Characterizing Dysarthric Speech with Transfer Learning

by

Michael Saxon

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2020 by the
Graduate Supervisory Committee:

Sethuraman Panchanathan, Co-Chair
Visar Berisha, Co-Chair
Hemanth Venkateswara

ARIZONA STATE UNIVERSITY

May 2020

ABSTRACT

Speech is known to serve as an early indicator of neurological decline, particularly in motor diseases. There is significant interest in developing automated, objective signal analytics that detect clinically-relevant changes and in evaluating these algorithms against the existing gold-standard: perceptual evaluation by trained speech and language pathologists. Hypernasality, the result of poor control of the velopharyngeal flap—the soft palate regulating airflow between the oral and nasal cavities—is one such speech symptom of interest, as precise velopharyngeal control is difficult to achieve under neuromuscular disorders. However, a host of co-modulating variables give hypernasal speech a complex and highly variable acoustic signature, making it difficult for skilled clinicians to assess and for automated systems to evaluate. Previous work in rating hypernasality from speech relies on either engineered features based on statistical signal processing or machine learning models trained end-to-end on clinical ratings of disordered speech examples. Engineered features often fail to capture the complex acoustic patterns associated with hypernasality, while end-to-end methods tend to overfit to the small datasets on which they are trained. In this thesis, I present a set of acoustic features, models, and strategies for characterizing hypernasality in dysarthric speech that split the difference between these two approaches, with the aim of capturing the complex perceptual character of hypernasality without overfitting to the small datasets available. The features are based on acoustic models trained on a large corpus of healthy speech, integrating expert knowledge to capture known perceptual characteristics of hypernasal speech. They are then used in relatively simple linear models to predict clinician hypernasality scores. These simple models are robust, generalizing across diseases and outperforming comprehensive set of baselines in accuracy and correlation. This novel approach represents a new state-of-the-art in objective hypernasality assessment.

ACKNOWLEDGEMENTS

Thank you Mom and Dad for instilling the ambition and confidence to succeed. I think you nailed pressuring me only slightly too much. I know that balance can be hard to pull off; I both deeply grateful for and impressed by your deft parenting.

Thanks to my many friends who have helped me along the way with your encouragement and support, your genuine camaraderie and occasional commiseration. This is not a comprehensive list, but thank you Abhik, Alex, Alexander, Calvin, Chris, Evan, Gabrielle, Gamal, Jacob, Jaylia, John, Lewis, Meredith, Pranav, Rohit, Sami, Samarth, Todd, and Vaish for the myriad ways you shaped my experience.

Thank you to the many faculty who played an instrumental role in my growth. Thank you Dr. Holman, Dr. Zhang, Dr. Chakrabarti, and Dr. Karam for particularly informative and impactful courses.

Thank you Dr. Yu, for introducing me to the joy of research as a high school student, and for mentoring projects from freshman year FURI to senior capstone. Thank you Dr. Liss for your guidance, kindness, and clutch paper edits.

Thank you Dr. Venkateswara, for always giving me time to ask questions about my vague machine learning ideas. Thank you Dr. McDaniel, for bringing me in to CUbiC, for teaching me the ropes of user studies, and for always being a kind and understanding mentor. Thank you Dr. Panchanathan for your mentorship, for making CUbiC a great environment, and for all the opportunities you sent my way.

Finally, thank you Dr. Berisha, for your impressive patience, for your measured and helpful criticism, and for the many ways you have helped me kickstart my career, from guiding early projects, giving me a sense of promising directions, and holding me accountable, to reviewing my GRFP essay drafts, talking through Ph.D. program options and career paths, sending me to conferences, and everything in between.

TABLE OF CONTENTS

Page

LIST OF TABLES v

LIST OF FIGURES vi

CHAPTER

1 INTRODUCTION 1

1.1 Existing Work 4

1.2 Thesis Outline 8

1.3 Previously Published Work 8

2 DATA 9

2.1 Healthy Speech Corpus 9

2.2 Dysarthric Speech Corpus 10

2.3 Cleft Palate Speech Corpus 12

2.4 Wisconsin Microbeam Dataset 12

3 HYPERNASALITY-CORRELATED FEATURES 14

3.1 Data Pre-Processing 15

3.2 Unvoiced Phoneme Articulatory Precision (AP) Features 15

3.2.1 Acoustic Model 16

3.2.2 Feature Computation 16

3.3 Plosive Nasal Cognate Distinctiveness (NCD) Features 17

3.3.1 Feature Computation 18

3.4 Voiced Phoneme Nasalization (N) Feature 19

3.5 Analysis 21

4 CLINICIAN HYPERNASALITY SCORE ESTIMATION 25

4.1 Linear Models 27

4.2 Baselines 28

CHAPTER Page

4.2.1 SSP: Formant Feature Models 28
4.2.2 ML: Deep Neural Networks 29
4.2.3 Human Baseline 31
4.3 Results 31
4.3.1 Individual Feature Contributions 35
4.4 Relationship Between Articulatory Precision and Hypernasality 36
4.5 Effectiveness of Forced Alignment 38
5 DISCUSSION AND CONCLUSIONS 40
5.1 Feature-Rating Correlations 40
5.2 Performance of Estimation Models 42
5.3 Conclusions and Summary 44
BIBLIOGRAPHY 47
APPENDIX
A PERMISSION STATEMENTS FROM CO-AUTHORS 54

LIST OF TABLES

Table Page

2.1 Clinical Characteristics of the 75 Speakers 11

4.1 Summary of NCD-Based Linear Hypernasality Prediction 32

4.2 Hypernasality Prediction Model Comparison Table 34

LIST OF FIGURES

Figure		Page
2.1	Hypernasality Score Distribution by Disease	11
3.1	High-Level Overview of the NCD Feature System	18
3.2	Selected NAP Features for Low and High Hypernasality Speakers	22
3.3	NCD Distribution Box Plots	23
3.4	Plots of the Most Prominent AP and N Features	24
4.1	High-Level Overview of the NAP-Based Hypernasality Prediction System	26
4.2	Predicting Hypernasality from NCD	32
4.3	Predicting Hypernasality with SSP, ML, and Novel Models	33
4.4	Marginal Improvement Plot for NAS-Based Hypernasality Prediction	35
4.5	Marginal Improvement Plot for Prediction with Clinician AP + NAS	36
4.6	Forced Alignment Error Plot	39

Chapter 1

INTRODUCTION

Neuromuscular disorders tend to present early, detectable symptoms in speech due to the precise fine motor control required to properly render phonemes and assemble words. Changes to speech can even be the only evidence of or the only significant impairment resulting from neurological disease (Duffy (2000)). Thus, speech analysis is a promising route to improve diagnostic speed and quality in a variety of brain diseases (Harel *et al.* (2004)). Typically, this analysis is performed perceptually, by trained speech language pathologists (SLPs), who provide opinion assessments on a variety of speech qualities that allow for systematic insights into the underlying production failures taking place. Such analysis performed over time can provide unique insights into the progression of a disease and drive treatment decisions. Unfortunately, high-fidelity tracking of symptom progression over time using the opinions of trained specialists is difficult for two main reasons. First, clinician perceptual assessment is costly and time-consuming. Second, some speech parameters are particularly challenging to consistently assess objectively, either due to low inter- or intra-rater reliability. Therein lies the motivation to develop better automated pathological speech diagnostic metrics.

Hypernasality is one such diagnostically relevant quality of speech that is particularly difficult for clinicians to assess. Hypernasality refers to the perception of excessive nasal resonance in speech, caused by velopharyngeal dysfunction (VPD), an inability to achieve proper closure of the velum, the soft palate regulating airflow between the oral and nasal cavities. It is a common symptom in motor-speech disorders such as Parkinson's Disease (PD) (Theodoros *et al.* (1995)), Huntington's Disease

(HD) (Novotny *et al.* (2016)), amyotrophic lateral sclerosis (ALS) (Duffy (1995)), and cerebellar ataxia (Poole *et al.* (2015)), as rapid movement of the velum requires very precise motor control. It is also the defining perceptual trait of cleft palate speech, (Kuehn and Moller (2000)). Reliable detection of hypernasality is useful in both rehabilitative (e.g. tracking the progress of speech therapy) and diagnostic (e.g. early detection of neurological diseases) settings, as demonstrated by Carrow *et al.* (1974), and Theodoros *et al.* (1993). Because of the promise hypernasality tracking shows for assessing neurological disease, there is interest in developing measurement strategies for it that are robust to the limitations of existing work.

Detecting and assessing hypernasality are complex tasks that require inferring the ratio of resonances across the pharyngeal, oral, and nasal cavities. A disproportionately high amount of nasal resonance is regarded as atypical and hypernasal. This presents a challenging estimation task, vulnerable to co-modulating variables including word choice, the particular geometry of an individual's resonating cavities, and other covarying dysarthria symptoms (e.g. vocal quality). This results in a highly nonlinear and complex mapping between the percept and the actual acoustic nasal resonance (Bettens *et al.* (2018), de Stadler and Hersh (2015)).

Current techniques for measuring velopharyngeal function in-clinic employ perception, imaging, and instrumentation. The current state of the art is SLP perceptual assessment (Kummer and Lee (1996)), however there is a growing body of work suggesting clinical perception is susceptible to the co-modulating variables mentioned above and listener expertise (S. Paal and Schuster (2005)). Reliable perceptual measures of hypernasality require evaluation from multiple clinicians (Scarmagnani *et al.* (2014)) or intensive training according to specific protocols (Brunnegard *et al.* (2012)). Some potential approaches to mitigating these shortcomings include direct imaging of the velopharyngeal closing mechanism using X-Ray or multiview videofluoroscopy

(Woo (2012)), or nasometry, analysis using a specialized head-mounted apparatus (Pentax (2016)). However, the imaging techniques are invasive and uncomfortable, and nasometry requires specialized equipment. Neither approach is scalable or common practice in-clinic.

An ideal machine hypernasality assessment technique would assess the symptom directly from the speech signal without using any specialized equipment, and accurately and objectively model the most consistent opinion scores of SLPs. Such a system would enable the remote tracking of neurological disease progression, for example through a smartphone application, greatly reducing the burden of regular testing for patients and cost for clinics.

Previous work toward assessing hypernasality directly from the speech signal can be categorized broadly in two groups: engineered features based on statistical signal processing (SSP) (Rafael Orozco Arroyave *et al.* (2012)) and supervised methods based on machine learning (ML) (Hegde *et al.* (2018)). The SSP approaches tend to be explainable and demonstrate some effectiveness in measuring hypernasality, but the complex spectral signature of nasalization is difficult to capture with a simple representation and there is a great deal of person-to-person variability (Lohmander and Olsson (2004)). However, the more complex ML-based metrics are fundamentally dependent on the small, disease-specific datasets on which they are trained. These metrics tend to generalize poorly across diseases; it is not clear if black-box models are fitting to the true perceptual qualities of hypernasality or to other co-modulating variables. Furthermore, collecting good clinicial speech data is costly and time-consuming, this scarcity of training data for means the more sophisticated ML methods are particularly prone to overfitting.

This thesis represents an approach to hypernasality assessment that falls between the SSP and ML groups, inspired by transfer learning. In short, all versions of this

approach involve training two separate models. First, a more robust acoustic model than can be achieved through traditional SSP hypernasality techniques is trained on a large corpus of healthy speech. This model can then be used to extract "correct phoneme/realized phoneme" likelihood ratios from the speech of a neurological patient reading a known passage aloud. These likelihood ratios are used as input to the simpler, second model, which estimates the clinician-rated hypernasality score. These novel hypernasality assessment systems are evaluated against a set of competing approaches, representing the best of the SSP and ML groups, as well as novel extensions on the newest in pre-trained deep speech representations.

This evaluation is performed on a dataset of 75 English speaking patients with Parkinson's disease, Huntington's disease, ALS, or cerebellar ataxia, in two cross validation conditions—leave one speaker out and leave one disease out—to assess both best-case performance and robustness to disease-specific confounders. In these testing conditions the models are compared using MAE and PCC to assess consistency across similar severity levels and trend capture reliability, as both of these qualities are necessary to fulfill the purposes of the ideal system described above.

Against all of these baselines, our novel set of assessment systems presented herein achieves state-of-the-art performance in SLP-rated hypernasality score prediction.

1.1 Existing Work

Clinician perceptual assessment is the gold-standard technique for assessing hypernasality (Extence and Cassidy (2017)). However, this method has been shown to be susceptible to a wide variety of error sources, including stimulus type, phonetic context, vocal quality, articulation patterns, and previous listener experience and expectations (Kent (1996)). Additionally, these perceptual metrics have been shown to erroneously overestimate severity on high vowels when compared with low vowels

(Kuehn and Moon (1998)), and vary based on broader phonetic context (Lintz and Sherman (1961)). Although these difficulties may be mitigated by averaging multiple clinician ratings, this further drives up costs associated with hypernasality assessment and makes its use as a trackable metric over time less feasible.

Various instrumentation-based hypernasality assessment systems have been proposed to mitigate these shortcomings in perceptual assessment, but have not managed to supplant SLP perception. These direct assessment techniques visualize the velopharyngeal closing mechanism using videofluoroscopy (Henningsson and Isberg (1991)) or magnetic resonance imaging (MRI) (Kao *et al.* (2008)) and provide information about velopharyngeal port size and shape (Bettens *et al.* (2014)). These methods are invasive and may cause pain and discomfort to the patients. As an alternative, nasometry seeks to measure *nasalence*, the modulation of the velopharyngeal opening area, by estimating the acoustic energy from the nasal cavity relative to the oral cavity. This is done by measuring the acoustic energy from two microphones separated by a plate that isolates the mouth from the nose (Pentax (2016)). In some cases, nasalance scores yield a modest correlation with perceptual judgment of hypernasality (Brancamp *et al.* (2010); Watterson *et al.* (1993)), however there is considerable evidence that this relationship depends on the person and the reading passages used during assessment (Watterson *et al.* (1993)), (Sinko *et al.* (2017)). Because of this, the clinician's perception of hypernasality is often the de-facto gold-standard in clinical practice (Chapman *et al.* (2016)). Furthermore, properly administering the evaluation requires significant training and it cannot be used to evaluate hypernasality from existing speech recordings.

