Title

Perceptually Grounded Modeling and Modification of Speaker Identity

ISBN

9798189278914

Author

Netzorg, Robin

Publication Date

2026-05-01

Peer reviewed|Thesis/dissertation

Perceptually Grounded Modeling and Modification of Speaker Identity

By

Robin Netzorg

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Gopala Krishna Anumanchipalli, Chair
Professor Bin Yu
Professor Alane Suhr
Professor Keith Johnson

Spring 2026

Perceptually Grounded Modeling and Modification of Speaker Identity

Abstract

Perceptually Grounded Modeling and Modification of Speaker Identity

by

Robin Netzorg

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Gopala Krishna Anumanchipalli, Chair

Speech technology operates on representations of speaker identity that are either too high-level to be actionable (e.g., "masculine" or "old") or too high-dimensional to be interpretable (e.g., learned embeddings). How can we bridge this gap to enable interpretable analysis and modification of what makes a voice sound the way it does? I present research addressing this question in three stages. First, I establish that perceptual voice qualities, which are standardized descriptors of voice from speech pathology, speech forensics, and gender-affirming voice training, are reliably perceivable by non-experts and exhibit disentangled correlations with acoustic and physical measurements of the vocal tract. Second, I present the PQ-Encoder, a model that predicts these qualities from speech, and demonstrate that a compact 9-dimensional perceptual vector captures meaningful information about speaker identity, gender, and age. Finally, I introduce StyleMod, a conditional flow matching model that modifies speaker identity along perceptual voice quality axes, enabling interpretable and continuous manipulation of perceived gender and age while preserving non-targeted speaker attributes. Together, this work demonstrates that grounding speech technology in human perception yields representations that are both scientifically meaningful and practically useful for controllable voice generation.

To the (un)heard and (un)hearable,
and to those who live in the noise.

Contents

List of Figures

List of Tables

Acknowledgments

It would not be possible to make it through seven years of a PhD without some support, or emotional baggage. There were many times when I wanted to quit, when only support from others or teeth-grinding spite saw me to the end.

At the end of it all, I'm happy (and surprised) to say that my thesis turned out exactly how I would have wanted it to (ignoring perfectionist tendencies). When I first walked into my advisor Gopala Anumanchipalli's office in 2022, I expressed a desire to work on "technology that models and changes a person's voice." At the time, I knew nothing of signal or speech processing. Even so, he took me on as a PhD student, funding and supporting my research despite its sometimes circuitous route. Without him, this research would not have been possible. I cannot express my gratitude and thanks enough.

While Gopala supported my speech endeavours, my co-advisor Bin Yu shaped me into the statistician and scientist I am today. Her philosophy of collaborating with experts to pursue domain-relevant scientific questions directly led me to the collaborative work with gender-affirming voice coaches, speech language pathologists, and speech forensic scientists. Her philosophy of never trusting your own models without extensive perturbation analysis led me to always question my results.

My labmates in the Berkeley Speech Group made the daily work possible and enjoyable: Nicholas Lee, Cheol Jun Cho, Jiachen Lian, Kaylo Littlejohn, Tingle Li, Olorundamilola Kazeem, Akshat Gupta, and Bohan Yu. Conversations and collaborations throughout the years jumpstarted my career in and knowledge of speech processing. In particular, Louis Liu's signal processing expertise and contributions to speech synthesis made much of this work possible.

I cannot overstate the contribution of the trans and gender-diverse speech community to this work. Alyssa Cote, Klo Vivienne Garoute, and Sumi Koshin graciously lent hours and hours of their time to collaborate on our work together. Their expertise and willingness to share their voices with a stranger made this research not only possible but meaningful. Ruchi Kapila, for our conversations and their excited support.

I'd like to thank my broader committee. Alane Suhr made me excited about language technology again. Keith Johnson helped me perform the phonetics research present in this work.

My collaborators pushed this work in directions I could not have found alone. Juliana Francis turned my dream of making speech technology people could use into a reality. Shm Garanganao Almeda taught me how to moonlight as an HCI researcher. William Agnew, for being a staunch supporter. Dr. Alexander Lin and Jaclyn Kukoff, for our work together on nasal technology. I'd like to thank my collaborators at SRI. Sarah Bakst, in particular, whose excitement motivated me to finish our modification work, and Erica Gold, who provided a much-needed perspective on speech forensics.

The master's students and undergrads I worked with throughout the years deserve special thanks. Andrea Guzman, who made many of the papers in this thesis possible. Naomi Carvalho, without whom the perception studies could not have happened. Xavier Yin, whom

I would gladly talk phonetics with at any point. David Babazadeh, who let me pretend to do music processing research. Faith Qiao, who always impressed me with her insight and drive.

I'd like to thank the broader speech science and technology community. Juliana Francis (again), Fran Pessanha, Cliodhna Hughes, Nina Markl, and Éva Székely for being largely responsible for making a space for queer and trans speech research internationally. Shaomei Wu, who broadened my perspective on what inclusive speech technology means. Maxwell Hope and his collaborators for doing the awesome data collection work that helped power my research. Odette Scharenborg, who, with a simple passing question at a conference, made me feel like I had a home in speech science. Nicholas Cummins, for making me feel like my objections to speech technology were not made in a vacuum. Nicole Holliday, for her cutting insight into the problems with speech technologies.

The Yu group shaped me as a person and statistician. But, largely as a person. Zach Rewolinski, Tiffany Tang, Abhineet Agarwal, Spencer Frei, Aliyah Hsu, Raaz Dwivedi, Rebecca L. Barter, Angela Zhou, Briton Park, James Duncan, Corine Elliot, Yan Shuo Tan, and Chandan Singh.

My friends at Berkeley made the entire experience unforgettable. Alok Tripathy, who taught me that it doesn't matter how well you do at trivia as long as you show up. Nicholas Tomlin, without whom this thesis would have taken at least another year or two. Rudy Corona, without whom my PhD would have been much, much less fun and tolerable. Frances Ding, Zoë Bell, Samantha Robertson, who were there for me during one of the most vulnerable times in my life. Serena Caraco, a found-sister and staunch, staunch friend. Riley Hayes, for all the road trips and solicited advice. Samantha Johnson, 'cause we made a home out of a hotel.

Old friends, who kept me tethered to the world outside of Berkeley. Andrea Floersheimer, for the random trips and calls and memories. Sofia Schembari, for the years of support. Hamed Nilforoshan, for reaching out randomly when I least expect it. Ryan Chakmak and Katie O'Neill, for being a port to come home to whenever I felt lost. William Neme-Micula. Hopefully now that this thesis is done, we can spend more time together.

Seven years is a long, long time. There are so many people I have to thank that I may have forgotten, so many people whom I wish to thank, but some things are best kept out of the light. Thank you all for the stolen moments.

Thank you to my loved ones. Always there for me, making the impossible possible.

Thank you to my sister. A grounding force of nature in a turbulent world.

Thank you to my Mom, who taught me how to love.

Thank you to my Dad. Things, as always, can only get better.

We may have to wait for speech science to provide better answers to some basic questions. The first point to note is that the acoustic theory of speech production, whether simplified or made complex by the introduction of better models of the larynx, trachea, and source-filter interactions, is not intended to be a model of the parameters directly controlled when we speak, nor of the parameters directly involved in the perceptual decoding of speech. The theory is a description of the acoustic behavior of a mechanical system. Therefore, efforts to relate observed spectral data from real female talkers to formant frequencies and other acoustic parameters of the theory have no a priori reason to succeed, and actually stand a good chance of failure, in part because there are too many model parameters compared with available spectral details (especially for talkers with high fundamental frequencies). Are we in a situation where it is possible to collect spectral data, yet be unable to relate it unambiguously to the underlying generation process, or to the processes of speech perception or articulation?

If this characterizes the present state of speech science, and I think it does, then the real bottleneck is the absence of a satisfactory perceptual theory to account for listeners' behavior in terms of observable spectral or waveform details. That we are far from such a theory is obvious, but how to go about attaining one is less clear. Attempts to mimic the steps believed to occur during the encoding stages of peripheral auditory processing are attractive as a first step, but it is unlikely that this encoding alone will be able to explain all of the fundamental perceptual skills that come naturally to humans, but not to speech recognition devices.

– Daniel Klatt, *Review of text-to-speech conversion for English* (1987)

Chapter 1

Introduction

Do you like the way your voice sounds? Maybe your voice is too feminine, or not feminine enough. Maybe it sounds too old, or too young. Maybe you simply want to understand what it is about your voice that makes it sound the way it does—and whether you could change it if you wanted to. These are not idle questions. For trans and gender-diverse individuals, the ability to modify one's voice is a matter of safety, identity, and daily life. For aging speakers or those with voice pathologies, understanding vocal change is a matter of health. And for all of us, the voice is the instrument we are born with, underpracticed by most and understood by few.

Experts across disciplines, such as phonetics, speech-language pathology, speech forensics, and gender-affirming voice care, have developed rich vocabularies for describing what makes a voice sound the way it does [1, 2, 3]. A speech-language pathologist hears *strain* or *breathiness*. A voice teacher hears *resonance* or *weight*. These perceptual voice qualities (PQs) are the tools by which experts diagnose, teach, and modify voices [4, 5]. They are low-dimensional, interpretable, and grounded in human perception. But they have remained largely outside the reach of speech technology until recently [6, 7, 8].

Speech processing systems, by contrast, represent the voice as statistics over waveforms and spectrograms [9]—or, increasingly, as high-dimensional learned embeddings from self-supervised models [10, 11]. These representations are powerful: speaker verification systems achieve sub-1% equal error rates, and voice conversion systems produce natural-sounding speech in unseen voices [12, 13]. But they offer no interpretability. If you wanted to make your voice less feminine, such a system might tell you to move 0.3 units along the 2nd principal component of a 768-dimensional embedding. How would you do that with your body? What would it sound like? The representations that enable machine performance are precisely the ones that prevent human understanding.

This thesis is built on a single conviction: the right level of abstraction for understanding and modifying speaker identity already exists—not in acoustics, not in deep learning, but in human perception. Perceptual voice qualities occupy exactly the intermediate level needed. They are low-dimensional enough to be interpretable, perceptually grounded enough to be meaningful, and acoustically and physiologically substantive enough to drive modification. The question is whether we can build computational systems that operate at this level while

preserving the aspects of identity that PQs do not capture.

We study the perceptual grounding of speaker identity: how the voice is produced, how it is perceived, and how it can be modified along interpretable axes. The approach is empirical and interdisciplinary, drawing on voice science, crowdsourced perception, and generative modeling. What unifies the work is a commitment to representations that are simultaneously meaningful to humans and useful to machines. The thesis proceeds in three stages:

1. **Production and Perception (Chapter 2).** We present two datasets: the Versatile Voice Dataset (VVD) and the Disentangled Source-Filter Dataset (DSFD). These datasets capture the flexibility of the human voice via expert gender-affirming voice modification, demonstrating that single speakers can produce voices spanning the full perceptual space of gender. Through crowdsourced studies, we establish that perceptual voice qualities are reliably perceivable by non-experts, that gender perception is stable across task types, and that age perception is ordinally reliable under pairwise comparison.

2. **Modeling Voice Quality and Speaker Identity (Chapter 3).** We demonstrate that state-of-the-art speaker identity models fail systematically under intra-speaker vocal modification, establish disentangled acoustic correlates of perceptual qualities on a single speaker, and present the PQ-Encoder: a model that predicts a compact 9-dimensional PQ-vector from speech. We show that this representation captures meaningful information about speaker identity, gender, age, and affect—occupying an interpretable subspace of what high-dimensional embeddings encode opaquely.

3. **Interpretable Modification of Speaker Identity (Chapter 4).** We introduce StyleMod, a conditional flow matching model that modifies speaker identity along perceptual voice quality axes. StyleMod enables interpretable and continuous manipulation of perceived gender and age while preserving non-targeted speaker attributes such as accent, demonstrating that perceptual grounding yields not only better understanding but practical control.

Together, this work argues that grounding speech technology in human perception yields representations that are both scientifically meaningful and practically useful. The voice is not a fixed property of a speaker. The voice is a flexible instrument, and perceptual voice qualities are the language for describing how it is played.

Chapter 2

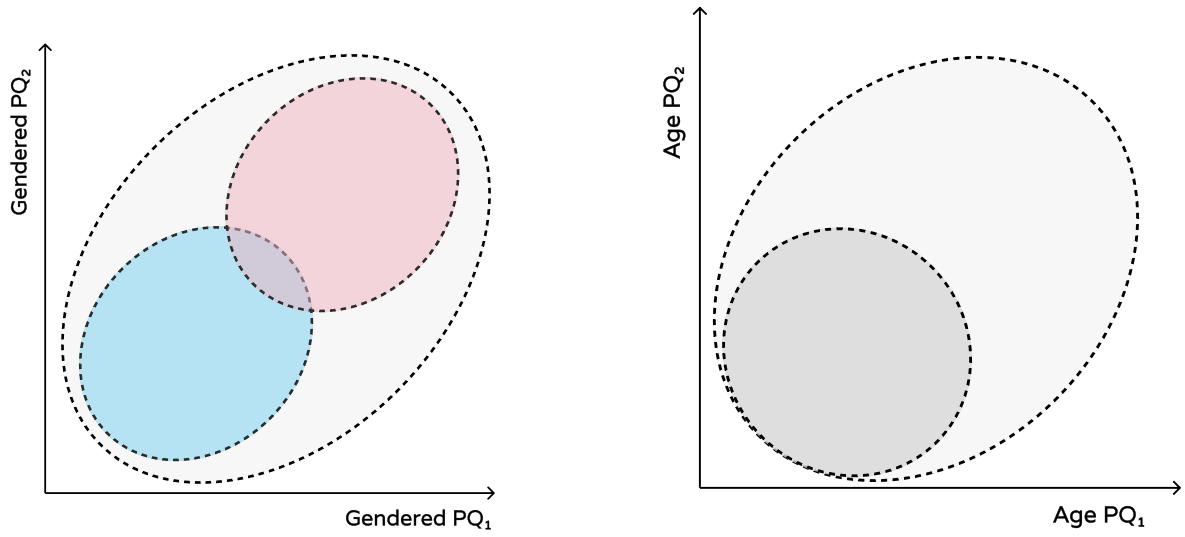
Production and Perception

2.1 Introduction

Speech processing views the human voice as a source-filter system. The vocal folds produce a quasi-periodic source signal, which is then shaped by the resonances of the vocal tract [14]. The acoustic theory of speech production describes this mechanical system in detail, yet—as Klatt observed nearly four decades ago—the theory is not intended to be a model of the parameters directly controlled when we speak, nor of the parameters directly involved in the perceptual decoding of speech. The gap between what can be measured acoustically and what listeners actually hear remains one of the central challenges of speech science. Nowhere is this gap more evident than in the perception of speaker identity. Listeners make rapid, reliable judgments about who a speaker is and what demographic categories they belong to [15], yet the acoustic basis of these judgments, particularly for gender, age, and vocal health, remain understood [5].

Voice quality, or perceptual voice quality (PQ) as we refer to it in this work, offers a bridge between production and perception. Drawn from speech pathology [2], speech forensics [1], and gender-affirming voice training [5], PQs are standardized descriptors of how a voice *sounds*–not the acoustic measurements themselves, but the subjective impressions that listeners form from those measurements. We consider nine PQ dimensions, hypothesized to span two primary subspaces of vocal variation. The first (pitch, resonance, weight, and creak) capture gendered dimensions of voice, corresponding to modifications of the source (vocal fold mass, fundamental frequency) and filter (vocal tract shape). The second (strain, roughness, breathiness, loudness, and nasality) capture dimensions related to vocal health and aging.

Figure 2.1 illustrates the hypothesized structure of these subspaces. Typical voices cluster in predictable regions defined by demographic norms (masculine voices being low in pitch, resonance, and high in weight while feminine voices high in pitch, resonance and low in weight in these qualities) while atypical voices, including those of gender-diverse individuals, speakers with voice pathologies, and elderly speakers, occupy regions that current speech processing systems do not well represent. A central goal of this chapter is to validate this



(a) Gendered PQ Space

(b) Age PQ Space

Figure 2.1: Hypothesized voice type spaces separated by Gender and Age/Health. Typical voices occupy the darker regions; atypical voices (gender-diverse, vocal health) extend into the lighter regions.

hypothesis: that PQs vary systematically across demographics and that both expert and non-expert listeners can reliably perceive these variations.

