the training set has utterances from 36 speakers and the development and testing sets have 5 speakers each (3 males, 2 females). None of the training, development and test sets have overlapping speakers and hence all the models were trained in a 'speaker-independent' fashion. All the models were implemented with a TensorFlow-Keras machine learning framework. ADAM optimizer with a starting learning rate of 1e-3 and an exponential learning rate scheduler was used. Both the models with HuBERT

and MFCCs were trained with an early stopping criteria (patience=5) monitoring the 'validation loss' on the development set. To choose the best starting 'learning rate', we did a grid search on [1e-3, 3e-4, 1e-4], whereas to choose the training batch size, we did a similar grid search on [16,32,64,128]. Based on the validation loss, 1e-3 and 32 were chosen as the learning rate, and batch size, respectively, for the model with HuBERT features and 1e-3 and 64 for the model with MFCCs.

3.5.4 Using reconstructed TV targets to extend the size of the dataset

In this subsection, we explore the idea of using reconstructed articulatory data to extend the available articulatory dataset. The SI system is trained on the original uncorrupted XRMB dataset ('small dataset': 4 hours of speech) and an 'extended dataset' (5.3 hours of speech). The 'extended dataset' is created by adding reconstructed articulatory data based on the TV reconstruction methodelogy proposed in Attia and Espy-Wilson (2023).

The SI systems are then trained on the 'small dataset' and the 'extended dataset' independently as shown in Table 3.11. The trained SI systems are evaluated on the same test split and the results are reported as Pearson Product Moment Correlation (PPMC) between the predicted and ground truth TVs.

The first part of Table 3.11 shows the performance of the SI system with MFCC input features. It can be seen that training on the extended dataset improves performance across the board for the small dataset. However, the overall performance of the SI system has improved only by 0.74%.

Similarly, the second part of the Table 3.11 shows the performance of the SI system with the

HuBERT SSL input features when trained on the 'smaller' and 'extended' datasets. The PPMC scores again show the slightest of improvement asserting the fact that incorporating additional reconstructed TVs to increase the size of the dataset did not significantly improve the performance of the SI system.

3.5.5 SSL features vs MFCCs for Speech Inversion

In this subsection, we discuss the effect of using HuBERT speech representation as the input to the SI system juxtaposed with MFCCs. The two model architectures are discussed in section 3.5.1 and section 3.5.2.

Training on the small dataset, HuBERT representation lead to a tangible improvement in the tongue TVs, namely TBCL, TBCD, TTCL and TTCD, with slight improvement in LA and LP. On average, using HuBERT representations lead to a 2.3% improvement in PPMC scores. Similarly, with the extended dataset, HuBERT representations gain a close to 2.0% average improvement over the MFCCs as input features. Overall, when using the extended dataset and HuBERT features, the SI system has gained a close to 2.6% average improvement in PPMC scores with respect to the baseline SI system trained with the smaller dataset and MFCC features.

Table 3.11: PPMC between predicted and ground truth TVs for SI systems trained on datasets according to each geometric transformation model, with the MFCCs and HuBERT input features.

Training Dataset	LA	LP	TBCL	TBCD	TTCL	TTCD	Average
MFCC Input Features							
Small Dataset	0.8603	0.7104	0.7426	0.7754	**0.7422**	0.8981	0.7881
Extended Dataset	**0.8697**	**0.7250**	**0.7508**	**0.7847**	0.7407	**0.9019**	**0.7955**
HuBERT Input Features							
Small Dataset	0.8779	**0.7243**	**0.7430**	0.8089	0.7865	**0.9248**	0.8109
Extended Dataset	**0.8902**	0.7142	0.7361	**0.8180**	**0.8032**	0.9229	**0.8141**

3.5.6 Estimated TVs with best performing SI systems

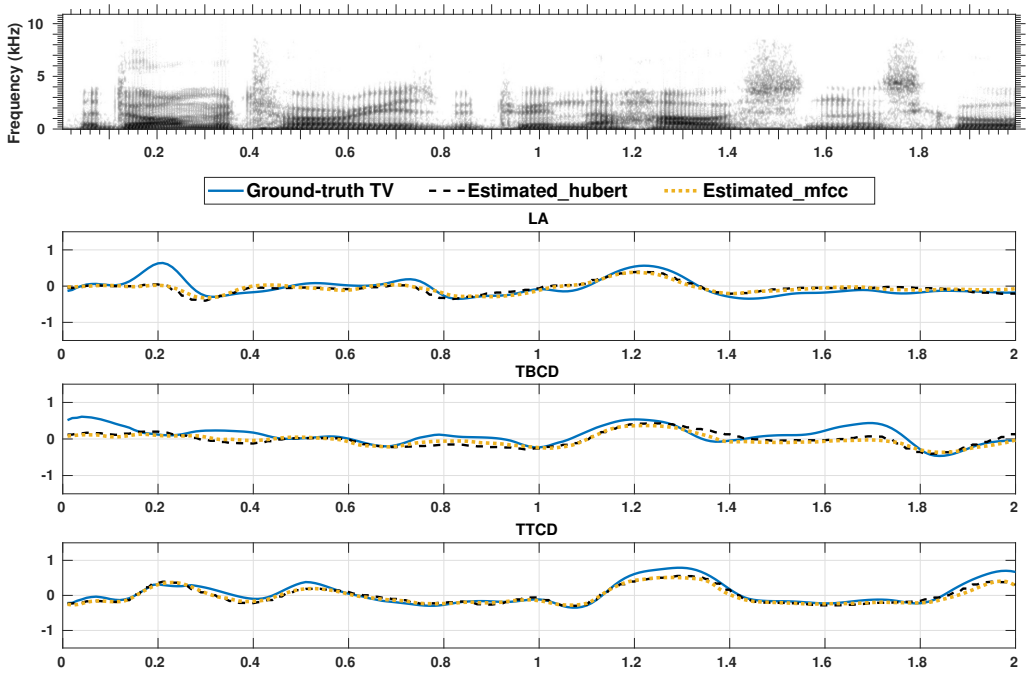

Figure 3.11: LA and constriction degree TVs for the utterance 'The dormitory is between the house and the school' estimated by the model trained with HuBERT embeddings (estimated_hubert) and the model trained with MFCCs (estimated_mfcc). Solid blue Line - ground truth, black dotted line - predictions by the HuBERT based model, yellow dotted Line - predictions by MFCC based model.

Figure 3.11 shows the estimated LA and constriction degree TVs for an utterance in the test set, by the two SI systems trained with HuBERT and MFCC features. Both the systems have been trained with the 'extended dataset'. As seen in the figure, for the random sample considered, the differences between the TV estimates by the two models are subtle and mostly indistinguishable.

3.6 Final Comparison on the Performance of best SI systems

This section presents the final results of the best performing SI systems discussed so far in the chapter. Since the two articulatory datasets used in our experiments (XRMB and HPRC) have their own unique TVs , separate SI systems are trained independently on the two articulatory datasets. The best performing SI systems are therefore discussed separately in the following sections.

3.6.1 SI systems trained with XRMB and estimating 6 TVs

Until recently, the SI systems were trained with 6 TVs generated by a TV transformation model (set of geometric transformations), which were first used in Sivaraman et al. (2019). However, a recent work in Attia et al. (2023a) proposed a new TV transformation model which made slight enhancements on a selected set of TVs. In the same work, the train, development and test splits were extended with additional data by fixing the corrupted and discarded articulatory ground-truth data in the original XRMB dataset (Attia and Espy-Wilson, 2023). The SI systems developed in section 3.2 (BiGRNN-dataAug) and section 3.4 (TCN-SF-Audspec) are trained to estimate the TVs with old geometric transformations and the SI systems developed in section 3.5 (BiGRNN-SSL) are trained to estimate the TVs with new geometric transformations. Hence, we

discuss the two different SI systems separately in he following sections.

Models trained to estimate TVs with old geometric transformations

Table 3.12: PPMC scores (Mean and Variance across 5 trials of training) for best performing SI systems

Model	**Acoustic Features**	**Average PPMC**
BiGRNN-dataAug	MFCC	0.7960 (0.011)
TCN-SF-Audspec	Auditory Spectrograms	**0.8770 (0.031)**
BiGRNN-SSL	HuBERT-large	0.8057 (0.003)

Models trained to estimate TVs with new geometric transformations

Table 3.13: PPMC scores (Mean and Variance across 5 trials of training) for best performing SI systems

Model	**Acoustic Features**	**Average PPMC**
BiGRNN-dataAug	MFCC	0.7955 (0.015)
TCN-SF-Audspec	Auditory Spectrograms	0.8112 (0.024)
BiGRNN-SSL	HuBERT-large	**0.8141 (0.002)**

3.6.2 SI systems trained with HPRC and estimating 9 TVs

Table 3.14 summarizes the best performing SI systems trained on the HPRC dataset and evaluated on a more challenging test split with 'normal' and 'fast' rate speech in the HPRC dataset (different from the test split in section 3.3). Here the BiGRNN-MTL model is based on the SI system trained with a multi-task learning framework in section 3.3. The TCN-SF-Audspec model is based on the SI systems in section 3.4. The BiGRNN-MTL and BiGRNN-SF-MTL trained with self-supervised HuBERT features are first discussed here, but follows the same architecture used

in section 3.3. Here the BiGRNN-SF-MTL model is trained by incorporating additional source features following the work in section 3.4.

Table 3.14: PPMC scores (Mean and Variance across 5 trials of training) for best performing SI systems

Model	**Acoustic Features**	**Average PPMC**
BiGRNN-MTL	MFCC	0.7118 (0.021)
BiGRNN-MTL-SSL	HuBERT-large	0.7571 (0.004)
BiGRNN-SF-MTL-SSL	HuBERT-large	**0.7591 (0.002)**
TCN-SF-Audspec	Auditory Spectrograms	0.7573 (0.021)

3.7 Summary

Results in this chapter highlights incremental improvements done to the SI framework to improve its generalizability and robustness in estimating articulatory representations. Section 3.6 summarizes all the best performing SI systems with respect to the two articulatory datasets used, and compares and contrasts their architectural differences in terms of model implementations, input acoustic features used and the nature of the target articulatory representations. A key observation of the results from the SI systems trained with XRMB and HPRC datasets is that the TCN-SF-Audspec SI system has performed comparatively better with both the datasets. However, it should also be noted that BiGRNN based SI system using HuBERT SSL features has slightly edged over the TCN-SF-Audspec SI system in both XRMB datasets with new TV transformations, and the HPRC dataset. This is a prime example to show that DNN model architectures are susceptible to domain differences, and a single model architecture may not fit every training dataset with their own unique distributions.

Section 3.4 presented the current state-of-the-art SI system trained on XRMB dataset (with old TV transformations) and is based on a TCN model and uses auditory spectrograms as the input speech representation. The same model is capable of estimating source features (aperiodicty, periodicity and pitch) along with 6 TVs. Section 3.5 on the other hand, presented the best performing BiGRNN SI system trained to estimate the enhanced TVs with the self-supervised HuBERT speech embeddings.

The availability of phoneme alignments with the HPRC dataset resulted in the multi-task learning based models which leverage on learning an acoustic-to-phoneme recognition task parallel to the SI task. This MTL based model performed comparatively better with respect to the single-task, conventional SI system. As shown in table 3.14, adding source features as additional targets and using HuBERT input acoustic features further improved the performance of the SI system. Hence, this will be used as the best performing SI system to estimate the articulatory representations with the HPRC dataset.

Chapter 4: Extending Speech Inversion Systems to Estimate Velar and Glottal Activity

4.1 Overview

Speech is produced by the coordinated movement of articulators such as tongue, velum, and lips that shape the acoustic signal produced by the larynx, forming alternations of vocal tract constriction (for consonants) and opening (for vowels) (Stevens, 2000). These movement patterns can differ according to the language, dialect, abilities, and habits of the speaker, but the fact that the movements themselves overlap in time means that the evidence of their movement in the acoustic signal can be compressed, scattered across time, and sometimes obscured by co-occurring events. The result is that many linguistic phenomena that are hard to express in acoustic terms are more readily explained by differences in the timing and degree of vocal tract constriction (Cho and Keating, 2009; Krivokapić, 2014). Systems that do speech inversion rely on ground truth articulatory variables; by using extracted acoustic features such as Mel Frequency Cepstral Coefficients (MFCCs), Mel-spectrograms, or the waveform itself as the input speech representation, the system can learn a mapping to the articulatory variables. However, none of the publicly available articulatory speech corpora have direct articulatory level data capturing the velar and glottal constrictions (Tiede et al., 2017; Westbury, 1994a). Therefore, most of the

available SI systems (trained on these datasets) are limited to estimating the articulatory level information pertaining to lip and tongue constrictions (Illa and Ghosh, 2018; Shahrebabaki et al., 2020; Siriwardena and Espy-Wilson, 2023; Siriwardena et al., 2023a; Udupa et al., 2021).

Acoustic-to-articulatory speech inversion inspired by Articulatory Phonology (Browman and Goldstein, 1992) maps the acoustic speech signal to the kinematic state of each constriction synergy (lips, tongue tip, tongue body, velum, and glottis) by its corresponding constriction degree and location coordinates, which are called vocal tract variables (TVs). In this work, we extend a speech inversion system based on TVs to estimate the activity of the velar constriction by collecting a dataset that can be effectively used in training a speaker-independent SI system. We choose 'Nasalance' as the ground-truth to capture nasality for two reasons. First, it is a non-invasive measure and can be easily collected from a larger population, which will be beneficial in building a more generalizable, speaker-independent SI system. However, nasalance measures the ratio of acoustical energy between the nasal and oral tract. Accordingly, as a variable it is dependent on the amount of energy flowing through the glottis, and thus has only an indirect relationship with VP articulation (Kochetov, 2020; Rong et al., 2011). Hence, the far reaching goal of the proposed SI system is not aimed at deriving aerodynamic relationships (such as nasalance) from the acoustic signal, but rather aimed at deriving VP articulatory movements. Our approach to achieve this goal was twofold. To investigate if nasalance is an accurate representation of velar constriction degree (Browman and Goldstein, 1992), we validated it with a more direct, invasive and accurate measure of VP activity called high-speed nasopharyngoscopy (HSN). To the best of our knowledge, this is the first time a SI system has been developed to estimate a proxy for a velar constriction degree TV, that will, in essence, capture the nasality in speech.

The second reason for using nasalance derives from this susceptibility to glottal source

effects. Learning a mapping from an acoustic representation that is rich with source level information (eg. Melspectrograms, auditory spectrograms) to nasalance, along with source features (eg. voicing and pitch) may positively influence the SI system performance for nasality prediction. To investigate such effects of using source features, Electroglottography (EGG) was synchronously collected to extract a voicing parameter, and aperiodicity, periodicity and pitch extracted from an aperiodicity, periodicity and pitch (APP) detector (Deshmukh et al., 2005) are also used as additional targets to further improve nasality prediction.

The content in the following sections are organized as follows. In section 4.2, we discuss the details of the dataset and explain the steps used to extract and validate the ground-truth nasalance parameter. In section 4.3, we highlight the details of the proposed SI system and the importance of using source features to estimate nasality. Finally in section 4.4, we discuss the key conclusions drawn from the experiments and possible future directions. The work in this chapter has been published in Siriwardena et al. (2023b).

4.2 Dataset

This work is based on a subset of data from an ongoing, collaborative data collection. The complete dataset, once collected, will be made public (subject to standard open source licensing agreements). One of the main goals of this dataset is to develop a speaker-independent speech inversion system to accurately estimate velar and glottal activity. The current dataset has been collected from 8 subjects (5 Female, 3 Male), and the demographic details of the speakers are listed in Table 4.1.

Table 4.1: Dataset Description. SW: South-west, C: Central, W: White, B: Black, H: Hispanic, NH : Non-Hispanic

Subject	**Gender**	**Language**	**HSN status**	**Age (years)**	**Ethnicity/Race**
1	M	English(SWOhio)	HSN	28	W, NH
2	F	English(STexas)	No HSN	24	W, H
3	F	English(SWOhio)	No HSN	31	W, NH
4	F	English(SWOhio)	No HSN	40	W, NH
5	F	English(CKentucky)	No HSN	28	B, NH
6	F	English(SWOhio)	HSN	34	W, NH
7	M	English(SWOhio)	No HSN	23	W, NH
8	M	English(SWOhio)	No HSN	35	W, NH

4.2.1 Ground-truth Nasalance Parameter

Background and Procedure for Data Collection

Since a direct observation of the velic constriction is difficult, several methods have been used in literature to estimate the VP port's activity in speech production. The simplest method is to measure "Nasalance". This is the relative proportion of nasal vs. oral acoustic output from two microphones (mic) mounted to the top and bottom of a separation plate located between the nose and upper lip to create an acoustic barrier. While the nasalance procedure is a simple, well-known, non-invasive and reliable technology for tracking velopharyngeal constriction, it is dependent on the amount of energy flowing through the glottis, and thus has only an indirect

relationship with velopharyngeal articulation (Kochetov, 2020). To our knowledge, the degree to which Nasalance data is sensitive to small changes in constriction degree and timing has not been previously determined. So as a secondary means of assessing velopharyngeal movement, we used a subset of speakers (subject 1 and subject 6) to synchronously collect a more direct, but invasive, measure of velopharyngeal constriction called High Speed Nasopharyngoscopy (HSN).

Figure 4.1 shows the setup used to collect the HSN and audio measurements to compute the nasalance parameter. For the HSN, a flexible scope (outer diameter: 2.2 or 3.6 mm) was connected to a video camera (MIRO 310; Vision Research, Inc., Wayne, New Jersey), and the images were captured at a rate of 1000 frames/second using 304 $\times$ 256 pixel resolution. To collect the audio data, 2 microphones (1/4", Type 4958, Bruel and Kjær, Duluth, Georgia) were connected to the top and the bottom of the separation plate made of aluminum. Windscreens were used to cover the microphones to prevent interference from airflow directed toward the microphones. The separation plate was placed against the participant's upper lip to create an acoustic barrier between the oral and nasal audio recordings. The acoustic data from the microphones were captured at 51.2 kHz using a data acquisition system (NI 9234, National Instruments, Austin, Texas) and customized LabVIEW code that digitized and converted the data to a ".wav" audio file. The initiation of the audio recording and imaging data (from the HSV nasopharyngoscopy) was synchronized using an input/output module (NI 9402; National Instruments) (Oren et al., 2020)

Using this setup, approximately 10 minutes of speech material per subject was recorded. This consisted of a mixture of short and long sentences and short paragraphs. For example, for nasality, the full set of prosodic nasal contrasts from Krakow et al. (1988) was included, including e.g. "hoe me" vs. "home E", "seam ore" vs. "Seymour". For voicing, sentences contrasting words such as "Dodd" vs. "Todd" in a carrier phase were included. Sentences illustrating consonant

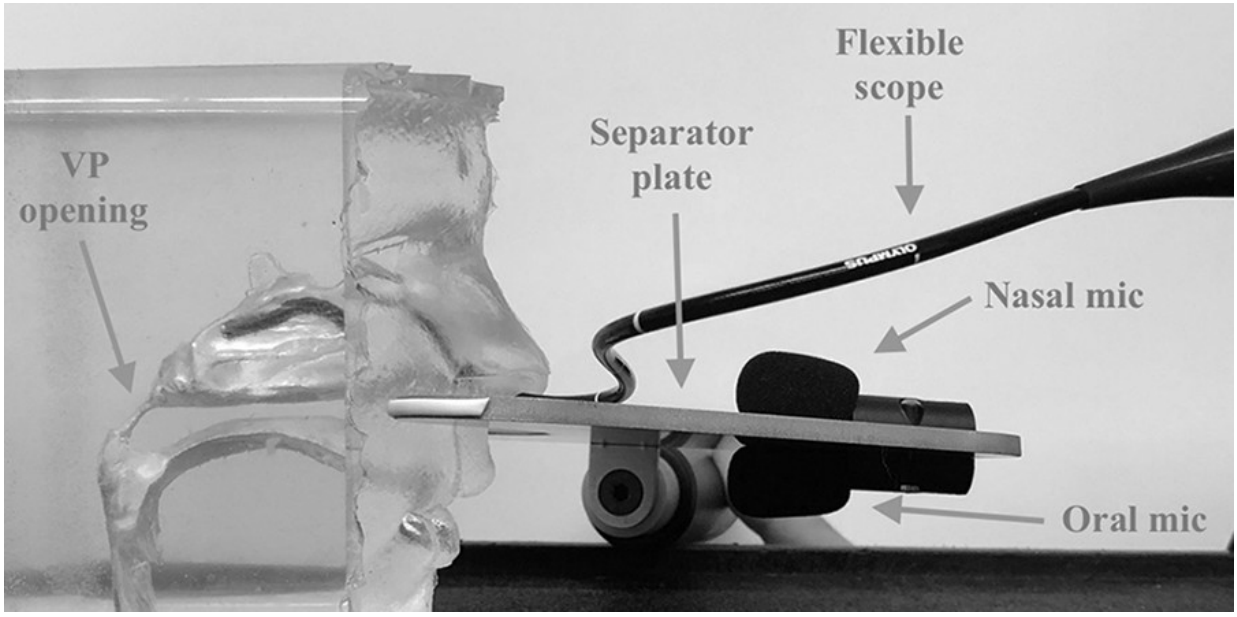

Figure 4.1: Illustration of the experimental setup. HSN measurements were taken by connecting a flexible scope to a high-speed video camera (not shown). The figure is taken from Oren et al. (2020) in The Cleft Palate-Craniofacial.

cluster articulatory patterns were drawn from (Zsiga, 1994; Zsiga and Nitisaroj, 2004). For cross-dataset comparison, we also included some sentences from speech materials used in the U.W.XRMB corpus (Westbury, 1994a).

Nasalance Parameter

Oral and nasal mic signals collected from the nasometer set-up were used to compute the nasalance parameter. The baseline wander was first removed from the two signals using a high pass filter (cutoff around 0.1Hz). The Root Mean Square (RMS) signals were then computed for both oral and nasal signals separately. During the RMS signal generation, both the squared signals were smoothed out using a moving average filter with a window size of 1000 ($\sim$ 20 ms) samples. Then a nasalance parameter (Nasalance$_{raw}$) was computed using the equation 4.1 based on Bunton and Story (2011). The Nasalance$_{raw}$ parameter was then downsampled to 100Hz and smoothed using a window of 10 samples (using Matlab function 'Fastsmooth' by (O'Haver, 2017)). The

final nasalance parameter was then normalized to [-1,1] range to be used as the ground-truth for the speech inversion system.

$$Nasalance_{raw} = \frac{RMS_{nasal}}{RMS_{nasal} + RMS_{oral}} \tag{4.1}$$

4.2.2 Validating Nasalance with HSN

HSN was synchronously collected from subject 1 and subject 6 to assess the accuracy and agreement with the computed nasalance parameter. Here the temporal dynamics of the VP port is captured by summing the light intensity in the images (intensity of pixels) of the high-speed video (HSV) data. The resulting intensity trace has been shown to be an accurate measure for capturing the velum (Oren et al., 2020). Figure 4.2 shows a sample HSV intensity trace and the corresponding HSV images at different points in time. Here an open VP port would be overall characterized by darker regions that come from the cavity of the VP port. On the other hand, a closed VP port would be characterized with brighter regions because of the increased amount of light reflecting off the tissue. It should be noted that the HSN parameter shows a trough (i.e. lower values) for nasal sounds in speech. This is in contrast to nasalance, which shows the opposite pattern of a peak.

The HSV data has a sampling rate of 1kHz and the nasalance parameter as discussed earlier is sampled at 100Hz. To match the number of samples to compute the cross correlations, the nasalance parameter is linearly interpolated to match with the HSV intensity trace. The Pearson correlation coefficients are then computed for each sample data from the subject. The average correlation coefficients across the samples for subject 1 and subject 6 are -0.6081($p<0.001$)

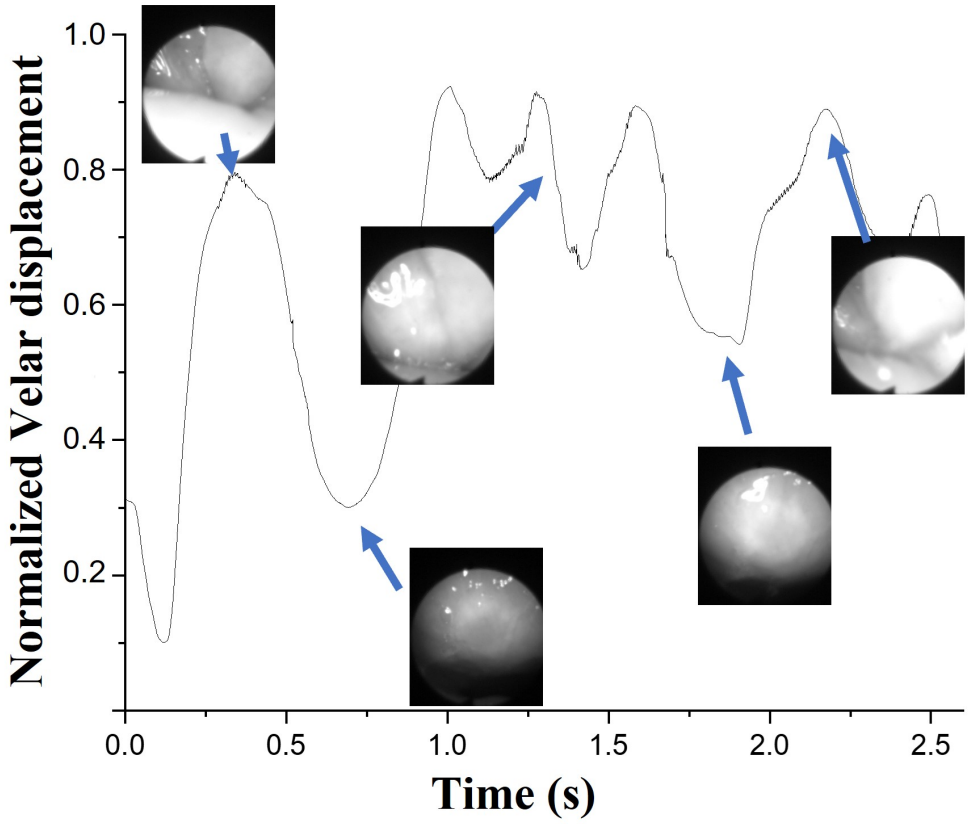

Figure 4.2: HSV intensity trace for a male native speaker of American English from Cincinnati, OH producing 'It's a see more, Sid. It's a seam ore, Sid'. Images of the VP port at key time points are indicated by arrows.

and -0.5136($p<0.001$) respectively. These statistically significant negative correlations give an important validation on the accuracy of the computed nasalance parameter with respect to HSN.