Spectral analysis of speech is a potentially effective method to analyze hypernasality. Acoustic cues based on spectral flattening, amplitude reduction, and bandwidth increases that accompany nasalization (Tarun *et al.* (2007)), formant F1 and F2 am-

plitudes (Kozaki-Yamaguchi *et al.* (2005), Yu and Barkana (2009)), $1/3^{\text{rd}}$ octave band analysis (Kataoka *et al.* (2001)), spectral peak shifts (Hawkins and Stevens (1985)), the introduction of low-frequency resonances (Vijayalakshmi *et al.* (2007)), pole/zero pairs (Glass and Zue (1985), Vijayalakshmi *et al.* (2009)) and changes in the voice low tone/high tone ratio (Lee *et al.* (2006), Lee *et al.* (2009), Tsai *et al.* (2012)) have been proposed to detect or evaluate hypernasal speech. These spectral modifications in hypernasal speech will have an impact on articulatory dynamics, thereby affecting speech intelligibility. Statistical signal processing methods that seek to reverse these cues, such as suppressing the nasal formant peaks and then performing peak-valley enhancement, have demonstrated improvement in the perceptual qualities of cleft palate and lip-caused hypernasal speech (Vikram *et al.* (2016)), further demonstrating the connection between these cues and intelligibility. The large variability of speech degradation patterns across neurological disease or injury challenges simple features that are based on domain expertise (Orozco-Arroyave *et al.* (2015)). Overall, these simple features are not robust to the complicated acoustic patterns that emerge in hypernasality, and are prone to high false positive and negative error rates in out-of-domain test cases.

In response, data-derived representations of hypernasality that combine more elemental speech features and supervised learning have been proposed. Mel-frequency cepstral coefficients (MFCCs) and other spectral transformations (Rah *et al.* (2001), He *et al.* (2014), Orozco-Arroyave *et al.* (2015), Rendón *et al.* (2011), Nikitha *et al.* (2017), Dubey *et al.* (2016), Dubey *et al.* (2018a), Vogel *et al.* (2009), Kataoka *et al.* (1996)), glottal source related features (jitter and shimmer) (Castellanos *et al.* (2006), Dubey *et al.* (2018b)), difference between the low-pass and bandpass profile of the Teager Energy Operator (TEO) (Cairns *et al.* (1996)), Maier *et al.* (2008), and nonlinear features (Orozco-Arroyave *et al.* (2012), Orozco-Arroyave *et al.* (2013)) have all

been proposed as model input features. Gaussian mixture models (GMM), support vector machines, and deep neural networks have been used in conjunction with these features for hypernasality evaluation from word and sentence level data (Nieto *et al.* (2014), Golabbakhsh *et al.* (2017), Cairns *et al.* (1996)). Recently, end-to-end neural networks taking MFCC frames as input and producing hypernasality assessments as output have also been proposed (Vikram *et al.* (2018)).

These methods rely on supervised learning and are trained on small data sets. For this application they run the risk of overfitting to the data by focusing on associated disease-specific symptoms rather than the perceptual acoustic cues of hypernasality itself.

Features based on automatic speech recognition (ASR) acoustic models targeting articulatory precision have been used in nasality assessment systems Maier *et al.* (2008). A particularly important technique in articulatory precision assessment comes from the related area of accent analysis in computer aided language learning. This approach, called Goodness of Pronunciation (GoP) involves using the ASR acoustic model to produce a likelihood ratio (Witt (1999), Witt and Young (2000)). We refer back to the GoP in Chapter 3.

A promising approach toward developing portable, generalized neural representations of speech was proposed as the "problem-agnostic speech encoder" (PASE) by Pascual *et al.* (2019). In this work a neural model is pretrained on a series of tasks such as speaker identification and speech recognition in a transfer learning framework similar to the one deployed in the production of word embeddings and generalized pretrained image recognition models. This work will motivate parts of Chapters 3 and 4.

1.2 Thesis Outline

The thesis is structured as follows.

Chapter 2 introduces the data used for training and evaluating the models presented afterward, in particular the large healthy ASR speech corpus LibriSpeech, the novel 75-speaker pathological speech corpus, a small 10-speaker evaluation cleft palate evaluation corpus, and the Wisconsin Microbeam articulatory inversion corpus which is used to supervise one of the competing neural models in Chapter 4.

Chapter 3 is about building features that are well-correlated with hypernasality using acoustic modeling, presents the nasalization + articulatory precision (NAP) and nasal cognate distinctiveness (NCD) families of features, and discusses the correlation between these novel features and clinician hypernasality rating.

Chapter 4 brings the introduced features together as input to clinician hypernasality score predicting models, describes the baselines in more detail, and presents the results of the comparison between the novel systems and the baselines.

Chapter 5 concludes the thesis with a summary of the findings and discussion of future research directions.

1.3 Previously Published Work

The contents of Chapters (2) and (3) include material adapted from previously published work, "Objective Measures of Plosive Nasalization in Hypernasal Speech," Saxon *et al.* (2019). Material from Chapters (2-5) has been publicly released as a preprint, "Robust Estimation of Hypernasality in Dysarthria with Acoustic Model Likelihood Features," Saxon *et al.* (2020), and is currently in peer review.

Chapter 2

DATA

This thesis focuses on an approach to modeling hypernasality that falls between the two extremes of simple hand-engineered statistical signal processing (SSP) features and sophisticated supervised machine learning (ML) models. To achieve this, more sophisticated perceptual modeling features are trained on abundantly available healthy speech, and the disordered speech is saved for training simpler models atop the feature model representations.

This approach requires multiple corpora of data: a large healthy corpus to train the perceptual models, and a pathological speech corpus collected from neurologically disordered individuals with clinician-assessed hypernasality severity labels. For these two purposes the publicly available LibriSpeech dataset and a clinical dataset collected by Profs. Liss and Berisha are used, respectively.

For auxiliary tasks, two other datasets are used. First, a publicly available dataset cleft palate speech (CLP) is used to validate the hypernasality modeling features, as individuals with cleft palate exhibit extreme hypernasality but otherwise healthy and normal speech. Finally, to fine-tune existing pre-trained generalized neural speech representations the publicly available Wisconsin Microbeam dataset is used to prove an articulatory inversion supervising task.

2.1 Healthy Speech Corpus

LibriSpeech is a public domain corpus of transcript-labelled healthy English utterances. It contains roughly 1000 hours of speech sampled at 16kHz. The speech consists of 1,128 female and 1,210 male speakers reading book passages aloud. It

contains "clean" samples, which have been carefully segmented and aligned, as well as "other" samples which are more challenging to use Panayotov *et al.* (2015). It is freely available for download at `openslr.org`. This corpus was employed in training all three acoustic models presented in Chapter 3.

2.2 Dysarthric Speech Corpus

The database consists of recordings from 75 speakers (40 male and 35 female) of varying levels of hypernasality. The corpus contains data from speakers diagnosed with several different neurological disorders: 38 patients have Parkinson's disease (PD), 6 patients have Huntington's disease (HD), 16 patients have cerebellar Ataxia (A), and 15 patients have amyotrophic lateral sclerosis (ALS).

All individuals read the same set of five sentences, capturing a range of phonemes. Reading is an ideal stimulus for this task because it controls for phonetic distributional variations that would be present in more spontaneous speech and enables for consistency between speakers and between assessments in-time, ideal qualities for a clinical measure.

The perceptual evaluation of hypernasality from recorded samples was carried out by 14 different speech language pathologists on a scale of 1 to 7. The average hypernasality score for each speaker was used as the ground truth. The inter-rater reliability of the SLPs was moderate, with a Pearson Correlation Coefficient of 0.66 and an average inter-clinician mean absolute error of 1.44 on the 7-point scale. The sentences spoken were:

1. The supermarket chain shut down because of poor management.
2. Much more money must be donated to make this department succeed.
3. In this famous coffee shop they serve the best doughnuts in town.

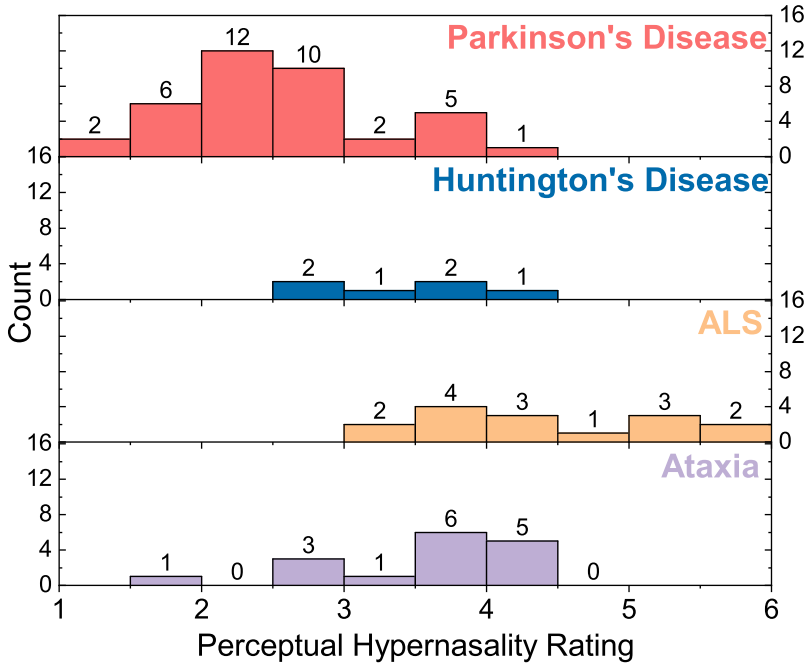


Figure 2.1: Distribution of clinician-rated hypernasality score by disorder.

4. The chairman decided to pave over the shopping center garden.

5. The standards committee met this afternoon in an open meeting.

The speech recordings were carried out in sound treated room using a microphone. Table 2.1 shows the breakdown of clinical characteristics of the subjects and the statistics of the nasality score (NS) subsets. S.D. denotes standard deviation. Figure 2.1 contains the clinician hypernasality score histograms for each disorder population.

Table 2.1: Clinical characteristics and nasality scores of the subjects.

Disease	Male	Female	Mean Age	S.D. Age	Mean NS	S.D. NS
PD	20	18	71.06	9.62	2.55	0.75
A	6	10	62.47	14.05	3.58	0.68
ALS	8	7	59.54	13.23	4.41	0.85
HD	6	0	58.40	13.20	3.31	0.59
Total	40	35	65.80	12.67	3.20	1.04

2.3 Cleft Palate Speech Corpus

Cleft lip and palate (CLP) also gives rise to hypernasal speech. However, unlike individuals with neuromuscular disease, CLP speakers produce otherwise healthy speech without the other perceptual changes (slurring, generalized articulatory imprecision) that also arise in dysarthria (Kuehn and Moller (2000)). Thus, CLP speech is invaluable in validating that a model trained on pathological speech is indeed capturing the perceptual qualities of hypernasality.

I use a corpus of 6 child and 12 adult CLP speakers with different levels of hypernasality severity, that span the hypernasality range (from normal to extreme) in equal intervals (Kuehn *et al.* (2002)) to demonstrate that our model chiefly captures hypernasality rather than any associated neurologically disorded speech symptoms. These CLP speakers are otherwise healthy and exhibit no other co-modulating symptoms such as imprecise articulation resulting from other motor impairments. Because the hypernasality assessments for these speakers were performed by different clinicians than our dysarthric data, I focus on correlation alone to evaluate the performance of our hypernasality evaluation system on this speech.

2.4 Wisconsin Microbeam Dataset

One of the baseline neural feature models, the problem agnostic speech encoder (PASE) (Pascual *et al.* (2019)) is trained using a series of "worker tasks" to learn generalized speech representations to enable simple models trained on downstream tasks. Among the tasks not included in PASE is "articulatory inversion," the problem of inferring the time series of articulator positions (tongue, lips, jaw, etc) from the speech signal they produced. Shah *et al.* (2019) demonstrate how articulatory inversion targets can provide useful constraints that improve the performance of models

on seemingly unrelated tasks such as spoken emotion recognition. As neurologically pathological speech fundamentally is an inability to adequately control the articulators, I was interested in using articulatory inversion to attempt to similarly improve the performance of PASE as an input to hypernasality estimation. To train this auxilliary task the Wisconsin Microbeam Dataset was used.

The Wisconsin Microbeam Dataset consists of the recorded speech of 57 native speakers of American English, 32 female and 25 male, as defined in Westbury (1994). Each speaker recorded a variety of passages, phonations, and words while being measured by an X-Ray microbeam recorder, an apparatus that tracks a set of pellets glued to the articulators in their mouths to provide real-time measurement of the articulator motion. Because the muscular control issues resultant from speech disorders are fundamentally a lack of control of the articulators, articulatory inversion, the estimation of this articulator data from the speech signal, is a sensible task for supervision of generalized models. I use this dataset in Chapter 4 to provide an articulatory inversion task for a model estimating nine "tract variables" (Sivaraman *et al.* (2019)) that define the vocal tract shape formed by the articulators in a relatively speaker-invariant manner.

Chapter 3

HYPERNASALITY-CORRELATED FEATURES

This chapter will introduce the two families of acoustic model features that represent the first half of the "train features on healthy speech, models on disordered speech" framework that drives the thesis. The feature families are

1. Nasal cognate distinctiveness (NCD), an acoustic model likelihood ratio inspired by GoP that models the clinician observation of "nasal cognates pairs" plosives that are co-located with a nasal sonorant.