To this end, we present two datasets: the Versatile Voice Dataset (VVD) and the Disentangled Source-Filter Dataset (DSFD). These datasets capture the flexibility of the human voice via expert gender-affirming voice modification, demonstrating that single speakers can produce voices that span the PQ space along gendered axes. We then establish that perceptual voice qualities are reliably perceivable through crowdsourced studies on the PVQD+ dataset. Third, we explore listener perception of gender and age, finding gender judgments remarkably stable across task types and age judgments reliable when framed as pairwise comparisons. Together, these production and perception results provide the data and empirical foundation for the computational modeling and modification work in Chapters 3 and 4.

2.2 Background

Perceptual Voice Quality

Defined as the acoustic "coloring" of an individual's voice, perceptual voice quality, hereby referred to as perceptual quality (PQ), has long been studied in speech language pathology

and processing [16, 17]. From the mood of a voice to vocal fry and breathiness, perceptual quality consists of the subjective perceptions of a voice. Perceptual qualities have long been noted as being important to spoken language processing, with prior work noting that voices with uncommon or pathological perceptual qualities lead to poor performance for spoken language processing systems if not taken into consideration [17].

Of particular interest is how perceptual quality can be used to describe an individual's voice. In speech language pathology (SLP), experts will use vocal quality to perform initial diagnosis of the health of an individual's voice [2, 18]. Non-surgical treatment of a voice involves a patient performing exercises to bring certain perceptual qualities into healthy levels [2]. PQs that are highly correlated with dysphonic speech are particularly useful for voice rehabilitation, but others, such as vocal fry, timbre/resonance, and weight, see application in general voice modification as well. Musical vocal teachers, SLPs, or voice teachers specializing in voice feminization/masculinization will use these other PQs to guide students towards target voices [19].

As the speech processing community often uses voice quality to mean the quality of synthetic audio, and perceptual approaches to measure voice quality are the gold standard, we refer to voice qualities as Perceptual Voice Quality (PQ). Three types of PQ research exist: Acoustic, Articulatory, and Perceptual. There are limited studies on PQ from a sociolinguistic perspective, with most research coming from speech language pathologists and clinicians. That said, PQ is *highly* idiosyncratic, with users' particular configurations resulting in perceptually different versions of a voice.

Prior work in automatic assessment of dysphonia [20] and emotion recognition [21] has explored the predictability of individual perceptual qualities. Deep neural network approaches to predicting perceptual qualities have been conducted [22, 23], but these have primarily focused on labeled audio clips of sustained vowels with the goal of predicting dysphonic voices from spectral features and the waveform directly.

We note that there are a variety of limitations to existing work on PQ, primarily due to the sheer amount of research that remains to be done. Overlap in existing methods' PQs, like Voice Profile Analysis, has been noted in prior work. The amount of time needed to analyze a voice is also unclear. Proposed acoustic or "objective" connections to Voice Quality fail to generalize in an interpretable manner.

Gender Perception

The gender of a speaker is one of the first assumptions a listener makes, often in less than 200ms [24]. Understanding the acoustic correlates that contribute to gender perception, such as F_0 and formant measures, has long been a standing field of research in phonetics, especially in trying to understand how listeners separate content from speaker information [25].

Interest in gender-diverse voice science and care has caused a resurgence of research in this field, whereby trans and queer speakers alter features of their voice to align with a desired gender presentation [3]. While scalar and categorical perceptual evaluations of

gender have been shown to be stable across gender-diverse and cis-gender individuals for "unambiguous" voices [26], gender-diverse and cis-gender individuals have been shown to perceive "gender-ambiguous" voices differently, with queer individuals utilizing more diverse language and broader gender ratings to describe voices [26, 27]. Although true for cis-gender voices, coupled with gender-diverse samples, F_0 and formant information alone has been shown to be unable to fully account for speaker gender [25, 28], especially for speakers within a gender ambiguous range [29].

As teaching tools, Pitch, Resonance, and Weight have been incredibly important for gender-affirming voice modification [5, 30], allowing single speakers to produce voices that span the entirety of the gender perception space [31]. Control and understanding of these PQs and others, like breathiness and harshness, are of utmost importance to gender-diverse people [32, 33] and expressive speech modeling more generally [34, 35, 6]. Prior work has primarily focused on inter-speaker acoustic metrics or synthetic manipulation to study acoustic metric correlation with gender perception, often finding average F_0 values to be most predictive of gender perception [36, 28]. To the best of our knowledge, there is no work which systematically explores gender perception along a range of diverse voices physically produced by a single-speaker.

Naturalness Perception

Expert listeners and speaker verification systems are adept at determining when someone is "mimicking" another individual's voice [37, 1], and when someone is "disguising" their voice [38, 39]. Natural voices modified along Pitch, Resonance, and Weight have been shown to fool automatic speaker recognition systems [31], leaving open the question of how listeners perceive the "naturalness" or "realness" of expert intra-speaker modification. Queer voices have been identified by cis-gender listeners as being more "unnatural" [40], which the GAVT community has hypothesized as being due to atypical configurations of Pitch, Resonance, and Weight [5, 33, 41], and additionally misalignments of intonation, speaking rate, and articulation with expected prototypical values [40, 36].

The Trans and Gender-Diverse Community

Despite its importance, gender-affirming voice modification is a notoriously difficult task within the trans and gender-diverse community [33], with experienced clinicians being rare [26] and gender-affirming voice training only just starting to be studied by the scientific community [4]. Due to these limitations, many trans individuals seek training from informal online sources, such as tutorial videos or private coaches [33], which we will group together as the Informal Trans Voice Training Community (IVTC). Spanning the Internet, the IVTC is a broad community across Discord servers, YouTube channels, and Reddit forums. These sources primarily consist of private individuals, many of whom are trans themselves, teaching techniques they developed or adapted from other sources for their own voice training, often without institutional or formal training [5]. Each subcommunity uses similar language to

describe perceptual qualities of voice, such as resonance and weight, but lack of formalized research leads to disagreement upon the precise importance and definitions of these qualities.

Understanding the similarities and differences between trans, gender-expansive and cis-normative speech has gained interest in recent years. Prior work has noted that trans and gender-expansive individuals tend to use more diverse vocabulary and varied scalar ratings when describing the gender of an individual's voice [26, 27]. These differences in perception are primarily attributed to trans and gender-expansive individuals' use of voice to present their gender identities.

It is important to note that many teachers in the IVTC highlight different aspects of voice than those typically emphasized in speech language pathology. Rather than an emphasis on oral cavity resonance and pitch as in speech language pathology [4], many in the IVTC choose to focus on pharyngeal resonance, and use the concept of weight to balance pitch modification [5].

2.3 Limitations of Existing Approaches

Just as gender-diversity cannot be limited to a single label of "cisgender", "non-binary", or "transgender" (as notions of gender differ across cultures and time [42]), categories of "man", "woman", and "non-binary" do not capture the full complexity of voice. Binary or categorical gender information, while helpful for defining a broad distribution over possible voices, results in large amounts of ambiguity concerning the specific texture that makes a voice unique.

By summarizing vocal diversity with a small number of identity-based categories, gender obscures the richer acoustic space that affects voice perception. While helpful for modeling many voices, categorical notions of gender can result in unreliable and possibly harmful classifications to those who deviate from prototypical voices. Knowing that a voice is a woman's, for example, does not accurately identify the unique texture of that voice. Moreover, systems based on categorical gender can be manipulated by deliberate behavioral changes to voice.

Datasets that focus on exploring the acoustic properties of trans and gender-expansive speech are only just starting to be developed. A Palette of Voices for Transmasculine Individuals [43] and the Mid-Atlantic Gender-Expansive Speech Corpus [44] are, to the best of our knowledge, the only two publicly available datasets of trans and gender-expansive speech currently available. While more exist, they are usually regional and privately held to protect the identity of trans and gender-expansive individuals [28]. Analyses of trans and gender-expansive datasets focus on exploring inter-speaker acoustic variability across acoustic measures that correlate with gender, such as pitch or formants. While prior datasets are designed to explore inter-speaker variability, we are primarily interested in exploring intra-speaker variability, as seen through the lens of gender-affirming voice modification.

2.4 Perceptual Voice Quality

PQ Space

We consider nine PQs, summarized in Table 2.1. Five (strain, loudness, roughness, breathiness, and pitch) are drawn from the CAPE-V protocol [2]. These primarily capture information about vocal atypicality and aging, but do not distinguish gendered characteristics of voice. We supplement the CAPE-V set with three qualities from the gender-affirming voice training literature: resonance, weight, and creak. Resonance corresponds to the perceived size of the vocal tract, with larger tracts producing darker timbres due to lower resonant frequencies [14]. Weight corresponds to the perceived vibratory mass of the vocal folds, correlating with the open quotient and spectral slope [45]. These two qualities, along with pitch, are the primary axes of behavioral gender modification [5, 31, 46]. Creak, defined as aperiodic low-frequency vibrations of the vocal folds, is included as a ninth dimension due to its documented role in perceived gender and speaker affect [47]. Nasality, the final quality, captures audible nasal airflow and is relevant to both clinical voice assessment and certain speaker characteristics [48].

This nine-dimensional PQ space spans the major axes of vocal variation relevant to speaker identity: gender (pitch, resonance, weight, creak), age and health (strain, loudness, roughness, breathiness), and anatomical characteristics (nasality). While the space of possible voice qualities is vast, these nine dimensions provide sufficient coverage to represent the primary perceptual distinctions among adult voices and to enable the demographic-level modifications demonstrated in later sections.

2.5 Disentangled Production

Prior work has discussed the importance of Pitch, Resonance, and Weight on perceived speaker gender [31, 5], and provided expected or high-level correlations with acoustic or physical phenomena. Resonance is often assumed to be associated with the overall size of the vocal tract, but prior work has not demonstrated this correlation with a single metric [36, 31, 6]. Weight, believed to be associated with the thickness of the vocal fold vibratory mass and the spectral shape of the glottal flow [49], has been hypothesized to influence gender perception via glottal closure and Loudness, but has yet to be strongly correlated with concrete metrics [30, 50, 31]. The connections between individual acoustic and physical metrics and perceptual qualities, as measured on the DSFD, are explored in Chapter 3.

Pitch, Resonance, and Weight

As these qualities have been found to be extremely important in gender perception [51], pitch, resonance and weight are the most commonly used concepts in the IVTC to teach behavioral voice modification. Connecting the concepts to acoustic and physical phenomena,

PQ	Correlate	Description
Pitch	Gender	Deviation in pitch
Resonance	Gender	Sound quality of the size of the vocal tract
Weight	Gender	Sound quality of the vocal fold vibratory mass
Creak	Gender	Aperiodic vibrations of the vocal folds
Strain	Age/Health	Perception of excessive vocal effort (hyperfunction)
Loudness	Age/Health	Deviation in loudness
Roughness	Age/Health	Perceived irregularity in the voicing source
Breathiness	Age/Health	Audible air escape in the voice
Nasality	Health	Audible air escape through the nasal cavity

Table 2.1: The complete list of Perceptual Voice Qualities considered in this thesis. Resonance and weight are taken from gender-affirming voice training, and the other qualities are taken from Speech Language Pathology and Speech Forensics.

pitch is understood as the fundamental frequency, resonance as the acoustic qualities of the overall vocal tract shape (pharyngeal, oral, and nasal) [14], and weight as the sound quality associated with the vocal fold vibratory mass, which correlates with the closed quotient [5].

Perceptual resonance corresponds to the amount of space above the vocal folds in the vocal tract. More space causes lower resonant frequencies to be amplified [14], resulting in a deeper timbre, even at high pitches. Perceptual weight corresponds to the vibratory mass of the vocal folds, which is correlated with the open quotient (the proportion of the glottal cycle during which the vocal folds are open) and spectral slope (the decline in amplitude from the first to the Nth harmonic) [45]. These ideas are also often used in singing lessons for vocalists, but the focus of transgender voice training on gender perception ties resonance and weight to a representation of speaker identity. Current pedagogies in transgender voice modification are particularly concerned with these two primary PQs, as they correspond to two physiological differences [51] that distinguish male and female voices.

While pitch, resonance, and weight as concepts are mostly agreed upon across the IVTC, there is more ambiguity in these terms when it comes to precise definitions and perceptual

descriptions, especially for weight. Different voice teachers describe weight differently: some describe it as a "buzzy" sound, whereas others describe it as more of a "rumble", only sounding buzzy under high adduction, high resonance configurations. These disagreements serve as stepping stones for future perceptual studies.

The Versatile Voice Dataset (VVD)

To demonstrate the flexibility of the human voice, we present the Versatile Voice Dataset (VVD), a collection of three speakers modifying their voices along gendered axes [31]. Along self-interpreted configurations of low, medium, and high for pitch, resonance, and weight, 3 trans-feminine gender-affirming voice teachers produced up to 27 voices. With each voice configuration, the voice teacher read the six CAPE-V sentences aloud [2], for a total of 14.5 minutes of speech. A summary of speaker information is provided in Table 2.2.

Speaker Id	Pronouns	Accent	Voices Produced	Microphone
001	She/Her	Californian American	27	AT2020
002	Any	Australian	25	RZ19-03450100
003	She/Her	Californian American	27	AKG P120

Table 2.2: Speaker Information for the Three Speakers in the Versatile Voice Dataset.

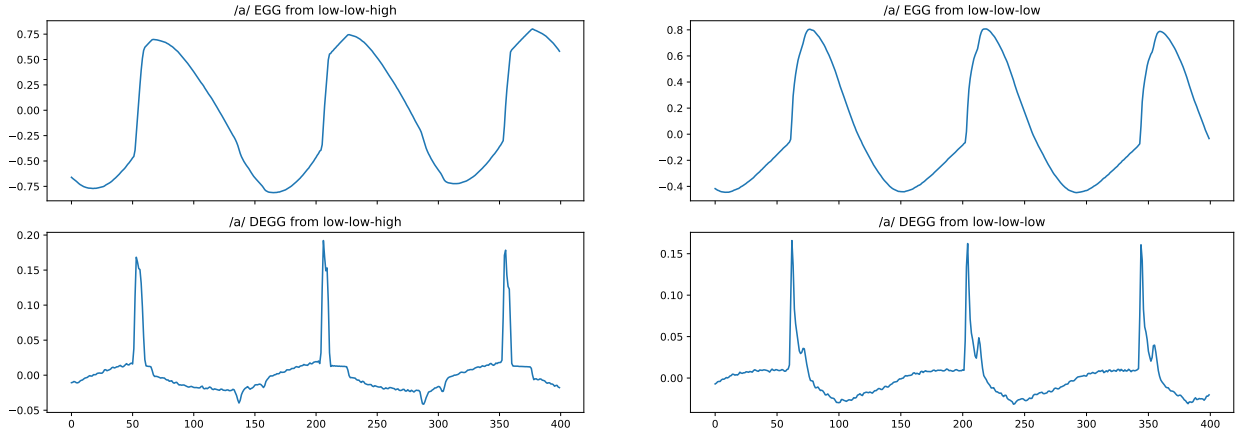
As the IVTC is broad, the three voice teachers had not interacted before the collection of this data and had not heard each other's voices. The ambiguous instructions of low, medium, and high configurations led to each teacher interpreting the terms and task slightly differently. Speaker 001, for example, speaks with a yawn-like voice in the low-low-low configuration, while Speaker 002 and 003 speak in more natural-sounding voices. Similarly, different interpretations of weight lead to voice variation, reflected in Speaker 002 omitting med-high-high and high-high-high configurations to avoid vocal strain. Without an external and consistent measurement of pitch, resonance, and weight, these discrepancies lead to difficulties in drawing conclusions on inter-speaker comparisons such as naturalness of speech across configurations. Samples from the VVD are provided online[1].

The Disentangled Source-Filter Dataset (DSFD)

Using self-interpreted axes of low, medium, and high for the PQs Pitch, Resonance, and Weight, similar to the VVD, we collected 45 minutes of read speech over 25 voice configurations from a single trans-feminine gender-affirming voice teacher with over 3 years of experience and training as a classical singer. To supplement collecting audio at 16kHz, we collect Electroglottograph (EGG) data at 16kHz, which provides a proxy for the physical movement of the vocal folds.

[1]https://berkeley-speech-group.github.io/VersatileVoiceDataset/

In line with datasets that collect articulatory and acoustic information in parallel [52], we utilize the set of 460 sentences from the MOCHA-TIMIT corpus [53], coupled with sustained phonation of /a/ and /i/ vowels. Due to time constraints, the voice teacher produced the first 30 sentences from the corpus. 25 out of the 27 possible configurations were collected, as two configurations (namely high pitch, medium/high resonance, and high weight) were excluded to avoid harmful vocal strain.