4.2.3 Patterns of timing for Nasality

A number of studies have shown that American English shows different patterns of velum raising and lowering (i.e. VP port constriction) according to syllabic organization (Krakow, 1999). As shown in Krakow (1999) an example of this pattern for "home E" vs "hoe me" is that the velum moves earlier and the VP port stays open longer when the /m/ is in the rime (home) than

when the /m/ is in the onset of the following word (me). The lip-velum coordination during the syllable-initial and -final nasal was also observed in Krakow (1999), where it has been noted that there is close temporal proximity between the end of velum lowering and the beginning of lip raising for the syllable-initial and a large offset between the end of velum lowering and the end of lip raising for syllable-final.

To see if the nasalance parameter will also showcase such patterns (word-initial vs word-final /m/) with respect to the HSN and lip movement, the words 'hoe me' and 'home e' were analyzed. Figure 4.3 shows the data for 'It's hoe me' and 'It's home e' collected from subject 1 in the dataset. To analyze the lip movement pattern, the lip aperture tract variable (LA TV) was extracted from the articulatory speech inversion system in Siriwardena and Espy-Wilson (2023). Both the HSN and nasalance patterns shown in Figure 4.3 replicate the timing patterns described in Krakow (1999) with respect to the LA TV. Data from a larger group of subjects is needed to further verify the pattern.

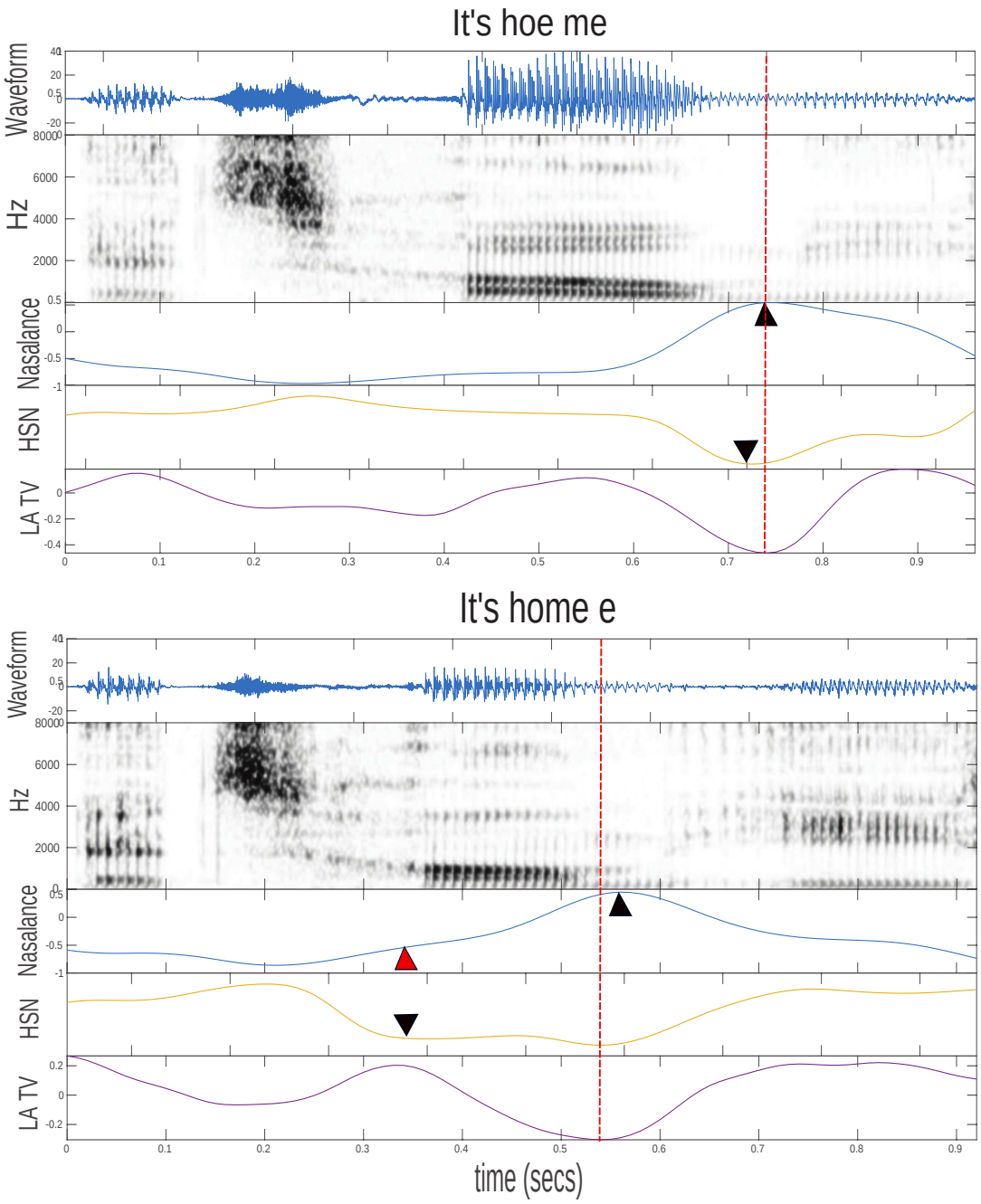

Figure 4.3: The vertical red dash lines in the top and bottom panels mark the onset of bilabial contact for the /m/. The black triangles in the HSN signal mark velum lowering onset. In the nasalance signals, the black triangle marks the observed velum lowering onset, whereas the red triangle in the bottom panel marks the actual velum lowering onset with nasalance.

4.2.4 Voicing parameter: EGG envelope

Electroglottography (EGG) is a well-established technology for tracking vocal fold oscillation, using the degree of electrical conductance across the glottal gap between electrodes placed on the two parallel outer sides of the throat. In this study, EGG data was also collected (from all the subjects) synchronously with the other HSN and audio measurements in section 4.2.1.

The EGG signal is sampled at 51.2 KHz, and to compute a parameter which can capture the voicing activity of speech, the envelope of the EGG signal was extracted. As with the nasalance parameter, we first high pass filtered the signal to remove the baseline wander. Then the magnitude of the Hilbert transform (Feldman, 2001) was computed as the envelope of the EGG signal. The envelope was downsampled to 100 Hz and smoothed and normalized the same way to the nasalance parameter to generate the final voicing parameter.

4.3 Speech Inversion System

4.3.1 Input Audio Representation

The audio recorded by the oral and nasal mic signals were mixed together to create a combined audio signal. The combined signal was then downsampled to 16kHz and segmented to 2 second long segments. The shorter, remaining segments were zero padded at the end. The segmentation was done mainly to increase the number of audio samples to train the DNN based SI system and to have input acoustic representations of fixed dimensionality to the input layer of the DNN model.

We used auditory spectrograms (Audspec) (Wang and Shamma, 1994) as the input speech

representation for the SI system. The auditory spectrograms have a logarithmic frequency scale and provide a unified multi-resolution representation of the spectral and temporal features likely critical in the perception of sound (Wang and Shamma, 1994).

4.3.2 Model Architecture and Training

Model Architecture

We developed a Temporal Convolution Network (TCN) based SI system inspired by the work in Siriwardena and Espy-Wilson (2023). The model was optimized using the Mean Squared Error (MSE) loss computed between the predicted parameters and the ground truth. The SI system was implemented in PyTorch with 1-D convolutional (CNN) layers. Figure 5.1 shows the proposed model architecture with its sub-modules used for pre-processing and dilated TCN. The pre-processing module contains two 1-D CNN layers with 1×1 kernels (C1, and C2), which have 128 filters each. The d1, d2 and d3 dilated CNN layers have a kernel size of 3 with 1,4 and 16 dilation rates respectively. Upsampling (window size 4) was done after C4 layer and average pooling (window size 5) was done after C5 layer along with BatchNorm layers after every CNN layer in the TCN network. The upsampling and average pooling operations take care of matching the time dimension of the input spectrograms to the target time dimension of TVs.

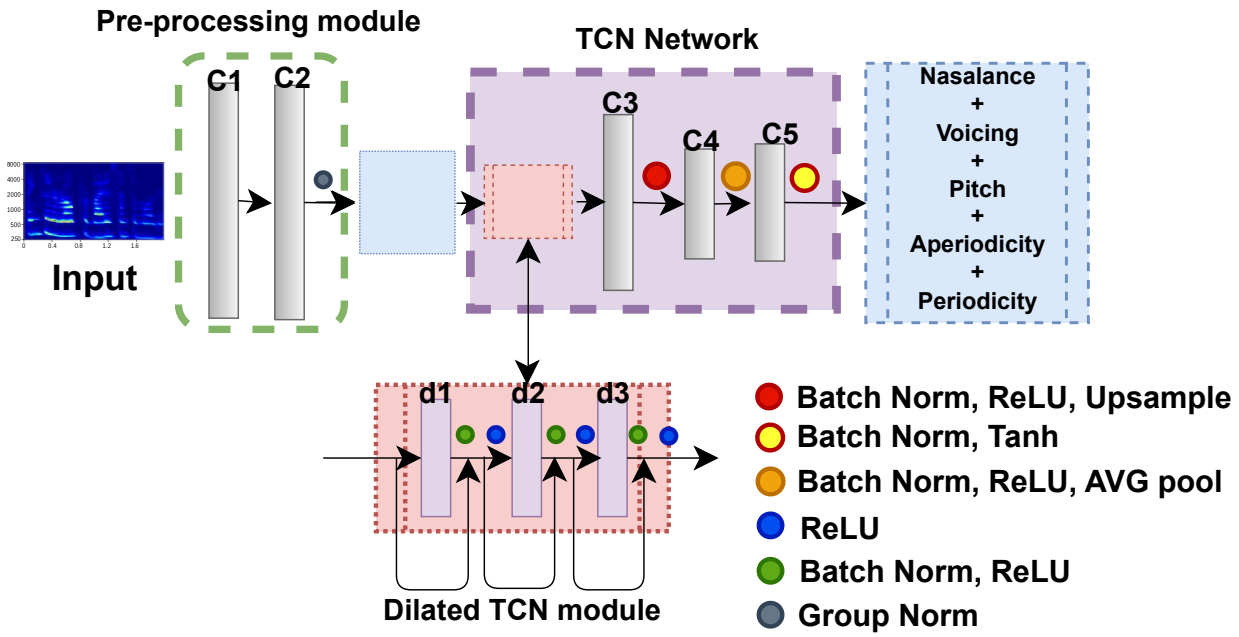

Figure 4.4: Model architecture (SI system). Here C1-C5 represent 1D-CNN layers and d1-d3 represent 1D dilated CNN layers

Model Training

All the model parameters were randomly initialized with a seed (=7) for reproducibility. Table 4.2 lists the hyper-parameters and the corresponding values considered to fine-tune the model. A grid search was performed when fine-tuning the hyper-parameters and the best parameters were chosen based on the validation loss. All the models were implemented with PyTorch machine learning framework and trained with NVIDIA TITAN X GPUs. The best performing model has around 1 million trainable parameters, takes around 8 minutes (± 2) to converge, and can be found in a Github repository[1]

[1] https://github.com/Yashish92/TCN-SI-tool-Nasality

Table 4.2: Hyperparameter Tuning for the TCN model

Parameter	**Possible Values**	**Chosen Values**
Learning Rate	[1e-4, 3e-4, 1e-3, 1e-2]	1e-3
Batch size	[16,32,64,128]	64
Optimizer	ADAM, RMSprop, SGD	ADAM
Rate scheduler	ExponentialLR, PolynomialLR	ExponentialLR

The dataset was divided into training, validation and testing splits, so that the training set has utterances from 6 speakers (4 females, 2 males). The validation and testing splits have data from 2 speakers (1 male, 1 female) with 1/2 of the data from each speaker in the validation split and the other half in the test split. None of the data from the speakers in the validation and test splits were included in the training split and hence all the models are trained in a 'speaker-independent' fashion. The splits also ensured that around 70% of the total number of utterances were present in training (1 hour of speech), and all the allocations were done in a completely random manner.

4.3.3 Results of Speaker-independent Speech Inversion

Two speech inversion systems were trained to estimate the nasalance parameter from the input auditory spectrograms. Pearson Product Moment Correlation (PPMC) score is used as the metric to evaluate the predictions by the SI systems. Table 4.3 shows the PPMC scores for correlations between the estimated and ground-truth nasalance parameter for the systems trained with additional source features as targets (SI-SF) and the one with nasalance parameter as the only target (SI-noSF).

Figure 4.5 shows sample nasalance estimation by the SI-SF and SI-noSF models for an

utterance in the test set. The utterance, 'Say tube again' contains a nasal consonant [n] around 1.15-1.25 seconds which is captured by both the SI systems. However, it is important to note that the nasalance parameter estimated by the SI-SF model has better agreement with the ground-truth compared to the SI-noSF model.

Table 4.3: PPMC scores (mean and .std across 8 trials) for the SI systems trained with and without source features as additional targets to estimate nasalance.

	Nasalance	**Voicing**	**Perio.**	**Aperio.**	**Pitch**	**Average**
SI-SF	**0.7341(0.02)**	0.80541(0.01)	0.9008(0.03)	0.8257(0.02)	0.7995(0.03)	0.8131(0.03)
SI-noSF	0.6967(0.02)	-	-	-	-	-

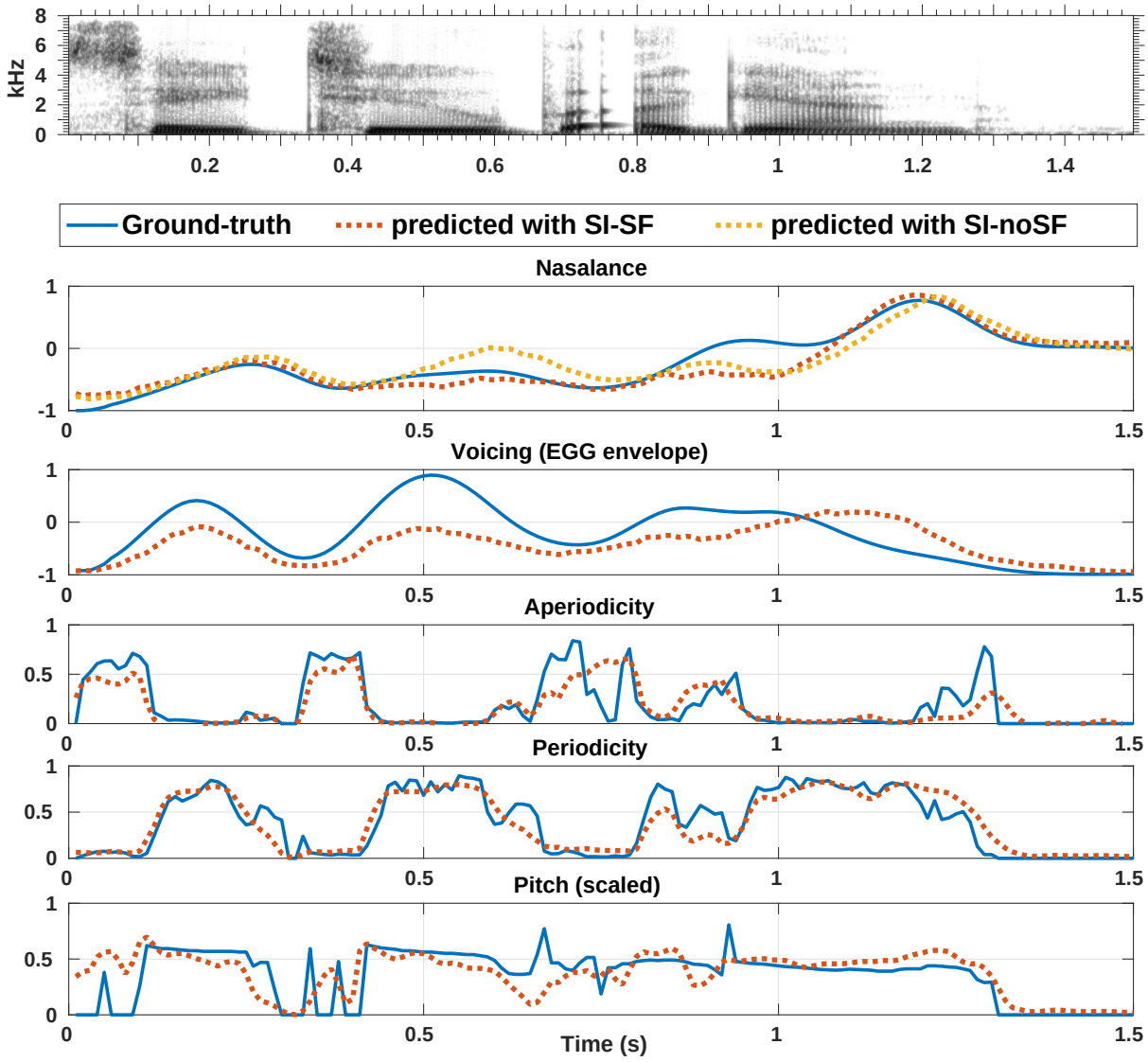

Figure 4.5: Nasalance and source features for the utterance 'Say tube again' estimated by the SI-SF model and nasalance estimated by the SI-noSF model with respect to the ground-truth . Solid blue Line - ground truth, red dotted line - predictions by the SI-SF, yellow dotted Line - predictions by SI-noSF.

4.4 Discussion

The results of correlation analysis in the section 4.2.2 gives a general, but an important validation for the nasalance parameter with respect to the more direct HSV intensity trace. The fact

that we found known patterns of timing for nasality (discussed in section 4.2.3) further supports the validity of using nasalance as a proxy variable for velopharyngeal constriction. This work highlights the performance of our SI system in estimating velopharyngeal movement dynamics for unseen speaker data. It also shows that incorporating source features as additional targets improves the estimation accuracy of the velopharyngeal movement parameter. This is consistent with the observations made in Siriwardena and Espy-Wilson (2023) with conventional acoustic-to-articulatory speech inversion, and could also suggest that the TCN model is particularly sensitive to source/VP interactions.

In future work, the authors plan to improve the performance and generalizability of the current SI system by training on data from a larger group of subjects (from the ongoing data collection). More emphasis will also be made on validating and fine tuning the nasalance parameter as a proxy to the velar TV. Further experiments will also be done to understand what the DNN models are actually picking as source-filter interactions that are ultimately helping the overall SI task.

4.5 Summary

To summarize, in this work we present the details on a dataset collected to estimate the velar and glottal activity in speech. We particularly looked into estimating a validated nasalance parameter (as a proxy to a velar TV) using a speaker-independent SI system. It should be noted, that having a SI system to estimate parameters directly related to the velar (and glottal) constrictions can be hugely beneficial, since it gives an almost complete articulatory level representation of speech which can be useful in diverse speech applications (eg. articulatory speech synthesis (Siriwardena

et al., 2023c; Wu et al., 2022)). An accurate, validated speech inversion system would also be a significant breakthrough for researchers with little or no ability to collect articulatory data directly, e.g. scholars without well-equipped phonetics laboratories, scholars doing field studies in dispersed communities. While speech inversion data is not equivalent to direct observation, it may enable hypothesis formation and testing that will motivate more targeted studies.

Chapter 5: MirrorNet: Learning Articulatory Representations inspired by Sensorimotor Interactions

5.1 Overview

This chapter discusses the application of an unsupervised learning algorithm called the 'Mirror Network' (MirrorNet), which is inspired by the existence of bidirectional interactions between the auditory and motor cortices of the human brain. This naturally motivated algorithm is first applied to learn control parameters to drive a parametric audio synthesizer, to synthesize a given input melody of notes. With the success made with learning control parameters from musical notes, which have a simpler spectra-temporal characteristics, the same algorithm is then experimented with learning articulatory representations for an arbitrary, continuous speech signal. To that end, a DNN based articulatory synthesizer (motor plant/vocal tract) is first developed and then integrated with the MirrorNet, to perform the acoustic-to-articulatory speech inversion. The work in this chapter has been published in Siriwardena et al. (2022a) and Siriwardena et al. (2023c)

5.2 Motivation for the MirrorNet

Most organisms function by coordinating and integrating sensory signals with motor actions to survive and accomplish their desired tasks. For instance, visual and auditory signals guide animals to navigate their surroundings (Keller et al., 2012; Wolpert and Ghahramani, 2000). Similarly, auditory and proprioceptive percepts are essential in skilled tasks like playing the piano or speaking. The difficulty of learning to perform these tasks is enormous. It stems from the fact that to control such actions, one needs to harmoniously close the loop between sensing and action. That is, it is necessary to map the desired sensory signals to the correct commands, which in turn produce exactly the desired sensory signals when executed. For example, a pianist guided by vision (sheet music) or audition (musical excerpt) can move his fingers over the keyboard to reproduce the desired music. Equally impressive and more common is our ability to articulate a desired speech utterance and reproduce it in our auditory brain with remarkable ease.

But to learn the necessary mappings and interactions between the perception and action domains, standard Artificial Intelligence (AI) methodology typically relies on creating large databases that map the input sensory data to their corresponding actions, and then train intervening Deep Neural Networks (DNN) to associate the two domains (Fu et al., 2019; Tai and Liu, 2016). For example, to teach an autonomous vehicle to move from point A to B, one needs to create a large array of trajectories that navigate around any potential obstacles and contingencies, and generate the controls and commands to accelerate, brake, and switch directions so as to accomplish the journey (Kiran et al., 2021). Humans and animals however never learn complex tasks in this way. For instance, human infants learn to speak by first going through a "babbling" stage as they learn the "feel" or the range and limitations of their articulatory commands. They also listen

carefully to the speech around them, initially implicitly learning it without necessarily producing any of it. When infants are ready to learn to speak, they utter incomplete malformed replica of the speech they hear. They also sense these errors (unsupervised) or are told about them (supervised) and proceed to adapt the articulatory commands to minimize the errors and slowly converge on the desired auditory signal. In other words, learning these complex sensorimotor mappings proceeds simultaneously and often in an unsupervised manner by listening and speaking all at once (Kuhl, 2004; Pagliarini et al., 2021; Shamma et al., 2020).

The inspiration of the MirrorNet also comes from the area of computational neuroscience and especially to learning and predictive processing. Our brain is able to extract strong relations between sensory stimuli and their corresponding motor parameters that enable children to learn to speak by mere passive exposure to speech without any proper external teaching. In addition, after learning to control their own vocal tract, adults can, without any additional training, produce sounds they hear even if the acoustic target is not reachable by their specific vocal tract. However, the brain is able to find a set of motor parameters that approximate well the target sound while being produced by the specific vocal tract. Such predictive mechanism can also be seen in music production when humans learn how to play an instrument by mapping the auditory stimulation to the motor commands to a specific instrument. Even music perception rely on similar predictive pathways where high-order cortical areas constantly predict activation in the auditory cortices in order to modulate attention and emotions (Di Liberto et al., 2021; Marion et al., 2021).

Finally, from an engineering perspective, the MirrorNet can solve problems where it is hard to find a reasonable number of examples to train a regular DNN network, or to learn from examples that may not be exactly similar to the motor-plant outputs, e.g., learning to synthesize a melody from naturally played music. We moreover believe that the MirrorNet can be generalized

to design algorithms that can control motor-plants far reaching as self-driving vehicles given various sensory data.

5.3 MirrorNet for learning audio synthesizer controls

Motivated by such learning of complex sensorimotor tasks, a new autoencoder architecture, referred to as the "Mirror Network" (or MirrorNet) was proposed in Shamma et al. (2020). The essence of this biologically motivated algorithm is the bidirectional flow of interactions ('forward' and 'inverse' mappings) between the auditory and motor responsive regions, coupled to the constraints imposed simultaneously by the actual motor plant to be controlled. In this work, we extend and demonstrate the efficacy of the MirrorNet architecture in learning audio synthesizer controls/parameters to synthesize a melody of notes using a commercial, widely available audio synthesizer (DIVA) developed by U-He[1].

The MirrorNet was initially proposed as a model for learning to control the vocal tract and is based on an autoencoder architecture. The structure of this network is shown in Figure 5.1 (Shamma et al., 2020), depicting the biological structures and experiments that motivated the network. The goal of the model is to learn two neural projections, an inverse mapping from auditory representation to motor parameters (Encoder), and a forward mapping from the motor parameters to the auditory representation (Decoder). For simplicity, we use auditory spectrograms (Wang and Shamma, 1994) generated from the audio streams as the input and output representations, but other representations may prove more versatile (e.g., cortical representations (Chi et al., 2005)). The "motor" parameters in this study are the parameters needed to synthesize the closest possible audio signals matching the inputs. The primary difference between this MirrorNet and the previously

[1] https://u-he.com/products/diva/

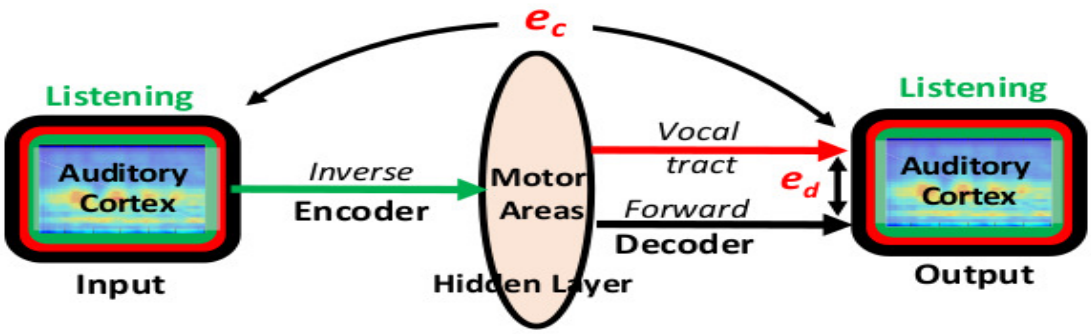

Figure 5.1: MirrorNet: Autoencoder Architecture

studied model in Shamma et al. (2020) is the use of the music synthesizer (DIVA) with its unique set of parameters.

As shown in Figure 5.1, the MirrorNet model is optimized simultaneously with two loss functions, namely the 'encoder loss'(e_c) and the 'decoder loss'(e_d). The encoder loss is the typical autoencoder loss - the Mean Squared Error (MSE) between the input auditory spectrogram and the reconstructed auditory spectrogram from the decoder (forward path). The decoder loss is the MSE between the auditory spectrograms generated by the DIVA (the motor plant path) and the decoder (forward path). It is the 'decoder loss' that constrains the latent space to converge to the expected control parameters while simultaneously reducing (e_c), and this is one of the distinctive features of the MirrorNet architecture.