2. Nasalization and Articulatory precision (NAP), a set of features that analyze voiced phonemes using a perceptual modeling "nasalization feature" (N), and unvoiced phonemes using an "articulatory precision feature" (AP), goodness of pronunciation (Witt (1999)).

The three features (NCD, N, AP) contained in these feature families are all evaluated on a phoneme-by-phoneme basis. When used in a clinical setting, patients would be asked to read from a phonetically rich script. Then, a forced alignment process matches the phoneme timestamps from the speech audio to the sequence of phonemes that compose the transcript (in this case, I use Viterbi decoding with ASR models). These transcript-derived phonemes compose the ground truth against which the likelihood of some test phoneme or phoneme class is assessed. The specifics of these feature computations follow.

Each feature is specific to a class of phonemes. NCD is only assessed on the plosives T, D, P, B, K, and G, for which the "nasal cognates" N, M, and NG are perceptibly substituted when a speaker is unable to appropriately close the velum. For all other

phonemes, NCD cannot be assessed. The NAP family is assessed on all phonemes, which are divided into the voiced and unvoiced classes. Within the unvoiced class, AP is assessed, because for unvoiced sounds, hypernasal speech results in imprecise consonant production—the characteristic insufficient closure of the velopharyngeal port renders the speaker unable to build sufficient pressure in the oral cavity to properly form plosives and fricatives, causing the air to instead leak out through the nose (Woo (2012)). N is assessed on the voiced phonemes, evaluating a likelihood ratio for whether the phoneme in question belongs to the "nasal" or "oral" class. Since hypernasal speech results in perceptible extra resonances at the lower frequencies in voiced sounds (Kummer and Lee (1996)), an incorrect high likelihood for "nasal" in a non-nasal voiced phoneme would be evidence of hypernasality.

3.1 Data Pre-Processing

Consider an utterance $x(t)$ with sampling rate F_s and a corresponding transcript of phonemes p_j, $\{p_1, p_2, \ldots p_{N_p}\}$. $x(t)$ is analyzed with a 20ms frame length and 10ms overlap. For a frame indexed by i, $x_i(t)$, extract a set of features, $\mathbf{x}_i$. The utterance $x(t)$ is force-aligned using the Montreal Forced Aligner[1] at the phoneme level (McAuliffe *et al.* (2019)). I denote the data feature matrix for all frames that are aligned to phoneme p_j by $\mathbf{X}^{p_j}$.

3.2 Unvoiced Phoneme Articulatory Precision (AP) Features

Although originally designed to aid second language learners as a component in computer aided language learning software, the Goodness of Pronunciation (GoP) articulatory precision (AP) feature has found use in the analysis of disordered speech Pellegrini *et al.* (2014). The implementation of GoP used in my work is based on an

[1]Section 4.5 addresses forced alignment performance on these dysarthric speech samples.

acoustic model trained using Kaldi as specified in Tu *et al.* (2018).

3.2.1 Acoustic Model

To train the acoustic model, I extract a set of observation feature vectors from each training speech sample. The input speech sampling rate is 16 kHz. I analyze the speech at a frame rate of 10 ms and denote the acoustic features for frame i by O_i. For our implementation I used a triphone model trained with a Gaussian Mixture Model-Hidden Markov Model on 960 hours of healthy native English speech data from the LibriSpeech corpus Panayotov *et al.* (2015). I use the Kaldi toolkit training scripts for training the model. The input features to the ASR model are 39-dimensional second order Mel-Frequency Cepstral Coefficient (MFCC) with utterance-level cepstral mean variance normalization and Linear Discriminant Analysis transformation (same approach as in Tu *et al.* (2018)).

3.2.2 Feature Computation

After training, the acoustic model can be queried using the Viterbi decoding algorithm for the posterior probability $P(\mathbf{X}|q)$ of a given set of acoustic feature frames $\mathbf{X}$ representing a realization of some ground-truth transcript-assessed phoneme q. For a "well-articulated" phoneme, no phoneme apart from the one intended by the speaker should maximize this posterior.

I use the acoustic model to assess articulatory precision as follows. Considering the set of phonemes Q in the language, I assess the log-likelihood ratio of the frames $\mathbf{X}^{p_j}$ from a given phoneme p_j, to the maximum log-likelihood across all phonemes,

$$AP(p_j) = \log\Big(\frac{P(\mathbf{X}^{p_j}|p_j)}{\max_{q\in Q} P(\mathbf{X}^{p_j}|q)}\Big)/|\mathbf{X}^{p_j}|, \tag{3.1}$$

where $|\mathbf{X}^{p_j}|$ represents the number of acoustic frames aligned to phoneme p_j.

This processing is performed after forced alignment to the transcript labels, and assessed for each unvoiced phoneme to permit by-phoneme analysis of precise articulation within the NAP model.

3.3 Plosive Nasal Cognate Distinctiveness (NCD) Features

A characteristic of hypernasal speech is the unintentional production of "nasal cognates," nasal sonorants sharing the same place of articulation as certain voiced plosives, when production of the corresponding plosive is intended. This transformation means that the voiced alveolar stop D will sound like the alveolar nasal N, the voiced bilabial stop B will sound like the bilabial nasal M, and the voiced velar stop G will sound like the velar nasal NG (Shriberg and Kent (1982)). Similarly, the unvoiced counterparts of these stops T, P, and K frequently are present in phonetic environments where they are proceeded or followed by vowels, or proceeded by nasal consonants (Giegerich (1992)), which means they also can share a propensity to be mapped to the same nasal cognates (Shriberg and Kent (1982)). Predictable phenomena such as this suggest that perceptually-motivated, phoneme-level objective measures of hypernasality are warranted.

The Goodness of Pronunciation algorithm discussed above assesses the pronunciation of a speaker on a phoneme-by-phoneme basis as the log ratio of the probability of the uttered phoneme segment given the correct phoneme from an aligned transcript to the maximum across all phonemes of the uttered segment given a phoneme. I substitute the max denominator term in GoP with a nasal cognate mapping function, so instead of looking for a "worst case" likelihood ratio that will reduce generally when a phoneme is poorly realized, NCD will instead specifically look for the degradation that uniquely occurs under velopharyngeal dysfunction.

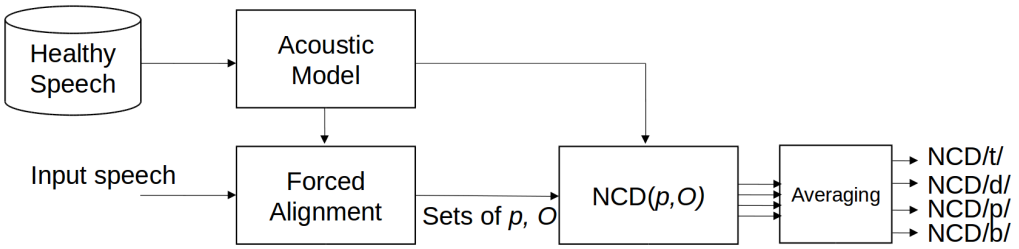


Figure 3.1: High-level overview of the NCD feature system.

3.3.1 Feature Computation

Fig. 3.1 is an overview of the NCD approach. I assume that I have an input speech segment and corresponding transcript for analysis. Furthermore, I assume that the input utterances have several instances of the phonemes of interest (P, B, T, D). Similar to the Goodness of Pronunciation feature, the Nasal Cognate Distinctiveness feature computation begins with an ASR acoustic model trained on healthy speech, as explained in Section 3.2.1. This acoustic model is used to both force-align the speech to the transcript to sample the plosives and estimate the likelihood ratios between the plosives and their nasal cognates with which the NCD features are computed. Finally, the individual instances of each phoneme are averaged to generate average NCD features.[2]

NCD is formulated for phoneme $p \in S = [\text{T,D,P,B,K,G}]$, frame $O_i \in O$, the observation corresponding with p based on forced alignment to the transcript,

$$NCD(p, O) = \Sigma_i \log\Big(\frac{P(O_i|p)}{P(O_i|\text{cog}(p))}\Big)/|O| \tag{3.2}$$

where $\text{cog}(p)$ is a "cognate mapping function" that maps the stops in the set S to their corresponding nasal cognate, and $|O|$ is the total number of frames in observation O.

[2]Code is available at `https://github.com/michaelsaxon/ncd`

The probabilities in the numerator and denominator of the formula are assessed using the Viterbi alignments in the ASR model. To assess the denominator probability the cog(p) function is called first, swapping the given plosive with its cognate in the triphone context.

Given a set of recordings of a speaker reading from a set of transcripts, the four NCD features are evaluated as follows. First, the transcripts are force-aligned at the phoneme level using the ASR model. With this alignment the $NCD(p)$ feature can be computed for each all phonemes $p \in S$ in the input utterances. This produces a set of many output NCD values, with each corresponding to an occurrence of one of the four phonemes in consideration in the transcripts. The NCD values are then averaged across all occurrences of the corresponding phonemes to return the output features, NCD for T, D, P, B, K, and G.

3.4 Voiced Phoneme Nasalization (N) Feature

The acoustic nasalization model is trained using healthy speech data from the LibriSpeech dataset. Two classes distributions of voiced phonemes are modeled. The "oral" non-nasal (ORL) class consists of all voiced oral consonants and all vowels from syllables where nasal consonants are not present. Similarly, the "nasal" class (NAS) is defined to contain the nasal consonants as well as half of adjacent vowels surrounding them. These rules were implemented after alignment; an illustrative example of the two classes is shown in the third tier of the aligned example in Fig. 4.1.

For this task, 100 hours of clean-labeled speech from the LibriSpeech dataset are used. First, forced phone-alignment to the transcript is performed as shown in Figure 4.1. I partition all phonemes into the NAS and ORL classes. For each frame in each phoneme, I extract 13 PLP coefficients, giving two feature matrices, $\mathbf{X}^{NAS}$

and $\mathbf{X}^{ORL}$, containing all frames of nasal PLPs in one, and non-nasal PLPs in the other. PLP features were chosen rather than MFCCs because they preserve acoustic cues that have been previously used to model hypernasality, including formant frequencies, bandwidths, and spectral tilt (Hermansky (1990)). To model the probability density functions, I use a 16-mixture Gaussian Mixture Model (GMM). The weight, mean, and covariance matrix for each of the GMM components is learned using the expectation maximization (EM) algorithm. The GMM for the nasal class is represented by $\lambda_{NAS} = \{\mu_{NAS}, \Sigma_{NAS}, \omega_{NAS}\}$, $i = 1, 2, ...16$. Here, μ_{NAS}, Σ_{NAS} and ω_{NAS} represent the mean, covariance matrix and weight of the i^{th} Gaussian, respectively. Similarly, for the non-nasal class the GMM components are given by $\lambda_{ORL} = \{\mu_{ORL}, \Sigma_{ORL}, \omega_{ORL}\}$, $i = 1, 2, ...16$.

After training on healthy speech, I provide a segmented dysarthric utterance to evaluate the likelihood from each of the two learned probability density functions. For an out-of-sample input, I estimate the likelihood, voiced phoneme by voiced phoneme. That is, for data feature matrix $\mathbf{X}^{p_j}$, the likelihood that this phoneme is nasalized is

$$f(\mathbf{X}^{p_j}|\lambda_{NAS}) = \prod_{i \in p_j} f(\mathbf{x}_i|\lambda_{NAS}), \tag{3.3}$$

where the notation $i \in p_j$ is shorthand notation for all 20ms frames aligned to phoneme p_j. Similarly for the ORL class, I have

$$f(\mathbf{X}^{p_j}|\lambda_{ORL}) = \prod_{i \in p_j} f(\mathbf{x}_i|\lambda_{ORL}). \tag{3.4}$$

I use the log-likelihood ratio test statistic as a continuous measure of nasalization. In particular, I define

$$N(p_j) = \log\left(\frac{f(\mathbf{X}^{p_j}|\lambda_{NAS})}{f(\mathbf{X}^{p_j}|\lambda_{ORL})}\right) / |\mathbf{X}^{p_j}|, \tag{3.5}$$

where $|\mathbf{X}^{p_j}|$ represents the number of acoustic frames aligned to phoneme p_j. This statistic is calculated for every voiced, non-nasal phoneme in the input utterance.

Thus, for a given speaker, a nasalization ratio is computed containing the log-likelihood ratios of nasalization of the voiced phonemes, (AA, AE, AH, AO, AW, AY, B, D, DH, EH, ER, EY, G, IY, JH, V, Z).

3.5 Analysis

For non-nasalized speech, the value of $N(\mathbf{X}^{p_j})$ should be low, whereas for nasalized speech, it should be high. For speakers who exhibit little hypernasality, $AP(\mathbf{X}^{p_j})$ should be high, whereas for hypernasal speakers, it should be lower. Figure 3.2 shows a comparison of the values of select NAP likelihood ratio features between a group of high hypernasality (> 4 perceptual rating) and low hypernasality (< 3 Perceptual rating) speakers. I average the hypernasality scores for the 4 most relevant voiced phonemes and 2 most relevant unvoiced phonemes for predicting hypernasality, as assessed in Section 4.3.1. As expected, there is an increase in the nasalization feature value and a decrease in the articulatory precision feature value corresponding to an increase in severity of hypernasality. Furthermore, I expect hypernasality to exhibit unique patterns in terms of affected and unaffected unvoiced phonemes, that are not general to dysarthria Saxon *et al.* (2019), making phoneme-level AP classification a valuable signal in quantifying hypernasality.

Figure 3.3 contains box plots for the four phoneme NCD features. The speakers were divided into four groups based on nasality severity for this analysis: control, mild, moderate, and severe. To perform the separation the real range of non-control assessed nasality was divided roughly in three, with the mild nasality $N \in [1.3, 2.7)$, moderate $N \in [2.7, 4.1)$ and severe $N \in [4.1, 5.6]$.