(a) Low Pitch (146Hz), Low Resonance (1027Hz), High Weight (54%)

(b) Low Pitch (156Hz) - Low Resonance (951Hz) - Low Weight (27%)

Figure 2.2: Electroglottograph (EGG) and its derivative (DEGG) visualized from Low Pitch ($\bar{F}_0$), Low Resonance (ΔF), and High/Low Weight (CQ) /a/ sustained phonation.

2.6 Perception of Perceptual Qualities

The Perceptual Voice Qualities Database

In clinical settings, perception assists greatly in the early stages of diagnosis of voice pathologies. Speech Language Pathologists (SLPs) will often use rating scales like the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) to provide early information on possible voice pathologies an individual may have [2]. SLPs undergo training to successfully identify the PQs: strain, loudness, roughness, breathiness, pitch, and severity [18].

While this data is difficult to collect, the Perceptual Voice Qualities Database (PVQD) serves as publicly available ratings of PQs from the CAPE-V scale [18]. The PVQD includes 296 audio files of around 30 seconds of audio, whereby a speaker follows the CAPE-V evaluation protocol, and reads six sentences and produces vowels /a/ and /i/ for 1-2 seconds. The authors behind the PVQD had each audio clip rated by three separate clinicians across two trials according to the CAPE-V scale. We examine five of the six PQs, excluding severity, which is an overall measure of vocal atypicality.

Rater	Res.	Wt.	Strain	Loud.	Rough.	Breath.	Pitch	Avg.
Non-Experts	0.68	0.76	0.81	0.67	0.79	0.87	0.79	0.77
Experts	0.91	0.77	0.83	0.87	0.79	0.83	0.86	0.84

Table 2.3: Correlation of Non-Expert Labels with Expert Labels. A measure of inter-rater agreement, Intra-class Correlation Coefficient, is reported for Expert Labels.

Collecting Gendered Perceptual Qualities

While the PVQD serves its purposes as a diagnostic tool of speech pathology, for the purposes of providing a general representation of voice, it is incomplete. Information on a voice's gender measures is missing. Attempting to perform manipulations that are common in speech processing tasks like voice conversion, such as converting a masculine voice to a feminine voice, would not be possible with the CAPE-V scale's ratings of deviation alone.

As such, we augment the labels with those provided by three voice teachers, who specialize in transgender voice training. As discussed in Section 2.5, current pedagogies in transgender voice modification are particularly concerned with two primary PQs: vocal resonance and vocal weight.

Three voice teachers listened to subsets of 100 audio clips from PVQD and provided a label of resonance and weight on a scale 1-100. For resonance, a value of 1 represented the darkest resonance possible and a value of 100 represented the brightest resonance possible. Similarly for weight, a value of 1 represented the lightest voice possible and a value of 100 represented the heaviest voice possible. Healthy feminine voices were given a resonance value of 90 and a weight value of 10, and the opposite for healthy masculine voices. We note that one voice teacher labeled the entire dataset, and the two other voice teachers overlap on 25 audio clips to allow for the calculation of averages and correlations. We call the extended dataset PVQD+.

Along with those reported in the original PVQD, we report the intra-class correlation (ICC) for expert ratings of resonance and weight in Table 2.3. ICC is a common metric for measuring inter-rater reliability [18]. We see that the expert ICC of both resonance and weight are the maximum and minimum of the ICCs for all PQs, respectively. The high inter-rater reliability for resonance is unsurprising, given that vocal tract size is highly correlated with gender. Weight is less clear. While an ICC of 0.77 still suggests high inter-rater agreement, weight having the lowest agreement among experts suggests that labelling weight is a more difficult task. Regardless, these results demonstrate the high agreement of experts on the gendered perceptual qualities resonance and weight.

To further assess expert agreement on the gendered PQs, we additionally asked two of the three voice teachers in the VVD to perform pairwise ranking comparisons across voice configurations: Given two clips differing in a PQ level, to identify which clip is higher in that quality. Voice teachers achieve pairwise ranking accuracies of 93.8% for resonance and

74.5% for weight. The high ranking accuracy for resonance, coupled with its high ICC of 0.91, suggests that resonance is a well-defined and consistently perceived quality. Weight, while still achieving reasonable ranking accuracy, shows both lower ICC (0.77) and lower ranking accuracy, reinforcing the observation that weight is the most difficult of the gendered PQs to consistently label.

Can Non-Experts Hear Perceptual Qualities?

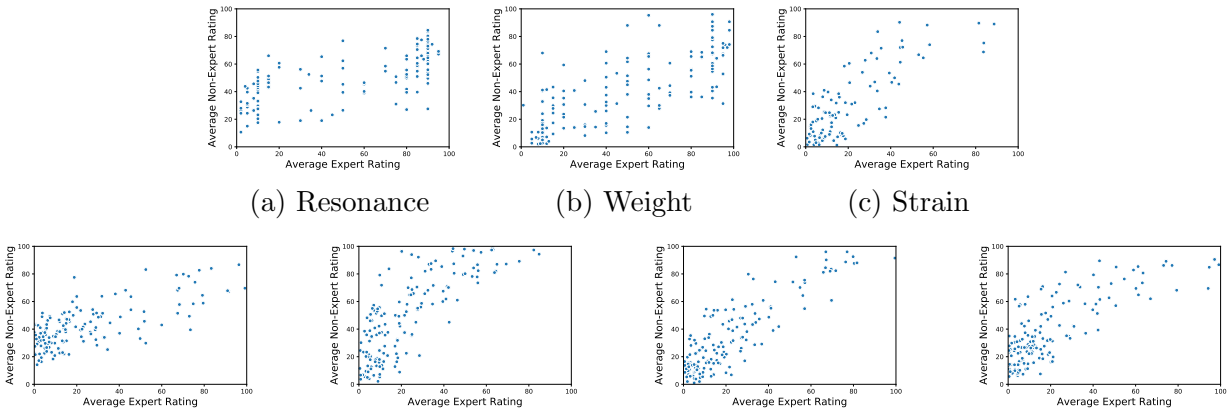
A common statement from voice teachers and speech language pathologists is that hearing perceptual qualities requires training [54, 5]. The obstacle of expertise calls into question the utility of perceptual qualities as a representation of speaker identity, since collecting additional labels would be a costly and time-consuming task [22]. Recent work has demonstrated that non-experts can accurately label the overall atypicality of a voice [55], but no work has explored the ability of non-experts to label specific perceptual qualities. If non-experts can accurately label PQs, collecting a large-scale and high-quality dataset of perceptual qualities and using perceptual qualities as an evaluation metric of speaker modification systems would be possible.

Using Amazon Mechanical Turk (AMT), we asked 6 workers with master's qualifications to rate the clips using the CAPE-V protocol, and the resonance and weight as described in Section 2.6. Workers were provided two examples for each perceptual quality, one example being low in that quality (ex. No Strain) and the other being high in that quality (ex. High Strain). Due to cost constraints, workers labeled only 150 audio clips of the 296 audio clips in PVQD.

The results of the AMT experiments are very promising for the ability of non-experts to hear perceptual qualities. As Table 2.3 demonstrates, average non-expert ratings achieve a correlation of 0.77 with average expert ratings, with the lowest correlations being 0.68 and 0.67, for resonance and loudness respectively (loudness was often conflated with audio clip volume, not inherent loudness of the voice). Other PQs, especially breathiness with a correlation of 0.87, are easier for non-experts to hear and rate. In terms of agreement with experts, non-experts achieve a surprising level of performance.

While the correlation is remarkably high between non-experts and experts, RMSE suggests a high level of deviance between non-experts and experts. While experts have an average standard deviation amongst themselves of 10.47, average non-expert RMSE with expert ratings is 23.45, slightly over twice that of experts.

For CAPE-V PQs, which are all measures of deviance, non-experts consistently overrate speech clips as having higher levels of deviance. But, for those clips that experts label as having high levels of deviance, non-experts will always label as having high levels of deviance as well. Regarding the gendered PQs of resonance and weight, higher levels of deviation between non-expert and expert ratings are observed, but similar trends hold. Non-experts appropriately label voices with high values of a perceptual quality with a high value, and vice versa.



(a) Resonance (b) Weight (c) Strain

(d) Loudness (e) Roughness (f) Breathiness (g) Pitch

Figure 2.3: Average Expert Rating (x-axis) vs. Average Non-Expert Rating (y-axis) across perceptual qualities.

Further training or better explanation is most likely required to improve non-expert and expert agreement on the gendered PQs. These results, however, have shown that the ability of non-experts to hear perceptual qualities with minimal prompting is remarkably high and the promise of collecting mass perceptual quality data is well within reach.

2.7 Perception of Demographics

Perception of Gender

Data Collection

Using crowdsourcing platform Prolific to recruit fluent English participants, we asked participants 4 questions on the perception of gender, naturalness, and realness in voice, an overview of which is provided in Table 2.4. Each of the individual trials focused on a different facet of perception from the DSFD across the first two sample sentences from MOCHA-TIMIT: "This was easy for us." (Sentence 1) and "Is this seesaw safe?" (Sentence 2). 660 participants rated 5 voice samples at an average pace of 25 seconds each. Each configuration and sentence sample was rated 15 times, resulting in 30 ratings per configuration.

Gender Scale vs. Gender Classification

Gender perception studies were split into two different tasks: scalar rating and classification. For scalar ratings, participants were prompted with a voice sample and asked to rate their perception of the gender of the voice on a scale from 1 to 10, where 1 is most masculine and 10 is most feminine. By contrast, classification studies were intentionally designed to elicit a

Question	Avg.	Std.
1 (Very Masc) to 10 (Very Femme)	6.68	1.96
Male (0) or Female (1)?	0.70	0.32
1 (Very Natural) to 10 (Very Unnatural)	5.26	0.48
Fake (0) or Real (1)?	0.63	0.10

Table 2.4: Overview of Perceptual Questions and the average and standard deviation of the responses.

split-second judgment from the user, asking them "Is this voice male or female?". Notably, gender classification tasks took 16 seconds less than gender scaling on average.

Despite their difference, correlating the scalar rating with the ratio of Female to Total Votes resulted in a 0.95 correlation, demonstrating a stability of gender perception across tasks. Accounting for listener demographics across age, gender, and ethnicity found no statistically significant difference in ratings between groups, implying stability of gender perception across demographics as well. While voices spanned the gender spectrum, samples from the DSFD skewed more feminine on average.

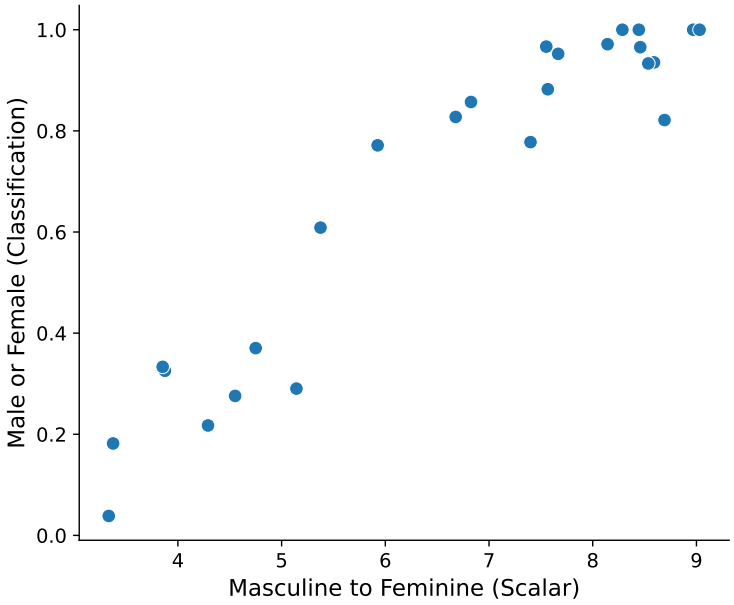


Figure 2.4: Gender classification (ratio of Female votes) vs. gender scalar rating across voice configurations, demonstrating a 0.95 correlation between the two perceptual tasks.

Perception of Age

Data Collection

While gender perception has received considerable attention in the voice perception literature, age perception from voice remains comparatively understudied, despite its relevance to speaker identity modeling. To investigate how reliably listeners can estimate a speaker's age from voice alone, we conduct a crowdsourced age estimation study following a similar methodology to the gender perception study in Section 2.7.

We selected 20 speakers spanning a wide age range: 10 from the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [56] and 10 from the Voice Cloning Toolkit (VCTK) [57]. Two datasets were used to ensure sufficient age diversity, as older speakers are underrepresented in most speech corpora. Speakers were selected to span ages from the 20s to the 70s, with roughly balanced gender representation. Using the crowdsourcing platform Prolific, we recruited 30 fluent English participants and asked each to estimate the age of a speaker on a scale from 5 to 80 years old after listening to a short speech clip. Each speaker was rated by all 30 participants, yielding 30 age estimates per speaker.

Age Instability

In contrast to gender perception, where scalar and classification tasks showed a 0.95 correlation (Section 2.7), scalar age estimation proves far less stable. Figure 2.5 plots the average estimated age against the ground truth age for each speaker. While a general positive trend is visible, the estimates exhibit substantial variance and systematic compression towards the middle of the age range: younger speakers tend to be overestimated while older speakers are slrightly underestimated. For all ages, inter-rater labels are highly variable per speaker, calling into question the usefulness of scalar age ratings.

This variance motivates the use of *pairwise ranking accuracy* as an alternative evaluation metric for age-related perceptual tasks. Rather than asking listeners to assign an absolute age value, which is prone to calibration errors and individual bias, ranking accuracy asks only whether one voice sounds older than another. This relative judgment is more robust. Listeners who systematically over- or underestimate age will still correctly rank pairs, provided their ordinal perception is intact. Converting the scalar estimates into pairwise rankings, we find that listeners achieve a ranking accuracy of 85.5% across all speaker pairs, rising to 94.1% when the age gap is at least 10 years (Table 2.5). The contrast between the noisy scalar estimates in Figure 2.5 and the better performing ranking accuracies confirms that listeners possess reliable relative age perception even when their absolute calibration is poor. Listeners' ability to perform ranking tasks directly motivates model training and evaluation methods to follow.

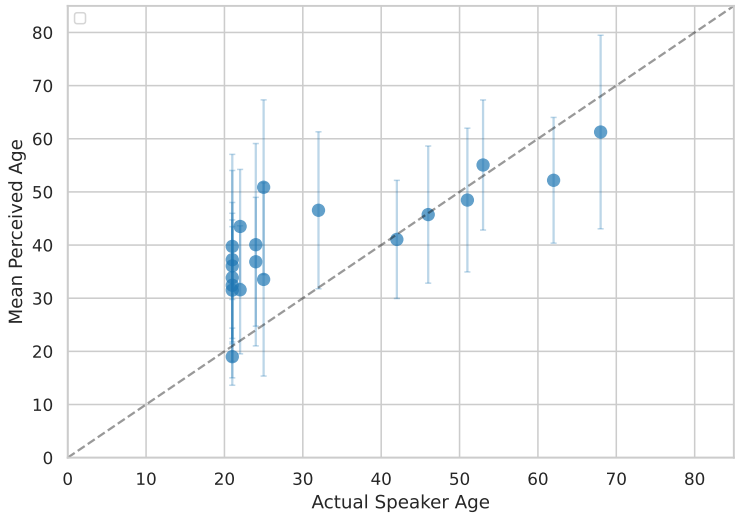


Figure 2.5: Ground truth age vs. estimated age across speakers from CREMA-D and VCTK. While a positive trend is visible, scalar estimates exhibit substantial variance and compression towards the middle of the age range.

Age Difference (Years)	Ranking Accuracy
> 0 Years	85.5%
≥ 10 years	94.1%

Table 2.5: Pairwise ranking accuracy of crowdsourced age estimates, conditioned on the ground truth age gap between speakers.

2.8 Discussion

This chapter has established three foundational results for the remainder of this thesis. First, through the VVD and DSFD, we demonstrated that single speakers can produce voices spanning the full perceptual space of gender via deliberate modification of pitch, resonance, and weight—a degree of intra-speaker flexibility that challenges assumptions underlying most speech processing systems. Second, through crowdsourced perception studies, we showed that both expert and non-expert listeners can reliably perceive perceptual voice qualities: experts achieve high inter-rater agreement (average ICC of 0.84), while non-experts achieve an average correlation of 0.77 with expert ratings given only minimal instruction. Third, we established that gender and age perception behave differently: gender perception is remarkably stable across scalar and classification tasks (0.95 correlation), while age perception is ordinally reliable (85.5–94.1% pairwise ranking accuracy) but poorly calibrated in scalar rating terms.