Figure 5.2 shows the role of the 'forward' path in the model, namely to back-propagate the errors computed to learn the 'inverse' mapping and hence the control parameters. In general, directly computing a vocal-tract or an audio synthesizer inverse is difficult, if not impossible because of its complexity, nonlinearity, and our incomplete knowledge of its workings. The MirrorNet in Figure 5.2 (bottom panel) solves this problem by adding the forward projection that serves as a parallel, "neural" model of the vocal tract or the audio synthesizer, or any motor-plant

to be used. The critical importance of this "neural" projection is that it readily provides a route for the e_c errors to back-propagate to the motor areas (latent space), enabling the training of the inverse mapping (Encoder).

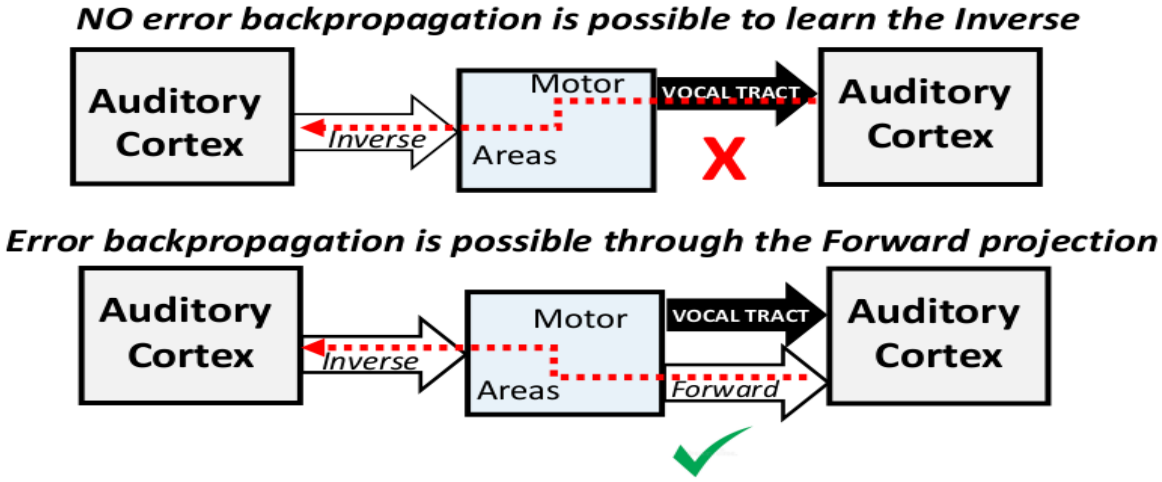

Figure 5.2: Role of the forward pass

5.3.1 Model implementation and training

The MirrorNet for audio synthesizer control is implemented in PyTorch with 1-D convolutional (CNN) layers modeling both the encoder and decoder. The complete network is inspired by the multilayered Temporal Convolution Network (TCN) (Lea et al., 2017). Figure 5.9 shows the complete DNN model architecture with its sub-modules used for pre/post processing and dilated TCN. The pre/post processing modules are symmetrically matched (C1≡C12, C2≡C11, C3≡C10) and have 128, 256 and 256 filters respectively with 1×1 kernels. d1, d2 and d3 dilated CNN layers have a kernel size of 3 with 1,4 and 16 dilation rates respectively. The CNN layers in the encoder and decoder are also symmetrically matched and the C4, C5 and C6 layers have 256, 128 and 7 filters respectively with 1×1 kernels. The latent space dimensions are chosen to match

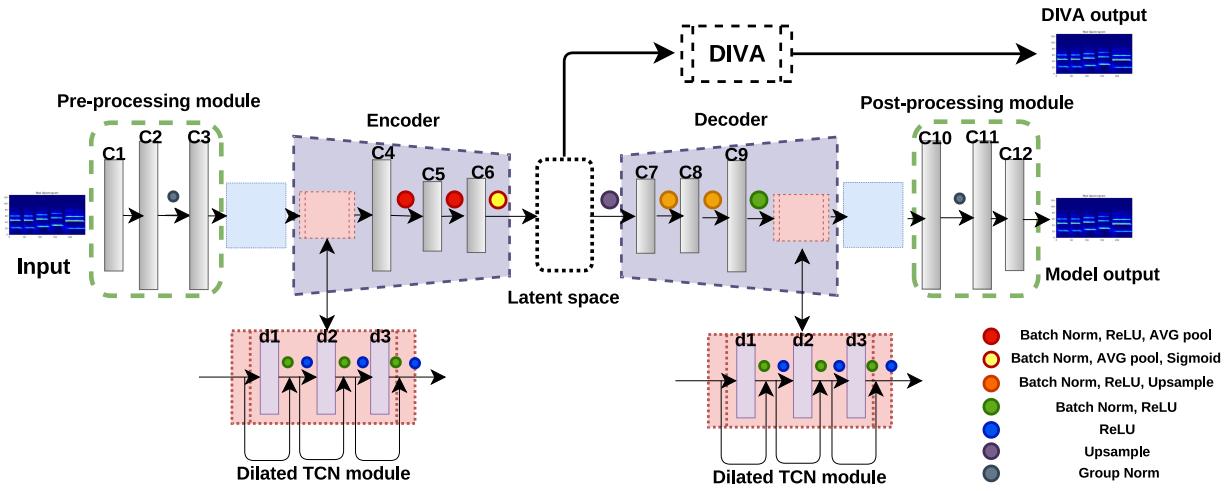

Figure 5.3: DNN architecture of the MirrorNet model. Here C1-C12 represent 1D-CNN layers where d1-d3 represent 1D dilated CNN layers.

with the number of parameters to be learned and the number of notes in each melodic segment. For example to learn 7 controls of the DIVA synthesizer to generate a melodic segment of 5 notes, we use a latent space of (7×5) dimensions. Average pooling is done after C4, C5 and C6 layers (window sizes of 5, 5 and 2 respectively) while upsampling is done before C7, C8 and C9 layers (window size of 2, 5 and 5 respectively). The auditory spectrograms used as inputs (and outputs) of the model are of dimension (128×250). We use auditory spectrograms which have a logarithmic frequency scale, simply because they provide a unified multi-resolution representation of the spectral and temporal features likely critical in the perception of sound (Chi et al., 2005; Wang and Shamma, 1994).

Unlike a regular autoencoder, the MirrorNet is trained in two alternating stages in each iteration. The decoder is trained first (to minimize e_d) for a chosen number of epochs. Then, the encoder is trained (to minimize e_c) for a given number of epochs and this alternation of training is continued until both losses converge to a minimum. Learning rates of 1e-2 and 1e-3 were used for the encoder and decoder networks, respectively. The best learning rates were determined based on

a grid search testing all the combinations from [1e-2, 1e-3, 1e-4, 3e-4] for both the encoder and decoder which result in the lowest training errors at convergence. The two objective functions were optimized using the ADAM optimizer with an 'ExponentialLR' learning rate scheduler and a decay (gamma) of 0.5. All the models were trained using NVIDIA Quadro P6000 GPUs and on average the models converged after around 32 hours of training. For further implementation information of the network, the PyTorch project is publicly available in GitHub [2]. Sample audio reconstructions can also be found in the supporting web page hosted in the GitHub repository.

5.3.2 DIVA audio synthesizer

We use DIVA, an off-the-shelf commercial synthesizer as our audio synthesizer for the MirrorNet model. DIVA has almost all its parameters MIDI-controlled. A python library called RenderMan [3] is used to batch-generate audio files using a fixed set of parameters. We built a software layer with RenderMan to drive DIVA to synthesize a melody of notes by concatenating individual notes synthesized by DIVA. All the melodies used in this paper are 2 seconds long and sampled at 44.1 kHz. The parameters are all continuous and normalized between [0,1]. Table 5.1 lists the set of parameters selected for the learning experiments with the MirrorNet, and the corresponding parameter labels from DIVA where applicable.

[2] https://github.com/Yashish92/MirrorNet-for-Audio-synthesizer-controls
[3] https://github.com/fedden/RenderMan

Table 5.1: Set of Audio controls/parameters used. Here MIDI note and MIDI duration are parameters set in RenderMan library to drive the synthesizer patch.

Parameter Name	**DIVA preset**
MIDI note (Pitch)	-
MIDI duration	-
Volume	OSC : Volume2
Band pass filter (center frequency)	VCF1: Frequency
Filter Resonance	VCF1: Resonance
Envelope Attack	ENV1: Attack
Envelope Decay	ENV1: Decay
Vibrato Rate	LFO1: Rate
Vibrato Intensity	OSC : Vibrato
Vibrato Phase	LFO1: Phase

5.3.3 Learning DIVA parameters from melodies synthesized with the same set of parameters (set 1)

In this first experiment, we use 400 melodies (set 1) to train the MirrorNet and test with 80 melodies, all originally synthesized by DIVA. The advantage of this set of melodies is that we have its ground-truth parameter values, and hence we can assess how accurately the MirrorNet rediscovers them and reconstructs the melodies. Each melody contains 5 notes and is 2 seconds long. The train and test set of melodies were synthesized by randomly sampling a total of 7 parameters (the first 7 parameters in Table 5.1) using a defined range and keeping a pre-defined

set of other parameters constant across all notes and melodies. The pre-defined set of parameters used for the experiments can be found in the GitHub repository of the project.

Figure 5.4 depicts auditory spectrograms of a given melody at various stages in the fully-trained MirrorNet. The spectrogram (b) suggests how well the decoder has learned to generate an identical spectrogram to the one generated with DIVA for the exact same controls. The spectrogram (d) suggests how well predicted DIVA controls are from the encoder to synthesize an identical melody to the input.

We performed preliminary statistical tests to evaluate the robustness of the MirrorNet in predicting the 7 parameters. Plot in Figure 5.5a validates that the predicted and ground truth parameters are significantly closer together than would result from a random set of values. A second test was performed to check how well the predictions of each parameter are compared to a random prediction. For that we performed a Levene's test that confirmed that all parameter predictions were significantly better than chance. Plot in Figure 5.5b shows the parameter difference distributions for the test set. The distributions also suggest that critical parameters like pitch, bandpass filter, filter resonance and duration are predicted with significant accuracy where as volume and envelope attack parameters are predicted with comparatively lower accuracy.

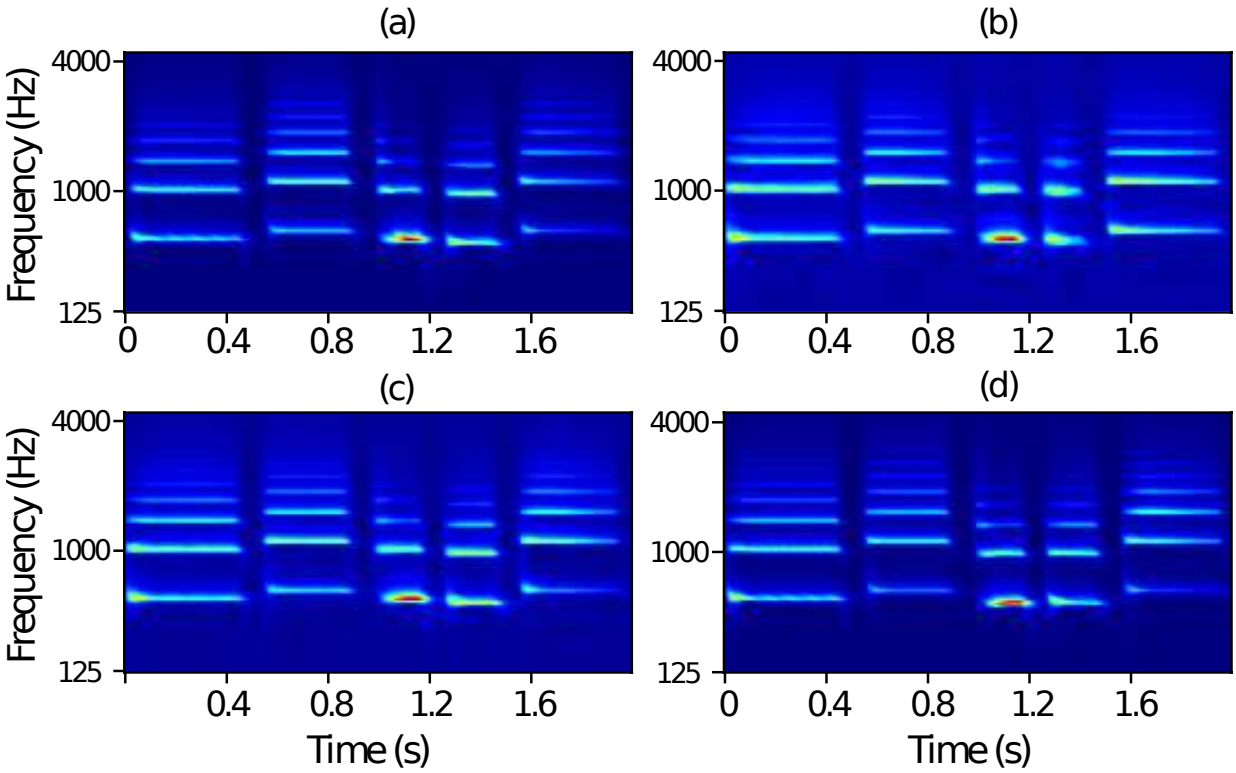

Figure 5.4: Auditory spectrograms from the model learned with DIVA synthesized melodies (set 1). (a) Input melody (b) Decoder output from true DIVA parameters (c) Final output from the decoder (d) DIVA output from the learned control parameters

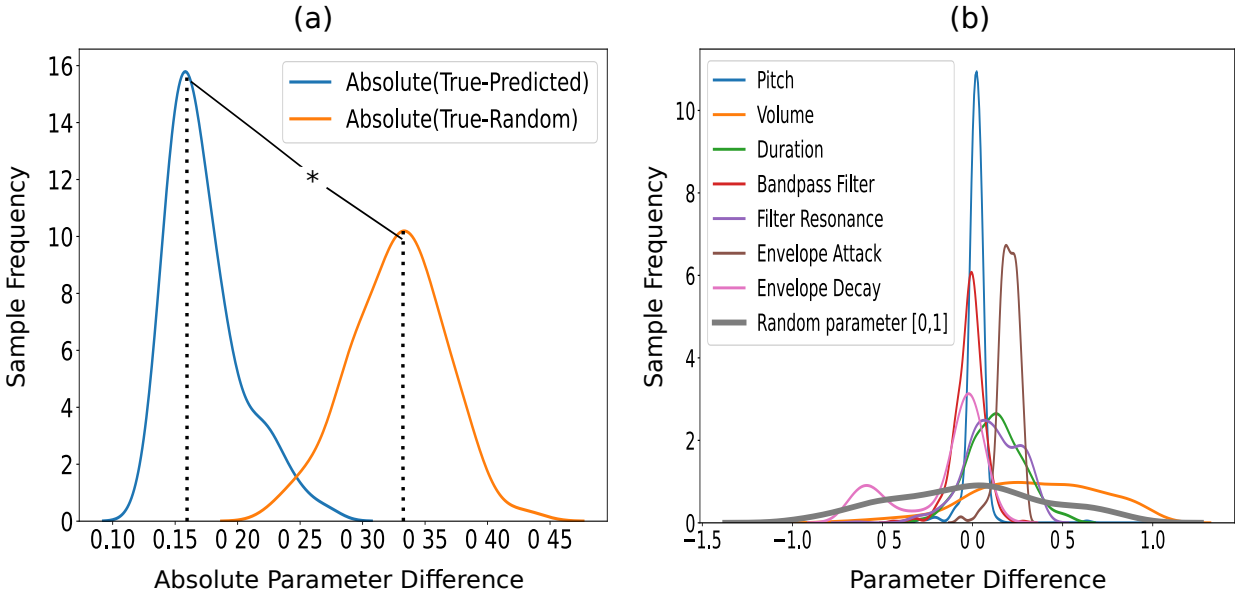

Figure 5.5: Evaluating statistical significance of the predicted DIVA parameters with respect to a set of random parameters on the test set (a) Distributions for absolute parameter differences across all parameters (b) Distributions of parameter differences (ground truth - predicted) for 7 parameters and the distribution for a random parameter difference (ground truth - random)

5.3.4 Learning DIVA parameters from melodies synthesized with extra unknown DIVA parameters (set 2)

In this experiment, we use a train set of 400 and a test set of 80, both DIVA generated melodies (set 2) which are synthesized in similar fashion to set 1 except for the fact that they now use all the 10 parameters in Table 5.1. The MirrorNet is still trained to predict 7 parameters as in previous experiment. The goal here is to demonstrate that the MirrorNet can approximate the input melodies even if they have additional sound/musical qualities that are impossible for the restricted set of 7 DIVA parameters to reproduce, e.g., vibrato in this case. The top panel in Figure 5.6 illustrates the original (vibrato) notes and the successfully regenerated melody captured with only 7 parameters (vibrato not included).

5.3.5 Learning DIVA parameters to synthesize melodies generated from other synthesizers

A fundamental advantage of the MirrorNet is its ability to discover the DIVA parameters corresponding to music generated by other sources and synthesizers by finding parameters that allow the DIVA output to be as close as possible, given the constraints of the number of parameters (here 7 are used), to the original input. The experiment utilized 400 5-notes long piano melodies of 2 seconds that are synthesized by a Fender Rhodes digital imitation (Neo-Soul Keys generated trough Kontakt 5). The network successfully reproduces accurate renditions of the piano music from unseen samples (test set of 80 samples) using the decoder/encoder mappings learned during the training. The bottom panel in Figure 5.6 shows such an example where the DIVA produces a

melody which closely resembles the input piano melody.

This idea opens up a whole new area of applications in music synthesis as it describes a tool to find parameters for an arbitrary synthesizer that maximally approximate an arbitrary sound without being necessarily capable to exactly reproduce it (reproduce a violin using a guitar for instance). It should also be noted that here we only tried synthesizing fixed duration melodies with a fixed number of notes, but it is a step in the right direction to synthesizing a piece of music which can have a variable number of notes in a fixed frame of audio.

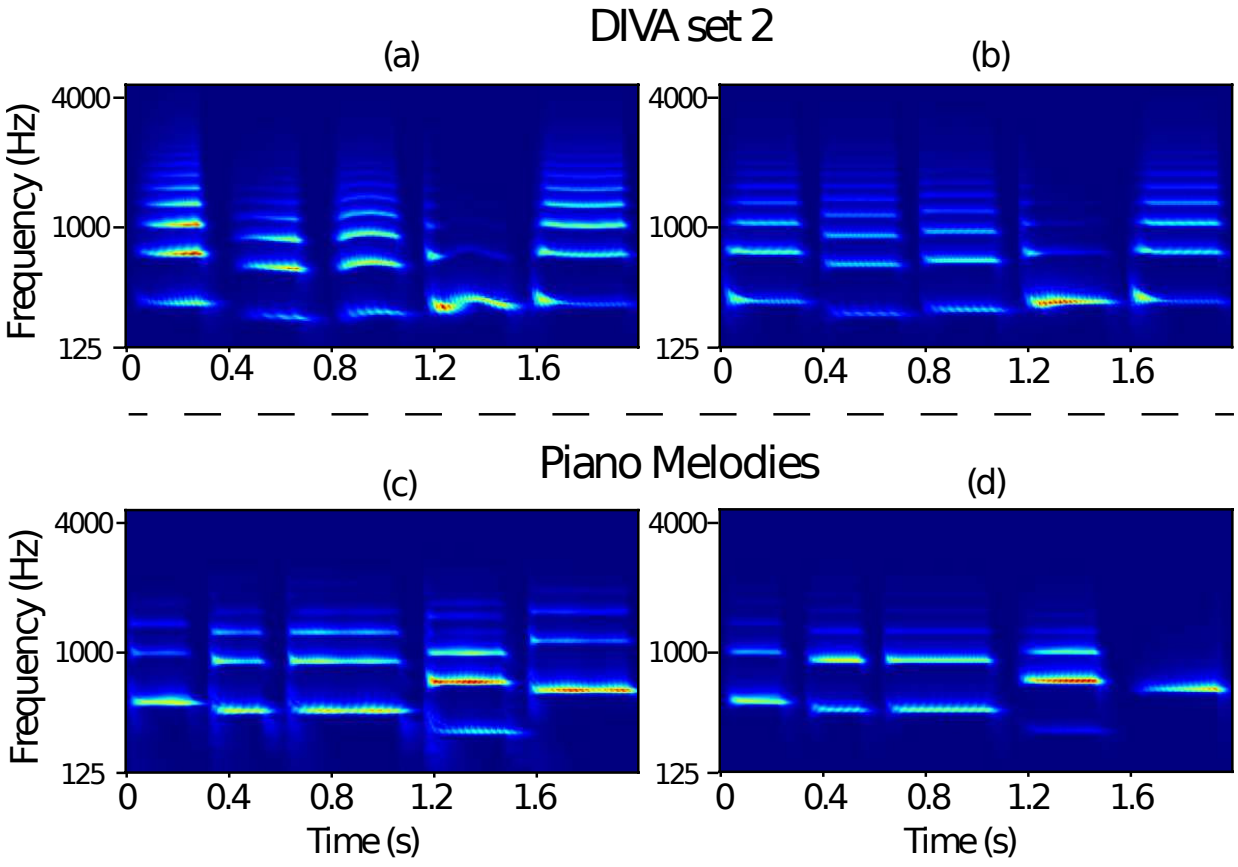

Figure 5.6: (Top panel) Auditory spectrograms from the model learned with DIVA synthesized melodies (set 2) (a) Input melody (b) DIVA output from the learned control parameters. (Bottom panel) Auditory spectrograms from the model learned with piano melodies. (c) Input melody (d) DIVA output from the learned control parameters.

Table 5.2: Mean and variance of Mean Square Errors (MSE's) across multiple model training runs

Input melody type	**Train/Test for Input vs DIVA(learned)**	**Parameter-Train**	**Parameter-Test**
DIVA melodies (set 1)	2.995±.21/3.596±.15	0.0666±.003	0.0671±.002
DIVA melodies (set 2)	6.380±.34/8.101±.20	0.0832±.007	0.0857±.004
Piano melodies	4.585±.25/4.751±.22	-	-

5.4 Acoustic-to-Articulatroy Speech Inversion with the MirrorNet

In the case of acoustic-to-articulatory speech inversion, the same MirrorNet algorithm can be applied to predict TVs as the latent space of the autoencoder architecture by using a vocal tract model or a TV based articulatory speech synthesizer as the motor plant. In this section, we first talk about how a TV based articulatory speech synthesizer is developed with a DNN based model and subsequently presents the results for carrying out acoustic-to-articulatory speech inversion by incorporating the designed synthesizer with the MirrorNet.

5.4.1 The articulatory synthesizer

Previous work has elaborated that articulatory parameters or gestures can be used to synthesize continuous, co-articulated and intelligible speech which can replicate realistic models of the vocal tract (Birkholz et al., 2006; Maeda, 1990). Articulatory based speech synthesizers can be mainly divided into two categories: (i) Geometrical approaches which model the geometry of the vocal tract (Story, 2013; Toutios et al., 2011), and, (ii) Machine learning based (Bocquelet et al., 2014a; Wu et al., 2022) which learn the non-linear relationships between articulatory representations and acoustic data.

Machine learning approaches typically need larger articulatory-acoustic datasets for model training where as physical models require high computational power and are comparatively slower than machine learning models at real-time synthesis (Bocquelet et al., 2016). However, the machine learning based articulatory synthesizers also have the limitation of speaker dependence where a synthesizer trained with one subject's articulatory data will not produce as good acoustic outputs for a different speaker (Bocquelet et al., 2016). Most of these synthesizers also do not learn a direct mapping from the articulatory variables to the acoustic waveform. But, learn an intermediate representation or an acoustic feature (MFCCs, Melspectrogram etc.) which will then be used as input to an off-the-shelf vocoder (parametric vocoders like World (MORISE et al., 2016) or neural vocoders like LPCNet (Valin and Skoglund, 2018)) to synthesize the final waveform (Bocquelet et al., 2016; Georges et al., 2020; Illa and Ghosh, 2019a). One key limitation of these vocoder based approaches is that the DNN model only learns a mapping from articulatory variables to the filter level parameters like MFCCs, and the source level information like pitch and aperiodicity (voicing) are directly extracted from the original waveform itself and then fed to the vocoder at the time of synthesis (Georges et al., 2020; Illa and Ghosh, 2019a).

In this work, we focus on developing a DNN based articulatory synthesizer trained and evaluated on a publicly available articulatory dataset. We developed a Temporal Convolution Network (TCN) based articulatory speech synthesizer to learn the mapping from 6 TVs, aperiodicity, periodicity and pitch to the audiotory spectrograms. The model is trained with the six ground truth TVs computed from the XRMB dataset along with aperiodicity, periodicity and pitch computed from the Aperiodicity, Periodicity and Pitch detector (Deshmukh et al., 2005). The model is optimized using the MSE loss computed between the predicted auditory spectrogram and the true auditory spectrogram of the input utterance.

The XRMB dataset was divided into training$_{full}$, development, and testing splits, so that the training$_{full}$ set has utterances from 36 speakers and the development and testing sets have 5 speakers each (3 males,2 females). A subset of the subjects (4 speakers) from the training$_{full}$ split is used to create an initialization split. The resulting training split, training$_{red}$ (after subtracting the data from 4 speakers) along with the initialization split are used in the experiments in sections 5.4.1.2 and 5.4.2.4. None of the training$_{full}$, development and testing sets have overlapping speakers and hence all the models are trained in a 'speaker-independent' fashion.

The TV based synthesizer is implemented in PyTorch with 1-D convolutional (CNN) layers. The complete network is inspired by the multilayered Temporal Convolution Network (TCN) (Lea et al., 2017) and figure 5.7 shows the complete DNN model architecture with its sub-modules used for post processing and TCNs. The learning rates for training the model were determined based on a grid search by testing all the combinations from [1e-2, 1e-3, 1e-4, 3e-4] which resulted in 1e-4 to be the best pick. The objective function is optimized using the ADAM optimizer with an 'ExponentialLR' learning rate scheduler and a decay (gamma) of 0.5. All the models were trained using NVIDIA Quadro P6000 GPUs by monitoring the validation loss, and on average the models converged after around 2 hours of training.