The feature trends very convincingly move for the voiceless phonemes T and P, with the control and mild nasality speakers exhibiting the highest values of Nasal Cognate Distinctness. The moderate nasality speakers then exhibit lower feature

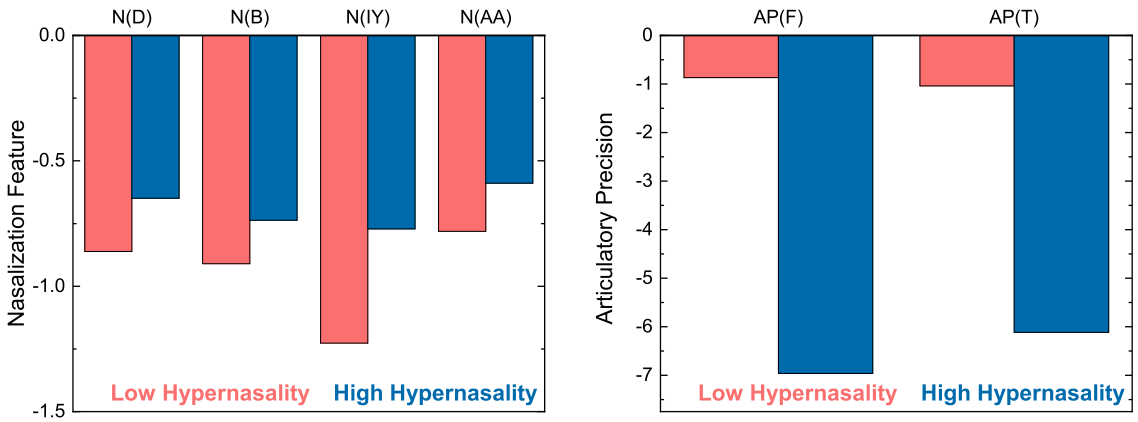
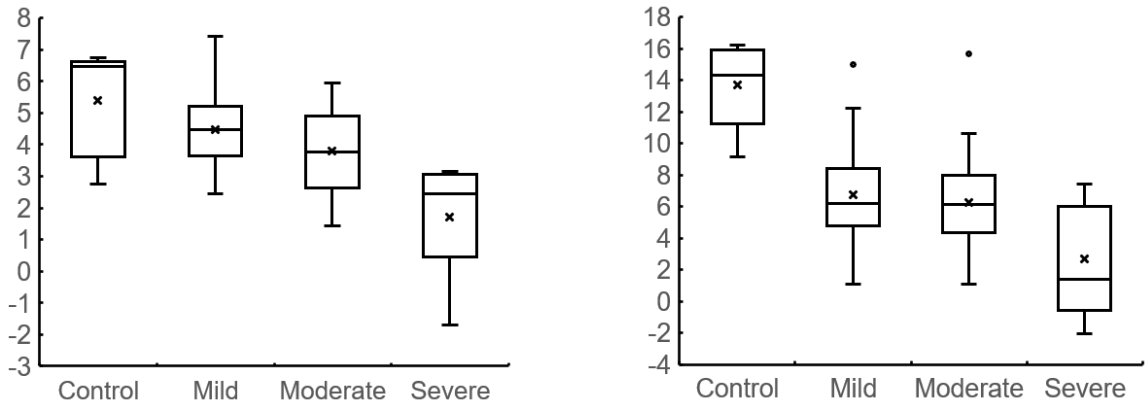


Figure 3.2: Bar charts of (left) N(D, B, IY, AA), and (right) AP(F, T) for low hypernasality and high hypernasality speakers.

values and the severe nasality speakers exhibit the lowest. The means, medians, and quartiles for all of the values decrease as nasality increases across groups. These expected trends are not all exhibited in the voiced phonemes D and B, however. For both phonemes the means, medians, and quartiles hardly move at all or do not move together between the mild and moderate nasality groups. For B, the moderate nasality NCD feature range even spans the entire range of values exhibited by all other groups. Despite these inconsistencies, for all phonemes the NCD score completely separates the control range from the severe nasality range.

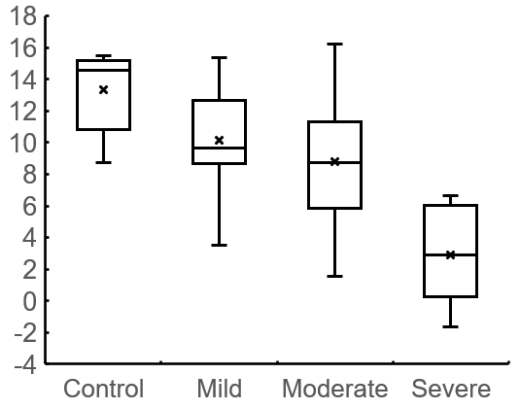
Figure 3.4 contains plots of various AP, N, and NCD features against the mean clinician hypernasality rating for all speakers. All demonstrate the expected trends (NCD and AP decrease with increasing hypernasality, N increases with increasing hypernasality), with linear correlation magnitude $|\text{PCC}| > 0.25$ The NAP features used in the figure are the same as selected from the forward feature selection process in Section 4.3.1, while the two NCD features plotted were chosen randomly.

Taken together, these plots demonstrate that these likelihood ratio features, AP, N, and NCD, assessed using acoustic models trained exclusively on healthy speech data, clearly capture the perceptual trends underlying hypernasality in speech.

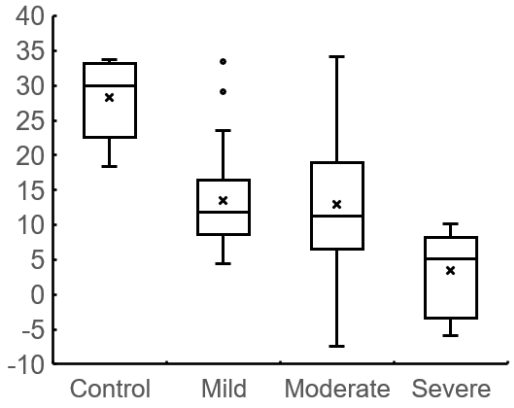


(a) NCD(T)

(b) NCD(D)



(c) NCD(P)



(d) NCD(B)

Figure 3.3: Box plots for the NCD feature distribution separated by nasality severity. The y-axis in each plot represents the NCD feature value for the phoneme under consideration.

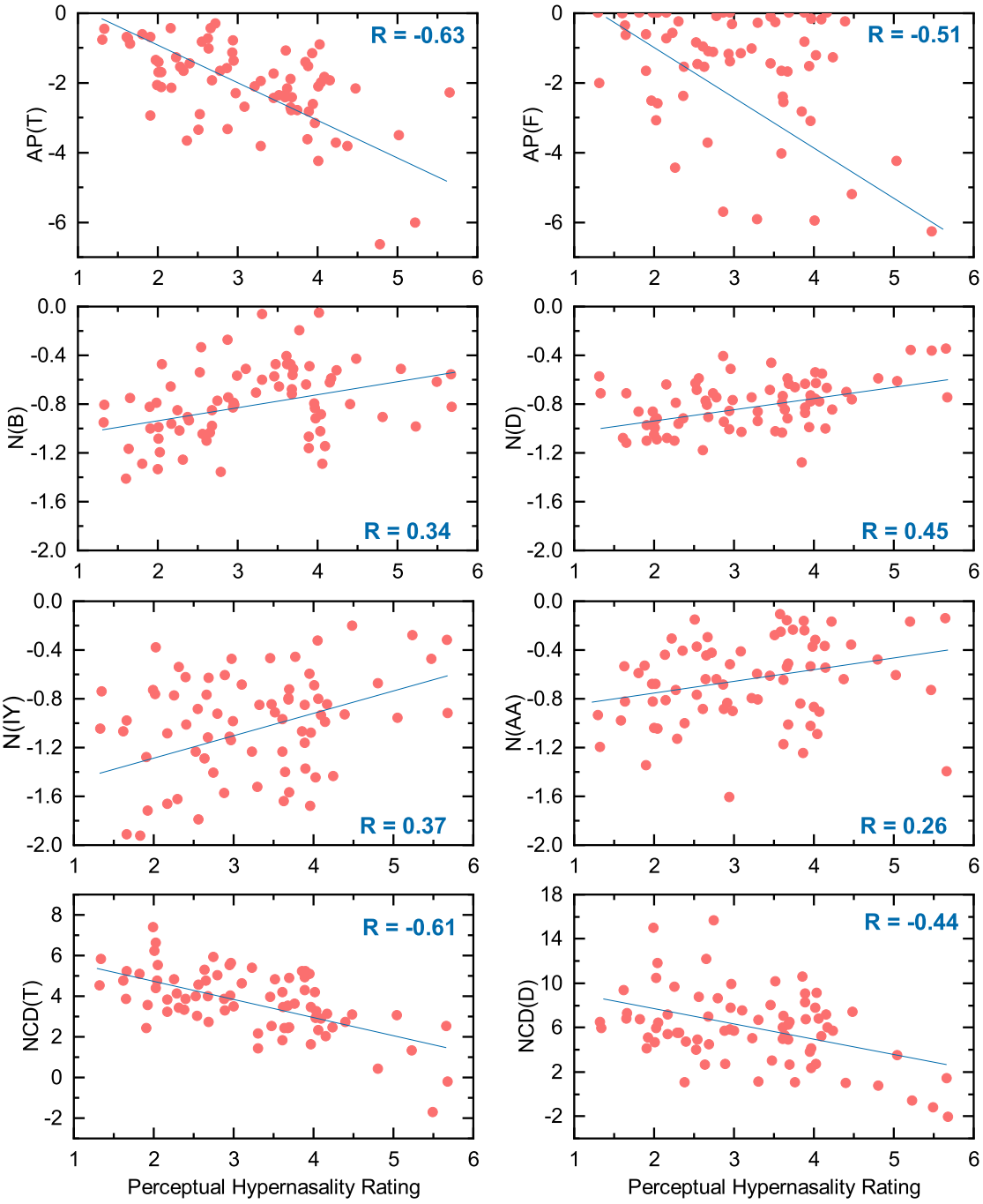


Figure 3.4: Plots of the two most prominent articulatory precision features (AP(T) and AP(F)), four of the most prominent nasalization features (N(D), N(B), N(IY), and N(AA)), and two nasal cognate distinctiveness features (NCD(T) and NCD(D)) against clinician-assessed hypernasality score.

Chapter 4

CLINICIAN HYPERNASALITY SCORE ESTIMATION

The central goal of this thesis is the prediction of ground-truth averaged clinician perceptual hypernasality scores from raw speech audio. This chapter presents the process for training simple models on the features presented in Chapter 3 as extracted from hypernasal speech, as well as the competing statistical signal processing (SSP) and machine learning (ML) baselines, and the final evaluation results.

All models considered in this chapter are evaluated on L1 loss (MAE) and Pearson's correlation coefficient (PCC) between the ground-truth clinician hypernasality scores and the model predictions. These metrics are assessed in across two different cross-validation strategies. Leave-one-speaker-out (LOSO) cross-validation provides a best-case look at how a given feature-model combination will generalize to a new patient, as is typically done in dysarthric speech processing studies. However, in light of the limitations of previous approaches, and the complexity of hypernasality modeling given the many co-modulating variables described in Chapter 1, the generalization across out-of-domain diseases is assessed using leave-one-disease-out (LODO) cross-validation, a condition in which a test model is trained on three of the speech disorder sets and evaluated on the fourth. In LODO, data for three of the neurological conditions is used for training and the fourth is used for testing.

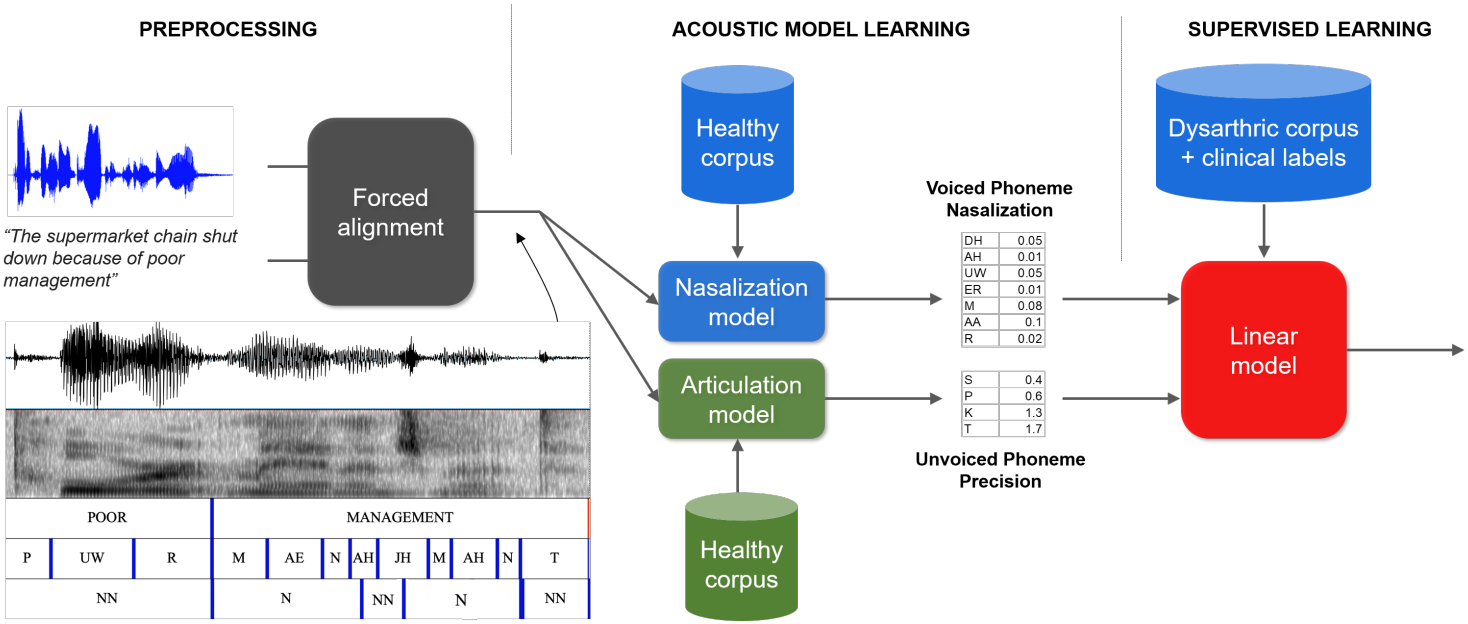


Figure 4.1: A high-level diagram of the NAP method. The leftmost pre-proccessing segment depicts the forced alignment of transcript to audio as well as the aligned word-phoneme-nasal class segmentation of the speech signal and spectrogram.