The perceivability of PQs by non-experts is a critical enabler for subsequent chapters. If perceptual qualities required years of clinical training to identify, their utility as a representation of speaker identity would be limited to expert-only settings. The finding that crowdsourced workers can rate PQs with high agreement, particularly for breathiness (0.87), strain (0.81), and pitch (0.79), opens the door to large-scale PQ data collection and the use of PQs as evaluation metrics for voice modification systems.

However, the results also reveal important asymmetries in the PQ space. Resonance emerges as the most consistently perceived gendered quality, with the highest expert ICC (0.91) and ranking accuracy (93.8%). Weight, by contrast, shows the lowest expert ICC (0.77) and ranking accuracy (74.5%) among the gendered PQs, suggesting that it is either less well-defined as a perceptual construct or more difficult to isolate from co-varying qualities like loudness and roughness. This asymmetry persists throughout the thesis: weight proves to be the most difficult quality to model acoustically (Chapter 3).

The lack of standardization across the GAVT community further complicates the picture. Different voice teachers in the VVD interpreted "low," "medium," and "high" configurations differently (Section 2.5), and disagreements on the precise definition of weight ("buzzy" versus "rumbly", Section 2.5) reflect the broader absence of formalized research on GAVT and gendered PQs. These disagreements motivate a shift from absolute scalar labels, which require a shared calibration that does not yet exist, toward pairwise comparison labels that ask only which of two voices is higher in a given quality. The success of pairwise ranking for both PQ assessment (93.8% for resonance) and age perception (94.1% for large gaps) validates this approach as more robust to individual differences in scale usage. This insight directly informs the training methodology of the PQ-Encoder in Chapter 3, which adopts a comparison-based objective for the gendered PQs.

Looking forward, the datasets and perceptual findings presented here provide the empirical foundation for the computational work that follows. The VVD and DSFD supply the intra-speaker variation that exposes the limitations of current speaker identity models. The PVQD+ provides expert PQ labels for training and evaluation. And the crowdsourced perception studies establish the ground truth against which model predictions can be validated. The question now shifts from whether PQs are perceivable and meaningful to whether they can be robustly predicted and leveraged for interpretable modification of speaker identity.

Chapter 3

Modeling Voice Quality and Speaker Identity

3.1 Introduction

Chapter 2 established that perceptual voice qualities are reliably perceivable, that expert and non-expert listeners agree on their relative levels, and that the human voice is far more flexible than commonly assumed. The natural next question is computational: how should we *model* the relationship between voice quality and speaker identity? Current speaker identity models treat voices as static points in a high-dimensional embedding space, optimized to separate speakers but not to explain what makes them sound different.

In this chapter, we examine how speaker identity has been modeled in the past, demonstrate the limitations of current approaches when faced with diverse voices, establish acoustic correlates of perceptual voice qualities, and explore how these qualities can ground speaker identity models in perception. The work presented here draws from and extends several prior publications: the VVD speaker identity analysis and gender perception study [31], the acoustic-perceptual correlations on the DSFD [46], and the multi-speaker probing experiments [6].

The chapter proceeds in four stages. First, we show that state-of-the-art speaker embeddings break down under intra-speaker vocal modification, failing at both gender classification and speaker verification (Section 3.3). Second, we establish disentangled acoustic correlates of pitch, resonance, and weight on a single speaker, and demonstrate that these correlates predict gender perception (Sections 3.4–3.5). Third, we probe multiple feature representations for their ability to predict perceptual qualities across speakers, motivating the choice of self-supervised features (Section 3.6). Finally, we present the PQ-Encoder, a model that predicts a compact 9-dimensional PQ-vector from speech, and validate that this representation captures meaningful information about speaker identity, gender, age, and emotion (Section 3.7).

3.2 Background

Speaker Identity Modeling

Modeling speaker identity is a fundamental task in speech processing, with applications to downstream tasks like voice conversion [58], speaker recognition [59], diarization [60], and anonymization [61, 62]. Voices are often treated as unique identifiers of speakers.

Historically, speaker identity modeling seeks to identify speakers across a wide range of variability. Unseen speakers, accents and dialects, and different microphone conditions can greatly affect the performance of speech systems across a variety of different tasks [63, 64]. In recent years, with the advent of large models, large foundation models have been developed to robustly capture differences between speakers, environments, and recording devices [13, 65]. In practice, this results in high-dimensional embeddings, such as the X-Vector [59] and ECAPA-TDNN [12].

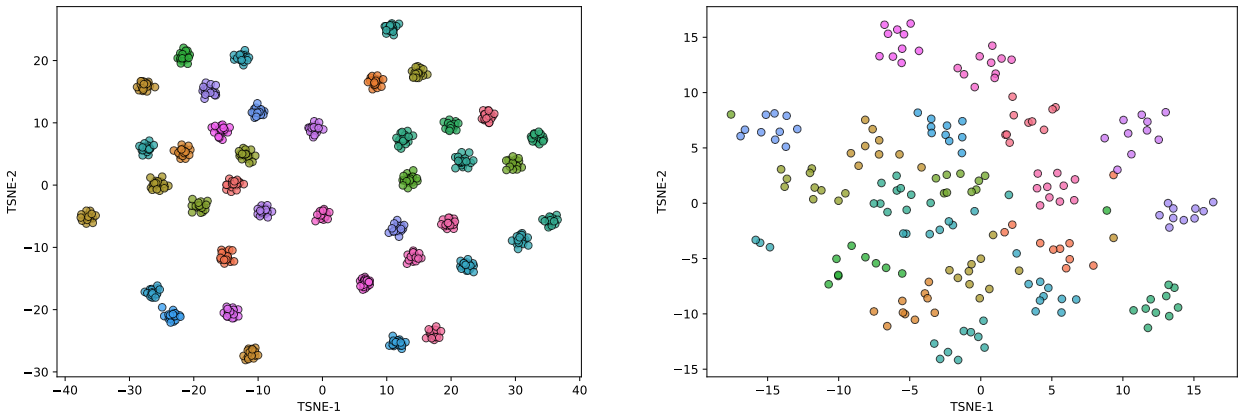
As accuracy across diverse tasks improves, the question naturally arises of whether these current models are sufficiently robust to the variability of the human voice. Descriptions of speaker identity are considered static, demographic categories such as 'Male/Female', 'Child/ Adult/Elderly', or 'Healthy/Pathological'. While there is much work to produce new voices and manipulate speaker or style embeddings in voice synthesis and modification [13, 66], study of intra-speaker variability has been constrained to "emotive" speech [67].

An example of running a 2-dimensional TSNE of SpeechBrain's ECAPA-TDNN model of 10 audio clips from 20 random speakers from VCTK and CREMA-D is provided in Figure 3.1. For VCTK, where a single speaker reads aloud neutrally, the underlying latent space consists of highly distinct clusters of individual speakers. For CREMA-D, where speakers are asked to vary the "emotion" in their voice along six axes (Disgust, Happy, Fear, Neutral, Sad, and Angry), that highly distinct clustering begins to break down. While speakers are still largely separable in the latent space, intra-speaker variation leads to the clusters not being as well-defined.

3.3 Intra-Speaker Variation and Speaker Identity Models

The CREMA-D results in Figure 3.1 hint at the fragility of speaker embeddings under intra-speaker variation, but emotional speech represents only modest vocal flexibility, primarily involving prosody and intensity modifications that maintain other habitual vocal qualities. The Versatile Voice Dataset (VVD) presents a far more extreme challenge: speakers deliberately modifying the physical configuration of their vocal tract along pitch, resonance, and weight, producing voices that span the perceptual space of gender.

Motivating this section, we see in Figure 3.2 the same TSNE analysis applied to the VVD. Where VCTK produced tight, well-separated clusters and CREMA-D produced diffuse but still separable clusters, the VVD produces a dramatically different picture: individ-



(a) VCTK (b) CREMA-D

Figure 3.1: TSNE of SpeechBrain ECAPA-TDNN speaker embeddings on 20 random speakers from VCTK and CREMA-D. Each color represents a distinct speaker. Intra-speaker variation along emotion axes complicates modeling speaker identity statically.

ual speakers produce broad, diffuse clusters that themselves can be broken into sub-clusters corresponding to different vocal configurations. Speaker 001, for instance, has a sub-cluster that lies nearer to Speaker 002's embeddings than to her own distant configurations. While the larger-scale separation between speakers may be attributable to recording differences (different microphones and environments, as noted in Table 2.2), the within-speaker fragmentation is driven by vocal modification. Here, we explore how this level of intra-speaker variation affects state-of-the-art ECAPA-TDNN speaker embeddings [12] from SpeechBrain [68] and NeMo [69] on two tasks in speaker modeling: Gender Classification and Speaker Verification.

Gender Classification

In many speech models, gender information is used as a proxy for vocal texture, converting the high-dimensional space of vocal acoustics into the categories of Male or Female. The utility of these categories is based on prior work on gendered differences in speech [5] and the ability of gender classification models to accurately infer gender categories from speech [70]. For typical voices, such models achieve high performances, seemingly justifying gender's use as a proxy. With Scikit-Learn's default parameter settings, for example, Random Forest models predicting speaker gender in VCTK [57] from either NeMo or SpeechBrain ECAPA-TDNN embeddings achieve high accuracies of 99.1% and 97.0% accuracy on the test split of VCTK, respectively.

However, gender begins to lose its predictive power with more vocal diversity. Using these Random Forest models to predict the gender of voices in the VVD, we see in Figure

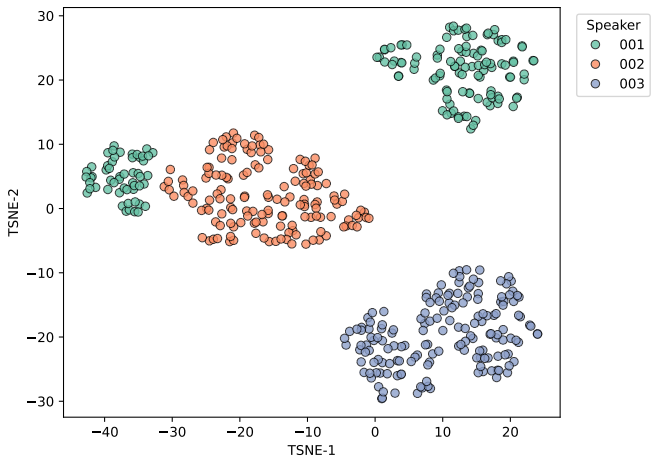


Figure 3.2: 2-dimensional TSNE of SpeechBrain ECAPA-TDNN speaker embeddings for VVD speakers. Each color represents a distinct speaker. Intra-speaker variation across vocal configurations causes significant overlap between speakers.

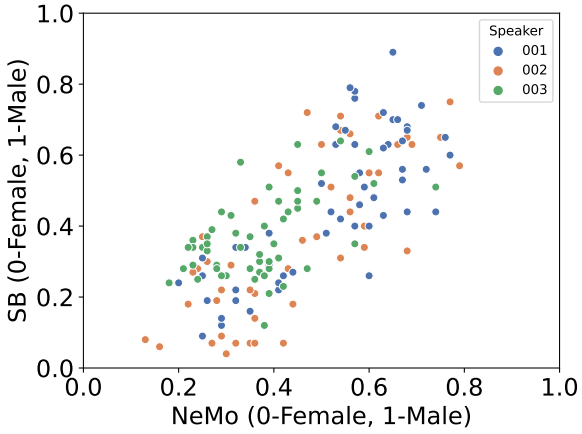


Figure 3.3: Predicted gender identity by NeMo's ECAPA-TDNN model compared with predicted gender identity by SpeechBrain's ECAPA-TDNN model across all speakers.

3.3 that these predictions span almost the entire range of 0 (Female) to 1 (Male) for both models and all speakers.

Furthermore, differences in predicted gender are not arbitrary, but the result of the deliberate modifications made by the three speakers. Shown in Table 3.1, pitch and resonance levels directly influence the gender prediction of the model, with both models predicting more feminine voices as both pitch and resonance increase. Importantly, multiple configurations produce similar predictions, revealing an ambiguity between vocal configuration and binary gender. Gender classes can delineate between voices, but cannot delineate *how* those voices

Pitch	Resonance	NeMo	SpeechBrain
low	low	0.61	0.61
	med	0.60	0.46
	high	0.51	0.38
med	low	0.58	0.56
	med	0.50	0.41
	high	0.40	0.29
high	low	0.52	0.55
	med	0.48	0.40
	high	0.32	0.25

Table 3.1: Probability of a given voice being Male averaged across configurations of Pitch and Resonance. Std. Error ≤ 0.02 for all entries, as estimated via bootstrap.

are different.

Speaker Verification

While gender classification reveals that categorical labels fail to capture intra-speaker variation, speaker verification poses a more fundamental challenge: determining whether two clips originate from the same speaker at all. If a single speaker can produce voices that span the gender spectrum, do state-of-the-art verification systems still recognize them as the same person? Here we explore speaker verification performance on the VVD, comparing automated systems with both expert and non-expert human listeners to understand where and how identification breaks down.

Equal Error Rate

The goal of speaker verification is to determine whether or not two voice clips are created by the same speaker. Many of these methods are trained on cross-entropy loss alongside losses that encourage inter-class diversity and intra-class compactness, such as additive angular margin loss (AAM-Softmax) [71]. Coupled with architecture design, state-of-the-art models like the ECAPA-TDNN models from SpeechBrain and NeMo are able to achieve low Equal Error Rates (EERs) on test sets, achieving EERs on VoxCeleb1 [72] of 0.80% and 0.92%, respectively, being able to identify clips as originating from different speakers with remarkable accuracy.

This modeling approach to speaker verification fundamentally assumes that a speaker's vocal texture is static, which is violated by the speakers in the VVD. Illustrated in Table 3.2, the EER of speaker verification on VVD is remarkably higher than typical test performance, with EER on VVD being 21.52% for NeMo and 29.00% for SpeechBrain. This is not merely

due to a distribution shift, as analysis on the test set of VCTK [57], which was not used in either of the models' training, achieves similar EER to the EER achieved on VoxCeleb1.

Dataset	NeMo	SpeechBrain
VoxCeleb1	0.80	0.92
VCTK	0.64	1.26
VVD	21.52	29.00

Table 3.2: Equal Error Rate (%) across speaker verification models and datasets. VoxCeleb1 is in-distribution while VCTK and VVD are both out-of-distribution.

Error Analysis and Difficulty

The high-level EER results make it clear that speaker verification systems perform worse on the VVD; however, these results do not indicate how the systems misclassify voices, or how they compare to human listeners. We explore these questions here.

A unique quality of the VVD is the ability to directly measure the distance between voices, and then compare the performance of both automated speaker verification systems and human identification across said distance. Converting the levels of low, medium, and high to $0, 1, 2$, and then taking the L1 distance between two audio clips' three dimensional vectors, we map all pairs of voices onto a discrete space distance ranging between 0 and 6. For example, the distance between low-low-low and high-high-high would be a distance of 6.

We performed two perceptual experiments with non-expert and expert listeners. The three voice teachers were unfamiliar with each others' voices. Therefore, for expert classifications, we asked each voice teacher (e.g. 002) to classify whether or not given pairs of clips from the other teachers (e.g. 001, 003) belonged to the same speaker. For non-expert listeners, we asked six workers with Master's Qualification on Amazon Mechanical Turk to perform the same task. Both non-experts and experts were given the same 200 tasks and instructions, whereby the clips in each task contained voices from a maximum of two speakers, and were balanced such that random guessing would result in a performance of 50% accuracy. To measure classification accuracy for the SpeechBrain and NeMo ECAPA-TDNN models, we use both libraries' default threshold for classification evaluated on all possible pairs.

Averaging classification across all pairs and speakers, the achieved accuracies are: 59% for non-experts, 63% for experts, 78.47% for SpeechBrain, and 77.89% for NeMo. For human labelers, even those that are aware of the flexibility of the vocal tract, speaker verification across diverse voices is a challenging task. Across pairs coming from both different and same speakers, non-expert and model accuracies correlate. For pairs coming from different speakers, accuracy increases as voices become distinct; for pairs coming from the same speaker, the opposite trend emerges where accuracies trend towards 0 as L1-Distance increases.

While low same-speaker performance is expected as voices become distinct, due to the fundamental ignorance of speech modification by the considered models, the expert vs. non-

(a) Different Speaker

(b) Same Speaker

Figure 3.4: Accuracy (±1 Std. Error) of Speaker Verification across L1 Voice Distance over voice clips produced by pairs of speakers.

expert performance comparison sheds light on the perception of voice. Across both types of pairs, even experts struggle to tell if two voice clips are produced by the same speaker, performing at random when voices are most distinct. Despite non-experts being informed that a total of two speakers produced all voice clips in this task, non-expert performance diverges from that of experts, consistently rating voices with an L1-Distance greater than 3 as belonging to different speakers. Non-experts' assumptions of voice flexibility are similar to that of models: the flawed assumption that vocal texture is fixed.