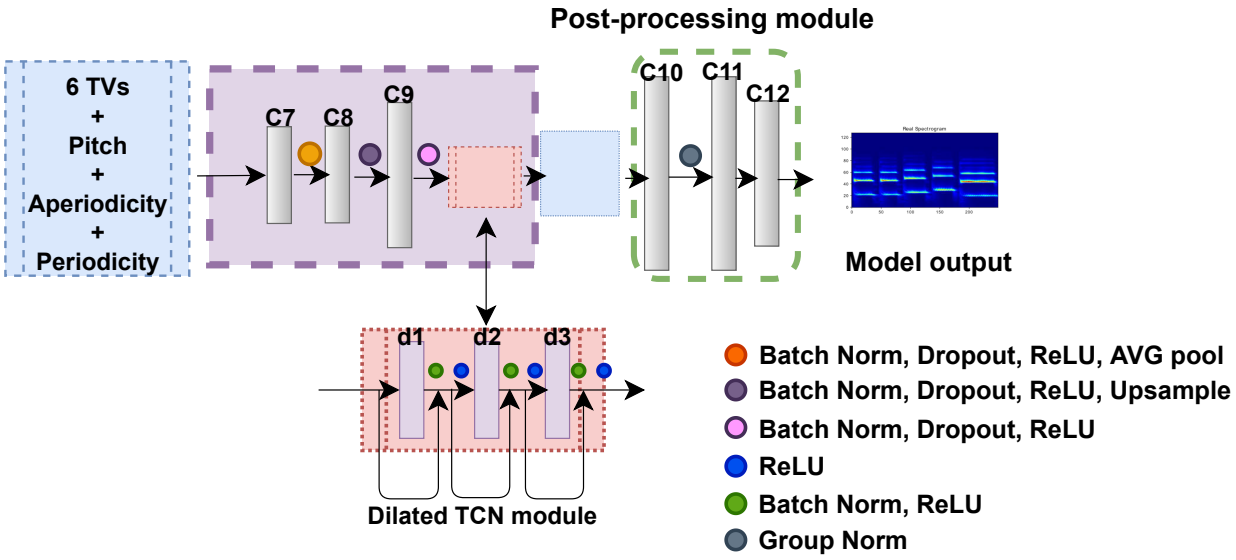

Figure 5.7: TV synthesizer model architecture

5.4.1.1 Importance of adding source level features

To investigate the importance of incorporating source level features for articulatory synthesis, we trained two synthesizers with one using 6 TVs and the other using 6 TVs + source features. The two synthesizers when evaluated on the test split have MSEs of 2.0607(0.31) and 1.6493(0.23), respectively. The auditory spectrograms (a), (b) and (c) in figure 5.8 corresponds to the original speech utterance, the synthesized spectrogram by the proposed articulatory synthesizer trained with 6 TVs + source features, and the one synthesized with only 6 TVs respectively. The auditory spectrograms clearly show that the synthesizer trained with source level features generates a better harmonic structure. Hence, for subsequent experiments, source features are used as inputs to the synthesizer.

5.4.1.2 Fully trained vs lightly trained synthesizer

Since the goal of the MirrorNet is to learn articulatory representations in a completely unsupervised or semi-supervised fashion with minimal exposure to ground-truth articulatory data, it can be expected that the articulatory synthesizer may also have to be trained with a limited amount of ground-truth data based on availability. To test the impact of using a fully trained articulatory synthesizer vs a lightly trained articulatory synthesizer as the vocal tract model in the MirrorNet, two versions of articulatory synthesizers were trained, (i) fully trained (FT): train$_{full}$ split (36 speakers, 3hours), (ii) lightly trained (LT): Initialization split (4 speakers, 30min). Here the FT synthesizer is trained with the same configurations in section 5.4.1 whereas the LT synthesizer has the same architecture and configurations except it converged better with a batch size of 64. Both the synthesizers were evaluated on the same original test split after training. Auditory spectrogram (d) in figure 5.8 shows the output from the LT synthesizer whereas (b) shows the auditory spectrogram from the FT synthesizer for the same input articulatory parameters. The two synthesizers when evaluated on the test split have mean MSEs of 2.1975(0.04) and 1.6493(0.02), respectively.

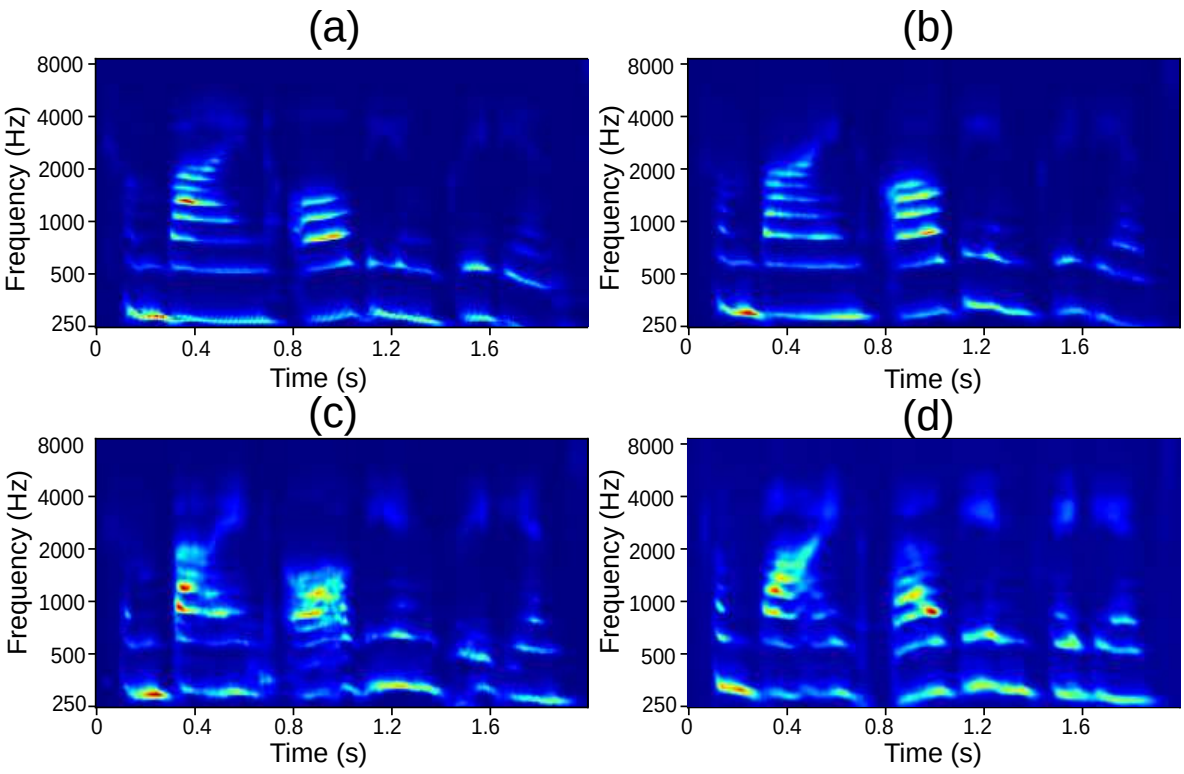

Figure 5.8: Auditory spectrogram outputs from the articulatory synthesizers. (a) Input speech utterance, (b) FT synthesizer with source features, (c) FT synthesizer 'without' source features, (d) LT synthesizer with source features

5.4.2 Learning articulatory representations with the MirrorNet

5.4.2.1 Model Architecture and Learning Phase

The MirrorNet was initially proposed as a model for learning to control the vocal tract and is based on an autoencoder architecture (Shamma et al., 2020). The goal of the model is to learn two neural projections, an inverse mapping (ϕ) from auditory representation to motor parameters (Encoder), and a forward mapping (f) from the motor parameters to the auditory representation (Decoder). The current MirrorNet implementation is an extension of the work in Siriwardena et al. (2022a) where the motor plant is replaced with a DNN based articulatory synthesizer (g). The

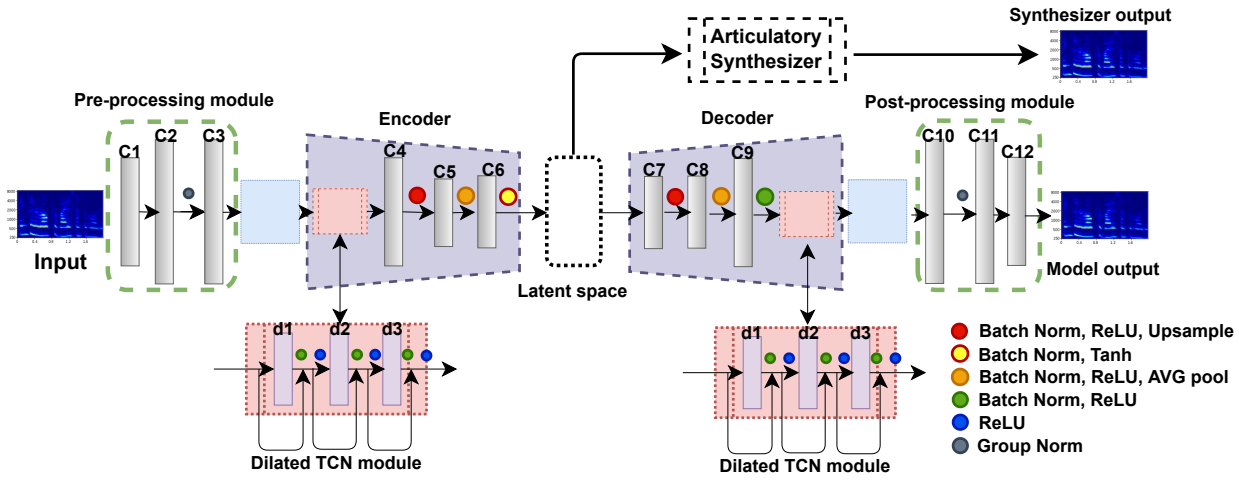

Figure 5.9: DNN architecture of the MirrorNet model

goal of the proposed network is to estimate articulatory representations as the latent space of the autoencoder. We use auditory spectrograms (Wang and Shamma, 1994) as the input and output representations of the utterances. The auditory spectrograms have a logarithmic frequency scale to provide a unified multi-resolution representation of the spectral and temporal features likely critical in the perception of sound (Wang and Shamma, 1994).

For a given input auditory spectrogram $x_i \epsilon R^{C \times L}$ where $C = 128$ and $L = 250$, the encoder generates a latent space, $\hat{l}$ such that $\hat{l} = \phi(x_i)$. The decoder then generates an auditory spectrogram x_d from the estimated $\hat{l}$ such that $x_d = f(\hat{l})$. The synthesizer which takes in the same $\hat{l}$ outputs an auditory spectrogram x_s such that $x_s = g(\hat{l})$. The MirrorNet model is optimized simultaneously with two loss functions during the 'learning phase', namely the 'encoder loss'(e_c) and the 'decoder loss'(e_d). Here $e_c = MSE(x_d, x_i)$ where $x_d, x_i \epsilon R^{C \times L}$ and $e_d = MSE(x_s, x_d)$ where $x_s, x_d \epsilon R^{C \times L}$. The encoder loss is the typical autoencoder loss whereas the 'decoder loss' constrains the latent space to converge to the expected articulatory representation while simultaneously reducing e_c.

5.4.2.2 Model Implementation and Training

The encoder and decoder of the MirrorNet is implemented in PyTorch with 1-D convolutional (CNN) layers. The complete network is inspired by the multilayered Temporal Convolution Network (TCN) (Lea et al., 2017). Figure 5.9 shows the complete DNN model architecture with its sub-modules used for pre/post processing and dilated TCN. The pre/post processing modules are symmetrically matched (C1≡C12, C2≡C11, C3≡C10) and have 128, 256 and 256 filters, respectively, with 1×1 kernels. The d1, d2 and d3 dilated CNN layers have a kernel size of 3 with 1,4 and 16 dilation rates respectively. The CNN layers in the encoder and decoder are also symmetrically matched and the C4, C5 and C6 layers have 256, 128 and 7 filters respectively, with 1×1 kernels. The dimensions of the latent space, $\hat{l}$ are chosen to match with the number of articulatory parameters to be learned and the length of the input speech utterance. For example to learn 9 articulatory parameters sampled at 100Hz for a 2 seconds long speech utterance, we use a latent space of (9×200) dimensions. Upsampling (window size 4) and average pooling (window size 5) are done after C4 and C5 layers, respectively, in the encoder, while upsampling (window size 5) and average pooling (window size 4) is done after C7 and C8 layers, respectively, in the decoder. The final model architecture has around 7.5 million trainable parameters.

Unlike a regular autoencoder, the MirrorNet is trained in two alternating stages in each iteration. The decoder is trained first (to minimize e_d) for a chosen number of epochs. Then, the encoder is trained (to minimize e_c) for a given number of epochs and this alternation is continued until both losses converge. The number of iterations of training is decided by monitoring the validation losses computed over the development split. Hyperparameter tuning was also done based on the validation losses at training. The best learning rates for the encoder and decoder, and

the batch sizes were determined with a grid search, testing all combinations from [1e-3, 1e-4, 3e-4, 1e-6] for learning rates and [16,32,64,128] for the batch sizes. The best performing models had a learning rate of 1e-6 (for both encoder and decoder) and a batch size of 16. The two objective functions were optimized using the ADAM optimizer with an 'ExponentialLR' learning rate scheduler and a decay of 0.5. All the models were trained using NVIDIA Quadro P6000 GPUs and took around 10-11 hours on average for convergence.

For further implementation information of the network, the PyTorch project is publicly available in GitHub [2]. Sample audio reconstructions can also be found in the supporting web page hosted in the GitHub repository.

5.4.2.3 MirrorNet with the Articulatory Synthesizer

We used the FT articulatory synthesizer in the MirrorNet to learn how to compute from any speech utterance the 6 TVs and source parameters needed for the synthesizer. These articulatory functions are estimated as the latent space of the MirrorNet. To evaluate how well the TVs are estimated by the MirrorNet, we calculate the Pearson Product Moment Correlation (PPMC) score between the estimated and ground truth TVs. During training of the MirrorNet, the already trained articulatory synthesizer weights are frozen and no error is back-propagated through the synthesizer.

[2] https://github.com/Yashish92/MirrorNet-for-speech

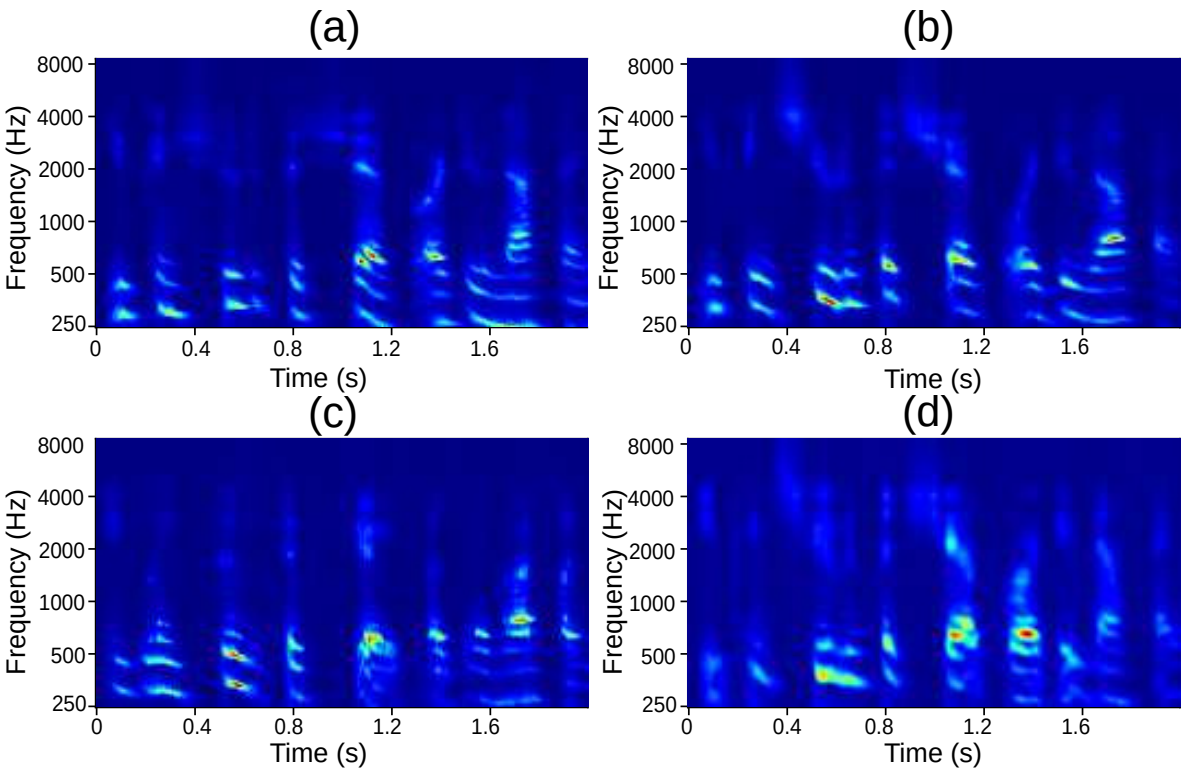

Figure 5.10: Auditory spectrogram outputs from the articulatory synthesizers (a) Input speech utterance, (b) Output of FT synthesizer from MirrorNet with init phase, (c) Output of FT synthesizer from MirrorNet without init phase, (d) Output of LT synthesizer from MirrorNet with init phase

5.4.2.4 Learning 'meaningful' articulatory representations

In the previous implementation of the MirrorNet with an audio synthesizer (Siriwardena et al., 2022a), the encoder and decoder weights are randomly initialized at the start of training. The latent space converges to a set of parameters (unsupervised) that will best synthesize the input audio (a melody of notes). The audio synthesizer used in Siriwardena et al. (2022a) is a parametric model and usually has a unique mapping from the synthesized audio to the corresponding parameters. Therefore, the latent space of the MirrorNet is more constrained and converges fairly easily to the expected parameters. However, the DNN based articulatory synthesizer learns a non-linear

mapping from articulatory variables to the auditory spectrograms and, therefore, when used in the MirrorNet could result in non-unique latent space representations. While the quality of the synthesized (final output) speech may well be excellent, it may often converge to non-physiological latent representations, and since the goal of this work is to learn interpretable and meaningful articulatory parameters, we explored here how to *initialize* the encoder and decoder learning to coax the MirrorNet learning to converge eventually to the physiological ranges expected from experimentally measured articulatory representations.

The 'initialization phase' of the MirrorNet is done by training the encoder and decoder independently with a small set of ground truth articulatory data. Here we used the initialization split (4 speakers, $\sim$30 min), a subset from the original train split (train$_{full}$) to initialize the network. The encoder loss $e_c^{init} = MSE(l, \hat{l})$, where $l, \hat{l} \epsilon R^{N \times k}$ and decoder loss $e_d^{init} = MSE(x_i, \tilde{x_d})$ where $x_i, \tilde{x_d} \epsilon R^{C \times L}$ is re-defined for the 'initialization phase'. Here e_c^{init} uses ground truth articulatory data $l \epsilon R^{9 \times 200}$ unlike in e_c in the 'learning phase' of the MirrorNet. Similarly e_d^{init} in 'initialization phase' is different from e_d in 'learning phase' in two ways, (i) e_d^{init} does not use the articulatory synthesizer generated auditory spectrograms, (ii) $\tilde{x_d} = f(l)$, and uses the ground truth articulatory representation l unlike x_d which only uses $\hat{l}$, the estimated articulatory representation by the encoder in 'learning phase'. The initialization phase also uses a larger learning rate for both the encoder and decoder (1e-3), whereas the learning phase uses a comparatively lower learning rate (1e-6). Both the learning rates were determined by a grid search.

Table 5.3: PPMC scores (mean and .std across 6 trials) for articulatory variable prediction. Here 'init' refers to 'initialization phase'. The TCN-SF-Audspec* is the state-of-the-art SI system trained in supervised fashion (Siriwardena and Espy-Wilson, 2023)

Model	**LA**	**LP**	**TBCL**	**TBCD**	**TTCL**	**TTCD**	**Ap.**	**Per.**	**Pitch**	**AVG. 6TVs**	**AVG. all**
MirrorNet(no init)	0.2033(0.02)	0.4907(0.01)	0.4967(0.03)	0.4760(0.02)	0.4735(0.03)	0.5114(0.04)	0.5354(0.02)	0.5143(0.01)	0.5930(0.03)	0.4420(0.02)	0.4771(0.04)
MirrorNet(init)	0.7701(0.11)	0.8078(0.03)	0.8132(0.05)	0.8258(0.02)	0.8696(0.06)	0.8783(0.05)	0.8970(0.01)	0.9045(0.03)	0.9125(0.02)	0.8286(0.02)	0.8540(0.03)
TCN-SF-Audspec*	0.8448	0.8640	0.8604	0.8818	0.9029	0.9005	0.9082	0.8860	0.9021	**0.8770**	**0.8834**

Table 5.3 shows the PPMC scores for the articulatory parameter estimation on the test split. The results are from the best performing MirrorNet models trained with and without the initialization phase. We also present results from a best performing SI system developed in Siriwardena and Espy-Wilson (2023) (also discussed in chapter 3), which is trained in 'completely supervised' fashion and evaluated on the same test split, as a baseline for comparison.

The Results in Table 5.3 clearly show the importance of the 'initialization phase' to estimate meaningful articulatory representations. This finding is further validated from the estimated articulatory trajectories shown in figure 5.11, where the trajectories predicted from the MirrorNet without initialization deviate significantly from the ground truth. However, these representations when fed to the DNN based articulatory synthesizer are producing auditory spectrograms (plot 5.10(c)) that closely resemble the input speech. As previously explained, this result is mainly due to the non-linear and non-unique nature of the DNN based articulatory synthesizer.

Another observation from Table 5.3 is that the MirrorNet trained with an 'initilaization phase' estimates TVs with comparatively low accuracy compared to the best performing SI system (TCN-SF-Audspec) in Siriwardena and Espy-Wilson (2023)). This is expected since the TCN-SF-Audspec model is trained in completely supervised fashion with access to all the ground-truth articulatory data, whereas the MirrorNet is mostly trained unsupervised. However, it is also important to note that except for the TCN-SF-Audspec SI system, the rest of the SI systems

discussed in chapter 3 and in Sivaraman et al. (2019) have all been outperformed by the MirrorNet in the task of speech inversion. This itself suggests the effectiveness of the MirrorNet algorithm in performing articulatory speech inversion, with minimal exposure to ground-truth articulatory data.

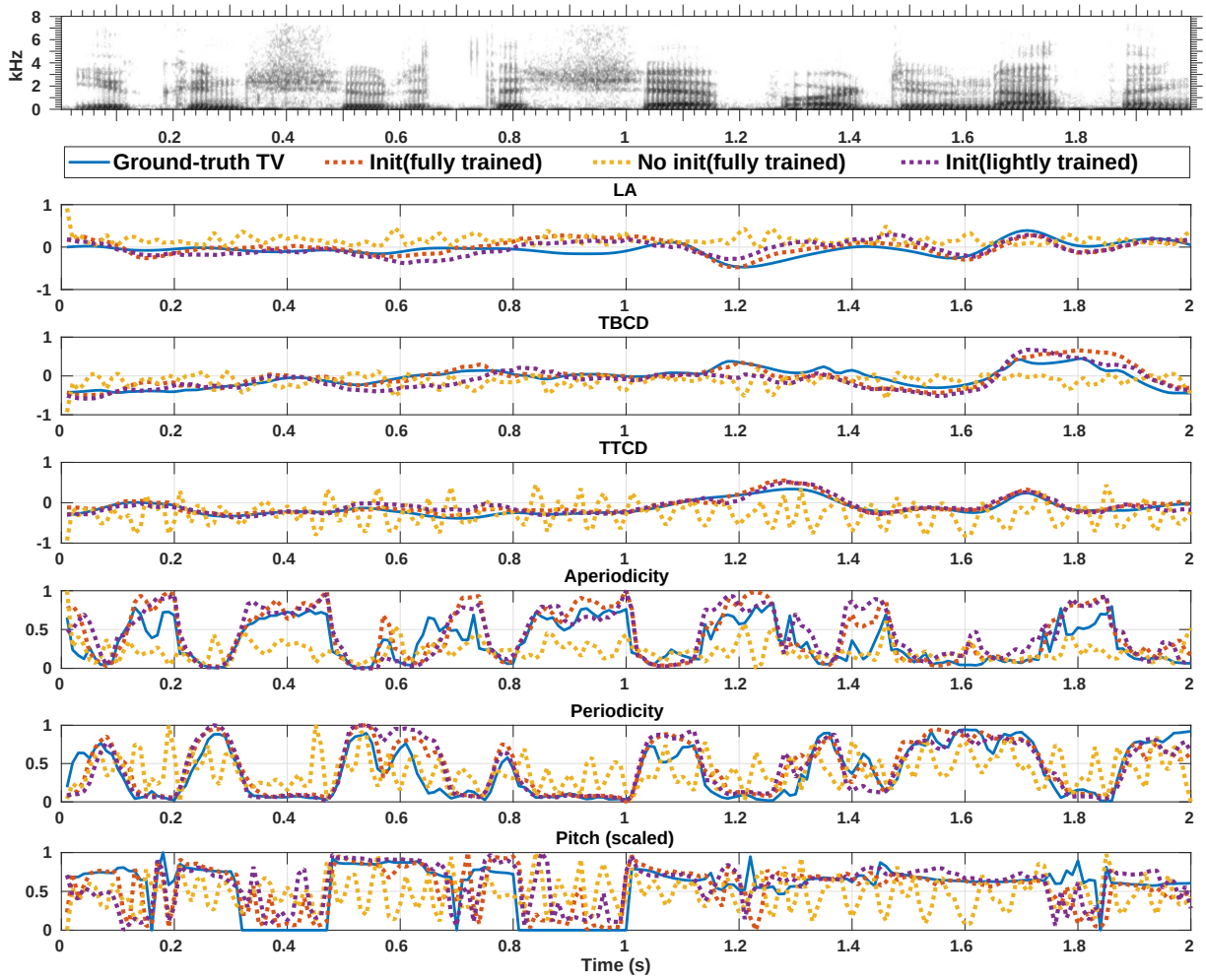

Figure 5.11: LA and constriction degree TVs + source features for the utterance 'You can shoot at the ship or do nothing' estimated by the MirrorNet. Solid blue Line - ground truth, red dotted line - estimated by the MirrorNet trained with init phase (with FT synthesizer), yellow dotted Line - estimated by the MirrorNet trained without init phase (with FT synthesizer), purple dotted line - estimated by the MirrorNet trained with init phase (with LT synthesizer)

5.4.2.5 Semi-supervised vs pseudo semi-supervised modeling

The MirrorNet model discussed so far was either trained in a completely unsupervised fashion (random initialization) or in a semi-supervised fashion (tailored initialization). To simulate an actual scenario of having limited ground truth articulatory data, the LT synthesizer in section 5.4.1.2 was used as the vocal tract model and the resulting network is trained with the 'initialization' phase. The same initialization split is used to initialize the MirrorNet, and has been used to train the LT version of the synthesizer. Table 5.4 shows the PPMC scores for the articulatory variable prediction from the MirrorNet using the FT synthesizer (pseudo semi-supervised) vs the LT synthesizer (semi-supervised). It is remarkable to see that the MirrorNet with the LT synthesizer is doing comparably well in terms of predicting the articulatory variables with respect to the model using the FT synthesizer. However, auditory spectrograms (b) and (d) in figure 5.10 shows that the FT synthesizer is producing better quality speech than the LT synthesizer.