4.1 Linear Models

The nasal cognate distinctiveness (NCD) and nasalization + articulatory precision (NAP) feature families are both intended for use on specific subsets of phonemes; NCD features may only be extracted from stops for which there are nasal cognates, while NAP partitions all non-nasal phonemes based on voicing to be processed by the nasalization (N) or articulatory precision (AP) acoustic model likelihood ratio feature extractor. In a comprehensive hypernasality rating system, a collection of these features are extracted for a subject's speech samples and then fed as input into a linear model, predicting the hypernasality score. In Figure 4.1, I provide a high-level overview of the proposed NAP-based hypernasality score estimation scheme.

In the interest of generalization and clinical interpretability, simple linear ridge regression models are used to estimate the nasality score using the phoneme-averaged NAP or NCD features as input.

Table 4.2 contains three linear models trained on NCD features, NAP features, or both. "NCD-Lin" uses the four NCD features for T, D, P, and B. "NAP-Lin" uses the full set of N and AP features for the voiced and unvoiced phonemes. "NAP+NCD" uses

Finally, "NAP+NCD" represents the best possible combination of all NAP and NCD features available. I expected some of the likelihood features are more salient to hypernasality prediction than others. To minimize the risk of overfitting, only a most-salient subset of the NAP or NCD features are used in this evaluation. Greedy correlation-based forward feature selection Hall (1999) was employed to choose which features to use, yielding the set of AP(T, S, F, SH), N(IY, D, AA), and NCD(T, G).

4.2 Baselines

To properly evaluate the novel approach, I implemented a set of baselines meant to represent both sides of the spectrum of automated hypernasality approaches. Broadly, this means a variety of traditional machine learning models ingesting a set of spectral formant-based features to represent SSP, and a variety of neural network architectures processing the raw audio or spectrograms to represent ML. Taken together, these can be considered to constitute the state of the art in engineered features and in supervised learning for hypernasality estimation.

4.2.1 SSP: Formant Feature Models

Styler (2015) presents a set of formant-based features that are maximally effective in and correlated with hypernasality. Formants are the spectral peaks in voiced speech that characterize vowel sounds, and are numbered Fx in an increasing series of overtones starting from the fundamental frequency of the phonation, $F0$.

The formant features (FF) used in this analysis included $F1$ formant amplitude, $P0$ nasality peak amplitude, and normalized and raw $A1-P0$ difference Chen (1997). The FFs were extracted using Praat source code provided by Styler (2013). All features were extracted for each vowel and used in a linear and non-linear model to estimate the clinician-assessed hypernasality labels.

The linear model is based on simple multiple regression whereas the non-linear models are based on additive regression and k-nearest neighbor regression. The results of this model are labeled FF-Lin, FF-Add, and FF-KNN in Table 4.2.

4.2.2 ML: Deep Neural Networks

Several neural network baselines were considered. First, I implemented the neural network proposed in Vikram *et al.* (2018), one of the first works exploring using neural networks in hypernasality estimation. The model consisting of three feed-forward layers with sigmoid activations. This network consumes a time-series of 39-dimensional Mel Frequency Cepstral Coefficients (MFCC), extracted from 20 ms windows with no overlap. The hidden layers are of size 100, and the output layer of size 1. The output value is averaged across all frames to provide a single nasality score estimation per speaker. The model is trained using L1 loss and the Adam optimizer (Kingma and Ba (2014)) for 50 epochs with a learning rate of 0.001. In Table 4.2 this model is referred to as MFCC-NN.

Because MFCCs are a very fundamental spectral feature, I was concerned that the 75 speaker nasality dataset was insufficiently large for good representations to be learned by the model. To alleviate this, the second neural approach employed the problem agnostic speech encoder (PASE) from Pascual *et al.* (2019). The hope that this neural speech encoder, pretrained on several supervising tasks including ASR and emotion recognition, would contain richer speech representations than MFCC frames with which MFCC-NN was trained. The PASE encodigns were fed through three feed-forward layers with ReLU activations, followed by a single LSTM layer, all with hidden size 250. The model is trained using L1 loss and the Adam optimizer for 50 epochs with a learning rate of 0.0001. After max-pooling in time, a final feed-forward layer projects the latent codes to the final hypernasality score estimation. The performance of this model was assessed both with the PASE encoding layers frozen as static feature extractors for the entire training process (PASE-NN in Table 4.2) and with unfreezing of the PASE encoder for fine-tuning after the 10th epoch

(PFT-NN in Table 4.2).

The third and final baseline neural architecture that was employed builds on the transfer-learning approach established in Pascual *et al.* (2019) with the supervisory task of articulatory inversion, a distinct task from any of the "worker tasks" used to train the PASE encoder, in which the underlying time-series of articulator positions (tongue, jaw, lips) is inferred from the speech signal. A neural network based on the architecture proposed by Sivaraman *et al.* (2019) is first trained on the Wisconsin Microbeam Dataset (Section 2.4, Westbury (1994)), and then adapted to the task of hypernasality estimation by replacing the output projection.

For the articulatory inversion task, PASE encodings are once again used as the input feature. The PASE encodings are first consumed by a single 1D convolutional layer in time with a filter size of 50 PASE frames. Each PASE frame is roughly 6.9 ms long, thus this convolution gives each sample in the resultant sequence a receptive field of roughly 340 ms. The output of this convolutional layer is then fed through four feed-forward layers of size 250, followed by Layer Normalization (Ba *et al.* (2016)) and a projection to size 9 to provide a time-sequence of tract variable (TV) positions corresponding to the articulators. Finally, to enforce the physiological constraints inherent to articulator motion, 1D Gaussian filtering in time is performed on the output sequence predictions, with a kernel of filter length 120 frames (832 ms) and sigma of 40 (274 ms). The model is trained using L1 loss, the Adam optimizer for 50 epochs with a learning rate of 0.0015. The articulatory inversion model achieves a correlation of 0.32 with the ground truth inversion data at the saturation point of training.

This trained articulatory inversion model is then adapted to hypernasality estimation by processing input speech through the PASE encoder, conv layer, and four feed-forward hidden layers to generate a time series of 250-dimensional latent codes.

The hypernasality estimation task then is performed with the same network architecture and hyperparameters as the PASE-based hypernasality estimator, but with the inversion model latent codes as input rather than PASE encodings, with no unfreezing of any pretrained elements. This model is referred to as AINV-NN in Table 4.2.

4.2.3 Human Baseline

Finally, to evaluate whether any of the approaches addresses the core problem of inter-rater non-reliability, I treat each of the 14 clinical hypernasality severity evaluators as an individual estimator, for which I can evaluate mean average error (MAE) and Pearson's Correlation Coefficient (PCC) from the average scores that are used as ground truth for training and evaluating the other models. For the LOSO evaluation I average the 14 human evaluator MAE and PCC scores across the 75 speakers, and then average these across the 14 evaluators to get an average human baseline MAE and PCC. Similarly, for the LODO conditions I evaluate only the evaluation disease subset. In other words, the reported "Human" MAE and PCC scores in Table 4.2 for the LODO conditions are evaluated on the "Test on" set for a given column.

4.3 Results

Multiple regression analysis was used to test if the NCD measure for the four nasal cognates predicted the average clinician nasality ratings. The results of the regression analysis indicated that the four predictors explained 47% of the variance ($R = 0.687, F(4, 79) = 16.798, p < 0.05$). It was found that the NCD for T significantly predicted the hypernasality rating ($\beta = -0.316, p < 0.05$), as did the NCD for P ($\beta = -0.278, p < 0.05$).

Factor	B	SE B	β	p
T	-0.225	.093	-0.316*	0.018
D	-0.061	.043	-0.201	0.163
P	-0.077	.030	-0.278*	0.013
B	0.000	.016	0.002	0.987
R^2		0.473		
F		16.798**		

Table 4.1: Summary of Linear Regression Analyses for Variables Predicting Clinical Hypernasality Scores from the NCD Measures (N=80). *$p < 0.05$, **$p < 0.001$

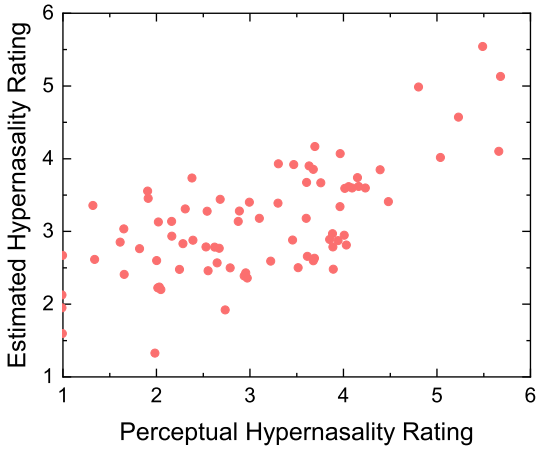
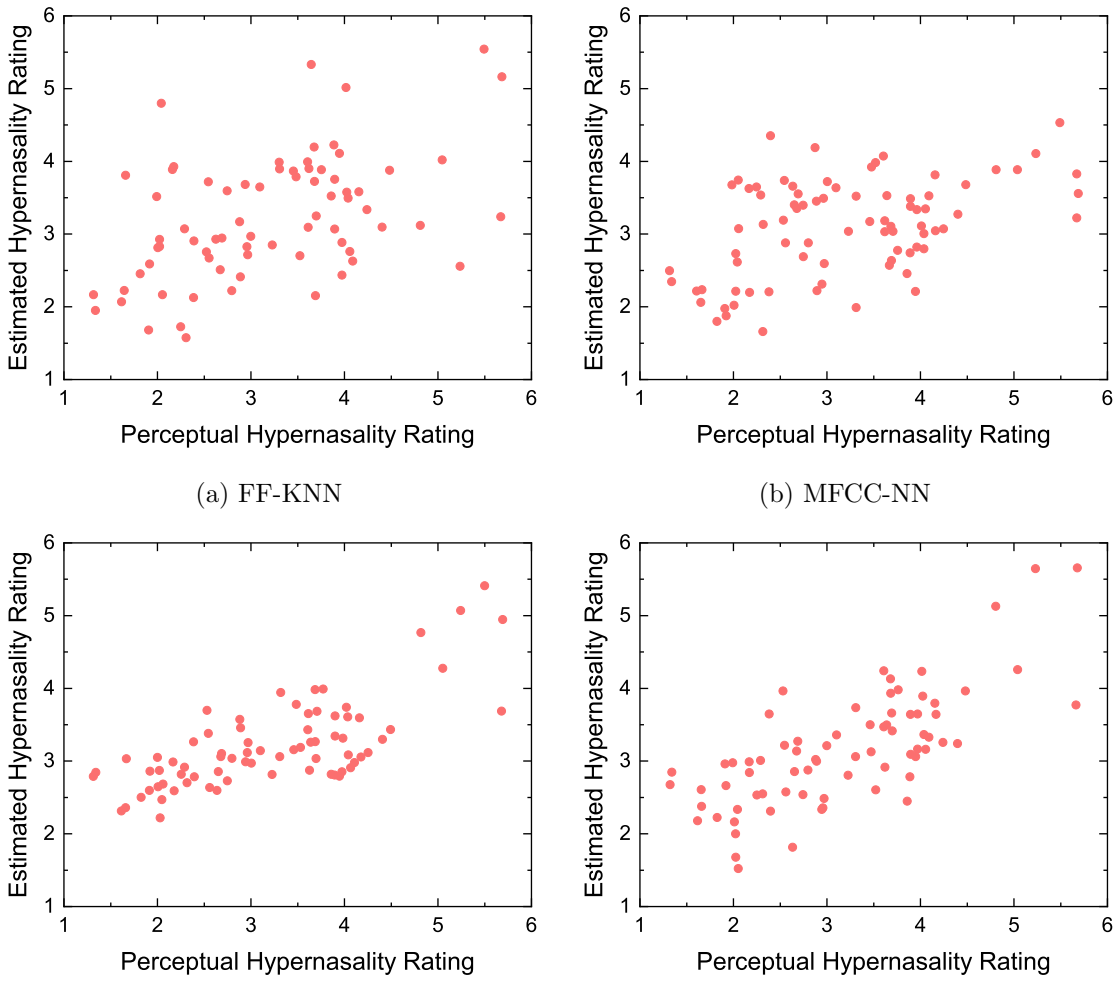


Figure 4.2: Output of the linear regression model predicting nasality using T, D, P, and B as shown in Table 4.1.

Figure 4.2 shows model-predicted nasality in Table 4.1 against the SLP-assessed clinical hypernasality measure.