Discussion

As seen in this section, existing speaker identity models fail to capture intra-speaker variation in a meaningful way. Single speakers can produce voices that span a broad perceptual space, violating the fundamental assumption of speaker identity models that treat voices as static. Gender classification models trained on typical voices achieve near-perfect accuracy ($\geq$97%), yet produce predictions spanning the full masculine-feminine range when faced with deliberate vocal modification (Figure 3.3). Speaker verification systems that achieve sub-1% EER on standard benchmarks degrade to 21–29% EER on the VVD (Table 3.2), not due to distribution shift (VCTK achieves comparable EER to VoxCeleb1), but because the assumption of static vocal texture is violated.

Critically, this degradation is not uniform but systematic: as the L1-Distance between voice configurations increases, same-speaker verification accuracy trends toward zero (Figure 3.4). The models do not fail randomly; they fail in proportion to the magnitude of vocal modification. Even human experts, who are aware of the flexibility of the vocal tract, perform at chance when configurations are maximally distinct. Non-experts fare worse still, sharing with the models the flawed assumption that vocal texture is fixed. Existing models tell us that voices are different, but not *why*, or *how*.

Before we can model inter-speaker variation, we must first model intra-speaker variation. Where high-dimensional embeddings collapse under vocal flexibility, a representation

grounded in the perceptual dimensions along which speakers actually vary may provide the interpretability and robustness that current approaches lack. To this end, we turn to voice qualities.

3.4 Acoustic Correlates of Perceptual Qualities on a Single Speaker

Utilizing the Disentangled Source-Filter Dataset (DSFD) introduced in Chapter 2, we now do a deep-dive into intra-speaker variation, taking on the task of linking perceptual voice qualities with individual acoustic and physical measurements.

Acoustic and Physical Metrics

For each voice configuration, we utilize a pretrained Convolutional REpresentation for Pitch Estimation (CREPE) model [73] to extract both F_0 and periodicity traces for each utterance, and then calculate the average F_0, denoted $\bar{F}_0$, and set of voiced regions $\mathcal{V}$. With these voiced regions, it is possible to calculate ΔF, a measure of the average interval between formants, which has been shown to correlate with Vocal Tract Length [25]. Given the set of voiced regions $\mathcal{V}$, we calculate ΔF in the following manner:

$$\Delta F = \frac{1}{4|\mathcal{V}|} \sum_{v \in \mathcal{V}} \sum_{i=1}^{4} \frac{F_i(v)}{i - 0.5} \tag{3.1}$$

where $F_i(v)$ is the average i-th formant over a given voiced region $v \in \mathcal{V}$. We found that calculating ΔF with CREPE-estimated voicing regions provided stronger correlations with PQs as compared to a version computed over forced alignment estimates of vowel regions, as is typically done in prior work.

We estimate the trace of the Contact Quotient (CQ) from the EGG by calculating the ratio between the time from peak to trough and the time from peak to peak in the derivative of the EGG (DEGG), which corresponds to the fraction of time the vocal folds are in contact during a single pitch period. We report two versions of the CQ: one calculated as the average over the sustained vowels, and one calculated as the average of the CQ trace over each voiced region $v \in \mathcal{V}$. An example EGG segment with corresponding F_0 and CQ estimates are provided in Figure 2.2 and online[1]. We also provide a version of the CQ that has been calculated using glottal flow estimation via inverse filtering of the waveform [74], which we report as CQ (IF).

In addition to the Contact Quotient, prior work has discussed vocal weight's connection with glottal spectral information such as spectral gravity, spectral slope, Loudness (average Spectral Energy), and harmonic-to-noise ratio (HnR) [49, 30, 31], which we include to compare with CQ.

[1] https://berkeley-speech-group.github.io/DisentangledSourceFilterDataset/

Acoustic-Physical-Perceptual Correlations

Here, we link the perceptual qualities of Pitch, Resonance, and Weight with individual acoustic and physical measurements. Ideally, we would be able to find a disentangled, linear correlation between an individual metric and PQ, whereby an increase in one individual measurement also corresponds to an increase in one individual perceptual quality, but not another.

Towards the end of measuring disentangled correlation, we employ pairwise ranking accuracy. For a given PQ and non-equal configurations, a pairwise ranking is deemed accurate if higher configurations have corresponding higher metric value. Using resonance and ΔF as an example, two voice clips, i, j, with configurations of high resonance and low resonance would be ranked accurately if $\Delta F_i > \Delta F_j$. We present the ranking accuracy of all metrics described in Section 3.4 in Table 3.3.

Metric	Pitch	Resonance	Weight
Avg. F_0	**99.3***	49.6	43.9
ΔF	61.5	**93.8***	55.6
CQ (Vowel)	41.3	47.2	**93.3***
CQ (Sentence)	66.2	34.4	84.0*
Loudness	57.6	37.7	92.1*
CQ (IF)	84.2*	55.1	67.5
Spectral Slope	50.0	79.1*	79.3*
Spectral Gravity	72.3	71.7	72.5
HnR	74.1*	21.7*	65.4

Table 3.3: Ranking Accuracy of Acoustic/Physical Metrics against Perceptual Voice Quality, demonstrating largely disentangled correlation between individual acoustic measures and a given PQ. (* = Statistically Significant Difference in Group Means via One-Way ANOVA)

Concerning Pitch and Resonance, the two corresponding metrics $\bar{F}_0$ and ΔF emerge as having largely disentangled correlations. With a pairwise ranking accuracy of 99.3%, $\bar{F}_0$ clearly delineates between pitch configurations, while performing near random chance for both Resonance and Weight. For Resonance, ΔF achieves a pairwise ranking accuracy of 93.8%, performing near random chance for Weight and slightly above random chance for Pitch. This suggests, for intra-speaker measurements, $\bar{F}_0$ and ΔF can serve as proxy measurements for Pitch and Resonance.

Weight, the most understudied of the PQs considered here, presents a more complicated story. Two metrics, namely CQ computed over sustained vowels and Loudness as computed via Avg. Spectral Energy, are able to rank Weight with high accuracies of 93.3% and 92.1%. Of the two, CQ (Vowel) achieves near random chance for both Pitch and Resonance, while Loudness achieves near random chance for Pitch and slightly below random chance for Resonance. Complicating disentanglement, the two metrics are highly correlated at 0.78,

Task	Model	RMSE	R^2	$\bar{F}_0$	ΔF	CQ	Loudness	UTMOS
Gender	Baseline	2.54 ($\pm$0.24)	-0.13 ($\pm$0.22)	-	-	-	-	-
Gender	LR	1.37 ($\pm$0.20)	0.66 ($\pm$0.12)	1.21	1.81	0.59	-1.04	-
Gender	RF	1.21 ($\pm$0.22)	0.73 ($\pm$0.13)	0.23	0.72	0.03	0.02	-
Realness	Baseline	0.17 ($\pm$0.03)	-0.12 ($\pm$0.15)	-	-	-	-	-
Realness	LR	0.15 ($\pm$0.03)	0.15 ($\pm$0.29)	0.019	-0.035	0.019	0.023	0.11
Realness	RF	0.14 ($\pm$0.02)	0.20 ($\pm$0.28)	0.073	0.126	0.081	0.078	0.48

Table 3.4: Overview of predictive model performance by Task and Model (Sample Avg. Baseline, Linear Regression (LR) or Random Forest (RF)), alongside model-specific feature importance (< 0.25 LR std., < 0.08 RF std.)

suggesting largely overlapping information. This aligns with expectations from the gender-affirming voice community and phonetics, which suggests that Weight modifications often lead to changes in Loudness and CQ [5], and that alternate combinations of CQ and Loudness can indicate other laryngeal voice qualities, such as creak or breathiness [1]. Weight might require CQ and Loudness both to be modeled independently of other PQs.

Furthermore, while we would ideally wish to model weight with acoustic metrics from the waveform, we note that every acoustic measure besides Loudness fails to accurately rank Weight independently of other perceptual qualities. While this suggests that glottal measurement is necessary for disentangled Weight modeling, the additional variability of read speech, let alone spoken speech, degrades the ranking accuracy of Contact Quotient, giving it a performance slightly worse than Loudness. It is unclear whether the robustness of the CQ computation algorithm or the variability due to speech is responsible for the degradation in performance, but we leave this to future work.

3.5 Modeling Demographics on a Single Speaker

Pitch, Resonance, and Weight vs. Gender

Following the approach in Section 3.4, we seek to evaluate the correlation between Pitch, Resonance, and Weight via predictive accuracy and feature importance. Utilizing $\bar{F}_0$, ΔF, CQ (Vowel), and Loudness as proxies for Pitch, Resonance, and Weight, we run a 100-fold cross validation on predicting the gender scalar rating, where train-test splits are chosen randomly. In place of a single train-test split, prior work has demonstrated that further perturbations along data splits lead to increased external validity of results [75, 76]. To compare the performance of the models, we also include a simple baseline which outputs the sample average as its prediction. We present these predictions using both Linear Regression and Random Forest Regression results in Table 3.4.

Across Linear Regression and Random Forest Regression, we see that both models achieve a low Root Mean Squared Error (RMSE) and moderately high R^2 score. While the LR

model has non-zero coefficients for all acoustic correlates, the RF model, which can account for non-linear dependencies of the features, reduces the contribution of CQ and Loudness to near 0, with $\bar{F}_0$ and ΔF accounting for the majority of feature importance. In the RF model, ΔF accounts for 3.13 times the feature importance of $\bar{F}_0$. While both contribute to gender perception, Resonance, as measured by proxy via ΔF, seems to account for the majority of the dependence. Despite interest in the GAVT community, it seems that Weight, as measured by proxy via CQ and Loudness, fails to account for much variation in gender perception.

Ambiguous Naturalness, Unpredictable Realness

To shed light on the relationship between Pitch, Resonance, and Weight on the perceived naturalness of a voice, we perform two similar tasks: a scalar rating task from 1 to 10, where 1 is very natural and 10 is very unnatural, and a classification task as to whether a voice is Fake or Real. To avoid biasing the listener, we provide no further definition of "natural" or "real."

Unlike gender, where scalar and classification tasks yielded largely stable results, naturalness and realness diverged greatly. The naturalness scalar rating task saw very little distinguishing variation, such that answers between 4 and 6 were the most frequent. The sample average baseline outperformed any predictive model of naturalness based on Pitch, Resonance, or Weight. The realness classification task saw a broader range of answers, but is nearly uncorrelated with naturalness and also saw a large dependence on the content of the audio clip. 73% of participants voted that audio clips Sent. 1 were real, compared to 53% of votes on audio clips of Sent. 2. The equivalent calculation for the gender classification task is 68.5% Female votes for Sent. 1, and 69.7% Female votes for Sent. 2.

As might be implied by this lack of stability, $\bar{F}_0$, ΔF, CQ, and Loudness fail to predict realness. Interestingly, including an estimate of Mean-Opinion-Score from automated MOS estimation system UTMOS [77] per utterance increases realness prediction and accounts for nearly half of the RF feature importance. Regardless, the models fail to perform substantially better than the simple sample average baseline. Realness, then, might be better modeled via alternative perceptual qualities like harshness or breathiness, or prosodic modeling which can account for variations in intonation.

Discussion

The results of Sections 3.4 and 3.5 establish a two-link chain from production to perception on a single speaker: acoustic and physical metrics can rank perceptual voice qualities with high accuracy (Table 3.3), and those same metrics, when used as PQ proxies, can predict crowdsourced gender perception with an R^2 of 0.73 (Table 3.4). Together, these findings validate the hypothesis that perceptual voice qualities mediate the relationship between vocal tract configuration and listener judgments of speaker identity.

Three observations deserve emphasis. First, the dominance of ΔF over $\bar{F}_0$ in predicting gender perception—accounting for over three times the feature importance in the Random Forest model—challenges the common assumption that pitch is the primary determinant of gendered voice. This aligns with the claims of the GAVT community that resonance plays a more central role than traditional speech science has acknowledged [5, 31]. Second, the near-zero contribution of CQ and Loudness (proxies for weight) to gender prediction is surprising given the community's emphasis on weight as a primary axis of gender modification. One explanation is that weight may contribute more to naturalness/realness than to gender, but our results do not validate that claim.

The failure to predict naturalness and realness from pitch, resonance, and weight alone reveals a boundary of the current PQ representation. While gender perception is stable across tasks (0.95 correlation between scalar and classification) and largely captured by three PQ axes, naturalness appears to require additional dimensions that are not represented in the acoustic proxy set used here.

These findings motivate the transition from single-speaker to multi-speaker modeling: while acoustic proxies suffice for a controlled single-speaker setting, scaling to diverse speakers requires learned representations that can capture PQ variation across the full range of vocal anatomy and behavior. The question then becomes which feature representations best generalize across speakers, which we explore in the following section.

3.6 Probing Feature Sets for Voice Quality

The preceding sections demonstrated that perceptual voice qualities can be reliably ranked by individual acoustic and physical metrics on a single controlled speaker (Section 3.4), and that these metrics predict crowdsourced gender perception (Section 3.5). However, single-speaker analyses on the DSFD benefit from controlled recording conditions and minimal confounding variables, which are introduced when modeling voice quality across diverse speakers, microphones, and speaking styles. To scale PQ modeling to the multi-speaker setting required for practical applications, we must identify which feature representations generalize across this variability.

Here, we probe three classes of speech representations: hand-crafted acoustic features, estimated articulatory features, and self-supervised learned features. We explore their ability to predict expert PQ ratings on the multi-speaker PVQD+ dataset. The goal is twofold: to determine whether objective features can predict subjective perceptual qualities across speakers, and to identify which feature set best serves as the foundation for the multi-speaker PQ modeling.

Can Objective Features Predict Subjective Qualities?

While prior work has demonstrated the capability of predicting perceptual qualities directly from waveform and mel-spectrograms [22, 23], here we explore the universality of perceptual

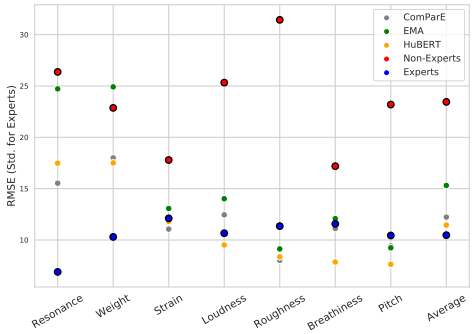


Figure 3.5: Test RMSE of various rating methods when compared with the average expert rating for each perceptual quality. Standard Deviation reported for expert ratings.

qualities across various representations of speech: acoustic, physical, and self-supervised.

Random Forest Regression

Given the highly non-linear relationship and high-dimensionality of various representations, we trained random forest regressors on a 60-20-20 train-validation-test split on PVQD+, finding that linear models such as Lasso resulted in lower performance. Hyperparameters for random forests were fine-tuned on the 20% validation set.

Feature Sets

Three feature sets are used in the random forest regression models. For the acoustic features, we use the ComParE 2016 feature set, which consists of a combination of functionals computed over prosodic, spectral, and sound quality-based features [9]. For the physical features, we use estimated Electromagnetic articulography (EMA) traces, which track movement of articulators with midsagittal x, y coordinates of jaw, lips, and tongue positions [78]. For the self-supervised features, we use the 7th layer of a pre-trained HuBERT model, which distills speech into a 1024-dimensional learned representation via training on masked audio samples [10].

Results

Across all three representations, several trends become clear. On average, all random forests across feature sets predict expert ratings with lower error than non-expert humans. Additionally, for CAPE-V PQs, ComParE and HuBERT features both achieve RMSE lower than inter-expert standard deviation, with the exception of Loudness for ComParE. EMA sees lower RMSE for two of the five CAPE-V PQs, and similar performance for both breathiness and strain.

While the models performed well for most of the CAPE-V PQs, the models failed to achieve similar performance on the gendered PQs. Considering the reliance of resonance and weight on laryngeal and source information, EMA's lack of such information justifies its poor performance. Given the excellent performance of ComParE and HuBERT features on the CAPE-V perceptual qualities, however, the performance drop for resonance and weight is surprising. While the performance is still better than that of non-experts (save for EMA and weight), we expect that this performance drop is in part due to the lack of labels for the entirety of the PVQD dataset.

To this end, we evaluate the HuBERT features' ability to rank intra-speaker variation of resonance and weight on the VVD. We show in Table 3.5 that this model is only able to rank weight marginally better than ranking via Praat's harmonics-to-noise ratio (HNR) estimation. The model is unable to accurately rank the resonance of clips, performing slightly below random, much worse than utilizing CREPE-estimated ΔF. While the model accurately fits typical voices in the PVQD+, the diversity of voices in the VVD leads to worse performance.