Table 5.4: PPMC scores (mean and .std across 6 trails) for MirrorNet using the fully trained vs lightly trained synthesizers

Model	**Average (6 TVs)**	**Average (all)**
pseudo semi-supervised	0.8286(0.02)	0.8540(0.03)
semi-supervised	0.8031(0.01)	0.8219(0.02)

5.5 Summary

In this chapter, an unsupervised learning algorithm inspired by cortical sensorimotor interactions, the MirrorNet, is first applied to learning control parameters with an audio synthesizer (DIVA). The first two experiments utilized DIVA generated melodies for training, and this allowed

the MirrorNet to be evaluated with ground truth parameters, e.g., to perform preliminary tests to validate the MirrorNet predictions of the synthesizer controls across all the training and test sets, as shown in Table 5.2. The MSE values for the test set compared to the train set in Table 5.2 also give an idea on how well the model generalizes for the unseen input melodies. In the following experiments, it was demonstrated that the MirrorNet could closely approximate a set of controls for DIVA to synthesize a set of piano melodies generated by a completely different synthesizer. This idea opens up a whole new area of applications in music synthesis as it describes a tool to find parameters for an arbitrary synthesizer that maximally approximate an arbitrary sound without being necessarily capable to exactly reproduce it (reproduce a violin using a guitar for instance). It should also be noted that this work in section 5.3 only discusses results in synthesizing fixed duration melodies with a fixed number of notes, but it is a step in the right direction to synthesizing a piece of music which can have a variable number of notes in a fixed frame of audio.

Following the success made with learning audio synthesizer controls with DIVA music synthesizer, the same MirrorNet architecture was adapted to learning meaningful articulatory representations from a given speech signal. A DNN based articulatory synthesizer was custom developed and trained in a speaker-independent fashion to be used as the vocal tract model in the MirrorNet. When incorporated with this articulatory synthesizer, the MirrorNet estimates the 6TVs and source features as the latent space in a semi-supervised fashion. Including an 'initialization phase' followed by conventional 'learning' procedures resulted in the best predictions of articulatory variables. 'Initialization' presumably constrained the MirrorNet's encoder and decoder coefficients to converge to the range of values expected from experimentally measured articulatory representations. This suggests that the semi-supervised approach is sufficient to ensure that the computed latent representations are physiologically meaningful. A lightly trained version

of the synthesizer was also simulated with the MirrorNet to explore the effects of limited availability of ground-truth data for estimating the articulatory representations. Results demonstrate that the MirrorNet can estimate articulatory representations with considerably better accuracy than previous approaches. Overall, this highlights the effectiveness and power of the MirrorNet's learning algorithm in enabling to solve the conventional acoustic-to-articulatory speech inversion problem with minimal use of ground-truth articulatory data.

Chapter 6: Application of Articulatory Representations for Detecting Schizophrenia and Child Speech Sound Disorders

6.1 Overview

This chapter explores two key applications where articulatory representations have been extremely useful as an effective speech representation. The first application builds upon prior research that shed light on the changes in speech (simpler articulatory coordination) in major depressive disorder. Here, our attention shifts to comprehending the changes in articulatory coordination seen in individuals exhibiting strong positive symptoms of schizophrenia, a potential potent indicator for the development of automated screening tools. The second part of the chapter discusses how articulatory representations estimated from a speech inversion system trained on adult speech is instrumental in predicting a prevalent speech sound disorder in children. The work in this chapter has been published in (Siriwardena et al., 2021a,b) and (Benway et al., 2023c).

6.2 Articulatory Representations for Schizophrenia Detection

6.2.1 Background on schizophrenia detection

Schizophrenia is a chronic mental disorder with heterogeneous presentations that affect around 60 million (1%) of the world's adult population (Kuperberg, 2010). Symptoms of

schizophrenia are broadly categorized as either positive, which are pathological functions not present in healthy individuals (e.g., hallucinations and delusions); negative, which involve the loss of functions or abilities (e.g., apathy, lack of pleasure, blunted affect and poor thinking); or cognitive (deficits in attention, memory and executive functioning) (Andreasen and Olsen, 1982; Demily and Franck, 2008). Previous studies have shown promising results in identifying the severity of depression by using coordination features derived from the correlation structure of the movements of various articulators (Espy-Wilson et al., 2019). Based on this, a preliminary study was done by Siriwardena et al. (2021b) to understand how positive symptoms of schizophrenia affect the articulatory coordination in speech. These findings are the impetus for the current investigation where more subjects and data are used to validate the fact that neuromotor coordination is altered in schizophrenia patients who are markedly ill and exhibit strong positive symptoms.

Time-delay embedded correlation (TDEC) analysis has shown promising results in assessing neuromotor coordination in Major Depressive Disorder (MDD), and the eigenspectra derived from the correlation matrices have been used effectively for classification of MDD subjects from healthy (Seneviratne et al., 2020; Williamson et al., 2014, 2019). Recently, new multi-scale full vocal tract coordination (FVTC) features generated with a dilated CNN have shown further improvement in classification for selected datasets of MDD subjects (Huang et al., 2020). The FVTC method addresses repetitive sampling and matrix discontinuity issues of TDEC analysis by introducing a new channel-delay correlation matrix. In this work, we compare both TDEC and FVTC methods for generating input correlation matrices for training a multimodal CNN with audio and video features. We also propose a model which uses both TDEC and FVTC correlation structures to classify subjects with strong positive symptoms in schizophrenia from healthy.

6.2.2 Database and low level features

Database

A database recently collected for a collaborative observational study conducted by the University of Maryland School of Medicine and the University of Maryland College Park has been used for this study (Kelly et al., 2020). The database contains video and audio data of free response assessments administered in an interview format. Data for this study was collected from 23 schizophrenia (SZ) patients and 20 healthy controls (HC). All of the subjects with schizophrenia were clinically diagnosed. Every subject participated in four interview sessions over a period of six weeks. Each interview session is 10-45 minutes long and every subject is assessed using standard depression severity measures and global psychopathology measures by a clinician and themselves. For this study, we used the clinician assessments based on the 18-item Brief Psychiatric Rating Scale (BPRS) (Hunter and Murphy, 2011), where we selected subjects based on the total BPRS score, and the subscores for psychosis (BPRS item11, item12, item4, item15) and activation (BPRS item6, item7, item17), and the Hamilton Rating Scale for Depression (HAMD) (Gonzalez et al., 2013).

Table 6.1 presents the information on the dataset used for the study. The 7 schizophrenia subjects (4 Males, 3 Females) are selected such that they are markedly ill (BPRS total $\geq$ 45), have higher sub scores for psychosis and activation, but are not depressed or only mildly depressed (HAMD between 0 and 14). The 11 healthy controls (5 Males, 6 Females) are chosen such that they are not depressed (HAMD $<$ 7) or schizophrenic (BPRS $<$ 32).

Table 6.1: Details of the dataset used

	SZ	**HC**
Number of subjects	7	11
Number of sessions	17	34
Mean session duration	35min	18min
Number of utterances	1208	1132
Hours of speech	10.0	9.43

The audio data were first diarized using transcripts which include the speaker ID and time stamps to separate out the utterances which correspond to the subject from the interviewer. The utterances which are longer than 40 seconds were then segmented into 40 second chunks. If the remaining amount was less than 5 seconds (the minimum length accepted), then it was added back to the last segment. Thus, for all the classification experiments, we used segments with a minimum length of 5 seconds and a maximum length of 45 seconds.

Facial Action Units (FAUs)

The video-based Facial Action Units (FAUs) provide a formalized method for identifying changes in facial expressions. We used the Openface 2.0: Facial Behaviour Analysis toolkit (Baltrusaitis et al., 2018) to extract seventeen FAUs from the recorded videos of the subjects during the interviews. The FAU features were sampled at a rate of 28 frames per second. We only analyzed those portions of the video when the subject was talking. Table 6.2 shows the list of FAUs extracted from the Openface tool kit.

For parallel delay CNN model in section 6.2.3 and FVTC model in section 6.2.4, we choose only 10 FAUs (FAU 6,7,9,10,12, 14,15,17,20 and 23 from FACS coding system (Prince

et al., 2015)) which are near the mouth area that can capture coordination of lip and near lip movements during the voice activity. One other reason to choose 10 FAUs is to handle the high dimensionality of the TDEC correlations structure which poses computational limitation when training multi-modal networks in section 6.2.5.3.

Table 6.2: List of FAUs extracted from Openface tool kit

FAU No	**FAU Name**	**FAU No**	**FAU Name**
1	Inner brow raiser	14	Dimpler
2	Outer brow raiser	15	Lip corner depressor
4	Brow raiser	17	Chin raiser
5	Upper lid raiser	20	Lip stretcher
6	Cheek raiser	23	Lip tightner
7	Lid tightner	25	Lips part
9	Nose wrinkler	26	Jaw drop
10	Upper lip raiser	45	Blink
12	Lip corner puller		

Vocal Tract Variables (TVs)

We used the SI system in (Sivaraman et al., 2016; Sivaraman, 2017) that maps the acoustic signal into 6 vocal tract variables (TVs). The 6 TVs are namely Lip Aperture (LA), Lip Protrusion (LP), Tongue Body Constriction Location (TBCL), Tongue Body Constriction Degree (TBCD), Tongue Tip Constriction Location (TTCL) and, Tongue Tip Constriction Degree (TTCD). These 6TVs capture the kinematic state of each constrictor by its corresponding constriction degree and location coordinates.

Seneviratne et al. (2020) in a recent study showed that incorporating glottal TVs generated by periodicity and aperidocity measures (by Aperiodicity, Periodicity and Pitch (APP) detector (Deshmukh et al., 2005)) improved the results of depression detection. Thus, in this study we use 6 TVs generated from the SI along with the 2 glottal TVs as the key audio features for the classification models.

Mel-Frequency Cepstral Coefficients (MFCCs)

Previous studies in depression prediction using speech (Ray et al., 2019; Ringeval et al., 2019) have shown the superiority of MFCCs over other audio based features like extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (Eyben et al., 2016) and DEEP SPECTRUM features (Amiriparian et al., 2017). Huang et al. (2020) showed with their depression classification study that coordination features computed from MFCCs perform better with respect to formants and eGeMAPS features. So to compare how robust and effective the TVs are for detecting schizophrenia, we chose MFCCs as the baseline audio features for our study. We extracted 13 MFCCs from the librosa python library using an analysis window of 20 ms with a 10 ms frame shift. Only 12 MFFCs were used for analysis by discarding the 1st coefficient.

6.2.3 Time-delay embedded correlation Analysis (TDEC)

Coordination among 10 FAUs, 6 TVs and the 12 MFCCs were estimated using the correlation structure features. The features are estimated by computing a channel delay correlation matrix using time delay embedding at two delay scales (Espy-Wilson et al., 2019; Williamson et al., 2013, 2019). The computed correlation matrix can be considered as a compact representation which provides rich information on the underlying mechanisms in articulatory coordination. Each

correlation matrix R_i has a dimensionality (MN x MN) where M = 10, 8 and 12 for FAUs, TVs and MFCCs respectively. N corresponds to the number of time delays per channel which is 15 for all the considered feature types.

From the correlation matrix R_i calculated for each sample i, the eigenspectrum is computed. The eigenspectrum generated for FAUs is a 150-dimensional vector which is rank ordered (in descending order of magnitude of eigenvalues) from index j=1,..,150. The eigenspectra generated from TVs and MFCCs are 120-dimensional and 180-dimensional, respectively.

Figure 6.1 shows the averaged eigenspectra (on left) computed for FAUs, TVs and MFCCs. The difference plots in Figure 6.1 (in right) are calculated by taking the difference between averaged eigenspectra curve for speech from schizophrenia subjects with respect to that of healthy controls. These eigenspectra and difference plots help us to understand the coordination complexity of the speech and facial gestures. The magnitude of each eigenvalue is proportional to the amount of correlation in the direction of their associated eigenvectors. The difference plots show that schizophrenic speech has smaller low-rank eigenvalues relative to the healthy controls, and the trend is reversed towards the high-rank eigenvalues. Therefore, schizophrenic speech needs a larger number of independent dimensions implying a more complex articulatory coordination than speech from healthy controls (Williamson et al., 2012, 2013). The same argument is true for facial gestures.

The averaged eigenspectra and difference plots in figure 6.1 shows that the low-rank eigenvalues are smaller for schizophrenia subjects relative to the healthy controls, and this trend is reversed towards the high-rank eigenvalues. A key observation associated with depression severity (Espy-Wilson et al., 2019; Williamson et al., 2014, 2019) is that low-rank eigenvalues are larger for MDD subjects relative to healthy controls where as they are smaller for high-rank

eigenvalues. The magnitude of high-rank eigenvalues indicates the dimensionality of the time-delay embedded feature space. Thus, larger values in the high-rank eigenvalues can be associated with greater complexity of articulatory coordination (Espy-Wilson et al., 2019; Williamson et al., 2013). Therefore we can infer that the schizophrenia subjects with strong positive symptoms have a higher complexity than the healthy controls and the MDD patients. These results are likely due to the negative symptoms of depression which results in psychomotor slowing (i.e., simpler coordination) and the strong positive symptoms of the schizophrenia patients such as activation that results in motor hyperactivity (i.e., complex coordination). Supporting our hypothesis, we see this effect from eigenvalues computed from the FAUs, TVs and MFCCs.

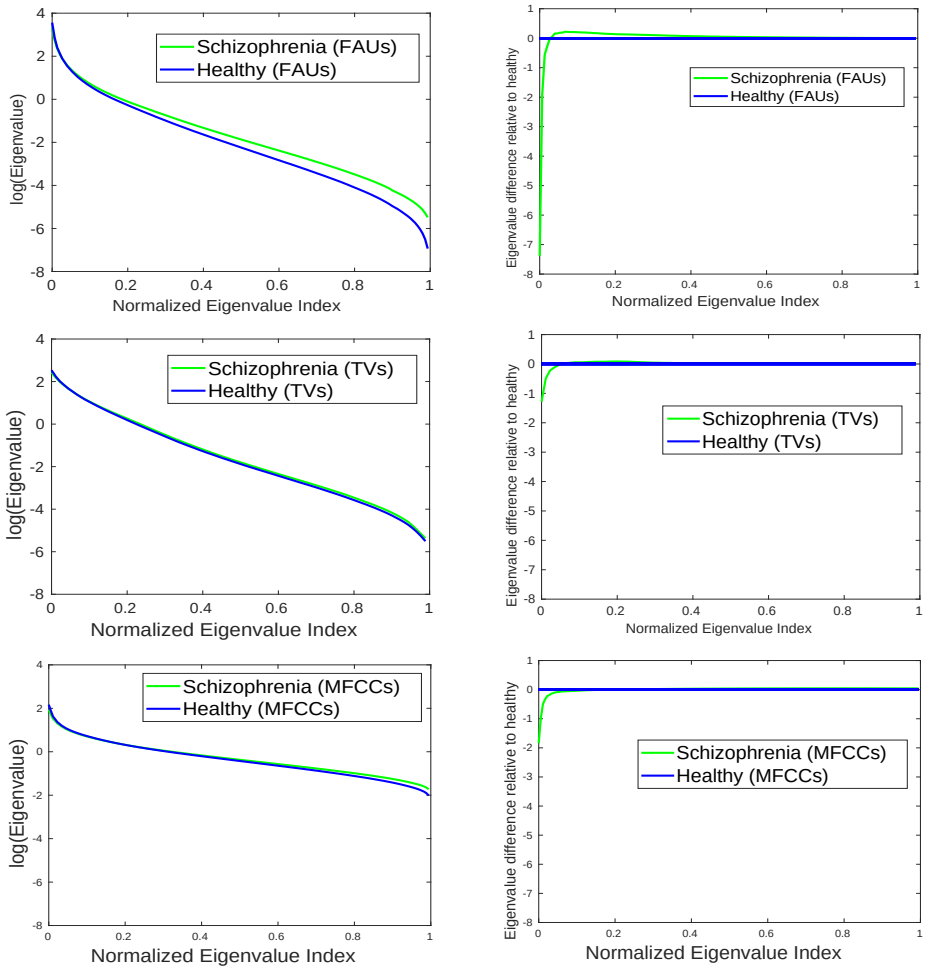

Figure 6.1: Average eigenspectra plots (left) and corresponding difference plots (right) for FAUs, TVs and MFCCs

6.2.4 Full vocal tract coordination (FVTC)

Huang et al. (2020) in a recent study with MDD introduces a new channel delay correlation method inspired by TDEC, which uses a different correlation structure with correlations starting from 0 to a delay of 'D' frames (a design choice). The delayed autocorrelations and cross-correlations across channels are stacked to form the FVTC correlation structure. FVTC includes every correlation within the considered D frames and also avoids the repetitive use of same correlations as in the TDEC correlation matrix. By learning each stream of correlations independently,

it also avoids the possible discontinuity issues in the edges of the TDEC correlation matrices. Dilated convolutions with rates of 1,3,7 and 15 are used to match multiple time scales used in the TDEC method.

6.2.5 Multi-modal systems

6.2.5.1 Parallel delay scale TDEC-CNN model (TDEC-CNN) : Model 1

We developed a CNN architecture which takes in multiple time-delay embedded correlation matrices with 2 delay scales in parallel as inputs for two 2D-CNN layers. The output from the 2 CNN layers are then concatenated and passed through another 2D-CNN layer. Batch normalization, max-pooling and dropout were applied afterwards. The flattened output is then fed to a fully connected layer with 64 neurons. 16 filters with kernel size (3,3) was used for every 2D CNN layer and every CNN layer has ReLU activation. We trained individual models for FAUs, TVs and MFCCs where 3 and 7 sample delay scales were used for FAUs and 7 and 15 sample delay scales were used for TVs and MFCCs.

6.2.5.2 FVTC CNN model (FVTC-CNN) : Model 2

We designed a CNN model inspired by the one in Huang et al. (2020) which takes the FVTC correlation matrix computed in section 6.2.4 as the input. To reduce the number of trainable parameters in the original model (Huang et al., 2020), we reduced the size of the two fully connected layers to 64 and 8. We used the same dilation rates 1,3,7,15 as in the original model. We chose 45 as the 'D' parameter for FAUs and 50 for TVs and MFCCs. The 'D' values were chosen from the set of (45,50,55) after doing a grid search on individual feature based systems.

6.2.5.3 Multimodal CNNs

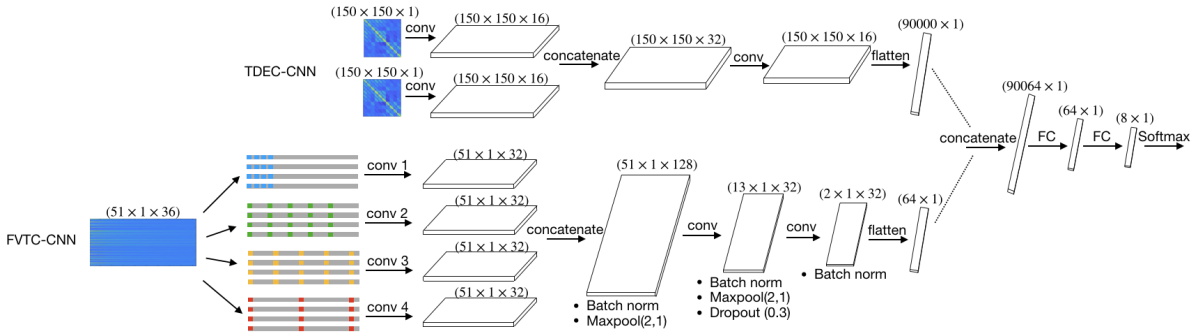

Figure 6.2: TDEC and FVTC combined multimodal architecture for best performing model in Table 6.4

One of the key contributions of this work is the development of mutimodal networks to fuse the audio and video based coordination features calculated from the TDEC and FVTC methods. 3 types of fusion models were designed namely the (Parallel delay TDEC-FVTC CNN), (FVTC-FVTC CNN) and (Parallel delay TDEC- Parallel delay TDEC CNN). The same model architectures developed in section 6.2.5.1 and section 6.2.5.2 were used and they are fused after passing through all the 2D-CNN layers by concatenating the flattened outputs from each model. The concatenated output then passes through 2 fully connected layers which have 64 and 8 neurons, respectively. That output is then fed to a softmax output layer to generate the final class probabilities (for schizophrenic and healthy classes).

To choose the learning rate and batch size for the individual and multimodal systems, we did a grid search from sets of (1e-4, 3e-4, 1e-5, 3e-5) and (32,64,128) for learning rates and batch sizes, respectively. 1e-4 for learning rate and 64 for batch size gave the best metrics in classification for the best performing model in Table 6.4. Every model was trained in leave one subject out cross validation fashion where both accuracy and F1 scores are calculated across 18 folds. This

ensured that every model is always trained with around 2000 sample points (17 subject's data) and then tested on the excluded subject (test fold). To come up with the subject level prediction label from the multiple segment level predictions, we use the best 25% of the total segments in the test fold which are predicted with the highest class probabilities (even if the model predicts the wrong class). Every model is trained for 200 epochs with early stopping based on validation loss with patience of 15 epochs. Figure 6.2 shows the architecture for the best performing multimodal system from Table 6.4.

Table 6.3: Unimodal results for FAUs, TVs and MFCCs. Best Model for each feature type is highlighted in bold

Features	**TDEC-CNN (Model1)**		**FVTC-CNN (Model2)**	
	Accuracy	F1(S)/F1(H)	Accuracy	F1(S)/F1(H)
FAU	**83.33%**	**0.80/0.86**	83.33%	0.77/0.87
TV	66.67%	0.57/0.73	**72.22%**	**0.62/0.78**
MFCC	61.11%	0.46/0.70	**72.22%**	**0.55/0.80**
MFCC+Glottal TVs	60.05 %	0.45/0.69	**72.22%**	**0.55/0.80**

Table 6.4: Multimodal classification results

Models	**Accuracy**	**F1(SZ)/F1(HC)**
FAU (Model2)+TV(Model2)	66.67%	0.67/0.67
FAU (Model1)+TV(Model1)	72.22%	0.67/0.76
FAU (Model2)+MFCC(Model2)	72.22%	0.62/0.78
FAU (Model1)+MFCC(Model2)	77.78%	0.67/0.83
FAU (Model1)+(MFCC+Glottal TVs)(Model2)	83.33%	0.73/0.88
FAU (Model1)+TV(Model2)	**88.89%**	**0.86/0.91**

Table 6.3 shows the average accuracy across the 18 folds and the F1 scores for classifying schizophrenic and healthy subjects by training individual models for FAUs, TVs and MFCCs based on TDEC-CNN and FVTC-CNN models. Results suggest that FAUs perform the best when compared to TVs and MFCCs in classification metrics. This could be due to the inclusion of a wider range of facial muscle movements which are not limited to only those around the speech articulators. Moreover, FVTC-CNN models trained with TVs and MFCCs perform the best when compared to TDEC-CNN models trained with the same features. With respect to F1 scores, TVs perform better than the MFCCs in both FVTC-CNN and TDEC-CNN models showing the robustness of TVs in capturing the articulatory changes in speech.

Table 6.4 shows results for the 6 multimodal systems trained by fusing video and audio features from TDEC and FVTC methods. It is interesting to note that the models with heterogeneous architectures perform the best when compared to models which use the same correlation structure for both audio and video features. To do a fair comparison between TVs and MFCCs, we also trained individual and multimodal networks by incorporating the 2 glottal TVs along with the 12

MFCCs, so that glottal source level information is also accounted. Even then, Table 6.4 shows that the TV based best performing model out performs the MFCC based multimodal systems. Furthermore, the multimodal system which uses TDEC and FVTC correlation structures for FAUs and TVs, respectively, outperforms the baseline model trained on FVTC for both FAUs and MFCCs by around 18% in terms of the mean F1 score. This result is interesting in the sense that the video features complement well with TVs over MFCCs in the proposed multimodal setting.

6.3 Articulatory Representations for Mispronunciation Detection of /ɹ/in Child Speech Sound Disorders

The experiments in this subsection are motivated by the recent work in Benway and Preston (2023) which effectively automated evidence-based treatment for speech sound disorders. While best-practice clinician-led motor-based interventions include both summary perceptual judgment feedback (i.e., knowledge of results; KR) and detailed auditory/somatosensory corrective feedback (i.e., knowledge of performance; KP) (Maas et al., 2008), no available automated treatment delivers clinical-grade KP. To date, automated systems deliver KR with clinician-delivered KP (McKechnie et al., 2020) or approximated KP using random selection of plausible clinical cues (often mimicking clinician-delivered KP due to obscured visualization of the tongue in the oral cavity, (Preston et al., 2020)). However, a mispronunciation detection system that leverages acoustic-toarticulatory speech inversion (SI) may someday automate true corrective KP that is not attainable with state-of-the-art formant-based systems, while also circumventing known issues in (child) formant estimation (Shadle et al., 2016).

6.3.1 Acoustics and articulation of American English /ɹ/

Articulatory configurations for /ɹ/

We focus on rhotic /ɹ/, the most commonly impacted sound in American English speech sound disorders persisting past age 8 (Ruscello, 1995). For fully rhotic ɹ, there is evidence for up to five quasidependent articulatory actions in the oral and pharyngeal cavities: (1) elevation of the tongue tip/blade, (2) lateral bracing of the tongue against the posterior molar teeth, (3) a low posterior tongue dorsum, (4) retraction of the tongue root into the pharynx, and (5) slight rounding of the lips (Preston et al., 2020). Insufficient vocal tract configurations will likely generate a more neutral spectral envelope that may be perceived as a "derhotic" /ɹ/. Derhotic vocal tract configurations commonly include a lower tongue tip, insufficient tongue root retraction, and, notably, a higher tongue dorsum (Boyce, 2015; Preston et al., 2020). Indeed, blade/dorsum relative displacement has been found to be salient for predicting perception of the syllable /aɹ/in speech sound disorders (Li et al., 2023)

Acoustic configuration of /ɹ/

Although there is variation in the vocal tract configurations that generate a perceptually-correct American English /ɹ/, these configurations generate a similar spectral envelope in a linear predictive coding (LPC) formant feature space. Formant features serve as the baseline condition for the present work because they are well-motivated by decades of acoustic phonetics investigation for /ɹ/and also meet reproducibility guidelines advocating low-dimension, validated feature sets in clinical speech technology systems (Berisha et al., 2022). Furthermore, recent work has demonstrated that age-and-sex normalized formants outperform cepstral representations in the binary classification of /ɹ/rhoticity in the context of speech sound disorders (Benway et al., 2023b).