In Table 4.2, I show the results of the evaluations (LOSO - leave one speaker out and LODO - leave one disease out for the four diseases) for seven different models. The results show that the linear model based on NAP features consistently outperforms the other two models, especially under the LODO conditions. The differences are also apparent when I analyze the individual LOSO correlation plots in Figure 4.3. These scatter plots relate the estimated hypernasality score for each speaker against the



(a) FF-KNN

(b) MFCC-NN

(c) NAP-Linear

(d) NAP+NCD-Forward

Figure 4.3: Leave one speaker out (LOSO) results from predicting the hypernasality score for the simple feature baseline with (a) the KNN classifier and simple formant features (FF-KNN), (b) the neural network baseline (MFCC-NN), (c) the NAP features with simple linear regression (NAP-Linear), and (d) the NAP+NCD optimized set.

actual hypernasality score. As is clear from the figures, the correlation of the baseline methods is largely driven by the samples with very high nasality scores. The NAP model exhibits a linear trend between the predicted and actual values throughout the hypernasality range.

Table 4.2: Comparative evaluation of statistical signal processing (SSP) and neural (ML) baselines against the (novel) hybrid features NCD, NAP, and their composition, and the human raters for predicting average clinician hypernasality score. Conducted using leave-one-speaker-out (LOSO) and leave-one-disease-out cross validation. Mean absolute error (MAE) on the 7-point scale and Pearson correlation coefficient (PCC) are reported. For each metric the **best overall model** is bold, and the *best non-novel model* is italicized.

	Train on	LOSO (74)		HD, PD, ALS		Ataxia, PD, ALS		Ataxia, HD, ALS		Ataxia, PD, HD	
	Test on	LOSO (1)		Ataxia		HD		PD		ALS	
	Model	MAE	PCC	MAE	PCC	MAE	PCC	MAE	PCC	MAE	PCC
SSP	FF-Lin	0.871	0.180	0.823	0.042	0.666	-0.751	1.316	0.351	1.426	-0.425
SSP	FF-Add	0.789	0.435	0.730	-0.123	0.693	-0.557	1.334	0.277	1.260	*0.429*
SSP	FF-KNN	*0.754*	*0.481*	0.781	*0.333*	0.567	0.381	1.218	0.402	*1.227*	-0.039
ML	MFCC-NN	0.884	0.458	0.904	-0.120	*0.429*	*0.568*	*0.800*	*0.457*	1.233	0.315
ML	PASE-NN	0.774	0.417	0.707	-0.204	0.433	0.237	1.150	0.163	1.407	0.176
ML	PFT-NN	0.896	0.135	0.995	0.074	0.489	-0.271	1.100	0.227	1.540	0.238
ML	AINV-NN	0.874	0.146	*0.594*	-0.474	0.515	-0.053	1.271	0.114	1.245	0.211
Novel	NCD-Lin	0.699	0.572	0.950	0.674	0.466	0.545	1.068	-0.025	1.108	0.296
Novel	NAP-Lin	0.587	0.722	**0.546**	**0.750**	0.559	**0.737**	**0.509**	**0.697**	0.597	0.527
Novel	NAP+NCD	**0.558**	**0.748**	0.619	0.496	**0.400**	0.550	0.767	0.523	**0.578**	**0.740**
	Human	0.832	0.725	0.871	0.476	1.256	0.550	0.746	0.636	0.979	0.601

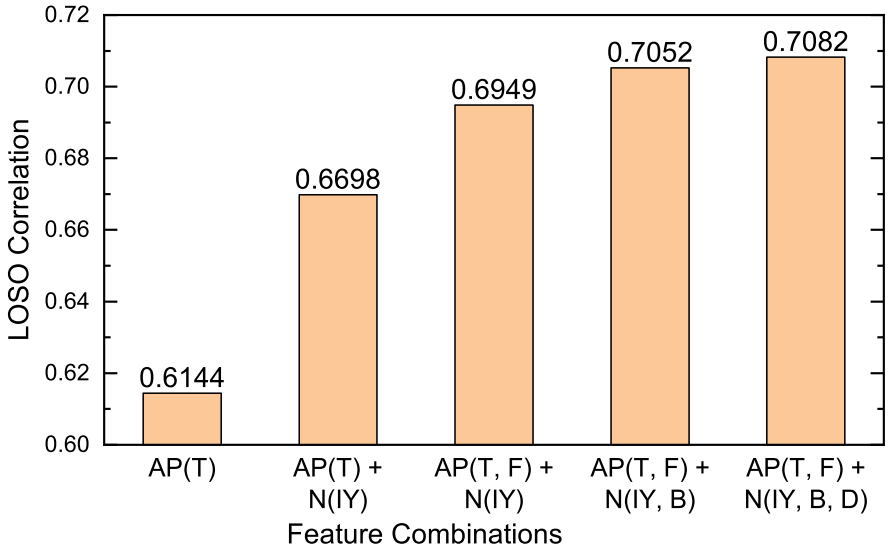


Figure 4.4: Cumulative marginal improvement plot of leave one speaker out correlation with the addition of the most optimal articulatory precision and nasalization features.

4.3.1 Individual Feature Contributions

I use a simple forward selection algorithm for the LOSO model to identify the most predictive NAP features. The algorithm identifies the subset of features that minimizes the cross-validation mean square error between the predicted hypernasality rating and the clinical hypernasality rating. Features are iteratively added until the cross-validation loss is no longer decreased. This procedure results in 6 non-redundant features selected for prediction. This includes the articulatory precision for T and F and the nasalization for D, B, IY, and AA. Figure 4.4 depicts the marginal improvement in LOSO correlation as features are added in by decreasing feature prominence.

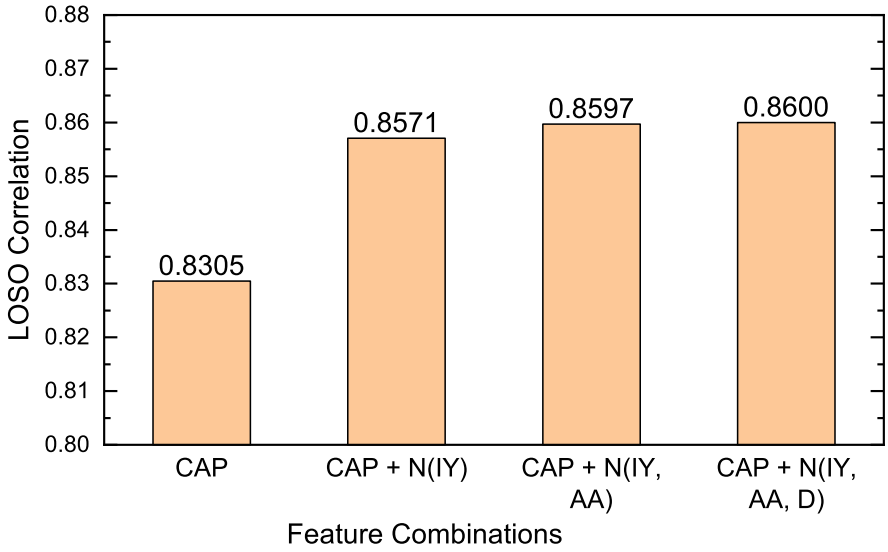


Figure 4.5: Cumulative marginal improvement plot of leave one speaker out correlation with the addition of the most optimal articulatory precision and nasalization features, and the clinician articulatory precision (CAP).

4.4 Relationship Between Articulatory Precision and Hypernasality

Articulatory precision and hypernasality are tightly linked. Hypernasal speech results in impaired articulatory precision. However, articulatory impairments can occur in motor-speech disorders for a variety of reasons. The neurological conditions I study herein impact several aspects of speech production including, respiration, voicing, resonance, and articulation. This brings up two important questions:

- Do the features capture changes related to hypernasality that go beyond changes in articulatory precision?
- Are the features sensitive to changes in articulatory precision that result from only hypernasality (and not other articulatory impairments resulting from dysarthria)?

In an attempt to decouple articulatory precision from hypernasality, I collect clinical articulatory precision ratings (in addition to the hypernasality ratings) from the same clinicians. The inter-rater reliability of the ratings was robust, with a Pearson

correlation coefficient of 0.75 and a mean absolute error of 1.01 on a 7-point scale.

To answer the first question above, and demonstrate that the NAP features capture information beyond changes in articulatory precision, I use a multiple linear regression model with clinician-rated articulatory precision alongside our six most predictive features (N(AA), N(IY), N(B), N(D), AP(T), AP(F)) as independent variables. The dependent variable is the clinical hypernasality rating. I once again use the forward selection algorithm on Pearson correlation coefficient to cumulatively select the most predictive features. The results are depicted in Figure 4.5. As expected, the subjective AP rating is most predictive as there is significant overlap with hypernasality, and it is selected first. In the presence of this generalized measure of articulatory precision, it makes sense that AP(T, F), features that are themselves estimating AP, would not be selected. This reinforces the rationale for their inclusion in the model. Three nasalization features, N(IY, AA, D), are able to further improve the correlation of the linear model predictions.

To answer the second question, and demonstrate that our features are sensitive to hypernasality alone, I evaluate a linear model trained on our full dataset of dysarthric speech using the six most predictive features predicting hypernasality scores for the 18 speech samples from individuals with cleft lip and palate in our CLP dataset. The linear hypernasality model trained on our dysarthric speech corpus achieves a PCC of 0.89 for predicting the adult hypernasality severity, and 0.82 for predicting the hypernasality level of the children. This provides additional evidence that our features capture the perceptual quality of hypernasality and not other co-modulating symptoms.

4.5 Effectiveness of Forced Alignment

The features I have proposed herein rely on force aligning known transcripts to dysarthric speech (Yeung *et al.* (2015)). This can be problematic as coarticulation, blending, missed targets, distorted vowels, and poor articulation present in severely disordered speech (Green *et al.* (2003)) may interfere with the appropriate matching of dictionary phoneme-word pairs to the realized sounds (Knowles *et al.* (2018)).

I directly evaluate the prevalence of alignment errors generated by our forced alignment methodology using manually aligned transcripts. Two annotators produced word- and syllable-level aligned transcripts using the same spelling and phoneme-word conventions employed in the acoustic model dictionary for all utterances in the dataset. For each speaker in the dataset, I count word- and phone-level alignment errors based on the position of the center point of a word or phoneme t'_c as assessed by the forced aligner and the beginning and end of the corresponding word or syllable, t_{min}, t_{max} as assessed by the human transcriber. For each word or phoneme, the error is counted as $t_e = \max(0, t_{min} - t'_c, t'_c - t_{max})$. This error measure returns 0 if the center of the phoneme falls within the syllable; otherwise it returns the maximum error between the center of the automatically aligned phoneme and the start and end of the manually-aligned syllable. In Figure 4.6 I show the alignment error (in seconds) against the hypernasality rating to show how alignment error rates progress as hypernasality increases. The results clearly show that for all but the most severely hypernasal speakers forced alignment works effectively.

These results also indicate that our objective hypernasality ratings for the most imprecise speakers are not reliable. While this is a limitation of the approach, it is not severely limiting. In most cases, clinicians are more concerned with evaluating speakers in the mild-moderate end of the scale where they can monitor disease progress

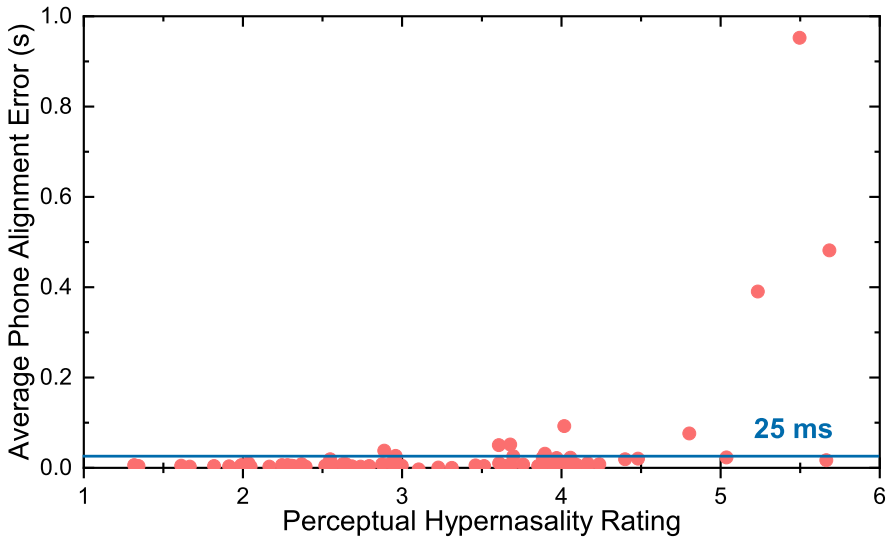


Figure 4.6: Plot of average alignment errors per speaker (s) against clinician-rated articulatory precision at the phone level. Dashed line indicates an average alignment error of 25 ms.

early or evaluate the effects of an intervention. This is less common for later stages of disease.

It is interesting to note that, while the alignment is poor, the model still yields high hypernasality scores for imprecise speakers. Precise alignment for speakers in this range is simply not possible, manually or otherwise. It's likely that the poor hypernasality ratings predicted by the model is driven by the poor alignment itself (Green and Carmichael (2004)).

Chapter 5

DISCUSSION AND CONCLUSIONS

5.1 Feature-Rating Correlations

The NCD features are formulated as a log probability ratio between the expected class of a given transcript plosive and its nasal cognate. Increasingly positive values correspond with a higher confidence that the speaker has correctly articulated the plosive rather than its nasal cognate, and values closer to zero or negative represent plosives that sound more like their nasal cognate than the intended stop. This directionality is exhibited as expected in the phoneme-by-phoneme analysis of the feature, where across speaker classes the NCD of a given phoneme decreases as nasality increases.

Figure 3.3 shows that the NCD features very clearly separate the control and severe nasality groups with all phonemes. However, for the voiced phonemes D and B the moderate and mild means and medians are close and the quartile ranges overlap considerably. Performance is worst for B, which exhibits both high cross-group overlap of the quartile ranges and insignificance as a predictor of the subjective hypernasality scores in the multiple regression analysis.