-	Resonance	Weight
Representation (HuBERT)	49.5%	**62.5%**
ΔF-HNR	**90.8%**	62.3%

Table 3.5: Accuracy of correctly ranking pairs of audio clips with different resonance or weight levels, as measured by the Hubert-based Representation, ΔF, and HNR.

Discussion

The probing results reveal a clear hierarchy among feature sets for PQ prediction. Self-supervised features (HuBERT) achieve the lowest RMSE across nearly all PQs, matching or exceeding inter-expert agreement for the CAPE-V qualities. ComParE acoustic features perform comparably on CAPE-V PQs but fall short on the gendered qualities, while EMA articulatory features struggle with resonance and weight entirely. These results motivate the use of a self-supervised speech model as the input representation for robust PQ modeling. SSL features encode information relevant to both health-related and gendered PQs in a single learned representation, without requiring explicit acoustic or articulatory measurements.

However, the HuBERT-based model's failure to generalize to intra-speaker ranking on the VVD (Table 3.5) exposes a critical limitation. A model trained on multi-speaker PVQD+ data learns to associate PQ levels with inter-speaker variation, but does not necessarily capture the finer-grained intra-speaker variation produced by deliberate vocal modification. This gap motivates two design choices in the PQ-Encoder that follows: (1) the use of WavLM's CNN feature extractor, which provides noise-robust, low-level features amenable to learning both inter- and intra-speaker PQ variation, and (2) a pairwise comparison training objective for the gendered PQs, which directly supervises the model to rank voices within and

across speakers rather than predicting unstandardized absolute scale values that may not generalize.

More broadly, these results confirm that perceptual voice qualities occupy a meaningful position in the representational landscape of speech. The information encoded in subjective PQs has an objective basis, present across hand-crafted acoustic features, estimated articulatory trajectories, and self-supervised learned representations alike. PQs provide what other speech representations lack: a low-dimensional, interpretable level of abstraction whereby any listener, given minimal training, can hear the particular aspects of each quality. The challenge is building a model that predicts these qualities robustly across the full diversity of speakers and speaking conditions.

3.7 Robust Multi-Speaker Voice Quality Modeling

A central premise of this work is that voice qualities provide a low-dimensional, interpretable intermediate representation of speaker identity. In this section, we test this hypothesis and describe the PQ-Encoder, which extends the modeling of Section 3.6 to a broader set of nine PQ dimensions, which we call the PQ-vector. We end by presenting evidence that PQs capture meaningful information about speaker identity, demographics, and affect.

PQ Modeling

To scale beyond expert-labeled data, we train a PQ-Encoder g that predicts PQ-vectors from speech. As Section 3.6 demonstrated that self-supervised embeddings had the overall best performance, the PQ-Encoder takes as input features from the CNN feature extractor of WavLM, chosen for its noise-robustness and compact architecture suitable for deployment on small devices. The PQ-Encoder produces a 9-dimensional output corresponding to the PQs in Table 2.1, trained via two complementary objectives depending on the voice quality.

Comparison-based qualities (Pitch, Resonance, Weight, Loudness): For these qualities, where absolute rating scales are either unreliable or nonexistent [46], we train using pairwise comparisons. Given two audio clips a_1 and a_2, a label $y \in \{-1, 0, 1\}$ indicates which clip is higher in a given PQ (or whether they are equal). For resonance and weight, comparison labels are derived from human annotations. For pitch and loudness, comparison targets are computed on-the-fly during training in a self-supervised fashion: the pitch target is derived from the sign of the difference in median F_0 between the two clips, while the loudness target is derived from the sign of the difference in mean waveform energy [46]. Training utilizes a pairwise cross-entropy loss derived from the Bradley-Terry model [79]:

$$p_1 = \frac{e^{g(a_1)}}{e^{g(a_1)} + e^{g(a_2)}}, \qquad p_2 = \frac{e^{g(a_2)}}{e^{g(a_1)} + e^{g(a_2)}} \tag{3.2}$$

$$\mathcal{L}_{\text{comp}}(a_1, a_2, y) = -\mathbb{E}\big[\mu_1(y) \log p_1 + \mu_2(y) \log p_2\big] \tag{3.3}$$

where $\mu_1(y) = \mathbf{1}[y = -1] + \frac{1}{2}\mathbf{1}[y = 0]$ and $\mu_2(y) = \mathbf{1}[y = 1] + \frac{1}{2}\mathbf{1}[y = 0]$. This formulation naturally handles ties and produces a scalar score per clip that reflects relative ordering rather than absolute magnitude.

Rating-based qualities (Breathiness, Roughness, Strain): For these CAPE-V qualities, expert ratings on a 1–100 scale are available from the PVQD [18]. The PQ-Encoder is trained with a standard mean squared error loss against the average expert rating for each clip.

Creak and Nasality: For creak, we leverage the Creapy creaky voice detection tool [80], which provides frame-level creak probabilities that are averaged over an utterance to produce a scalar creak score. For nasality, we extend the approach of Mathad et al. [48], where we use frame-level probability of a nasal phoneme to derive utterance-averaged nasality scores. These two automatic scores are concatenated with the seven PQ-Encoder outputs to form the complete 9-dimensional PQ-vector p_x.

Dataset Overview

We draw from eight datasets, each contributing complementary coverage of the PQ space:

1. **LibriTTS** [81]: A large-scale multi-speaker corpus of English audiobook recordings, providing broad coverage of typical adult voices and serving as the primary source of speaker diversity during training.

2. **PVQD** [18]: The Perceptual Voice Qualities Database, originally designed for clinical voice assessment, provides expert-labeled ratings of CAPE-V PQs across 296 speakers with varying degrees of vocal health, anchoring the PQ labels and providing coverage of atypical voices.

3. **VVD** [31]: The Versatile Voice Dataset, introduced in Chapter 2, provides critical coverage of intra-speaker variability and gender-diverse voices via three trans-feminine voice teachers producing up to 27 voice configurations each.

4. **DSFD** [46]: The Disentangled Source-Filter Dataset, also from Chapter 2, supplements the VVD with additional sentence-level diversity and physical measurement via Electroglottograph data.

5. **MAGES** [44]: The Mid-Atlantic Gender-Expansive Speech Corpus, which includes speakers across the gender spectrum, extending demographic coverage beyond binary gender categories.

6. **Palette of Transmasculine Voices** [43]: A dataset of transmasculine speakers, providing coverage of voices underrepresented in existing speech datasets.

7. **VCTK** [57]: A multi-speaker English corpus of 110 speakers with various accents, providing broad acoustic and demographic diversity.

Table 3.6: PQ-Encoder validation with train/test splits. Rating-based qualities report Pearson r against expert CAPE-V ratings on PVQD. Comparison-based qualities report intra-speaker ranking accuracy on VVD and DSFD.

PQ	Metric	Train	Test
Breathiness	Pearson r	0.972	0.744
Roughness	Pearson r	0.970	0.438
Strain	Pearson r	0.965	0.697
Pitch	Ranking Acc.	0.953	0.975
Resonance	Ranking Acc.	0.969	0.816
Weight	Ranking Acc.	0.906	0.937

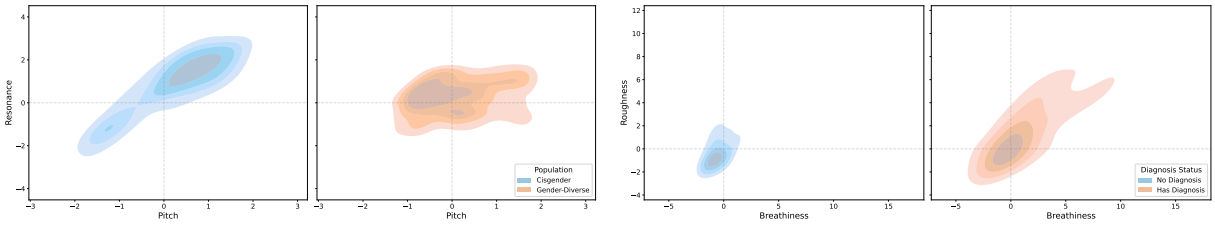
8. **CREMA-D** [56]: The Crowd-sourced Emotional Multimodal Actors Dataset, containing emotional speech from 91 actors across diverse demographics, extending coverage of expressive speech variation.

For each dataset, PQ labels are obtained either from existing expert annotations (as in the PVQD), additional crowdsourced labels (Breathiness, Roughness, and Strain labels for a random subset of 100 speakers across all datasets), or semi-supervised labels from the PQ-Encoder g during training. This combination of expert, non-expert, and model-estimated PQ labels allows StyleMod to leverage the scale of large unlabeled corpora while remaining grounded in perceptual judgments, achieving generalization across diverse speakers and recording environments.

PQ-Encoder Validation

To validate the PQ-Encoder, we evaluate its predictions against ground-truth labels on two datasets, reporting both train-split and held-out test-split performance. For the rating-based qualities (breathiness, roughness, strain), we measure Pearson correlation between the PQ-Encoder's predictions and expert CAPE-V ratings on the PVQD, split into 194 train speakers and 83 held-out test speakers. For the comparison-based qualities (pitch, resonance, weight), we measure intra-speaker ranking accuracy on the VVD [31], where ground-truth orderings are defined by the controlled recording conditions. Table 3.6 reports the results.

On the train split, the rating-based qualities achieve Pearson correlations above 0.96, indicating near-perfect agreement with expert CAPE-V ratings for seen speakers. On the held-out test split, correlations drop to 0.44–0.74, with roughness exhibiting the largest generalization gap (0.970 $\rightarrow$ 0.438). Test MAE values are 13.64 (breathiness), 16.39 (roughness), and 17.75 (strain) on the 1–100 CAPE-V scale. For context, Chapter 2 reports non-expert RMSE of 17.19 (breathiness), 31.44 (roughness), and 17.79 (strain), with inter-expert standard deviations of 11.57, 11.35, and 12.11, respectively. The PQ-Encoder's held-out MAE



(a) Gender Space (b) Health Space

Figure 3.6: Visualization of the PQ state space across different populations. (a) The gender-correlated dimensions (pitch, resonance, weight) show distinct clustering by speaker group, with gender-diverse speakers occupying intermediate and atypical regions. (b) The health-correlated dimensions (strain, roughness, breathiness) show separation between healthy and pathological voices.

is thus comparable to or lower than non-expert error for all three qualities, while remaining above inter-expert variation.

For the comparison-based qualities on the VVD, train and test ranking accuracies are more balanced: pitch ranking actually improves on the test speaker (0.879 $\rightarrow$ 0.943), while resonance shows a larger drop (0.924 $\rightarrow$ 0.747). For comparison, Chapter 3 reports ranking accuracies of acoustic proxy metrics on the DSFD: 99.3% for avg. F_0 vs. pitch, 93.8% for ΔF vs. resonance, and 93.3% for contact quotient vs. weight. The PQ-Encoder's test ranking accuracies (94.3% pitch, 74.7% resonance, 81.7% weight) are lower than these acoustic proxies, which is expected: the acoustic metrics are computed directly from the signal and evaluated on a single highly-controlled speaker, whereas the PQ-Encoder must generalize across speakers from learned WavLM features. The gap is largest for resonance, suggesting that formant-derived ΔF remains a stronger single-speaker resonance measure than the learned representation.

Overall, the PQ-Encoder achieves ranking accuracies of 0.75–0.94 on held-out speakers and Pearson correlations of 0.44–0.74 on held-out PVQD speakers, establishing it as a reasonable measurement tool for the evaluations that follow while also highlighting the challenge of generalizing perceptual quality predictions to unseen speakers.

PQ State Space and Voice Types

The PQ-Encoder defines a nine-dimensional, continuous manifold over which voices can vary. In practice, not all regions of this space are equally populated, and different speaker styles reside in distinct subregions. Figure 3.6 visualizes this structure by projecting speakers from the training corpora onto the gender-correlated and health-correlated PQ subspaces, as hypothesized in Figure 2.1.

Along the gender-correlated axes (Figure 3.6a), cisgender voices typically group into two clusters, corresponding to conventionally masculine and feminine configurations of pitch,

Model	EER
PQ-Vector	19.48%
PQ-Vector (Dist.)	19.30%
StyleStream	9.14%
Resemblyzer	3.01%

Table 3.7: Equal Error Rate (EER) on the test set of VCTK.

resonance, and weight. Gender-diverse voices, particularly those of individuals who have undergone gender-affirming voice modification [31], occupy intermediate regions and configurations that break the correlations typical of cisgender speakers—such as high pitch combined with low resonance.

Along the health-correlated axes (Figure 3.6b), healthy typical voices cluster in a region of low strain, roughness, and breathiness, while atypical voices—those with pathological conditions or advanced age—extend into higher-valued regions of these CAPE-V dimensions. The PVQD speakers, who include individuals with a range of voice disorders, span a much wider region of this subspace than the healthy corpora.

This heterogeneous coverage motivates the use of diverse training data drawn from clinical, gender-diverse, and typical speech corpora to ensure that the PQ-Encoder and StyleMod encounter the full range of the PQ state space during training, and can model a broad range of voice types.

PQ and Speaker Identity

A key claim of this work is that PQs capture substantial information about speaker identity. To quantify this, we evaluate the ability of PQ-vectors to perform speaker verification on the VCTK corpus [57], comparing against two learned speaker embedding models: StyleStream and Resemblyzer.

Table 3.7 reports Equal Error Rates (EER) on the VCTK test set. The PQ-Vector achieves an EER of 19.48%, substantially higher than StyleStream (9.14%) and Resemblyzer (3.01%). This gap is expected: a 9-dimensional perceptual representation cannot capture the full speaker-discriminative information present in a 768-dimensional learned embedding. Many speakers share similar PQ profiles, meaning that PQs alone are insufficient to uniquely identify a speaker.

However, while PQs do not uniquely identify speakers, the distances in PQ space are strongly correlated with distances in learned embedding spaces. Table 3.8 reports correlations between PQ-Vector similarity measures and the cosine similarity of StyleStream and Resemblyzer embeddings. The PQ-Vector cosine similarity correlates at 0.76 with StyleStream cosine similarity, and PQ-based distance correlates at -0.80, indicating that speakers who are close in PQ space tend to also be close in learned embedding space. This relation-

Model 1	Model 2	Cos. vs. Cos.	PQ-Dist vs. Cos.
PQ-Vector	StyleStream	0.76	-0.80
PQ-Vector	Resemblyzer	0.43	-0.62
PQ-Vector (Dist.)	StyleStream	0.63	-0.74
PQ-Vector (Dist.)	Resemblyzer	0.42	-0.55
StyleStream	Resemblyzer	0.72	–

Table 3.8: Correlation between PQ-Vector Cosine Similarity and PQ-Distance with Cosine Similarity of Speaker Embeddings.

	VCTK		PVQD	
Model	Age MAE	Gender Acc.	Age MAE	Gender Acc.
PQ-Vector	1.266	99.3%	24.15	84.3%
PQ-Vector (Dist.)	1.505	99.9%	24.31	42.4%
StyleStream	1.085	100.0%	23.75	44.0%
Resemblyzer	1.164	97.7%	24.04	89.4%

Table 3.9: Age regression (MAE in years) and gender classification accuracy of Random Forest models trained on VCTK, evaluated on VCTK test and cross-dataset on PVQD.

ship confirms that PQs capture a meaningful, identity-relevant subspace of the full speaker representation.

Age and Gender Prediction

To further validate that PQs encode demographic information, we train Random Forest classifiers and regressors to predict speaker gender and age from PQ-vectors, StyleStream embeddings, and Resemblyzer embeddings. Models are trained on the VCTK corpus and evaluated both in-distribution (VCTK test set) and cross-dataset (PVQD).

Table 3.9 reports the results. On the in-distribution VCTK test set, all models achieve near-perfect gender classification ($\geq$ 97.7%) and low age MAE ($\leq$ 1.5 years), reflecting the relative homogeneity of the VCTK corpus. The cross-dataset evaluation on PVQD is more revealing: all models exhibit substantially higher age MAE ($\sim$24 years), confirming that age prediction from short speech segments is inherently difficult and does not transfer well across corpora. For gender, the PQ-Vector achieves 84.3% accuracy on PVQD, outperforming both StyleStream (44.0%) and PQ-Vector (Dist.) (42.4%), and approaching Resemblyzer (89.4%). The strong cross-dataset gender performance of the PQ-Vector suggests that the gendered PQ dimensions (pitch, resonance, weight) generalize across recording conditions better than the full learned embeddings, whose gender-relevant features may be confounded with corpus-specific characteristics.