Prior acoustic phonetics investigations show that vocal tract configuration for a correct, fully rhotic /ɹ/is marked by a relatively high second formant (F2) (DELATTRE and FREEMAN, 1968) and a relatively low third formant (F3) (Espy-Wilson et al., 2000). This results in the average rhotic F3-F2 distance being much narrower than that of a neutral vocal tract. LPC estimation, however, requires speaker customization and can be error-prone (Burris et al., 2014; Kent and Vorperian, 2018), particularly for samples taken from (child) speakers with high fundamental frequencies (Shadle et al., 2016). Errors in estimation may be speaker-specific (Derdemezis et al., 2016) or due to population-general traits such as wider bandwidths in children (Kent and Vorperian, 2018). It is also believed that manually correcting formant estimates in the context of /ɹ/also indicates that near merging of F3 and F2 in hyper-rhotic tokens is an additional source of estimation error.

Motivation and contributions

Speech analysis systems that estimate the learner's articulatory trajectory have the potential to improve performance of stateof-the-art formant-based rhoticity classifiers. Such systems could eventually automate the delivery of KP feedback at or above the level possible by human clinicians, due to the low visual salience of the vocal tract during for /ɹ/. Because a ground truth articulatory dataset is not available for children with rhotic speech sound disorders, de novo training of SI for this use case is not possible at this time. Therefore, in this study we test the performance of an adult-trained SI system on utterance-normalized child speech data to improve over a formant baseline for the binary classification of /ɹ/rhoticity in children with speech sound disorders.

We offer two contributions. Our first research question shows that tract variable estimates generated from adult-trained models are able to index clinician perceptual judgment of rhotic errors, particularly for tongue body constriction location (d =.39). Our second research question shows that a bidirectional LSTM trained on tract variable output from SI meets or exceeds

the performance of a comparable system trained on age-and gender normalized formants when predicting clinician judgment of rhoticity ($\bar{x}_{F1-score} = .90$ σ_x= .05; med = .92, n = 6).

6.3.2 Articulatory representations with SI systems

Acoustic-to-articulatory SI is tasked with retrieving articulatory dynamics from a speech signal (Sivaraman et al., 2019). Attempts at recovering articulatory movements from the continuous speech signal have a long history (Papcun et al., 1992a), but have generally limited to tracking a specific set of articulators like upper and lower lip, tongue tip, velum closure, etc. However, it is important to derive not just the main effect of individual articulator movements, but the interactions among articulators; for example, lips and jaw as individual articulators work together to achieve a desired vocal tract shape (Saltzman and Munhall, 1989). Hence, general SI systems focus on recovering the vocal tract constriction, estimating the constriction degree and location of functional tract variables (TVs), rather than the movement of individual articulators. During SI, acoustic features extracted from the speech signal are used to predict the TVs. The inverse mapping is learned by associating these features through training on a corpus consisting of matched acoustic and directly observed articulatory data.

Here, we use the SI system developed in Siriwardena and Espy-Wilson (2023), which is a Temporal-Convolution Network (TCN) trained on 36 (adult) speakers from the U. Wisconsin X-Ray microbeam (XRMB) corpus (Westbury, 1994b), and evaluated with no speaker overlap in training, development, and test splits. Apart from the 6 TVs (LA: Lip Aperture, LP: Lip Protrusion, TTCL: Tongue Tip Constriction Location, TTCD: Tongue Tip Constriction Degree, TBCL: Tongue Body Constriction Location, TBCD: Tongue Body Constriction Degree), the system predicts 3

source features: aperiodicity, periodicity, and pitch.

6.3.3 Dataset and feature extraction

This study is a binary classification experiment seeking to predict listener judgment of rhoticity (i.e., fully rhotic/ "correct" vs derhotic/ "incorrect") in a subset of the open access PERCEPT-R audio Corpus (Benway et al., 2022) collected during a clinical trial of biofeedback interventions (Benway et al., 2021). The PERCEPT subset selected for reanalysis are 3,210 word-level utterances from the 6 speakers providing consent/assent for future use of study audio (publicly available at (Benway et al., 2022); see (Benway et al., 2023a) for additional corpus audio and ground-truth label details which include multilistener average perceptual ratings of rhoticity). Original data collection received ethics approval from the Biomedical Research Alliance of New York. Speech data were lab collected by research clinicians during word-reading probes using head-mounted microphones. Although a small sample, these data were chosen for this exploration for several reasons. Firstly, they are clinically-valid group of children with speech sound disorder. Secondly, the range of speaker ages in this dataset (Table 6.5) allows for preliminary exploration of the impact of child age on model performance. Thirdly, because these data come from a treatment study in which some participants experienced statistically and clinically significant gains in /ɹ/production, these data contain tongue shape variation between pre-treatment to post-treatment time points within the same speaker. Lastly, a subset of the audio has hand placed, within word segmentation boundaries for the rhotic target.

Table 6.5: Participants in the current investigation

PERCEPT-R ID	Study ID	Age	Formant Ceiling	Number of Utterances
33	6102	15.7	6000Hz	543
34	6103	14.9	4500Hz	638
35	6104	9.3	6000Hz	560
36	6108	14.5	5000Hz	692
37	3101	9.8	5500Hz	337
38	3102	11.8	500Hz	440

6.3.3.1 Formant extraction

Time series estimates of F1, F2, and F3 were extracted from the utterance using custom Python scripts and the Praat "To Formant (Robust)" command with default settings except for participant-specific Formant Ceiling settings. Formant ceilings were customized using the procedure in Benway et al. (2021). Formant transform time series (F3-F2 distance and F3-F2 deltas were also included in the feature set. Age-and-sex norming was completed as in Benway et al. (2023b) using a third-party dataset (Lee et al., 1999).

6.3.3.2 Tract variable extraction

We extract 6 TVs from the SI system to capture the degree and location of the lip, tongue body and tongue tip constrictions. The extracted TVs (in the range of [-1,1]) are z-normalized, utterance-wise, to generate the final 6 TV feature set. We additionally estimate glottal activity by extracting three source level features (aperiodicity, periodicity and pitch) (Deshmukh et al., 2005).

Similar to the 6 TVs, the 3 source features are also z-normalized, utterance-wise. The 6 TVs and 3 source features together comprise the 9 TV feature set.

Rhotic Segment Boundary Estimation

Research assistants who were trained in speech signal segmentation manually reviewed each pre- and post-treatment utterance to mark the onset and offset of the rhotic phoneme using Praat TextGrids. Only pretreatment and posttreatment files were selected for this analysis because data collection at these timepoints elicited citation speech, rather than intreatment speech that may be marked by unnaturally long or unstable articulations. Within each utterance, research assistants selected the segment corresponding to the perception of the (target) rhotic phoneme. Coarticulatory transitions between target rhotics and neighboring segments were wholly assigned following the sonority hierarchy, so transitions were included with the more sonorous segment. In other words, rhotic-vowel transitions were included with the vowel, while rhotic-consonant transitions were included with the rhotic. Boundary decisions were made based on visual breakpoints in F2 slope and confirmed perceptually during file playback. If F2 breakpoints were not discernable, F3 was used, and then F1. We used these rhotic boundary timestamps to extract all rhotic associated TVs, which we then averaged into 10 bins to facilitate visual inspection for our first research question (only).

6.3.4 Prediction of Clinician Perceptual Judgment using Leave One Participant Out Cross Validation

The differentiation of derhotic and rhotic segments by formants and tract variables was assessed with Deep Neural Network (DNN) based models, which can be extremely effective in

binary classification tasks. We specifically experimented with Recurrent Neural Networks (RNN) due to the timeseries nature of the data.

Ground truth labels

The binary class label for the audio files present investigation was derived from the listener-average binary perceptual rating in the PERCEPT-R Corpus (i.e., 1 = fully rhotic/"correct" vs 0 = derhotic/"incorrect"). For files with non-unanimous listener ratings, .66 served as the floor for class 1 (the fully rhotic class) to reflect the lack of full agreement between expert raters in the context of Residual Speech Sound Disorders (RSSD) (Klein et al., 2013). All utterances with a listener-average rating $<$ .66 were assigned to class 0 (the derhotic class).

Data preprocessing

The five age-and-sex normalized formants and transforms described in 6.3.3.1 and the TVs described in 6.3.3.2 were used independently as input features for model training. The features were segmented or padded to generate 2-second-long input embeddings. Each input embedding was matched to a groundtruth label for rhoticity from the PERCEPT-R Corpus, representing the average binary listener rating (0 = derhotic, 1 = fully rhotic). The heuristic for discretizing the average rating to binary ground-truth in the present investigation was .66.

Model architecture and training

We experimented with Bidirectional gated RNN (BiGRNN) and Bidirectional LSTM (BiLSTM) models with different architectures. Figure 6.3 shows the best performing BiLSTM model architecture for the classifier.

All model parameters were randomly initialized with a seed (=7) for reproducibility of the results. All the models were trained using leave-one-participant-out cross validation. Models in each training split were trained with features from $\sim$2300 utterances ($\sim$85% of non-test data). The

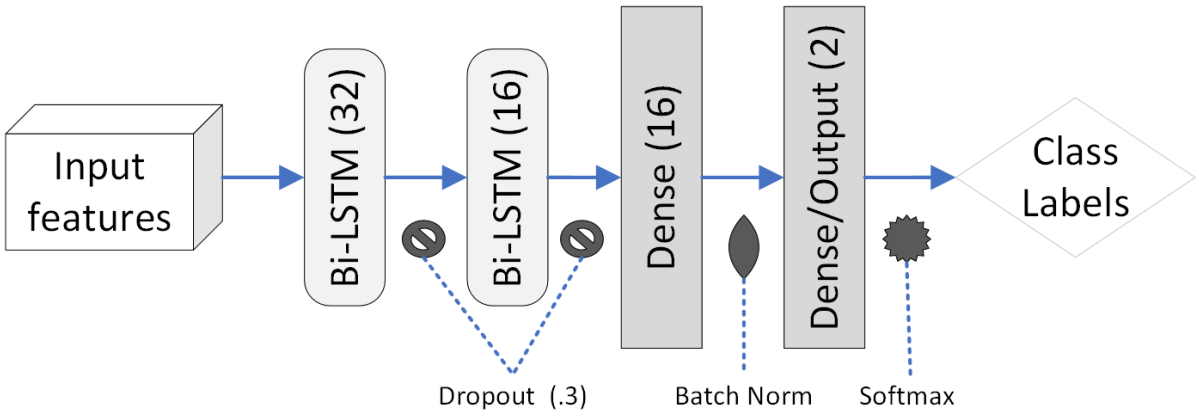

Figure 6.3: Best-performing BiLSTM model architecture

validation set was chosen at random and included ∼15% of non-test data. Models were trained for 50 epochs with an early stopping criterion and a patience of 5 epochs. Since the dataset is imbalanced (favoring derhotic samples, as would be expected in speech sound disorder clinical trial data), a weighted cross entropy loss function was used to optimize the models.

Hyperparameter fine-tuning

Table 6.6 lists candidate hyperparameters and corresponding values for the best performing BiLSTM model. Hyperparameters were fine-tuned with grid search optimizing validation loss. All models were implemented with the keras/TensorFlow framework and trained with NVIDIA TITAN X GPUs. The best performing model included ∼123k trainable parameters, and converges in ∼7 min. for each validation split.

Table 6.6: Tuned hyperparameters (bold) with candidate values

Parameter	**Candidate Values**
Learning Rate	[**1e-4**, 3e-4, 1e-3, 1e-2]
batch Size	[16, 32, **64**, 128]
Optimizer	**ADAM**, RMSProp, SGD
BiGRNN/BiLSTM Layers	[1,**2**,3 4]
Dense Layers	[1, **2**, 3]
Dropout	[**.3**, .5, .7]

Effect Size Calculation for Tract Variables

Our first research question analyzed the univariate ability of (time-binned) tract variables to distinguish between child speech productions with a ground truth label corresponding to "derhotic" and "fully rhotic". We quantified the separation of the derhotic and fully rhotic tract variable means using Cohen's d. Mean separation was "negligible" for lip aperture (d = -.06), lip protrusion (d = .11), tongue tip constriction location (d = .13), and tongue tip constriction degree (d = .08). Mean separation was "small" for tongue body constriction location (d = .39; 95%CI [.33, .44];) and for tongue body constriction degree (d = -.22; 95%CI [-.27, -.16]), as seen in Figure 6.4. The lower values for tongue body constriction location imply a more anterior constriction in fully rhotic tokens than derhotic tokens

Predicting clinician judgment of rhoticity in utterances from children with RSSD

Our second research question analyzed the performance of age and-sex normalized formant and tract variable feature sets. Results from BiGRNN and BiLSTM models are shown in Table 6.7. BiLSTM improved participant-specific F1-score (weighted) over BiGRNN. BiLSTM performance

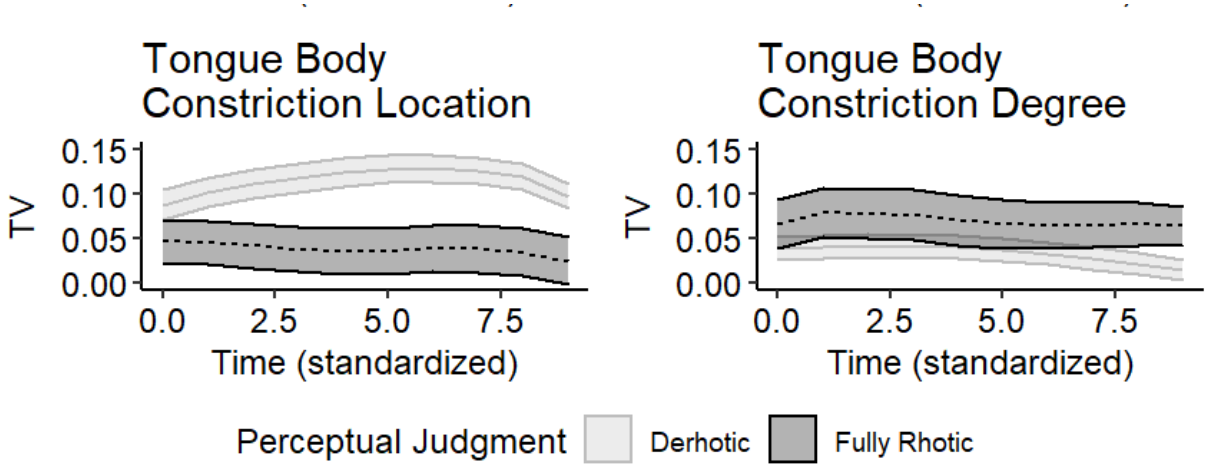

Figure 6.4: Univariate differentiation of perceptual judgment for (binned) TVs. Ribbons: 95% confidence intervals of the mean.

was comparable for formant and 9 TV feature sets, except for the notable increase in AUROC in the context of 9 TV features (Figure 6.5)

Table 6.7: Mean (standard deviation) of participant specific performance. 9 TVs include 3 source features

Model	**Feature Set**	**F1-Score**	**Precision**	**Recall**	**AUROC**
Bi-GRNN	Formants	.83(.05)	.90(.05)	.79(.06)	.82(.05)
Bi-GRNN	6 TVs	.72(.10)	.89(.05)	.67(.10)	.76(.06)
Bi-GRNN	9 TVs	.81(.06)	.91(.04)	.77(.07)	.80(.04)
Bi-LSTM	Formants	.89(.03)	.92(.04)	**.88(.03)**	.79(.17)
Bi-LSTM	6 TVs	.82(.08)	.91(.04)	.77(.11)	.81(.13)
Bi-LSTM	**9 TVs**	**.90(.05)**	**.94(.04)**	.87(.08)	**.87(.07)**

The interactions between participants and feature sets (Figure 6.5) naturally raised the question of combined performance, so we trained one Bi-LSTM with a combined feature set of formants and 9 TVs. Performance ($\bar{x}_{F1-score} = .82$ σ_x= .05) was lower than in models trained on

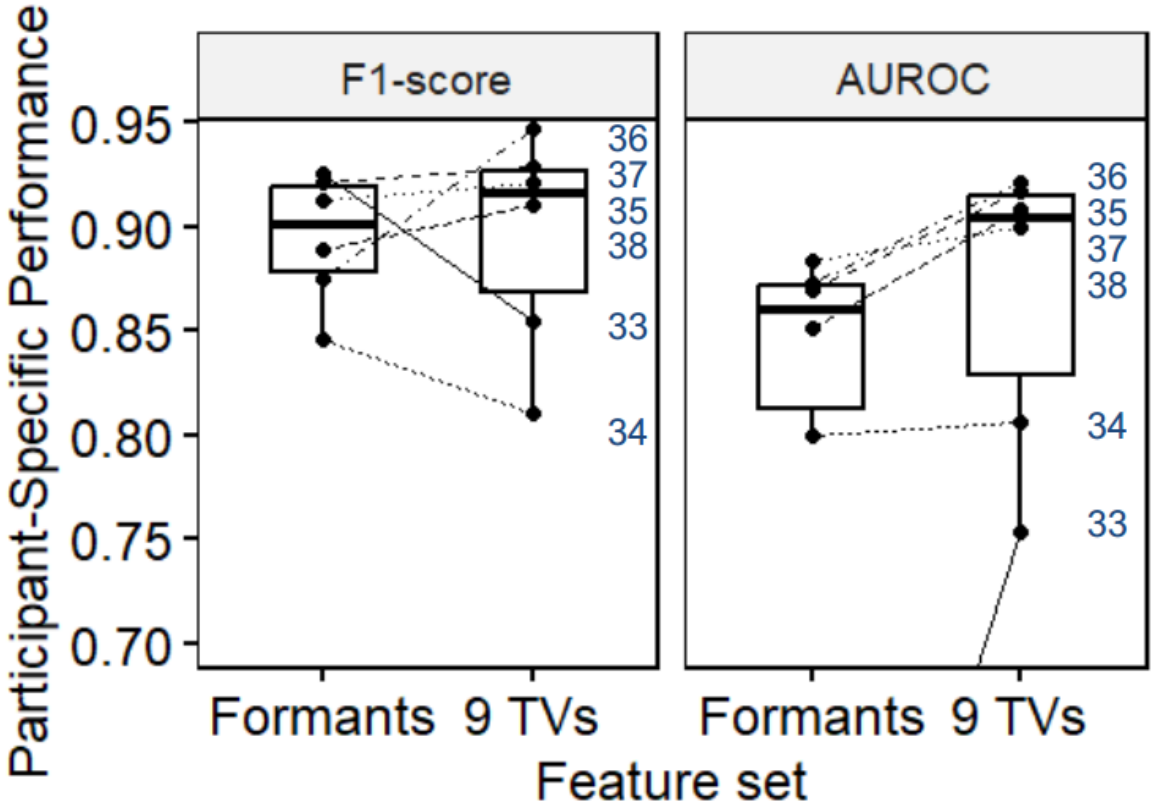

Figure 6.5: Bi-LSTM performance for individual participants (labels on the right). Not shown: formant AUROC for 33 (.44).

these features individually.

Although the number of participants in this analysis was limited, we explored the correlation of ranks to see if participant age was associated with AUROC. We selected AUROC for exploration as its participant-specific scores showed greater variance than F1-score in our results. Spearman's ρ was not significant in this sample (ρ = -.54; p = .30, n = 6); this is supported by visual inspection of Figure 6.6 which illustrates that children aged 9-14 achieved AUROC $>$ .9. Notably, the lowest performing participants in this sample had vocal tracts that, presumably, shared the most age-based similarities with the adult participants that the SI system was originally trained on.

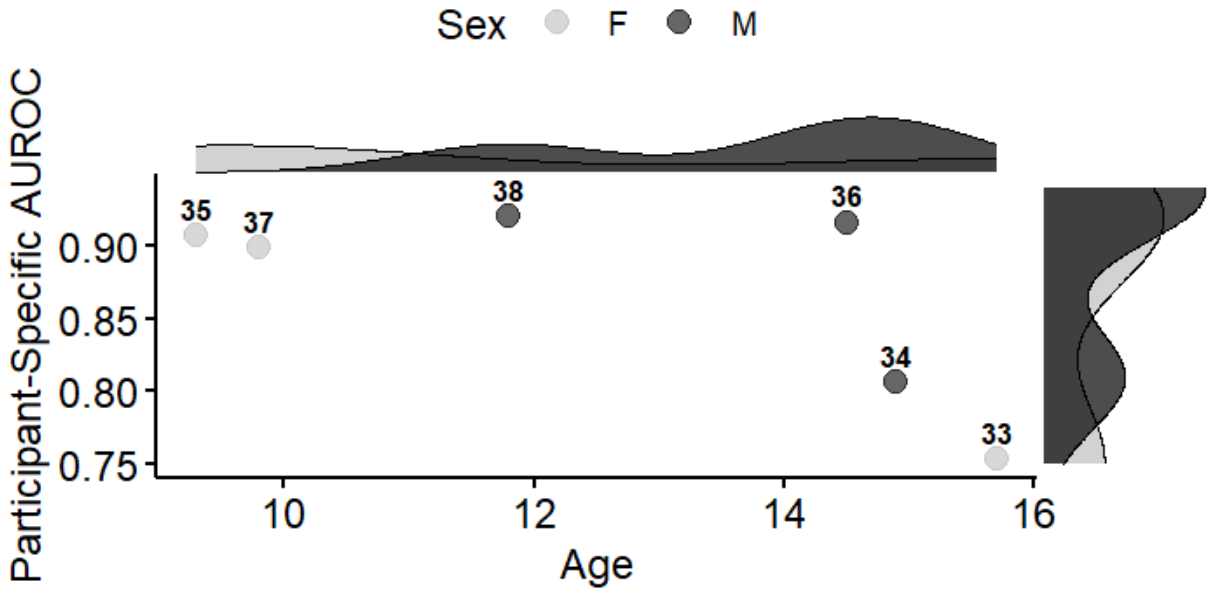

Figure 6.6: Participant-specific Bi-LSTM performance (AUROC) by participant age and sex. Labels = PERCEPT IDs

6.3.5 Discussion on Results with Mispronunciation Detection of /ɹ/in Child Speech Sound Disorders

The present classification study predicts binary listener perceptual judgment of rhoticity in child speakers with speech sound disorder impacting /ɹ/. We found that BiLSTM performance was largely comparable when trained on (1) handcrafted age-and-sex normalized formant features and (2) tract variable + source feature (9 TV) outputs generated from an adult-trained acoustic-to-articulatory SI system. In this sample, however, AUROC was notably higher for 9 TVs than formant features. The comparison of classifier performance in 6 TV and 9 TV feature sets evidences the importance of modeling source features as additional input, consistent with observations of TV performance in depression (Seneviratne et al., 2020). Notably, univariate separation between classes for individual TVs indicates that tongue dorsum dynamics may be an

important signal for clinical /ɹ/, which is corroborated by magnetic resonance/ultrasound imaging of American English /ɹ/perceived as clinically incorrect (Boyce, 2015; Preston et al., 2020) and supportive of recent modeling work elsewhere (Li et al., 2023). Future reanalysis of ultrasound data in Benway et al. (2021) could investigate if this finding might associate with direct instruction of a retroflexed (versus bunched) configuration.

Even though this TV estimator was trained on adult speech, there was no evidence in the present, small, sample that TV performance was systematically related to age. However, it was surprising that the oldest participants in this sample had the (relatively) poorest performance. Informal observations during (Benway et al., 2021) indicate that participant 34's improvements in /ɹ/resulted in a consistently hyper-rhotic articulatory pattern with a merged F3-F2 by treatment session 3, and participant 33 had a large F3-F2 distance not due to a high F3, but because of a low F2. Future investigation can quantify the extent that poorer performance may be due to speaker-specific vocal tract dynamics.

Although this investigation is limited by a small sample size, it motivates larger investigation of how adult-trained TV estimates perform for child speech. This motivation arises from the 1) lack of obvious age effects for younger speakers using 9 TVs, 2) advantage for 9 TVs in AUROC, and 3) lower amount of participant-specific customization required by TVs in feature preprocessing. Larger samples can also investigate model explainability, as well as evaluate if lower performance herein on the fused formant + 9TV feature set is evidence of true performance or perhaps due to the high feature dimensionality relative to the amount of training data for the 6 participants. Lastly, the salience of TVs in this investigation lay the foundation for future research modeling interpretable KP feedback for speech sound learners from the 9 TVs. In addition to predicting perceptual judgment of /ɹ/, visualizations of tongue shape interpolated from TVs have

the potential to provide similar benefits to ultrasound biofeedback while circumventing some of the barriers associated with widespread ultrasound use (i.e., system cost and training).

6.4 Summary

This chapter discusses how articulatory representations extracted from an acoustic-to-articulatory speech inversion system can be effectively used for applications in detecting mental health disorders like schizophrenia and detecting a common child speech sound disorder.

The section on schizophrenia detection highlights a multimodal approach to classify subjects with strong positive symptoms in schizophrenia from healthy. Results show that the video based features are more effective in identifying articulatory coordination changes while also asserting the fact that fusing with audio based TVs significantly boost the performance in detection of positive symptoms.

The section on mispronunciation detection of /ɹ/in children's speech shows that BiLSTM models trained individually on hand-crafted age-and-sex normalized formants and 9 tract variables (predicted from an adult-trained acoustic-to-articulatory speech inversion system) have comparable performance when predicting clinical perceptual judgment of rhoticity in child speech sound disorders. However, improvements in AUROC and lower customization needs for tract variable generation, as well as the potential for clinically interpretable KP feedback, motivate future development of tract variable-based mispronunciation detection for /ɹ/. This work further motivates the collection of ground-truth articulatory data from children to validate tract variables for child clinical speech sound technologies.

Chapter 7: Conclusions and Future Directions

7.1 Conclusions

The contributions of this dissertation can be discussed under three main aspects. The first one is towards improving the generalizability and robustness of the acoustic-to-articulatory speech inversion systems to estimate lip and tongue related articulatory representations. Supplementary to the same work, an unsupervised learning algorithm based on learning and predictive processing in the human brain is applied to learn articulatory representations, with minimal exposure to ground-truth articulatory data. The second contribution comes with the development of a dedicated speech inversion system to estimate the velar and glottal activity of speech in a speaker-independent fashion. Through both the first and second contributions, this dissertation tries to achieve a rare feat of building a framework that could estimate an almost complete articulatory representation of speech (lips, tongue, velum and glottis articulators) which could be a game changer in the field of speech analysis and understanding. The final and the third contribution of the dissertation highlights two distinct applications where the articulatory representations have shown to outperform the conventional speech representations in their respective tasks. This also asserts the quality of the estimated articulatory representations and thereby validates the accuracy of the developed SI systems.