When considering these inter-phone performance inconsistencies, differences in the phonetic environments in which the test phonemes appear are noteworthy. In the five test sentences, T appears 18 times, D 9 times, P 6 times and B only 3 times. Of these appearances, T has 8 word-internal appearances, 7 word-final appearances, and 3 word-initial appearances. The phoneme D has 5 word-initial appearances, 3 word-internal appearances, and 1 word-final appearance. The phoneme P has 4 word-

internal appearances and 2 word-initial appearances, and B exclusively has 3 word-initial appearances. It is likely that these disparities in overall occurrence and word-internal occurrence play an important role in explaining the performance disparity.

Additionally, it is important to note that the NCD features are intended to assess a physical phenomenon, the realized allophones, not the underlying phonemes themselves. The phonetic transcriptions provided for HMM-based ASR systems fall somewhere between broad phonetic transcriptions and allophonic narrow transcriptions, allowing for possible confusion scenarios. For example, a T may be realized in English as [t], [ɾ], or [ʔ] depending on phonetic environment. All three could be compared to [n] (N) in the NCD model even though the glottal stop [ʔ] is unaffected by VPD and shares no place of articulation with [n].

The NCD features tend to be high-variance because they require reliable phoneme-level alignment to compute; higher frequency phonemes exhibit reduced variability through averaging. Accordingly, in this study the more frequent phonemes are more useful predictors of hypernasality. This suggests that future datasets to evaluate methods like NCD should include a higher frequency of plosive consonants balanced across the categories, in consistent environments in which the correct allophones are reliably produced.

The feature-level analyses of the NAP features behave as expected, with the nasalization log likelihood of the phonemes increasing as hypernasality increases, while the articulatory precision decreases as hypernasality increases (Figure 3.4). Analysis of Eqns (3) and (4) shows that this makes sense. As hypernasality increases, the voiced phonemes become more and more like the N class in the acoustic model in Section 3.4. Similarly, as hypernasality increases, the acoustics of the unvoiced phonemes become less and less like the intended target, therefore the ratio in Eqn (4) decreases.

During the feature selection analysis for NAP in Section 4.1 certain consonants

appeared prominently. In particular, the nasalization feature for phonemes D, and B, as well as the articulatory precision of T and F were prominent. T, B, and D are referred to as a "nasal cognates" in Saxon *et al.* (2019), as the bilabial consonant B shares a place of articulation with the bilabial nasal M, the lingua-alveolar consonants T and D share a place of articulation with the lingua-alveolar nasal N. Leakage through the nasal cavity will interfere with the production of all of these phonemes, and in the voiced case, they will sound like their corresponding nasal phonemes. It is not surprising that the nasalization model is most sensitive to these phonemes since that model is trained on healthy speech, where the N class consists mostly of instances in M and N and surrounding vowels.

Through the same analysis, the most prominent vowels selected were AA and IY. AA is the most open and back vowel in English, whereas IY is the most closed and fronted. It may be the case that these extreme ends of the vowel chart exhibit more noticeable patterns of nasalization, either on a perceptual level or just in their PLP-nasalization feature realization.

5.2 Performance of Estimation Models

The novel models based on NAP and NCD features outperform all estimation baselines, across all conditions, on both the MAE and PCC metrics. Furthermore, the NAP and NCD-based hypernasality estimation systems outperform even the trained human SLP annotators in estimating their own average scores, speakerwise. In the LOSO condition the MFCC-NN approach outperforms the simpler formant features in PCC, while the formant feature model does achieve a lower MAE it seems to be a result of largely predicting the mean, with only a very modest upsloping trend in Figure 4.3 (a) as opposed to Figures 4.3 (b) and (c) which clearly show upward-sloping trends.

Although the novel approaches consistently outperform the baselines, it is noteworthy that neither the ML- nor SSP-based techniques consistently outperform the other across various pairs of cross validation condition and metric. For example, the articulatory inversion-based neural model AINV-NN achieves the best (lowest) MAE of the baselines in the test-on-Ataxia LODO condition, but also has the worst (lowest) PCC of the baselines on the same validation split. Furthermore, of the 10 combinations of validation-metric evaluations performed, the best baseline performance is achieved by SSP-based 5 times, and by ML-based methods the other 5 times, making an even split. It is noteworthy that in such a volatile problem space, one family of approaches is able to consistently outperform the others.

In the LODO conditions, the SSP and MFCC-NN models perform unpredictably. On some disease classes, MFCC-NN outperforms FF, while the opposite is true for others. By comparison, the NAP achieves consistent performance across all LODO classes. This suggests that these features are a robust measure of hypernasality, relatively invariant to the disease-specific co-modulating variables that hinder the performance of the baselines on the same task. The nasalization features in the NAP, by virtue of being trained on a large corpus of healthy speech, and targeting a specific perceptual quality are simultaneously more robust to both the disease-specific overfitting expected from NN methods such as Vikram *et al.* (2018) and speaker-to-speaker variances discussed in the design of the formant-based A1P0 and related features in Styler (2015), Chen (1997). Articulatory precision features are robust in a similar way.

One of the added benefits of the proposed approach over the baseline methods is the direct interpretability of the individual NAP and NCD features. While it is not immediately clear how MFCC features or formant-based features are expected to change with different hypernasality levels, the proposed features are easy to interpret;

this stems directly from their design as likelihood ratios comparing known classes of phonemes.

In spite of its robustness the NAP and NCD likelihood ratio technique has limitations. Most limiting is its reliance on aligned transcripts to perform the estimation. The results shown in this paper were based on forced alignment. This is always possible when the ground truth transcript is known but is not feasible for spontaneous speech. The robustness of the model comes from the fact that it is trained on a large corpus of healthy speech; however, this training also induces a bias in the model. As the feature selection results show, the model is adept at detecting hypernasal speech from phonemes that look similar to nasals in healthy speech; however it is impossible to capture nasalization acoustic patterns for unvoiced speech since these sounds never occur in healthy speech (and hence cannot be captured in our model). As a result, I use articulatory precision as a proxy for nasalization for these sounds. Increased hypernasality typically implies reduced articulatory precision, but the converse is not necessarily true. As such, it is possible for speakers to exhibit reduced precision for other reasons than hypernasality. As I showed with the CLP speech experiments, when the reduction in articulatory precision is due to hypernasality, the model generalizes out-of-disease quite well.

5.3 Conclusions and Summary

In this thesis, I proposed a set of acoustic modelling features as an objective and noninvasive proxy for hypernasal speech. All seek to model various explainable aspects of perceptual hypernasality. They leverage a data-driven approach to learning expert-designed features on healthy speech that capture perceptible elements of velopharyngeal dysfunction in hypernasal speech.

The NCD features are motivated by the simple observation that alveolar stops T

and D map to the alveolar nasal N and the bilabial stops P and B map to bilabial nasal M when the energized column of air is shunted into the nasal passage during speech production. The feature is measured by first training an acoustic model on healthy speech and, for a test speaker, evaluating the likelihood ratio between the plosives and their respective nasal cognates. For healthy speakers that exhibit no signs of hypernasality, this ratio is large and decreases with increasing levels of hypernasality. This is confirmed on speech samples from 75 speakers diagnosed with different dysarthria subtypes and exhibiting varying levels of hypernasality. The results show that the features are strongly correlated with clinical perception. The variability of the NCD features, driven by the differences in representation of B, D, and P, motivated the design of a more robust feature set that could analyze the full set of phonemes present in the phonetically rich standard read stimuli provided to the 75 speakers.

This more robust system became the Nasalization-Articulation Precision (NAP) features. I demonstrated that these features, when evaluated on disordered speech, track the expected trends in perceptual hypernasality ratings, and can be used with ridge regression to estimate a clinician-rated hypernasality score more accurately than several representative baseline methods. Additionally, I demonstrated that the NAP algorithm predictions for hypernasality rating generalize across diseases with significantly less loss in accuracy than existing approaches. This implies that the NAP features are a robust method for estimating hypernasality in dysarthria.

A limitation of this approach, and articulatory precision estimation techniques more generally, is a reliance on known transcripts with which alignment may be performed. Neural models for directly assessing articulatory precision from raw speech audio is a promising future research direction—such models could provide the simultaneous identification of and precision assessment of phonemes on the fly, and provide

downstream representations that could drive characterization of hypernasality without relying on reading as a stimulus, or known transcripts for assessment.

However, in spite of these limitations, the NAP and NCD-based linear hypernasality estimation models significantly outperform all evaluated baselines, including the human annotators themselves. In other words, this approach to robustly and objectively modeling hypernasality achieves superhuman performance in both accuracy and difference assessment, and achieves state-of-the-art results in generalization across both machine learning- and statistical signal processing-based baselines. Systems applying these principles may one day drive the deployment of objective, scalable, and interpretable speech-based telemedicine metrics. These metrics stand to improve diagnostic performance and clinical outcomes for patients exhibiting the early stages of neurological disease.

BIBLIOGRAPHY

Ba, J. L., J. R. Kiros and G. E. Hinton, "Layer normalization", arXiv preprint arXiv:1607.06450 (2016).

Bettens, K., L. Bruneel, Y. Maryn, M. De Bodt, A. Luyten and K. M. Van Lierde, "Perceptual evaluation of hypernasality, audible nasal airflow and speech understandability using ordinal and visual analogue scaling and their relation with nasalance scores", J Commun Disord 76, 11–20 (2018).

Bettens, K., F. L.Wuyts and K. M. Lierde, "Instrumental assessment of velopharyngeal function and resonance: A review", Journal of Communication Disorders (2014).

Brancamp, T. U., K. E. Lewis and T. Watterson, "The relationship between nasalance scores and nasality ratings obtained with equal appearing interval and direct magnitude estimation scaling methods", The Cleft Palate-Craniofacial Journal 47, 6, 631–637 (2010).

Brunnegard, K., A. Lohmander and J. van Doorn, "Comparison between perceptual assessments of nasality and nasalance scores", Int J Lang Commun Disord 47, 5, 556–566 (2012).

Cairns, D., J. Hansen and J. Riski, "A noninvasive technique for detecting hypernasal speech using a nonlinear operator.", IEEE Transactions on Biomedical Engineering (1996).

Carrow, E., V. Rivera, M. Mauldin and L. Shamblin, "Deviant Speech Characteristics in Motor Neuron Disease", JAMA Otolaryngology–Head & Neck Surgery 100, 3, 212–218, URL `https://doi.org/10.1001/archotol.1974.00780040220014` (1974).

Castellanos, G., G. Daza, L. Sánchez, O. Castrillón and J. Suárez, "Acoustic speech analysis for hypernasality detection in children", in "2006 International Conference of the IEEE Engineering in Medicine and Biology Society", pp. 5507–5510 (IEEE, 2006).

Chapman, K. L., A. Baylis, J. Trost-Cardamone, K. N. Cordero, A. Dixon, C. Dobbelsteyn, A. Thurmes, K. Wilson, A. Harding-Bell, T. Sweeney *et al.*, "The Americleft Speech Project: a training and reliability study", The Cleft Palate-Craniofacial Journal 53, 1, 93–108 (2016).

Chen, M. Y., "Acoustic correlates of English and French nasalized vowels", The Journal of the Acoustical Society of America 102, 4, 2360–2370 (1997).

de Stadler, M. and C. Hersh, "Nasometry, videofluoroscopy, and the speech pathologist's evaluation and treatment", Adv. Otorhinolaryngol. 76, 7–17 (2015).

Dubey, A. K., S. M. Prasanna and S. Dandapat, "Zero time windowing based severity analysis of hypernasal speech", in "2016 IEEE Region 10 Conference (TENCON)", pp. 970–974 (IEEE, 2016).

Dubey, A. K., S. M. Prasanna and S. Dandapat, "Pitch-adaptive front-end feature for hypernasality detection", Proc. Interspeech 2018 pp. 372–376 (2018a).

Dubey, A. K., A. Tripathi, S. R. M. Prasanna and S. Dandapat, "Detection of hypernasality based on vowel space area", Journal of Acoustical Society of America (2018b).

Duffy, J., *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management* (Mosby, 1995).

Duffy, J. R., "Motor speech disorders: clues to neurologic diagnosis", in "Parkinson's Disease and Movement Disorders", pp. 35–53 (Springer, 2000).

Extence, H. and S. Cassidy, "The role of the speech pathologist in the care of the patient with cleft palate", in "Maxillofacial Surgery (Third Edition)", edited by P. A. Brennan, H. Schliephake, G. Ghali and L. Cascarini, pp. 1014 – 1023 (Churchill Livingstone, 2017), third edition edn., URL `http://www.sciencedirect.com/science/article/pii/B978070206056400071X`.

Giegerich, H. J., *English Phonology: An Introduction*, Cambridge Textbooks in Linguistics (Cambridge University Press, 1992).

Glass, J. R. and V. W. Zue, "Detection of nasalized vowels in American English.", Proceedings of ICASSP, volume 4, 1569–1572. (1985).

Golabbakhsh, M., F. Abnavi, M. Kadkhodaei Elyaderani, F. Derakhshandeh, F. Khanlar, P. Rong and D. Kuehn, "Automatic identification of hypernasality in normal and cleft lip and palate patients with acoustic analysis of speech", Journal of Acoustical Society of America (2017).

Green, P. and J. Carmichael, "Revisiting dysarthria assessment intelligibility metrics", in "Eighth International Conference on Spoken Language Processing", (2004).

Green, P., J. Carmichael, A. Hatzis, P. Enderby, M. Hawley and M. Parker, "Automatic speech recognition with sparse training data for dysarthric speakers", in "Eighth European Conference on Speech Communication and Technology", (2003).

Hall, M. A., "Correlation-based feature selection for machine learning", (1999).