Model	Accuracy	F1-Score
PQ-Vector	40.6%	0.383
PQ-Vector (Dist.)	37.8%	0.345
StyleStream	67.0%	0.664
Resemblyzer	51.9%	0.503

Table 3.10: Accuracy and F1 Score on CREMA-D Emotion Prediction.

Emotion Prediction

Beyond demographics, PQs also encode information about affective state. Using an 80-20 train-test split on the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [56], we evaluate the ability of PQ-vectors to predict emotional categories, following the approach of Narain et al. [8].

Table 3.10 reports the results. The PQ-vector achieves 40.6% accuracy, well above the random baseline of 16.7% for six emotion classes, but substantially below StyleStream (67.0%) and Resemblyzer (51.9%). This gap reflects that emotion is conveyed through prosodic dynamics that are not captured by the utterance-level PQ-vector. PQs capture some of the timbral and quality-based correlates of emotion, but, overall, PQs miss the temporal structure that distinguishes emotion.

Limits of the PQ-vector

The evaluations above reveal both the strengths and limitations of PQs as a speaker representation. On the one hand, PQs capture meaningful information about speaker identity through demographic information (gender, age), and affect. On the other hand, a 9-dimensional PQ-vector cannot uniquely identify a speaker (EER of $\sim$19%) and misses temporal and prosodic information relevant to accent and emotion.

These limitations are a benefit of the representation for the purposes of voice modification. Because PQs capture a perceptually relevant *subspace* rather than the full speaker embedding, modifying PQ values while holding other factors constant could enable interpretable, targeted modifications, thus changing how a voice sounds along perceptual axes without destroying the aspects of identity that PQs do not capture. The gap between PQ-based and embedding-based speaker verification (Table 3.7) quantifies exactly the "non-PQ" information that the inversion-based approach of Section 4.3 aims to preserve.

Chapter 4

Interpretable Modification of Speaker Identity

4.1 Introduction

In Chapters 2 and 3, we established that perceptual voice qualities (PQs) provide an interpretable, low-dimensional representation of speaker identity grounded in human perception. We demonstrated that non-experts can reliably perceive PQs, that PQs exhibit disentangled correlations with acoustic and physical measurements, and that PQ-based representations capture meaningful information about speaker identity despite their compactness. The natural next question is whether this perceptual grounding can be leveraged not just for *analysis*, but for *modification*: can we change how a voice sounds along interpretable perceptual axes while preserving the aspects of identity that we do not wish to alter?

In this chapter, we present StyleMod, a conditional flow matching model that learns to generate and modify speaker embeddings conditioned on target PQ-vectors. We demonstrate that analysis and synthesis along PQ axes can generate new voices that span the demographic categories of age and gender, while enabling precise, single-axis modifications of individual voice qualities (Figure 4.1).

The approach rests on three components: (1) a PQ-Encoder that maps speech to a 9-dimensional PQ-vector, as described in Chapter 3; (2) a synthesis backbone (StyleStream [82]) that disentangles content from speaker information and provides a fixed-dimensional speaker embedding; and (3) StyleMod itself, a conditional flow matching model that learns to modify the PQ space via X-Prediction [83] with PQ consistency regularization. Through inversion-based editing and heuristic demographic manipulation, StyleMod enables interpretable and continuous modification of perceived gender and age while preserving non-targeted speaker attributes.

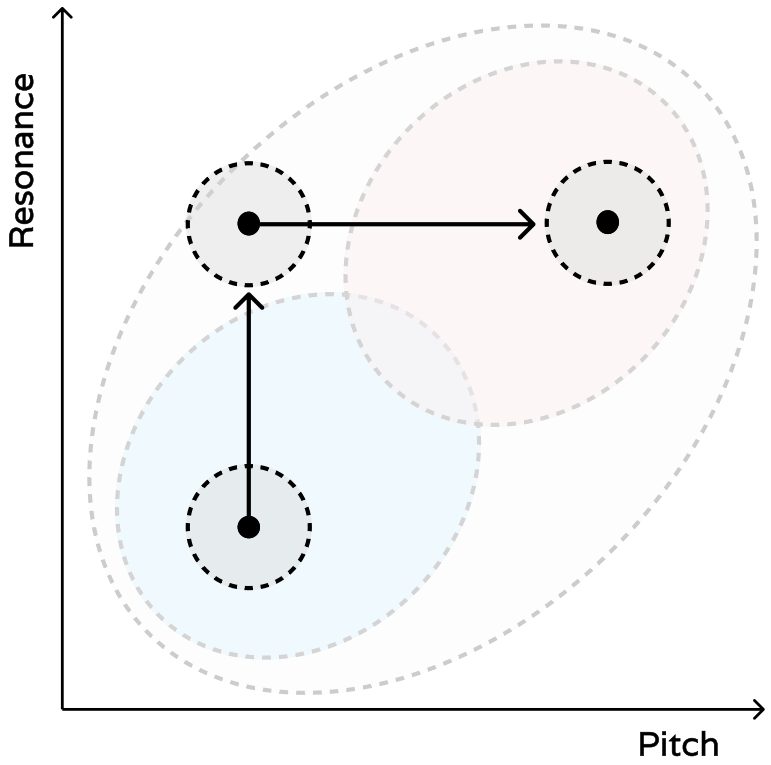


Figure 4.1: Overview of PQ-based speaker identity modification. Given a source speaker embedding and a target PQ-vector, StyleMod learns to traverse the PQ space via conditional flow matching, producing a modified embedding that reflects the desired perceptual qualities while preserving non-targeted attributes.

4.2 Background

The preceding chapters established the perceptual and representational foundations for this work. Here, we review the technical landscape of voice conversion, controllable synthesis, and generative modeling that situates StyleMod within the broader speech processing literature.

Voice Conversion and Controllable Synthesis

Voice conversion systems aim to transform the identity or style of a speaker while preserving linguistic content. Lian et al. [58] propose a disentangled variational approach for zero-shot voice conversion, separating content from speaker information in a learned latent space.

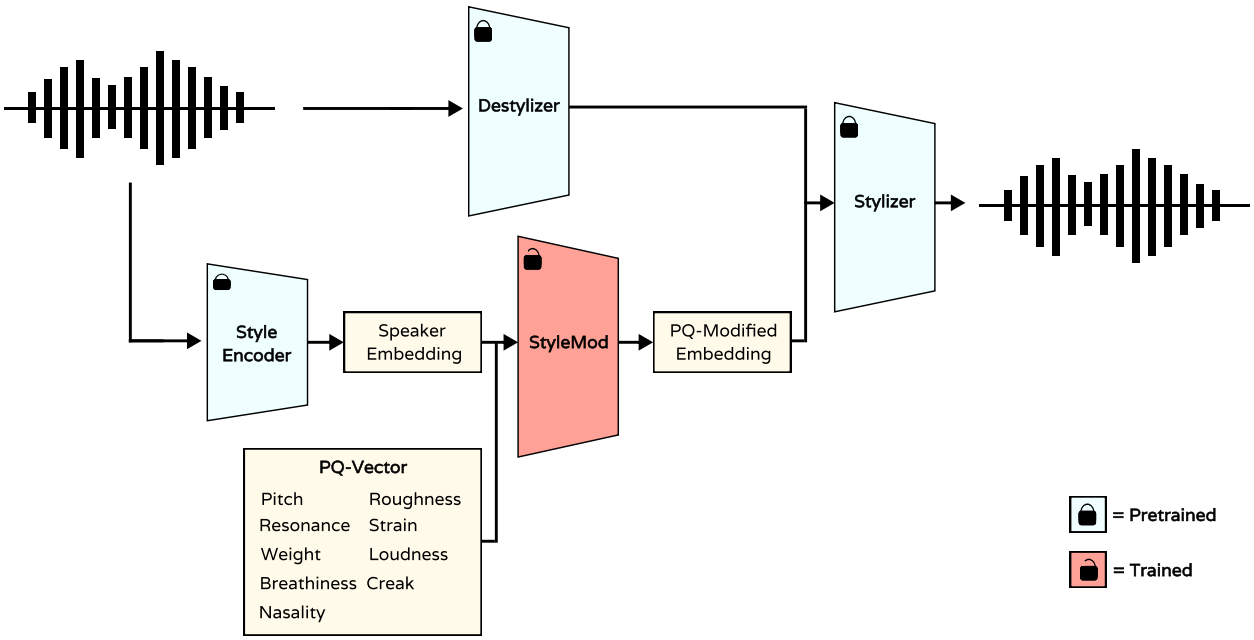


Figure 4.2: System overview of the proposed interpretable speaker identity modification framework.

Large-scale synthesis systems such as NaturalSpeech 3 [13] and Audiobox [65] achieve high-fidelity multi-speaker generation, but treat speaker identity as a holistic embedding without exposing interpretable axes of control. AudioBox and additional models like PromptTTS [66] enable text-described control over speaker attributes, but the mapping from natural language descriptions to acoustic properties remains coarse.

Controllable synthesis along specific perceptual dimensions has received growing attention. Zhou et al. [21] demonstrate control over emotion intensity in voice conversion, while Hu et al. [84] learn self-supervised prosody representations for style-controllable TTS. Li et al. [34] propose GTR-Voice, which uses articulatory phonetics to enable fine-grained expressive control. Rautenberg et al. [7] synthesize speech along perceptual voice quality dimensions, demonstrating that inversion-based editing can modify specific qualities while preserving others, which directly inspires the inversion procedure in this work.

The synthesis backbone used in this chapter, StyleStream [82], provides an analysis-by-synthesis framework that disentangles content from style via a quantized ASR destylizer and a jointly trained style encoder. StyleMod operates on the speaker embedding space of this system, learning to traverse the PQ space via conditional flow matching rather than modifying the synthesis model itself.

Diffusion-based Generative Models for Speech

Diffusion and flow-based generative models have become a dominant paradigm across modalities, from image generation [85] to protein design [86]. In speech, denoising diffusion probabilistic models [87] and score-based approaches [88, 89] have been applied to waveform and spectrogram generation, with many models such as AudioBox [65] and NaturalSpeech3 [13] relying on diffusion architectures, and models such as CosyVoice3 [90] relying on flow-matching architectures. Tasks extend beyond generation as well, such as applying latent diffusion to audio editing [91].

StyleMod adopts the X-Prediction formulation from Li and He [83], which parameterizes a flow matching model to predict the clean data point directly rather than the velocity field. This formulation is advantageous for our setting because it allows auxiliary losses to be applied directly to the model's output, enabling tighter control over the perceptual properties of generated speaker embeddings.

Gender, Age, and Demographics in Speech

The perception of speaker demographics from voice is explored in detail in Chapters 2 and 3, where we establish that pitch and resonance are the primary perceptual correlates of gendered voice variation. Unlike prior work that primarily analyzes gender and age as fixed speaker attributes, StyleMod treats demographic categories as configurations in PQ space, enabling continuous and interpretable modification of perceived speaker demographics, as discussed below.

4.3 StyleMod: Interpretable Multi-Speaker Synthesis

Having established that PQs provide an interpretable intermediate representation of speaker identity, we now describe StyleMod: the general recipe for modifying speaker identity along PQ axes. The recipe consists of three components: diverse training data spanning demographic categories, a synthesis backbone capable of reconstructing speech from a speaker embedding, and a conditional flow matching model that learns to traverse the PQ space. We detail each component and describe three modes of modification: unconditional sampling, inversion-based editing, and demographic-level manipulation via heuristic search.

Data Considered

A key requirement of StyleMod is training data that spans the space of perceptual voice qualities across diverse speakers and speaking conditions. The training data used by StyleMod is the same as the training data used in the PQ-Encoder, as described in Section 3.7.

Models Considered

StyleMod operates on the speaker embedding space of an analysis-by-synthesis system. In principle, any system that disentangles content from speaker information and provides a fixed-dimensional speaker embedding is compatible with our approach. We consider two such systems:

1. **SPARC Vocoder** [92]: A vocoder that provides a source-filter decomposition of speech. By explicitly relying on predicted electromagnetic articulography (EMA) traces, SPARC vocoders disentangle content and speaker information. We use a fast, DDSP-based implementation [93].

2. **StyleStream** [82]: An analysis-by-synthesis system that disentangles content and style information in speech.

StyleStream serves as the primary backbone for the experiments in this chapter. Given a speech segment a, StyleStream extracts content information via a "Destylizer" block, which uses a quantized ASR head to produce a content representation stripped of speaker and style information. A jointly trained "Style Encoder" then learns to reinsert speaker and style information into the reconstructed output. Prior work on StyleStream demonstrates a high degree of controllability and disentanglement of style attributes such as accent, emotion, and speaker timbre from the content output of the Destylizer.

For StyleMod, we utilize a variant of StyleStream that conditions reconstruction solely on a 768-dimensional learned embedding, unlike the original architecture which conditions on both a learned embedding and a Mel-spectrogram. This choice constrains all speaker and style information to pass through a single, fixed-dimensional bottleneck, making the embedding space a natural target for PQ-conditioned generation.

Architecture

StyleMod is a conditional flow matching model that learns to generate speaker embeddings conditioned on a target PQ-vector. We adopt the X-Prediction formulation from [83], which parameterizes the model to directly predict the clean data point $\tilde{x}$ rather than the velocity field. This has two practical advantages: first, it simplifies the training objective; second, because the model output $\tilde{x}$ lives in the same space as the speaker embedding x, auxiliary losses can be applied directly to the output, enabling PQ consistency regularization. An overview of the system is provided in Figure 4.2, with precise training and sampling architectures given in Figure 4.3

We define the following notation used throughout this section:

1. $f(x, t, p_x)$: The flow matching model, parameterized as a neural network.

2. $x \in \mathbb{R}^D$: A $D = 768$ dimensional speaker embedding from StyleStream.

3. $t \in [0, 1]$: The flow matching timestep, where $t = 0$ corresponds to pure noise and $t = 1$ corresponds to clean data.

4. z_t: The noisy interpolant at timestep t, defined as $z_t = (1 - t)z_0 + tx$ for noise sample $z_0 \sim \mathcal{N}(0, I)$.

5. $p_x \in \mathbb{R}^K$: A $K = 9$ dimensional PQ-vector describing the perceptual voice qualities of embedding x.

6. $\tilde{x}$: The output of the flow matching model, i.e., the predicted clean embedding.

7. $g(x) = p_x$: The PQ-Encoder, a pretrained model that maps from a speaker embedding to its corresponding PQ-vector.

Training

Training follows a classifier-free guidance (CFG) scheme [94], whereby the PQ conditioning vector p_x is randomly dropped (set to $\emptyset$) with some probability during training. This dropout encourages the model to learn both unconditional and conditional generation, enabling stronger conditional generation at inference time through guidance scaling.

The training objective consists of two losses. The first is the X-prediction flow matching loss, which encourages the predicted velocity to match the ground truth:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\|\tilde{v} - v\|^2 \tag{4.1}$$

where the predicted velocity $\tilde{v}$ is derived from the model's output $\tilde{x}$:

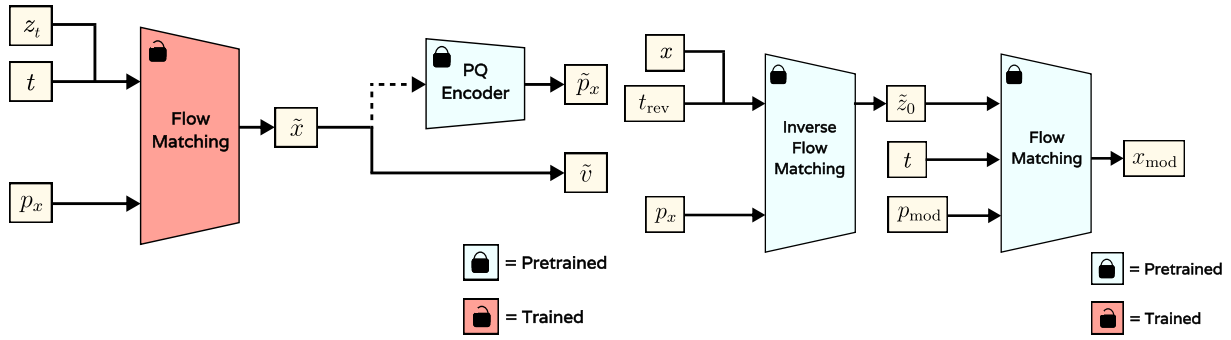
$$\tilde{v} = \frac{\tilde{x} - z_t}{1 - t} \tag{4.2}$$

for $t \neq 0$, and $v = x - z_0$ is the ground truth velocity.