As discussed above, chapter 3 highlights the incremental improvements done with improving

the generalizability and robustness of the SI systems trained independently on the XRMB and HPRC datasets. The best performing SI system trained on the XRMB dataset estimates 6 TVs with an average PPMC score of 0.8770, which is close to a 9% improvement from the previous SI system proposed in Sivaraman et al. (2019). The proposed SI system is based on a TCN network and uses auditory spectrograms as input speech representation. The same SI system is trained to learn three additional source features as targets along with the 6 TVs and is one of the key design changes that contributed to a significant improvement in performance. The best performing SI system can be readily used and is hosted in a github repository at `https://github.com/Yashish92/TCN_SI_tool`. Section 3.5 discusses the best performing BiGRNN SI system trained to estimate the enhanced TVs with the self-supervised HuBERT speech embeddings. The model trained with enhanced TVs and HuBERT features is hosted in a github repository at `https://github.com/Yashish92/SSL-SI-tool`. Moreover, with the HPRC dataset and its unique 9 TVs, a multi-task learning based BiGRNN network trained with self-supervised HuBERT features performed the best. The same repository `https://github.com/Yashish92/SSL-SI-tool` also holds the current best performing MTL based SI system trained independently on the HPRC dataset.

Creating an accurate articulatory dataset is an incredibly demanding task in terms of time, effort, and the tools involved. This is precisely why the development of speaker-independent (SI) systems was initially pursued. However, the dependence of DNN systems on more ground-truth data begs the question as to find alternative training mechanisms that could build more generalizable SI systems. The MirrorNet training algorithm discussed in chapter 5 was experimented to circumvent the exact same issue, so that an inverse mapping to estimate the articulatory representations can be learned with a minimal amount of ground-truth data. The MirrorNet

architecture once trained in unsupervised fashion is capable of estimating articulatory representations as the latent space of the autoencoder network. The estimated articulatory representations have comparable accuracy to that estimated by the state-of-the-art SI systems trained in completely supervised fashion.

None of the publicly available articulatory datasets contain any ground-truth measures to capture the velar and glottal activity of speech due to obvious reasons. However, velar and glottal activity of speech adds rich, complementary information that could make significant improvements to speech applications. This motivated a data collection which is described in chapter 4 to primarily derive a parameter which can ideally act as the velar constriction degree TV. Once the derived parameter (Nasalance) was validated with a more invasive velar activity measure (HSN), a novel SI system was trained to estimate the derived velar constriction degree TV in speaker-independent fashion. Since EGG data was synchronously collected, a voicing parameter was also derived as a proxy to capture the glottal activity and was also estimated from the same SI system. The proposed system also benefited from the estimation of additional proxy source features as targets. The particular data collection described in this chapter is still ongoing and further improvements on the SI system will be made with more date to come.

Chapter 6 asserts the importance of using articulatory representations as an effective speech representation for two distinct speech applications. With the application of developing a multimodal system, the articulatory representation based ACFs played an important role in improving the performance of detecting subjects with strong positive symptoms in schizophrenia from healthy controls. In the same work, an important discovery was made with regard to understanding the articulatory coordination patterns displayed by subjects with strong positive symptoms. This distinctive articulatory coordination difference between the schizophrenia subjects

and the healthy controls is hypothesized to be a key differentiating factor in improving the classification performance of the developed multimodal systems. Finally, as the second application, articulatory representations were analyzed as a speech feature to understand if any of the six TVs could capture the mispronounciation of /ɹ/in child speech (fully rhotic vs derhotic). It was noticed that the tongue body related TVs have a significant difference between the fully rhotic and derhotic speech which motivated the development of a classification model to detect the rhoticity in child speech. Results in this experiment showed that the formant based speech features which have been used in the state-of-the-art classification systems can be simply outperformed by using articulatory representations with source features (aperiodicity, peridocity and pitch). While the study's findings are constrained by a small sample size, they serve as a catalyst for conducting more extensive research into the performance of adult-trained TV estimates for child speech.

Overall, this dissertation showcases machine learning, signal processing, and sensorimotor learning inspired techniques undertaken to improve and extend the speech inversion frameworks. Subsequently, the articulatory representations obtained from these speech inversion frameworks are put to practical use in enhancing real-world speech applications.

7.2 Future Directions

Following is a brief list of future directions that could be pursued based on the final findings of this dissertation.

7.2.1 Exploring effective evaluation metrics for SI systems

In the context of evaluating the performance of SI systems, PPMC scores have mostly been used. It is important that any SI system estimates the 'critical articulators' pertaining to the production of consonant sounds with better accuracy, compared to the non-critical articulators. However, PPMC calculates an average correlation score between the estimated and ground-truth articulatory parameters, and therefore can be less effective, especially when comparing closely performing SI systems. So it is imperative to explore effective means of automatically evaluating the SI systems with more emphasis on how well the critical articulators are estimated.

7.2.2 Extending the glottal source features to capture voice qualities with EGG data

The EGG data collected in the ongoing data collection discussed in chapter 4 has only been used to estimate an envelope parameter which is limited to capturing voicing in speech. However, by capturing different signal properties of the EGG signal (eg. rate for rise and fall of the glottal pulse, period duration), different voice qualities can be inferred (eg. breathy, creaky, modal). The ability to estimate such voice qualities from a trained SI system can indeed be helpful in a myriad of speech applications.

7.2.3 Experimenting the SI systems with accented English speech and other languages

The SI systems developed in this work are all trained with native American English and have not been extensively tested with estimating articulatory representations for accented English speech, or speech with other languages. The data collection described in chapter 4 is expected to be extended to collecting speech from non-native English speakers (collecting both accented English speech and speech from their native language), which will provide ground-truth articulatory data to evaluate the estimations from the current SI systems. Based on the evaluations, further fine tuning of the SI systems can be done to develop language specific SI systems.

7.2.4 MirrorNet to learn control parameters to drive a parametric vocal tract model

Rule based methods are currently used to generate articulatory control parameters to drive some of the parametric vocal tract models (Birkholz, 2013). This process is extremely time-consuming and needs prior knowledge with each control parameter in order to synthesize intelligible speech. MirrorNet could be experimented with learning to drive such complex, parametric vocal tract models in completely unsupervised fashion, so that the latent space could converge to the desired control parameters essential to re-synthesize a given speech target. To achieve this, more emphasis will also be needed in exploring faster training algorithms and powerful DNN model architectures (pre-trained transformers for encoder).

Bibliography

Ieee recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, 17(3):225–246, 1969. doi: 10.1109/TAU.1969.1162058.

Amber Afshan and Prasanta Kumar Ghosh. Improved subject-independent acoustic-to-articulatory inversion. *Speech Communication*, 66:1–16, 2015. ISSN 01676393. doi: 10.1016/j.specom.2014.07.005.

Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, M. Freitag, Sergey Pugachevskiy, Alice Baird, and B. Schuller. Snore sound classification using image-based deep spectrum features. In *INTERSPEECH*, 2017.

Nancy C. Andreasen and Scott Olsen. Negative v Positive Schizophrenia: Definition and Validation. *Archives of General Psychiatry*, 39(7):789–794, 07 1982. ISSN 0003-990X. doi: 10.1001/archpsyc.1982.04290070025006. URL `https://doi.org/10.1001/archpsyc.1982.04290070025006`.

Aravind Illa and Prasanta Kumar Ghosh. Speaker Conditioned Acoustic-to-Articulatory Inversion Using x-Vectors. In *Proc. Interspeech 2020*, pages 1376–1380, 2020. doi: 10.21437/Interspeech.2020-1222.

María Florencia Assaneo, Marcos A Trevisan, and Gabriel B Mindlin. Discrete motor coordinates for vowel production. *PloS one*, 8(11):e80373, 2013.

Ahmed Adel Attia and Carol Y Espy-Wilson. Masked autoencoders are articulatory learners. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Ahmed Adel Attia, Yashish M. Siriwardena, and Carol Espy-Wilson. Improving speech inversion through self-supervised embeddings and enhanced tract variables, 2023a.

Ahmed Adel Attia, Mark Tiede, and Carol Y Espy-Wilson. Enhancing speech articulation analysis using a geometric transformation of the x-ray microbeam dataset. *arXiv preprint arXiv:2305.10775*, 2023b.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL https://arxiv.org/abs/2006.11477.

R.J. Baken. *Clinical Measurement of Speech and Voice*. Singular Publishing Group, 1996a. ISBN 9781565938090. URL https://books.google.com/books?id=lpMhAQAAMAAJ.

R.J. Baken. *Clinical Measurement of Speech and Voice*. Singular Publishing Group, 1996b. ISBN 9781565938090. URL https://books.google.com/books?id=lpMhAQAAMAAJ.

T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018.

Gašper Beguš, Alan Zhou, Peter Wu, and Gopala K Anumanchipalli. Articulation gan: Unsupervised modeling of articulatory learning. *arXiv preprint arXiv:2210.15173*, 2022.

Nina Benway, Jonathan L. Preston, Elaine Hitchcock, Asif Salekin, Harshit Sharma, and Tara McAllister. PERCEPT-R: An Open-Access American English Child/Clinical Speech Corpus Specialized for the Audio Classification of //. In *Proc. Interspeech 2022*, pages 3648–3652, 2022. doi: 10.21437/Interspeech.2022-10785.

Nina R Benway and Jonathan L Preston. Prospective Validation of Motor-Based Intervention with Automated Mispronunciation Detection of Rhotics in Residual Speech Sound Disorders. In *Proc. INTERSPEECH 2023*, pages 4558–4562, 2023. doi: 10.21437/Interspeech.2023-1882.

Nina R Benway, Elaine R Hitchcock, Tara McAllister, Graham Tomkins Feeny, Jennifer Hill, and Jonathan L Preston. Comparing biofeedback types for children with residual // errors in american english: A Single-Case randomization design. *Am J Speech Lang Pathol*, 30(4): 1819–1845, July 2021.

Nina R Benway, Jonathan L Preston, Elaine Hitchcock, Yvan Rose, Asif Salekin, Wendy Liang, and Tara McAllister. Reproducible speech research with the artificial Intelligence-Ready PERCEPT corpora. *J Speech Lang Hear Res*, 66(6):1986–2009, June 2023a.

Nina R Benway, Jonathan L Preston, Asif Salekin, Yi Xiao, Harshit Sharma, and Tara McAllister. Classifying Rhoticity of // in Speech Sound Disorder using Age-and-Sex Normalized Formants. In *Proc. INTERSPEECH 2023*, pages 4563–4567, 2023b. doi: 10.21437/Interspeech.2023-312.

Nina R Benway, Yashish M Siriwardena, Jonathan L Preston, Elaine Hitchcock, Tara McAllister, and Carol Espy-Wilson. Acoustic-to-Articulatory Speech Inversion Features for Mispronunciation Detection of // in Child Speech Sound Disorders. In *Proc. INTERSPEECH 2023*, pages 4568–4572, 2023c. doi: 10.21437/Interspeech.2023-1924.

Visar Berisha, Chelsea Krantsevich, Gabriela Stegmann, Shira Hahn, and Julie Liss. Are reported accuracies in the clinical speech machine learning literature overoptimistic? In *Proc. Interspeech 2022*, pages 2453–2457, 2022. doi: 10.21437/Interspeech.2022-691.

Peter Birkholz. Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLOS ONE*, 8(4):1–17, 04 2013. doi: 10.1371/journal.pone.0060603. URL https://doi.org/10.1371/journal.pone.0060603.

Peter Birkholz, Dietmar Jackèl, and Bernd J Kroger. Construction and control of a three-dimensional vocal tract model. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE, 2006.

Florent Bocquelet, Thomas Hueber, Laurent Girin, Pierre Badin, and Blaise Yvert. Robust articulatory speech synthesis using deep neural networks for BCI applications. In *Proc. Interspeech 2014*, pages 2288–2292, 2014a. doi: 10.21437/Interspeech.2014-449.

Florent Bocquelet, Thomas Hueber, Laurent Girin, Pierre Badin, and Blaise Yvert. Robust Articulatory Speech Synthesis using Deep Neural Networks for BCI Applications. In *Interspeech 2014 - 15th Annual Conference of the International Speech Communication Association*, Proc. of the 15th Annual Conference of the International Speech Communication Association (Interspeech 2014), Singapour, Singapore, September 2014b. URL https://hal.archives-ouvertes.fr/hal-01228891.

Florent Bocquelet, Thomas Hueber, Laurent Girin, Christophe Savariaux, and Blaise Yvert. Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. *PLOS Computational Biology*, 12(11):1–28, 11 2016. doi: 10.1371/journal.pcbi.1005119. URL https://doi.org/10.1371/journal.pcbi.1005119.

Paul Boersma. Functional phonology: Formalizing the interactions between articulatory and perceptual drives. 01 1999.

Suzanne E Boyce. The articulatory phonetics of /r/ for residual speech errors. *Semin Speech Lang*, 36(4):257–270, October 2015.

Catherine P Browman and Louis Goldstein. Articulatory Phonology : An Overview *. *Phonetica*, 49:155–180, 1992.

Kate Bunton and Brad H Story. The relation of nasality and nasalance to nasal port area based on a computational model. *Cleft Palate Craniofac J*, 49(6):741–749, October 2011.

Carlyn Burris, Houri K Vorperian, Marios Fourakis, Ray D Kent, and Daniel M Bolt. Quantitative and descriptive comparison of four acoustic analysis systems: vowel measurements. *J Speech Lang Hear Res*, 57(1):26–45, February 2014.

Xingyu Cai, Jiahong Yuan, Renjie Zheng, Liang Huang, and Kenneth Church. Speech Emotion Recognition with Multi-Task Learning. In *Proc. Interspeech 2021*, pages 4508–4512, 2021. doi: 10.21437/Interspeech.2021-1852.

Rich Caruana. Multitask Learning. 1997. doi: 10.1023/A:1007379606734. URL https://doi.org/10.1023/A:1007379606734.

Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

Taijing Chen, Adam Lammert, and Benjamin Parrell. Modeling Sensorimotor Adaptation in Speech Through Alterations to Forward and Inverse Models. In *Proc. Interspeech 2021*, pages 3201–3205, 2021. doi: 10.21437/Interspeech.2021-1746.

Taishih Chi, Powen Ru, and Shihab A. Shamma. Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, 2005.

Chung-Ming Chien, Jheng-Hao Lin, Chien-yu Huang, Po-chun Hsu, and Hung-yi 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)*, pages 8588–8592, 2021. doi: 10.1109/ICASSP39728.2021.9413880.

Cheol Jun Cho, Peter Wu, Abdelrahman Mohamed, and Gopala K. Anumanchipalli. Evidence of vocal tract articulation in self-supervised learning of speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10094711.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL https://aclanthology.org/W14-4012.

Taehong Cho and Patricia Keating. Effects of initial position versus prominence in english. *Journal of Phonetics*, 37(4):466–485, 2009. ISSN 0095-4470. doi: https://doi.org/10.1016/j.wocn.2009.08.001. URL https://www.sciencedirect.com/science/article/pii/S0095447009000497.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

Jianwu Dang and Kiyoshi Honda. Construction and control of a physiological articulatory model. *The Journal of the Acoustical Society of America*, 115(2):853–870, 2004. doi: 10.1121/1.1639325. URL https://doi.org/10.1121/1.1639325.

PIERRE DELATTRE and DONALD C. FREEMAN. A dialect study of american r's by x-ray motion picture. 6(44):29–68, 1968. doi: doi:10.1515/ling.1968.6.44.29. URL https://doi.org/10.1515/ling.1968.6.44.29.

Caroline Demily and Nicolas Franck. Cognitive remediation: a promising tool for the treatment of schizophrenia. *Expert Review of Neurotherapeutics*, 8(7):1029–1036, 2008. doi: 10.1586/14737175.8.7.1029. URL https://doi.org/10.1586/14737175.8.7.1029. PMID: 18590474.

Ekaterini Derdemezis, Houri K Vorperian, Ray D Kent, Marios Fourakis, Emily L Reinicke, and Daniel M Bolt. Optimizing vowel formant measurements in four acoustic analysis systems for diverse speaker groups. *Am J Speech Lang Pathol*, 25(3):335–354, August 2016.

O. Deshmukh, C. Y. Espy-Wilson, A. Salomon, and J. Singh. Use of temporal information: detection of periodicity, aperiodicity, and pitch in speech. *IEEE Transactions on Speech and Audio Processing*, 13(5):776–786, 2005. doi: 10.1109/TSA.2005.851910.

Giovanni M. Di Liberto, Guilhem Marion, and Shihab A. Shamma. The music of silence: Part ii: Music listening induces imagery responses. *Journal of Neuroscience*, 41(35):7449–7460, 2021. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0184-21.2021. URL https://www.jneurosci.org/content/41/35/7449.

Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. Ddsp: Differentiable digital signal processing, 2020a.

Jesse Engel, Rigel Swavely, Lamtharn Hantrakul, Adam Roberts, and Curtis Hawthorne. Self-supervised pitch detection by inverse audio synthesis. 2020b.

Olov Engwall. Combining mri, ema and epg measurements in a three-dimensional tongue model. *Speech Communication*, 41(2):303–329, 2003. ISSN 0167-6393. doi: https://doi.org/10.1016/S0167-6393(02)00132-2. URL https://www.sciencedirect.com/science/article/pii/S0167639302001322.

Philippe Esling, Naotake Masuda, Adrien Bardet, Romeo Despres, and Axel Chemla-Romeu-Santos. Flow synthesizer: Universal audio synthesizer control with normalizing flows. *Applied Sciences*, 10(1), 2020. ISSN 2076-3417. doi: 10.3390/app10010302. URL https://www.mdpi.com/2076-3417/10/1/302.

Carol Espy-Wilson, Adam C. Lammert, Nadee Seneviratne, and Thomas F. Quatieri. Assessing Neuromotor Coordination in Depression Using Inverted Vocal Tract Variables. In *Proc. Interspeech 2019*, pages 1448–1452, 2019. doi: 10.21437/Interspeech.2019-1815.

Carol Y. Espy-Wilson, Suzanne E. Boyce, Michel Jackson, Shrikanth Narayanan, and Abeer Alwan. Acoustic modeling of American English /r/. *The Journal of the Acoustical Society of America*, 108(1):343–356, 07 2000. ISSN 0001-4966. doi: 10.1121/1.429469. URL https://doi.org/10.1121/1.429469.

F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016. doi: 10.1109/TAFFC.2015.2457417.

Sascha Fagel and Katja Madany. A 3-d virtual head as a tool for speech therapy for children. In *INTERSPEECH*, 2008.

M. Feldman. Hilbert transforms. In S. Braun, editor, *Encyclopedia of Vibration*, pages 642–648. Elsevier, Oxford, 2001. ISBN 978-0-12-227085-7. doi: https://doi.org/10.1006/rwvb.

2001.0057. URL https://www.sciencedirect.com/science/article/pii/B0122270851000576.

S G Fletcher and S D Frost. Quantitative and graphic analysis of prosthetic treatment for "nasalance" in speech. *J Prosthet Dent*, 32(3):284–291, September 1974.

A J Flint, S E Black, I Campbell-Taylor, G F Gailey, and C Levinton. Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. *J Psychiatr Res*, 27 (3):309–319, July 1993.

Joe Frankel and Simon King. Asr - articulatory speech recognition. In *INTERSPEECH*, 2001.

Yiwei Fu, Devesh K. Jha, Zeyu Zhang, Zhenyuan Yuan, and Asok Ray. Neural network-based learning from demonstration of an autonomous ground robot. *Machines*, 7(2), 2019. ISSN 2075-1702. doi: 10.3390/machines7020024. URL https://www.mdpi.com/2075-1702/7/2/24.

Lisa Furlong, Shane Erickson, and Meg E Morris. Computer-based speech therapy for childhood speech sound disorders. *J Commun Disord*, 68:50–69, June 2017.

Vittorio Gallese, Luciano Fadiga, Leonardo Fogassi, and Giacomo Rizzolatti. Action recognition in the premotor cortex. *Brain*, 119(2):593–609, 04 1996. ISSN 0006-8950. doi: 10.1093/brain/119.2.593. URL https://doi.org/10.1093/brain/119.2.593.

Marc-Antoine Georges, Pierre Badin, Julien Diard, Laurent Girin, Jean-Luc Schwartz, and Thomas Hueber. Towards an articulatory-driven neural vocoder for speech synthesis. In *ISSP 2020 - 12th International Seminar on Speech Production*, Providence (virtual), United States, December 2020. URL https://hal.archives-ouvertes.fr/hal-03184762.

Marc-Antoine Georges, Laurent Girin, Jean-Luc Schwartz, and Thomas Hueber. Learning Robust Speech Representation with an Articulatory-Regularized Variational Autoencoder. In *Proceedings Interspeech 2021*, pages 3345–3349, 2021. doi: 10.21437/Interspeech.2021-1604.

Marc-Antoine Georges, Julien Diard, Laurent Girin, Jean-Luc Schwartz, and Thomas Hueber. Repeat after me: Self-supervised learning of acoustic-to-articulatory mapping by vocal imitation, 2022. URL https://arxiv.org/abs/2204.02269.

Prasanta Kumar Ghosh and Shrikanth Narayanan. A generalized smoothness criterion for acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America*, 128(4):2162–72, 2010. ISSN 1520-8524. doi: 10.1121/1.3455847.

Laurent Girin, Thomas Hueber, and Xavier Alameda-Pineda. Extending the Cascaded Gaussian Mixture Regression Framework for Cross-Speaker Acoustic-Articulatory Mapping. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 25(3):662–673, 2017. ISSN 23299290. doi: 10.1109/TASLP.2017.2651398. URL http://ieeexplore.ieee.org/document/7814297/.

Jeffrey S. Gonzalez, Erica Shreck, and Abigail Batchelder. *Hamilton Rating Scale for Depression (HAM-D)*, pages 887–888. Springer New York, New York, NY, 2013. ISBN 978-1-4419-1005-9. doi: 10.1007/978-1-4419-1005-9_198. URL https://doi.org/10.1007/978-1-4419-1005-9_198.

J F Greden and B J Carroll. Psychomotor function in affective disorders: an overview of new monitoring techniques. *Am J Psychiatry*, 138(11):1441–1448, November 1981.

Frank H. Guenther, Carol Y. Espy-Wilson, Suzanne E. Boyce, Melanie L. Matthies, Majid Zandipour, and Joseph S. Perkell. Articulatory tradeoffs reduce acoustic variability during American English /r/ production. *The Journal of the Acoustical Society of America*, 105(5):2854–2865, apr 1999. ISSN 0001-4966. doi: 10.1121/1.426900. URL http://asa.scitation.org/doi/10.1121/1.426900.

Frank H H Guenther, Frank H. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language.*, 96(3), 2006-03-01. ISSN 0093-934X.

Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402, 2007.

Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan. Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm. In *INTERSPEECH*, 2017.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460, oct 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3122291. URL https://doi.org/10.1109/TASLP.2021.3122291.

Zhaocheng Huang, J. Epps, and D. Joachim. Exploiting vocal tract coordination using dilated cnns for depression detection in naturalistic environments. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6549–6553, 2020.

Lu Hui, Liu Hui Ting, Swee Lan See, and Paul Yaozhu Chan. Use of electroglottograph (egg) to find a relationship between pitch, emotion and personality. *Procedia Manufacturing*, 3:1926–1931, 2015. ISSN 2351-9789. doi: https://doi.org/10.1016/j.promfg.2015.07.236. URL https://www.sciencedirect.com/science/article/pii/S2351978915002371. 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015.

Edward E. Hunter and Meghan Murphy. *Brief Psychiatric Rating Scale*, pages 447–449. Springer New York, New York, NY, 2011. ISBN 978-0-387-79948-3. doi: 10.1007/978-0-387-79948-3_1976. URL https://doi.org/10.1007/978-0-387-79948-3_1976.

Aravind Illa and Prasanta Kumar Ghosh. Low Resource Acoustic-to-articulatory Inversion Using Bi-directional Long Short Term Memory. In *Proc. Interspeech 2018*, pages 3122–3126, 2018. doi: 10.21437/Interspeech.2018-1843.

Aravind Illa and Prasanta Kumar Ghosh. An Investigation on Speaker Specific Articulatory Synthesis with Speaker Independent Articulatory Inversion. In *Proc. Interspeech 2019*, pages 121–125, 2019a. doi: 10.21437/Interspeech.2019-2664.

Aravind Illa and Prasanta Kumar 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)*, pages 5931–5935, 2019b. doi: 10.1109/ICASSP.2019.8682506.

R. Gutierrez-Osuna P. Monroe P. McCabe J. McKechnie, B. Ahmed and K. J. Ballard. Automated speech analysis tools for children's speech production: A systematic literature review. *International Journal of Speech-Language Pathology*, 20(6):583–598, 2018. doi: 10.1080/17549507.2018.1477991. URL https://doi.org/10.1080/17549507.2018.1477991.

Navdeep Jaitly and E. Hinton. Vocal tract length perturbation (vtlp) improves speech recognition. 2013.

An Ji. *Speaker Independent Acoustic-To-Articulatory Inversion*. PhD thesis, Marquette University, 2014.

An Ji, Michael T Johnson, and Jeffrey J Berry. Parallel Reference Speaker Weighting for Kinematic-Independent Acoustic-to-Articulatory Inversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1865–1875, 2016.

Naoyuki Kanda, Ryu Takeda, and Yasunari Obuchi. Elastic spectral distortion for low resource speech recognition with deep neural networks. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 309–314, 2013. doi: 10.1109/ASRU.2013.6707748.

Georg B. Keller, Tobias Bonhoeffer, and Mark Hübener. Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron*, 74(5):809–815, 2012. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2012.03.040. URL https://www.sciencedirect.com/science/article/pii/S0896627312003844.

Christopher T Kello and David C Plaut. A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. *The Journal of the Acoustical Society of America*, 116(4):2354–2364, 2004.

Deanna L. Kelly, Max Spaderna, Vedrana Hodzic, Suraj Nair, Christopher Kitchen, Anne E. Werkheiser, Megan M. Powell, Fang Liu, Glen Coppersmith, Shuo Chen, and Philip Resnik. Blinded clinical ratings of social media data are correlated with in-person clinical ratings in participants diagnosed with either depression, schizophrenia, or healthy controls. *Psychiatry Research*, 294:113496, 2020. ISSN 0165-1781. doi: https://doi.org/10.1016/j.psychres.2020.113496. URL http://www.sciencedirect.com/science/article/pii/S0165178120331577.

Raymond D Kent and Houri K Vorperian. Static measurements of vowel formant frequencies and bandwidths: A review. *J Commun Disord*, 74:74–97, June 2018.

Suyoun Kim, Takaaki Hori, and Shinji 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)*, pages 4835–4839, 2017.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–18, 2021. doi: 10.1109/TITS.2021.3054625.

Harriet B Klein, Tara McAllister Byun, Lisa Davidson, and Maria I Grigos. A multidimensional investigation of children's /r/ productions: perceptual, ultrasound, and acoustic measures. *Am J Speech Lang Pathol*, 22(3):540–553, June 2013.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *INTERSPEECH*, 2015.

Alexei Kochetov. Research methods in articulatory phonetics ii: Studying other gestures and recent trends. *Language and Linguistics Compass*, 14(6):e12371, 2020. doi: https://doi.org/10.1111/lnc3.12371. URL https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12371.