Harel, B. T., M. S. Cannizzaro, H. Cohen, N. Reilly and P. J. Snyder, "Acoustic characteristics of parkinsonian speech: a potential biomarker of early disease progression and treatment", Journal of Neurolinguistics 17, 6, 439–453 (2004).

Hawkins, S. and K. Stevens, "Acoustic and perceptual correlates of the non-nasal–nasal distinction for vowels.", Journal of Acoustical Society of America (1985).

He, L., J. Zhang, Q. Liu, H. Yin and M. Lech, "Automatic evaluation of hypernasality and consonant misarticulation in cleft palate speech", IEEE Signal Processing Letters (2014).

Hegde, S., S. Shetty, S. Rai and T. Dodderi, "A survey on machine learning approaches for automatic detection of voice disorders", Journal of Voice URL http://www.sciencedirect.com/science/article/pii/S0892199718301437 (2018).

Henningsson, D. G. and D. A. Isberg, "Comparison between multiview videofluoroscopy and nasendoscopy of velopharyngeal movements", The Cleft Palate-Craniofacial Journal 28, 4, 413–418, URL https://doi.org/10.1597/1545-1569_1991_028_0413_cbmvan_2.3.co_2, pMID: 1742312 (1991).

Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech", the Journal of the Acoustical Society of America 87, 4, 1738–1752 (1990).

Kao, D. S., D. A. Soltysik, J. S. Hyde and A. K. Gosain, "Magnetic resonance imaging as an aid in the dynamic assessment of the velopharyngeal mechanism in children", Plastic and reconstructive surgery 122, 2, 572 (2008).

Kataoka, R., K.-I. Michi, K. Okabe, T. Miura and H. Yoshida, "Spectral properties and quantitative evaluation of hypernasality in vowels", The Cleft palate-craniofacial journal 33, 1, 43–50 (1996).

Kataoka, R., D. W. Warren, D. J. Zajac, R. Mayo and R. W. Lutz, "The relationship between spectral characteristics and perceived hypernasality in children", The Journal of the Acoustical Society of America 109, 5, 2181–2189 (2001).

Kent, R., "Some limits to the auditory-perceptual assessment of speech and voice disorders", American Journal of Speech Language Pathology (1996).

Kingma, D. P. and J. Ba, "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980 (2014).

Knowles, T., M. Clayards and M. Sonderegger, "Examining factors influencing the viability of automatic acoustic analysis of child speech", Journal of Speech, Language, and Hearing Research 61, 10, 2487–2501, URL https://lshss.pubs.asha.org/doi/abs/10.1044/2018_JSLHR-S-17-0275 (2018).

Kozaki-Yamaguchi, Y., N. Suzuki, Y. Fujita, H. Yoshimasu, M. Akagi and T. Amagasa, "Perception of hypernasality and its physical correlates", Oral Science International 2, 1, 21 – 35, URL http://www.sciencedirect.com/science/article/pii/S1348864305800047 (2005).

Kuehn, D. and J. Moon, "Velopharyngeal closure force and levator veli palatini activation levels in varying phonetic contexts", Journal of Speech Language and Hearing Research (1998).

Kuehn, D. P., P. B. Imrey, L. Tomes, D. L. Jones, M. M. O'Gara, E. J. Seaver, B. E. Smith, D. Van Demark and J. M. Wachtel, "Efficacy of continuous positive airway pressure for treatment of hypernasality", The Cleft palate-craniofacial journal 39, 3, 267–276 (2002).

Kuehn, D. P. and K. T. Moller, "Speech and language issues in the cleft palate population: the state of the art", The Cleft palate-craniofacial journal 37, 4, 1–35 (2000).

Kummer, A. and L. Lee, "Evaluation and Treatment of Resonance Disorders", Language, Speech, and Hearing in Schools 27, 271–281 (1996).

Lee, G.-S., C.-P. Wang and S. Fu, "Evaluation of hypernasality in vowels using voice low tone to high tone ratio", The Cleft Palate-Craniofacial Journal 46, 1, 47–52 (2009).

Lee, G.-S., C.-P. Wang, C. C. Yang and T. B. Kuo, "Voice low tone to high tone ratio: a potential quantitative index for vowel [a:] and its nasalization", IEEE transactions on biomedical engineering 53, 7, 1437–1439 (2006).

Lintz, L. B. and D. Sherman, "Phonetic elements and perception of nasality", Journal of Speech, Language, and Hearing Research (1961).

Lohmander, D. A. and M. M. Olsson, "Methodology for perceptual assessment of speech in patients with cleft palate: A critical review of the literature", The Cleft Palate-Craniofacial Journal 41, 1, 64–70, URL `https://doi.org/10.1597/02-136`, pMID: 14697067 (2004).

Maier, A., A. Reuß, C. Hacker, M. Schuster and E. Nöth, "Analysis of hypernasal speech in children with cleft lip and palate", in "International Conference on Text, Speech and Dialogue", pp. 389–396 (Springer, 2008).

McAuliffe, M., M. R. F. P. Danar, vannawillerton, michaelasocolof and A. Coles, "MontrealCorpusTools/Montreal-Forced-Aligner: Version 1.0.1", URL `https://doi.org/10.5281/zenodo.2630943` (2019).

Nieto, R. G., J. I. Marín-Hurtado, L. M. Capacho-Valbuena, A. A. Suarez and E. A. B. Bolaños, "Pattern recognition of hypernasality in voice of patients with cleft and lip palate", in "2014 XIX Symposium on Image, Signal Processing and Artificial Vision", pp. 1–5 (IEEE, 2014).

Nikitha, K., S. Kalita, C. Vikram, M. Pushpavathi and S. M. Prasanna, "Hypernasality severity analysis in cleft lip and palate speech using vowel space area.", in "INTERSPEECH", pp. 1829–1833 (2017).

Novotny, M., J. Rusz, R. Cmejla, H. Ruzickova, J. Klempir and E. Ruzicka, "Hypernasality associated with basal ganglia dysfunction: evidence from Parkinson's disease and Huntington's disease", PeerJ 4, e2530 (2016).

Orozco-Arroyave, J. R., J. D. Arias-Londoño, J. F. Vargas-Bonilla and E. Nöth, "Automatic detection of hypernasal speech signals using nonlinear and entropy measurements", in "Thirteenth Annual Conference of the International Speech Communication Association", (2012).

Orozco-Arroyave, J. R., E. A. Belalcazar-Bolanos, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, K. Daqrouq, F. Hönig and E. Nöth, "Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases", IEEE journal of biomedical and health informatics 19, 6, 1820–1828 (2015).

Orozco-Arroyave, J. R., J. Vargas-Bonilla, J. D. Arias-Londoño, S. Murillo-Rendón, G. Castellanos-Domínguez and J. Garcés, "Nonlinear dynamics for hypernasality detection in Spanish vowels and words", Cognitive Computation 5, 4, 448–457 (2013).

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

Pascual, S., M. Ravanelli, J. Serrà, A. Bonafonte and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks", arXiv preprint arXiv:1904.03416 (2019).

Pellegrini, T., L. Fontan, J. Mauclair, J. Farinas and M. Robert, "The goodness of pronunciation algorithm applied to disordered speech", in "INTERSPEECH", (2014).

Pentax, "Nasometer II: Model 6450", URL `https://www.pentaxmedical.com/pentax/en/99/1/Nasometer-II-Model-6450` (2016).

Poole, M., J. Wee, J. Folker, L. Corben, M. Delatycki and A. Vogel, "Nasality in Friedreich ataxia", Clin Linguist Phon 29, 1, 46–58 (2015).

Rafael Orozco Arroyave, J., J. Francisco Vargas Bonilla and E. Delgado Trejos, "Acoustic analysis and non linear dynamics applied to voice pathology detection: A review", Recent Patents on Signal Processing 2, 2, 96–107, URL `https://www.ingentaconnect.com/content/ben/rptsp/2012/00000002/00000002/art00003` (2012).

Rah, D. K., Y. I. Ko, C. Lee and D. W. Kim, "A noninvasive estimation of hypernasality using a linear predictive model", Annals of biomedical Engineering 29, 7, 587–594 (2001).

Rendón, S. M., J. O. Arroyave, J. V. Bonilla, J. A. Londono and C. C. Domínguez, "Automatic detection of hypernasality in children", in "International Work-Conference on the Interplay Between Natural and Artificial Computation", pp. 167–174 (Springer, 2011).

S. Paal, K. S.-S. E. N., U. Reulbach and M. Schuster, "Evaluation of speech disorders in children with cleft lip and palate", J Orofac Orthop 66, 4, 270–278 (2005).

Saxon, M., J. Liss and V. Berisha, "Objective measures of plosive nasalization in hypernasal speech", in "ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)", pp. 6520–6524 (2019).

Saxon, M., A. Tripathi, Y. Jiao, J. Liss and V. Berisha, "Robust estimation of hypernasality in dysarthria with acoustic model likelihood features", (2020).

Scarmagnani, R. H., A. Oliveira, A. Fukushiro, M. Salgado, I. Trindade and R. Yamashita, "Impact of inter-judge agreement on perceptual judgment of nasality", Codas 26, 5, 357–359 (2014).

Shah, M., M. Tu, V. Berisha, C. Chakrabarti and A. Spanias, "Articulation constrained learning with application to speech emotion recognition", EURASIP journal on audio, speech, and music processing 2019, 1, 14 (2019).

Shriberg, L. and R. Kent, *Clinical Phonetics*, Wiley Communications Series (Macmillan, 1982), URL `https://books.google.com/books?id=VFRDAAAACAAJ`.

Sinko, K., M. Gruber, R. Jagsch, I. Roesner, A. Baumann, A. Wutzl and D.-M. Denk-Linnert, "Assessment of nasalance and nasality in patients with a repaired cleft palate", European Archives of Oto-Rhino-Laryngology 274, 7, 2845–2854 (2017).

Sivaraman, G., V. Mitra, H. Nam, M. Tiede and C. Espy-Wilson, "Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion", The Journal of the Acoustical Society of America 146, 316–329 (2019).

Styler, W., "Using Praat for linguistic research", University of Colorado at Boulder Phonetics Lab (2013).

Styler, W., "On the acoustical and perceptual features of vowel nasality", (2015).

Tarun, P., E.-W. CY and S. BH, "Simulation and analysis of nasalized vowels based on magnetic resonance imaging data", Journal of Acoustical Society of America (2007).

Theodoros, D., B. Murdoch, P. Stokes and H. Chenery, "Hypernasality in dysarthric speakers following severe closed head injury: A perceptual and instrumental analysis", Brain Injury 7, 1, 59–69, URL `https://doi.org/10.3109/02699059309008157` (1993).

Theodoros, D., B. Murdoch and E. Thompson, "Hypernasality in Parkinson's disease: A perceptual and physiological analysis", J Med Speech-Lang Pathol 3, 2, 73–84 (1995).

Tsai, Y., C. Wang and G. Lee, "Voice low tone to high tone ratio, nasalance, and nasality ratings in connected speech of native mandarin speakers: a pilot study.", The Cleft Palate-Craniofacial Journal 49, 4, 437–46, URL `https://www.ncbi.nlm.nih.gov/pubmed/21740178` (2012).

Tu, M., A. Grabek, J. Liss and V. Berisha, "Investigating the role of L1 in automatic pronunciation evaluation of L2 speech", in "Proc. Interspeech 2018", pp. 1636–1640 (2018), URL http://dx.doi.org/10.21437/Interspeech.2018-1350.

Vijayalakshmi, P., R. MR and O. D., "Acoustic analysis and detection of hypernasality using a group delay function.", IEEE Transactions on Biomedical Engineering (2007).

Vijayalakshmi, P., T. Nagarajan and J. Rav, "Selective pole modification-based technique for the analysis and detection of hypernasality", in "TENCON 2009-2009 IEEE Region 10 Conference", pp. 1–5 (IEEE, 2009).

Vikram, C. M., A. Tripathi, S. Kalita and S. R. M. Prasanna, "Estimation of hypernasality scores from cleft lip and palate speech", in "Proc. Interspeech 2018", pp. 1701–1705 (2018), URL http://dx.doi.org/10.21437/Interspeech.2018-1631.

Vikram, M. C., N. Adiga and S. R. M. Prasanna, "Spectral enhancement of cleft lip and palate speech", in "INTERSPEECH", (2016).

Vogel, A., H. Ibrahim, S. Reilly and N. Kilpatrick, "A comparative study of two acoustic measures of hypernasality.", Journal of Speech, Language, and Hearing Research (2009).

Watterson, T., S. C. McFarlane and D. S. Wright, "The relationship between nasalance and nasality in children with cleft palate", Journal of Communication Disorders 26, 1, 13–28 (1993).

Westbury, J. R., "X-ray microbeam speech production database user's handbook", (1994).

Witt, S. and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning", Speech Commun. 30, 2-3, 95–108, URL http://dx.doi.org/10.1016/S0167-6393(99)00044-8 (2000).

Witt, S. M., *Use of speech recognition in computer-assisted language learning*, Ph.D. thesis, University of Cambridge (1999).

Woo, A., "Velopharyngeal dysfunction", Semin Plast Surg 26, 4, 170–177 (2012).

Yeung, Y. T., K. H. Wong and H. Meng, "Improving automatic forced alignment for dysarthric speech transcription", in "Sixteenth Annual Conference of the International Speech Communication Association", (2015).

Yu, L. and B. D. Barkana, "Classifying hypernasality using the pitch and formants", Proceedings of the 6th International Con ference on Information Technolog New Generations (2009).

APPENDIX A

PERMISSION STATEMENTS FROM CO-AUTHORS

Permission for including co-authored material in this dissertation was obtained from co-authors, Prof. Visar Berisha, Prof. Julie Liss, Dr. Yishan Jiao, and Ayush Tripathi.