R A Krakow, P S Beddor, L M Goldstein, and C A Fowler. Coarticulatory influences on the perceived height of nasal vowels. *J Acoust Soc Am*, 83(3):1146–1158, March 1988.

Rena A. Krakow. Physiological organization of syllables: a review. *Journal of Phonetics*, 27(1):23–54, 1999. ISSN 0095-4470. doi: https://doi.org/10.1006/jpho.1999.0089. URL https://www.sciencedirect.com/science/article/pii/S009544709990089X.

Jelena Krivokapić. Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658):20130397, 2014. doi: 10.1098/rstb.2013.0397. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2013.0397.

Bernd Kröger and Peter Birkholz. Articulatory synthesis of speech and singing: State of the art and suggestions for future research. volume 5398, pages 306–319, 01 2008. ISBN 978-3-642-00524-4. doi: 10.1007/978-3-642-00525-1_31.

Patricia K. Kuhl. Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, 5:831–843, 2004.

G. R. Kuperberg. Language in schizophrenia Part 1: an Introduction. *Lang Linguist Compass*, 4 (8):576–589, Aug 2010.

Gwendal Le Vaillant, Thierry Dutoit, and Sébastien Dekeyser. Improving synthesizer programming from variational autoencoders latent space. In *Proceedings of the 24th International Conference on Digital Audio Effects (DAFx20in21)*, September 2021.

Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468, 03 1999. ISSN 0001-4966. doi: 10.1121/1.426686. URL https://doi.org/10.1121/1.426686.

Sarah R. Li, Sarah Dugan, Jack Masterson, Hannah Hudepohl, Colin Annand, Caroline Spencer, Renee Seward, Michael A. Riley, Suzanne Boyce, and T. Douglas Mast. Classification of accurate and misarticulated /r/ for ultrasound biofeedback using tongue part displacement trajectories. *Clinical Linguistics & Phonetics*, 37(2):196–222, 2023. doi: 10.1080/02699206.2022.2039777. URL https://doi.org/10.1080/02699206.2022.2039777. PMID: 35254181.

Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In *INTERSPEECH*, 2019.

Zhen-Hua Ling, Korin Richmond, and Junichi Yamagishi. Articulatory control of hmm-based parametric speech synthesis using feature-space-switched multiple regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(1):207–219, 2013. doi: 10.1109/TASL.2012.2215600.

Andy T. Liu, Shang-Wen Li, and Hung-yi Lee. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29: 2351–2366, jul 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3095662. URL https://doi.org/10.1109/TASLP.2021.3095662.

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)*, pages 10434–10443, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.01045.

Edwin Maas, Donald A. Robin, Shannon N. Austermann Hula, Skott E. Freedman, Gabriele Wulf, Kirrie J. Ballard, and Richard A. Schmidt. Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology*, 17(3):277–298, 2008. doi: 10.1044/1058-0360(2008/025). URL https://pubs.asha.org/doi/abs/10.1044/1058-0360%282008/025%29.

Shinji Maeda. A digital simulation method of the vocal-tract system. *Speech Communication*, 1(3):199–229, 1982. ISSN 0167-6393. doi: https://doi.org/10.1016/0167-6393(82)90017-6. URL https://www.sciencedirect.com/science/article/pii/0167639382900176.

Shinji Maeda. Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech Production and Speech Modelling*, pages 131–149. Springer Netherlands, Dordrecht, 1990. ISBN 0-7923-0746-1. doi: 10.1007/

978-94-009-2037-8_6. URL http://www.springerlink.com/index/10.1007/978-94-009-2037-8{_}6.

Tanumay Mandal, K. Rao, and Sanjay Gupta. Identification of glottal instants using electroglottographic signal for vulnerable cases of voicing. *Healthcare Technology Letters*, 7: 132–138, 10 2020. doi: 10.1049/htl.2019.0085.

Guilhem Marion, Giovanni M. Di Liberto, and Shihab A. Shamma. The music of silence: Part i: Responses to musical imagery encode melodic expectations and acoustics. *Journal of Neuroscience*, 41(35):7435–7448, 2021. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0183-21.2021. URL https://www.jneurosci.org/content/41/35/7435.

Richard S. McGowan. Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests. *Speech Communication*, 14(1):19–48, 1994. ISSN 01676393. doi: 10.1016/0167-6393(94)90055-8.

Jacqueline McKechnie, Beena Ahmed, Ricardo Gutierrez-Osuna, Elizabeth Murray, Patricia McCabe, and Kirrie J Ballard. The influence of type of feedback during tablet-based delivery of intensive treatment for childhood apraxia of speech. *J Commun Disord*, 87:106026, July 2020.

Vikramjit Mitra, Hosung Nam, Carol Y. Espy-Wilson, Elliot Saltzman, and Louis Goldstein. Retrieving tract variables from acoustics: A comparison of different machine learning strategies. *IEEE Journal on Selected Topics in Signal Processing*, 4(6):1027–1045, sep 2010. ISSN 19324553. doi: 10.1109/JSTSP.2010.2076013.

Vikramjit Mitra, Hosung Nam, Carol Y. Espy-Wilson, Elliot Saltzman, and Louis Goldstein. Speech inversion: Benefits of tract variables over pellet trajectories. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5188–5191. IEEE, may 2011. ISBN 978-1-4577-0538-0. doi: 10.1109/ICASSP.2011.5947526. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5947526.

Vikramjit Mitra, Hosung Nam, Carol Espy-Wilson, Elliot Saltzman, and Louis Goldstein. Recognizing articulatory gestures from speech for robust speech recognition. *The Journal of the Acoustical Society of America*, 131(3):2270–2287, 2012. ISSN 0001-4966. doi: 10.1121/1.3682038. URL http://asa.scitation.org/doi/10.1121/1.3682038.

Masanori MORISE, Fumiya YOKOMORI, and Kenji OZAWA. World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, E99.D(7):1877–1884, 2016. doi: 10.1587/transinf.2015EDP7457.

Clément Moulin-Frier, Sao Mai Nguyen, and Pierre-Yves Oudeyer. Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. *Frontiers in Psychology*, 4, 2014. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.01006. URL https://www.frontiersin.org/article/10.3389/fpsyg.2013.01006.

Shrikanth Narayanan, Krishna Nayak, Sungbok Lee, Abhinav Sethy, and Dani Byrd. An approach to real-time magnetic resonance imaging for speech production. *The Journal of the Acoustical Society of America*, 115(4):1771–1776, mar 2004. ISSN 0001-4966. doi: 10.1121/1.1652588. URL http://asa.scitation.org/doi/10.1121/1.1652588.

Tom O'Haver. Fast smoothing function. *MathWorks*, 2017. URL https://www.mathworks.com/matlabcentral/fileexchange/19998-fast-smoothing-function.

Liran Oren, Michael Rollins, Srujana Padakanti, Ann Kummer, Ephraim Gutmark, and Suzanne Boyce. Using high-speed nasopharyngoscopy to quantify the bubbling above the velopharyngeal valve in cases of nasal rustle. *The Cleft Palate-Craniofacial Journal*, 57(5): 637–645, 2020. doi: 10.1177/1055665619894183. URL https://doi.org/10.1177/1055665619894183. PMID: 31867995.

Silvia Pagliarini, Arthur Leblois, and Xavier Hinaut. Canary Vocal Sensorimotor Model with RNN Decoder and Low-dimensional GAN Generator. In *2021 IEEE International Conference on Development and Learning (ICDL)*, pages 1–8, 2021. doi: 10.1109/ICDL49984.2021.9515607.

Ashutosh Pandey and DeLiang Wang. Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6875–6879, 2019. doi: 10.1109/ICASSP.2019.8683634.

G Papcun, J Hochberg, T R Thomas, F Laroche, J Zacks, and S Levy. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *J Acoust Soc Am*, 92(2 Pt 1):688–700, August 1992a.

George Papcun, Judy Hochberg, Timothy Thomas, François Laroche, Jeff Zacks, and Simon Levy. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *The Journal of the Acoustical Society of America*, 92:688–700, 09 1992b. doi: 10.1121/1.403994.

Raghavendra Pappagari, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velazquez, and Najim Dehak. Copypaste: An augmentation method for speech emotion recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6324–6328, 2021. doi: 10.1109/ICASSP39728.2021.9415077.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 2613–2617, 2019. doi: 10.21437/Interspeech.2019-2680.

Anja Kristina Philippsen, René Felix Reinhart, and Britta Wrede. Learning how to speak: Imitation-based refinement of syllable production in an articulatory-acoustic model. In *4th International Conference on Development and Learning and on Epigenetic Robotics*, pages 195–200, 2014. doi: 10.1109/DEVLRN.2014.6982981.

Jonathan L. Preston, Nina R. Benway, Megan C. Leece, Elaine R. Hitchcock, and Tara McAllister. Tutorial: Motor-based treatment strategies for /r/ distortions. *Language, Speech, and Hearing Services in Schools*, 51(4):966–980, 2020. doi: 10.1044/2020_LSHSS-20-00012. URL https://pubs.asha.org/doi/abs/10.1044/2020_LSHSS-20-00012.

Emily B. Prince, Katherine B. Martin, and D. Messinger. Facial action coding system. 2015.

Tarun Pruthi and Carol Espy-Wilson. Acoustic parameters for the automatic detection of vowel nasalization. pages 1925–1928, 08 2007. doi: 10.21437/Interspeech.2007-40.

Tarun Pruthi and Carol Y. Espy-Wilson. Acoustic parameters for automatic detection of nasal manner. *Speech Communication*, 43(3):225–239, 2004. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2004.06.001. URL https://www.sciencedirect.com/science/article/pii/S0167639304000573.

Chao Qin and Miguel Á Carreira-Perpiñán. An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping. *Interspeech*, pages 74–77, 2007.

Mirco Ravanelli, Titouan Parcollet, and Peter Plantinga et. al. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. Multi-level attention network using text, audio and video for depression prediction, 2019.

Manuel Sam Ribeiro, Joanne Cleland, Aciel Eshky, Korin Richmond, and Steve Renals. Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors. *Speech Communication*, 128:24–34, 2021. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2021.02.001. URL https://www.sciencedirect.com/science/article/pii/S0167639321000170.

Korin Richmond. A trajectory mixture density network for the acoustic-articulatory inversion mapping. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2, pages 577–580. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 9781604234497. doi: 10.1007/978-3-540-77347-4_23. URL http://link.springer.com/10.1007/978-3-540-77347-4{_}23.

Korin Richmond and Simon King. Smooth talking: Articulatory join costs for unit selection. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5150–5154, 2016. doi: 10.1109/ICASSP.2016.7472659.

Fabien Ringeval, Björn Schuller, Michel Valstar, NIcholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition, 2019.

Panying Rong, Ryan Shosted, and Christopher Carignan. The relationship between velopharyngeal opening and place of articulation: An aerodynamic and epg investigation. In *9th International Seminar on Speech Production (ISSP)*, 06 2011.

Martin Rothenberg and James J. Mahshie. Monitoring vocal fold abduction through vocal fold contact area. *Journal of Speech, Language, and Hearing Research*, 31(3):338–351, 1988. doi: 10.1044/jshr.3103.338. URL https://pubs.asha.org/doi/abs/10.1044/jshr.3103.338.

D M Ruscello. Visual feedback in treatment of residual phonological disorders. *J Commun Disord*, 28(4):279–302, December 1995.

Pramit Saha and Sidney Fels. Learning Joint Articulatory-Acoustic Representations with Normalizing Flows. In *Proceedings Interspeech 2020*, pages 3196–3200, 2020. doi: 10.21437/Interspeech.2020-2004. URL http://dx.doi.org/10.21437/Interspeech.2020-2004.

Elliot L. Saltzman and Kevin G. Munhall. A Dynamical Approach to Gestural Patterning in Speech Production. *Ecological Psychology*, 1(4):333–382, dec 1989. ISSN 1040-7413. doi: 10.1207/s15326969eco0104_2. URL http://www.tandfonline.com/doi/abs/10.1207/s15326969eco0104{_}2.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

Paul W. Schönle, Klaus Gräbe, Peter Wenig, Jörg Höhne, Jörg Schrader, and Bastian Conrad. Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, 31(1):26–35, may 1987. ISSN 10902155. doi: 10.1016/0093-934X(87)90058-7.

Celia Scully. *Articulatory Synthesis*, pages 151–186. Springer Netherlands, Dordrecht, 1990. ISBN 978-94-009-2037-8. doi: 10.1007/978-94-009-2037-8_7. URL https://doi.org/10.1007/978-94-009-2037-8_7.

Nadee Seneviratne and Carol Espy-Wilson. Generalized dilated cnn models for depression detection using inverted vocal tract variables, 2020. URL https://arxiv.org/abs/2011.06739.

Nadee Seneviratne, Ganesh Sivaraman, Vikramjit Mitra, and Carol Espy-Wilson. Noise Robust Acoustic to Articulatory Speech Inversion. In *Proc. Interspeech 2018*, pages 3137–3141, 2018. doi: 10.21437/Interspeech.2018-1509.

Nadee Seneviratne, Ganesh Sivaraman, and Carol Espy-Wilson. Multi-Corpus Acoustic-to-Articulatory Speech Inversion. In *Proc. Interspeech 2019*, pages 859–863, 2019. doi: 10.21437/Interspeech.2019-3168.

Nadee Seneviratne, James R. Williamson, Adam C. Lammert, Thomas F. Quatieri, and Carol Espy-Wilson. Extended Study on the Use of Vocal Tract Variables to Quantify Neuromotor Coordination in Depression. In *Proc. Interspeech 2020*, pages 4551–4555, 2020. doi: 10.21437/Interspeech.2020-2758. URL http://dx.doi.org/10.21437/Interspeech.2020-2758.

Christine H Shadle, Hosung Nam, and D H Whalen. Comparing measurement errors for formants in synthetic and natural vowels. *J Acoust Soc Am*, 139(2):713–727, February 2016.

Abdolreza Sabzi Shahrebabaki, Sabato Marco Siniscalchi, Giampiero Salvi, and Torbjørn Svendsen. Sequence-to-Sequence Articulatory Inversion Through Time Convolution of Sub-Band Frequency Signals. In *Proc. Interspeech 2020*, pages 2882–2886, 2020. doi: 10.

21437/Interspeech.2020-1140. URL http://dx.doi.org/10.21437/Interspeech.2020-1140.

Abdolreza Sabzi Shahrebabaki, Sabato Marco Siniscalchi, and Torbjørn Svendsen. Raw Speech-to-Articulatory Inversion by Temporal Filtering and Decimation. In *Proc. Interspeech 2021*, pages 1184–1188, 2021. doi: 10.21437/Interspeech.2021-1429.

Shihab Shamma, Prachi Patel, Shoutik Mukherjee, Guilhem Marion, Bahar Khalighinejad, Cong Han, Jose Herrero, Stephan Bickel, Ashesh Mehta, and Nima Mesgarani. Learning Speech Production and Perception through Sensorimotor Interactions. *Cerebral Cortex Communications*, 2(1), 11 2020. ISSN 2632-7376. doi: 10.1093/texcom/tgaa091. URL https://doi.org/10.1093/texcom/tgaa091. tgaa091.

Patrice Y. Simard, Yann A. LeCun, John S. Denker, and Bernard Victorri. *Transformation Invariance in Pattern Recognition – Tangent Distance and Tangent Propagation*, pages 235–269. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_17.

Yashish M. Siriwardena and Carol Espy-Wilson. The secret source : Incorporating source features to improve acoustic-to-articulatory speech inversion. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095630.

Yashish M. Siriwardena, Carol Espy-Wilson, Chris Kitchen, and Deanna L. Kelly. *Multimodal Approach for Assessing Neuromotor Coordination in Schizophrenia Using Convolutional Neural Networks*, page 768–772. Association for Computing Machinery, New York, NY, USA, 2021a. ISBN 9781450384810. URL https://doi.org/10.1145/3462244.3479967.

Yashish M. Siriwardena, Chris Kitchen, Deanna L. Kelly, and Carol Espy-Wilson. Inverted Vocal Tract Variables and Facial Action Units to Quantify Neuromotor Coordination in Schizophrenia. In *Proc. 12th International Seminar on Speech Production (ISSP 2020)*, pages 174–177, 2021b. URL https://issp2020.yale.edu/ProcISSP2020.pdf.

Yashish M. Siriwardena, Guilhem Marion, and Shihab Shamma. The mirrornet : Learning audio synthesizer controls inspired by sensorimotor interaction. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 946–950, 2022a. doi: 10.1109/ICASSP43922.2022.9747358.

Yashish M. Siriwardena, Ganesh Sivaraman, and Carol Espy-Wilson. Acoustic-to-articulatory Speech Inversion with Multi-task Learning. In *Proc. Interspeech 2022*, pages 5020–5024, 2022b. doi: 10.21437/Interspeech.2022-11164.

Yashish M Siriwardena, Ahmed Adel Attia, Ganesh Sivaraman, and Carol Espy-Wilson. Audio data augmentation for acoustic-to-articulatory speech inversion. In *European Signal Processing Conference (EUSIPCO)*. IEEE, 2023a.

Yashish M Siriwardena, Carol Espy-Wilson, Suzanne Boyce, Mark Tiede, and Liran Oren. Speaker-independent Speech Inversion for Estimation of Nasalance. In *Proc. INTERSPEECH 2023*, pages 4743–4747, 2023b. doi: 10.21437/Interspeech.2023-2352.

Yashish M Siriwardena, Carol Espy-Wilson, and Shihab Shamma. Learning to Compute the Articulatory Representations of Speech with the MIRRORNET. In *Proc. INTERSPEECH 2023*, pages 5137–5141, 2023c. doi: 10.21437/Interspeech.2023-562.

G. Sivaraman, V. Mitra, H. Nam, M. Tiede, and C. Espy-Wilson. Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 08-12-Sept, 2016. doi: 10.21437/Interspeech.2016-1399.

Ganesh Sivaraman. *Articulatory representations to address acoustic variability in speech*. PhD thesis, University of Maryland College Park, 2017.

Ganesh Sivaraman, Carol Espy-Wilson, and Martijn Wieling. Analysis of Acoustic-to-Articulatory Speech Inversion Across Different Accents and Languages. In *Interspeech 2017*, pages 974–978, Stockholm, aug 2017. ISCA. doi: 10.21437/Interspeech.2017-260. URL http://www.isca-speech.org/archive/Interspeech{_}2017/abstracts/0260.html.

Ganesh Sivaraman, Vikramjit Mitra, Hosung Nam, Mark Tiede, and Carol Espy-Wilson. Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion. *The Journal of the Acoustical Society of America*, 146(1):316–329, 2019. doi: 10.1121/1.5116130. URL https://doi.org/10.1121/1.5116130.

David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A Music, Speech, and Noise Corpus, 2015. arXiv:1510.08484v1.

Man Sondhi and Juergen Schroeter. A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(7):955–967, 1987.

Kenneth N. Stevens. On the quantal nature of speech. *Journal of Phonetics*, 17(1):3–45, 1989. ISSN 0095-4470. doi: https://doi.org/10.1016/S0095-4470(19)31520-7. URL https://www.sciencedirect.com/science/article/pii/S0095447019315207.

Kenneth N Stevens. *Acoustic phonetics*, volume 30. MIT press, 2000.

Brad H Story. Phrase-level speech simulation with an airway modulation model of speech production. *Computer speech & language*, 27(4):989–1010, 2013.

Yifan Sun, Qinlong Huang, and Xihong Wu. Unsupervised Acoustic-to-Articulatory Inversion with Variable Vocal Tract Anatomy. In *Proc. Interspeech 2022*, pages 4656–4660, 2022. doi: 10.21437/Interspeech.2022-477.

Lei Tai and Ming Liu. Deep-learning in mobile robotics - from perception to control systems: A survey on why and why not. *CoRR*, abs/1612.07139, 2016. URL http://arxiv.org/abs/1612.07139.

Mark Tiede, Carol Y. Espy-Wilson, Dolly Goldenberg, Vikramjit Mitra, Hosung Nam, and Ganesh Sivaraman. Quantifying kinematic aspects of reduction in a contrasting rate production task. *The Journal of the Acoustical Society of America*, 141(5):3580–3580, 2017. doi: 10.1121/1.4987629. URL https://doi.org/10.1121/1.4987629.

Ingo Titze, Tobias Riede, and Peter Popolo. Nonlinear source-filter coupling in phonation: vocal exercises. *J Acoust Soc Am*, 123(4):1902–1915, April 2008.

Tomoki Toda, Alan W Black, and Keiichi Tokuda. Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model. *Speech communication*, 50 (3):215–227, 2008.

Asterios Toutios and Konstantinos Margaritis. A rough guide to the acoustic-to-articulatory inversion of speech. In *in: 6th Hellenic European Conference of Computer Mathematics and its Applications, HERCMA-2003*, 2003.

Asterios Toutios, Slim Ouni, and Yves Laprie. Estimating the control parameters of an articulatory model from electromagnetic articulograph data. *The Journal of the Acoustical Society of America*, 129(5):3245–3257, 2011.

James Traer and Josh H. McDermott. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48): E7856–E7865, 2016. doi: 10.1073/pnas.1612524113.

Sathvik Udupa, Anwesha Roy, Abhayjeet Singh, Aravind Illa, and Prasanta Kumar Ghosh. Estimating Articulatory Movements in Speech Production with Transformer Networks. In *Proc. Interspeech 2021*, pages 1154–1158, 2021. doi: 10.21437/Interspeech.2021-1375.

Jean-Marc Valin and Jan Skoglund. Lpcnet: Improving neural speech synthesis through linear prediction, 2018. URL `https://arxiv.org/abs/1810.11846`.

Kuansan Wang and S. Shamma. Self-normalization and noise-robustness in early auditory representations. *IEEE Transactions on Speech and Audio Processing*, 2(3):421–435, 1994. doi: 10.1109/89.294356.

Tianrui Wang, Xie Chen, Zhuo Chen, Shu Yu, and Weibin Zhu. An adapter based multi-label pre-training for speech separation and enhancement. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, June 2023. doi: 10.1109/ICASSP49357.2023.10094883.

Anne Warlaumont, Gert Westermann, Eugene Buder, and D. Kimbrough Oller. Prespeech motor learning in a neural network using reinforcement. *Neural Networks*, 38:64–75, 01 2013. doi: 10.1016/j.neunet.2012.11.012.

John R Westbury. Speech Production Database User ' S Handbook. *IEEE Personal Communications - IEEE Pers. Commun.*, 0(June), 1994a.

John R Westbury. Speech Production Database User ' S Handbook. *IEEE Personal Communications* -, 0(June), 1994b.

Gert Westerman and Eduardo Reck Miranda. Modelling the development of mirror neurons for auditory-motor integration. *Journal of New Music Research*, 31(4):367–375, 2002. doi: 10.1076/jnmr.31.4.367.14166. URL `https://www.tandfonline.com/doi/abs/10.1076/jnmr.31.4.367.14166`.

James R. Williamson, Daniel W. Bliss, David W. Browne, and Jaishree T. Narayanan. Seizure prediction using eeg spatiotemporal correlation structure. *Epilepsy & Behavior*, 25(2):230 – 238, 2012. ISSN 1525-5050. doi: https://doi.org/10.1016/j.yebeh.2012.07.007. URL http://www.sciencedirect.com/science/article/pii/S1525505012004763.

James R. Williamson, Thomas F. Quatieri, Brian S. Helfer, Rachelle Horwitz, Bea Yu, and Daryush D. Mehta. Vocal biomarkers of depression based on motor incoordination. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13, page 41–48, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450323956. doi: 10.1145/2512530.2512531. URL https://doi.org/10.1145/2512530.2512531.

James R. Williamson, Thomas F. Quatieri, Brian S. Helfer, Gregory Ciccarelli, and Daryush D. Mehta. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, page 65–72, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450331197. doi: 10.1145/2661806.2661809. URL https://doi.org/10.1145/2661806.2661809.

James R. Williamson, Diana Young, Andrew A. Nierenberg, James Niemi, Brian S. Helfer, and Thomas F. Quatieri. Tracking depression severity from audio and video based on speech articulatory coordination. *Computer Speech & Language*, 55:40 – 56, 2019. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2018.08.004.

Stephen M. Wilson, Ayşe Pinar Saygin, Martin I. Sereno, and Marco Iacoboni. Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, 7(7):701–702, Jul 2004. ISSN 1546-1726. doi: 10.1038/nn1263. URL https://doi.org/10.1038/nn1263.

D. Wolpert and Zoubin Ghahramani. Computational principles of movement neuroscience. *Nature Neuroscience*, 3 suppl. 1:1212–1217, 2000.

Alan A. Wrench. A Multichannel Articulatory Database and its Application for Automatic Speech Recognition. *Proceedings of 5th Seminar of Speech Production*, pages 305–308, 2000.

Peter Wu, Shinji Watanabe, Louis Goldstein, Alan W Black, and Gopala Krishna Anumanchipalli. Deep Speech Synthesis from Articulatory Representations. In *Proc. Interspeech 2022*, pages 779–783, 2022. doi: 10.21437/Interspeech.2022-10892.

Matthew John Yee-King, Leon Fedden, and Mark d'Inverno. Automatic programming of vst sound synthesizers using deep networks and other techniques. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):150–159, 2018. doi: 10.1109/TETCI.2017.2783885.

Y. Yoshikawa, J. Koga, M. Asada, and K. Hosoda. Primary vowel imitation between agents with different articulation parameters by parrot-like teaching. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453)*, volume 1, pages 149–154 vol.1, 2003. doi: 10.1109/IROS.2003.1250620.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021. doi: 10.1109/TKDE.2021.3070203.

Xinhui Zhou, Carol Y Espy-Wilson, Suzanne Boyce, Mark Tiede, Christy Holland, and Ann Choe. A magnetic resonance imaging-based articulatory and acoustic study of "retroflex" and "bunched" American English /r/. *The Journal of the Acoustical Society of America*, 123(6):4466–81, jun 2008. ISSN 1520-8524. doi: 10.1121/1.2902168. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2680662{&}tool=pmcentrez{&}rendertype=abstract.

Elizabeth C. Zsiga. Acoustic evidence for gestural overlap in consonant sequences. *Journal of Phonetics*, 22(2):121–140, 1994. ISSN 0095-4470. doi: https://doi.org/10.1016/S0095-4470(19)30189-5. URL https://www.sciencedirect.com/science/article/pii/S0095447019301895.

Elizabeth C. Zsiga and Rattima Nitisaroj. Perception of thai tones in citation form and connected speech. *The Journal of the Acoustical Society of America*, 116(4 Supplement):2628–2628, 10 2004. ISSN 0001-4966. doi: 10.1121/1.4785480. URL https://doi.org/10.1121/1.4785480.