The second loss is a PQ consistency loss, which regularizes the generated embedding to preserve the target perceptual qualities as measured by the pretrained PQ-Encoder g:

$$\mathcal{L}_{\text{PQ}} = \mathbb{E}\|g(\tilde{x}) - p_x\|^2 \tag{4.3}$$

This consistency loss is made possible by the X-Prediction formulation: because $\tilde{x}$ is a direct prediction of the clean embedding, the PQ-Encoder can be applied to it without approximation. The total training loss is a weighted combination $\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda\mathcal{L}_{\text{PQ}}$, where λ controls the strength of PQ regularization. The entire training pipeline is illustrated in Figure 4.3a.



(a) Training (b) Sampling

Figure 4.3: Training and sampling procedures for the StyleMod architecture.

Sampling

Following recommendations from [83], sampling is performed with a simple Euler solver that integrates from $t = 0$ (noise) to $t = 1$ (clean data). At each integration step, classifier-free guidance is used to mix unconditional and PQ-conditioned predictions, steering generation towards the desired perceptual qualities. Given a guidance scale $w > 1$, the guided prediction is:

$$\tilde{x}_{\text{guided}} = (1 + w) \cdot f(z_t, t, p_x) - w \cdot f(z_t, t, \emptyset) \tag{4.4}$$

Higher values of w produce embeddings that more closely adhere to the conditioning PQ-vector, at the potential cost of reduced diversity. The sampling procedure is illustrated in Figure 4.3b.

Inversion-based PQ Modification

While unconditional sampling generates novel speakers from a target PQ-vector, it does not necessarily maintain speaker attributes not captured by the 9-dimensional PQ representation, such as accent, speaking rate, or emotional tone. As the PQ-vector does not uniquely identify a speaker, many distinct speakers can be generated from a single PQ-vector, meaning that general sampling will produce speakers that vary across these unmodeled attributes.

To enable precise modification of a specific speaker's perceptual qualities while holding other attributes constant, we adopt an inversion-based approach inspired by [7]. Rather than generating $\tilde{x}$ from random noise z_0 and a desired PQ-vector p_x, inverse flow matching first regresses from an input speaker embedding x backwards from $t = 1$ to $t = 0$, conditioned on the original PQ-vector $\tilde{p}_x = g(x)$. This produces an estimated noise vector $\tilde{z}_0$ that encodes the non-PQ aspects of the speaker's identity. With this estimated noise and a modified PQ-vector p_{mod}, forward integration from $t = 0$ to $t = 1$ generates a modified speaker

embedding x_{mod} that reflects the desired PQ changes while preserving accent, emotion, and other speaker-specific characteristics.

As we demonstrate in Section 4.4, this inversion-based approach enables precise modification of individual PQs such as pitch, resonance, and creak while holding other factors of a speaker's voice constant, providing a degree of control not achievable through unconditional sampling alone.

Hardcoded Gender Modification

Since StyleMod operates directly on the PQ space, higher-level modifications along demographic axes like gender require mapping from demographic targets to PQ configurations. Prior work has studied the gendered axes of pitch, resonance, and weight, demonstrating that typical values for cisgender individuals follow a roughly linear relationship across these three dimensions [5, 31, 46]. Building on these findings, we define a set of hardcoded PQ offsets that correspond to points along the masculinity-femininity spectrum:

1. Pitch = -5.0, Resonance = -5.0, Weight = 5.0: Extremely Masculine
2. Pitch = -2.0, Resonance = -2.0, Weight = 2.0: Moderately Masculine
3. Pitch = 0.0, Resonance = 0.0, Weight = 0.0: Androgynous (no modification)
4. Pitch = 2.0, Resonance = 2.0, Weight = -2.0: Moderately Feminine
5. Pitch = 5.0, Resonance = 5.0, Weight = -5.0: Extremely Feminine

These offsets are applied additively to a speaker's estimated PQ-vector, and the modified vector is used as the conditioning input for inversion-based modification. It is important to note that this linear parameterization is a simplification. StyleMod enables more complex generations along gendered axes by breaking the correlation between pitch, resonance, and weight. For example, a configuration of Pitch = -2.0, Resonance = 2.0, and Weight = -2.0 would produce a low-pitched feminine voice.

Grid-Search Age Modification

Unlike gender, where perceptual correlates are relatively well-studied and map onto a small number of PQ axes, the relationship between PQs and perceived age is more diffuse. Absolute age perception is inherently difficult for human listeners, and we hypothesize that listeners distinguish age in voice by cues associated with deviations in pitch, breathiness, roughness, and strain. As such, we adopt a data-driven, heuristic approach to age-based modification.

Utilizing the healthy speakers of the PVQD, which contain labels for speaker age, we train a Random Forest model that regresses from a PQ-vector p_x to a speaker's age in the range $[18, 83]$. This age predictor provides a differentiation-free proxy for how a given PQ configuration maps to perceived age.

Table 4.1: Speaker embedding reconstruction quality (cosine similarity, mean ± std) under naive and inverse Euler sampling across model variants ($n = 100$).

Model	Naive	Inverse
StyleMod V-Prediction	0.416 ± 0.078	0.975 ± 0.013
StyleMod X-Prediction	0.706 ± 0.064	0.998 ± 0.001
StyleMod X-Prediction + Consistency	0.704 ± 0.061	0.991 ± 0.002
SPARC X-Prediction + Consistency	0.408 ± 0.482	0.892 ± 0.138

At modification time, we estimate a given speaker's PQ-vector $\tilde{p}_x = g(x)$ and perform a greedy grid search over PQ dimensions. At each step, we randomly select one of the 9 PQ dimensions and perturb it by $\pm\alpha$, where $\alpha = 2.0$ is the step size. The perturbation is accepted if it moves the predicted age closer to a target age, and rejected otherwise. This process is repeated for a fixed number of steps, producing a modified PQ-vector that, when used as conditioning for inversion-based modification, generates a speaker embedding whose predicted age more closely matches the target. While simple, this greedy approach avoids the need for a differentiable age model and naturally respects the geometry of the PQ space by taking small, interpretable steps.

4.4 Results

Reconstruction Quality

We compare three model variants to validate the design choices underlying StyleMod: (1) *StyleMod V-Prediction*, which parameterizes the flow matching model to predict the velocity field directly; (2) *StyleMod X-Prediction*, which instead predicts the clean data point $\tilde{x}$; and (3) *StyleMod X-Prediction + Consistency*, which adds the PQ consistency loss $\mathcal{L}_{\text{PQ}}$ on top of the X-Prediction formulation. We also compare with an additional model (4) *DDSP-SPARC X-Prediction + Consistency*.

For each variant, we evaluate reconstruction quality via the cosine similarity between the original StyleStream speaker embedding and the model's output, measured under two sampling strategies: *naive* Euler sampling from random noise, and *inverse* sampling, which first inverts the original embedding to an estimated noise vector before regenerating it. Results are computed over $n = 100$ samples drawn from the development set.

Table 4.1 reports the results. Two trends are evident. First, switching from V-Prediction to X-Prediction substantially improves both naive sampling ($0.416 \rightarrow 0.706$) and inverse reconstruction ($0.975 \rightarrow 0.998$), confirming that predicting the clean embedding directly yields a more faithful generative model. Second, adding the consistency loss has negligible impact on reconstruction fidelity (naive: 0.704; inverse: 0.991), indicating that the PQ regularization does not degrade the model's ability to reconstruct speaker embeddings. The near-perfect

Model	PQ-Consistency Loss	UTMOS
StyleMod	13.81	3.90
StyleMod + Consistency	4.78	3.78
SPARC + Consistency	5.72	4.06

Table 4.2: PQ-Consistency Loss and UTMOS across model variants. Lower PQ-Consistency Loss indicates that the generated speaker embedding more closely matches the intended PQ target.

inverse cosine similarities across all three StyleMod variants validate the inversion procedure as a reliable mechanism for preserving speaker identity during PQ modification.

PQ-Consistency

A key requirement of controllable voice modification is that the generated output actually reflects the intended PQ targets. To measure this, we evaluate the *PQ-Consistency Loss*: for a random subset of speakers and audio files, we (1) randomly select an audio file and its associated PQ pair, (2) randomly sample a new target PQ-vector, (3) use the model to generate a modified speaker embedding conditioned on the new PQ target, and (4) measure the distance between the PQ values predicted from the generated embedding and the intended PQ target. A lower PQ-Consistency Loss indicates that the model more faithfully realizes the requested PQ modification.

Table 4.2 reports the results. The base *StyleMod* model achieves a PQ-Consistency Loss of 13.81, while *StyleMod + Consistency* reduces this to 4.78, a roughly 65% reduction. *SPARC + Consistency* achieves a PQ-Consistency Loss of 5.72, comparable to StyleMod + Consistency, while attaining the highest UTMOS score (4.06), suggesting that the DDSP-SPARC vocoder produces higher-fidelity speech despite operating in a different embedding space.

To provide a finer-grained view, we evaluate single-axis PQ modifications by measuring two error metrics per PQ dimension: *On-Axis MAE* (the error on the manipulated dimension) and *Off-Axis MAE* (the error on the remaining eight dimensions). Table 4.3 reports the per-dimension MAE results. *StyleMod + Consistency* achieves the lowest on-axis MAE (2.04) and off-axis MAE (1.39), confirming that the consistency loss improves both target fidelity and preservation of non-targeted dimensions. The base *StyleMod* exhibits higher on-axis error (2.82), particularly for breathiness (4.13) and nasality (3.41), reflecting its weaker adherence to PQ targets without the consistency regularizer. *SPARC + Consistency* achieves intermediate on-axis MAE (2.64) but the highest off-axis MAE (1.81), indicating that modifications in the DDSP-SPARC embedding space tend to produce more collateral perturbation to non-targeted qualities.

We additionally evaluate *Ranking Accuracy*: given three generated samples at the de-

Table 4.3: Per-PQ MAE for single-axis modifications: On-Axis MAE (error on the manipulated dimension, ↓) and Off-Axis MAE (error on the remaining dimensions, ↓).

	StyleMod		StyleMod + Consistency		SPARC + Consistency	
PQ	On↓	Off↓	On↓	Off↓	On↓	Off↓
Pitch	1.85	2.61	1.46	1.68	4.08	1.73
Resonance	2.86	1.64	1.31	1.50	1.94	2.08
Weight	2.50	1.65	2.37	1.64	3.04	1.97
Breathiness	4.13	1.56	2.61	1.46	1.84	1.89
Roughness	3.09	1.37	1.73	1.28	3.15	1.76
Strain	2.70	1.30	2.23	1.06	1.94	1.81
Loudness	2.73	1.40	2.14	1.01	3.38	1.64
Creak	2.07	1.66	1.75	1.83	2.23	1.66
Nasality	3.41	1.47	2.73	1.08	2.15	1.74
All	2.82	1.63	2.04	1.39	2.64	1.81

crease, middle, and increase levels of a PQ dimension, we check whether the predicted PQ values are correctly ordered. Table 4.4 reports pairwise accuracies (dec<mid, mid<inc, dec<inc) as well as the full triplet accuracy (all three orderings correct).

StyleMod + Consistency achieves the highest full ranking accuracy (41.5%), driven by strong dec<inc accuracy (92.9%) which indicates that the model reliably produces large-gap orderings even when adjacent comparisons are noisier. Notably, mid<inc accuracy is consistently the weakest comparison across all models, suggesting that distinguishing the "middle" condition from the "increase" condition is the hardest perceptual ordering task. Among individual PQs, resonance is the best-ranked dimension for *StyleMod + Consistency* (65.7% full), while pitch is the most difficult for *SPARC + Consistency* (16.0% full), likely due to the strong coupling between pitch and other source parameters in the DDSP-SPARC vocoder.

Speaker Verification Spoofing

We evaluate whether StyleMod can shift the perceived identity of a source speaker toward a target speaker, as measured by cosine similarity in two speaker-embedding spaces: StyleEncoder and Resemblyzer. All experiments in this subsection are conducted on the VCTK test set. Starting from a source utterance, we apply the interpolation procedure over a sequence of step indices, progressively moving the generated voice toward the target identity.

Table 4.4: Ranking accuracy for single-axis PQ modifications (↑). Columns report pairwise ordering accuracy (dec<mid, mid<inc, dec<inc) and full triplet accuracy (all three correct).

PQ	StyleMod				StyleMod + Consistency				SPARC + Consistency			
	dec <mid	mid <inc	dec <inc	Full	dec <mid	mid <inc	dec <inc	Full	dec <mid	mid <inc	dec <inc	Full
Pitch	.806	.552	.970	.388	.746	.582	1.00	.328	.250	.470	.520	.160
Resonance	.463	.657	.746	.373	.851	.791	.985	.657	.610	.780	.930	.460
Weight	.791	.612	.881	.493	.701	.731	.836	.448	.620	.590	.800	.400
Breathiness	.478	.224	.478	.194	.761	.507	.836	.433	.710	.700	.950	.460
Roughness	.507	.642	.761	.328	.970	.493	.985	.478	.690	.450	.830	.300
Strain	.776	.627	.910	.478	.940	.269	.940	.254	.710	.770	.920	.560
Loudness	.552	.657	.791	.418	.507	.806	.910	.403	.640	.370	.700	.290
Creak	.657	.716	.881	.493	.806	.478	1.00	.284	.500	.820	.920	.390
Nasality	.642	.537	.776	.358	.701	.657	.866	.448	.680	.850	.940	.550
All	.630	.580	.799	.391	.776	.590	.929	.415	.601	.644	.834	.397

Interpolation Trajectory

Figure 4.4 illustrates a representative example (source p259 → target p317). As the step index increases, cosine similarity to the source speaker decreases while similarity to the target speaker increases, confirming that the interpolation path across the speaker-embedding space moves in the intended direction. Both StyleEncoder and Resemblyzer exhibit this monotonic trend. A horizontal reference line indicates the mean intra-speaker similarity, providing an upper bound for what constitutes a plausible match.

Spoofing Effectiveness Across Pairs

To assess spoofing effectiveness at scale, we compute the cosine similarity between each generated utterance and its target speaker before and after interpolation for all source–target pairs in the VCTK test set. Figure 4.5 plots the final (post-interpolation) similarity against the baseline (pre-interpolation) similarity. Points above the $y = x$ line indicate pairs for which interpolation successfully increased similarity to the target. The majority of points lie above this line for StyleStream, but the manipulations have a much less pronounced effect for Resemblyzer.

Spoofing Rate at Verification Thresholds

Figure 4.6 shows the percentage of source-target pairs whose cosine similarity exceeds a given threshold, comparing the baseline condition (unmodified source) against the spoofed

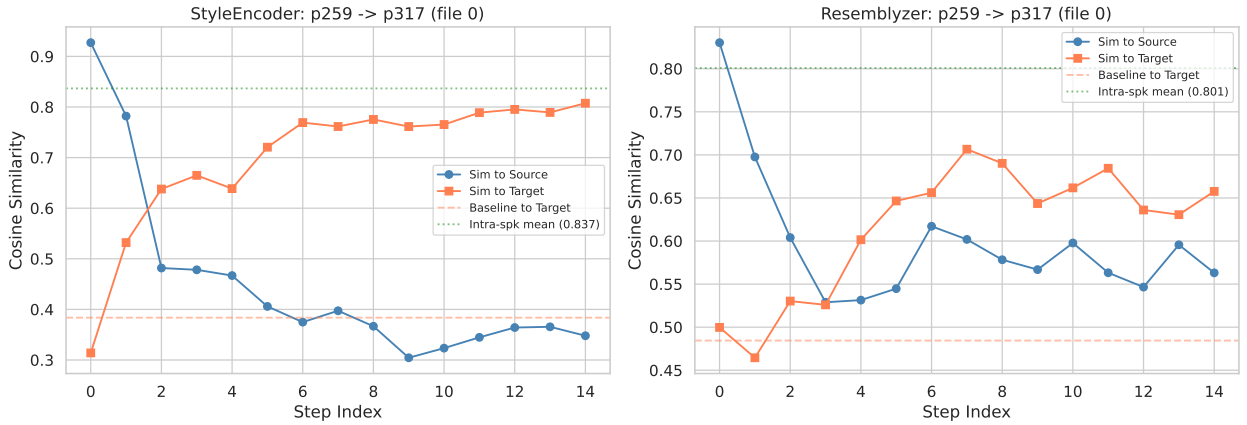


Figure 4.4: Speaker-embedding cosine similarity as a function of interpolation step for source speaker p259 → target speaker p317. Left: StyleEncoder. Right: Resemblyzer. Dashed lines indicate the baseline similarity to the target and the mean intra-speaker similarity.

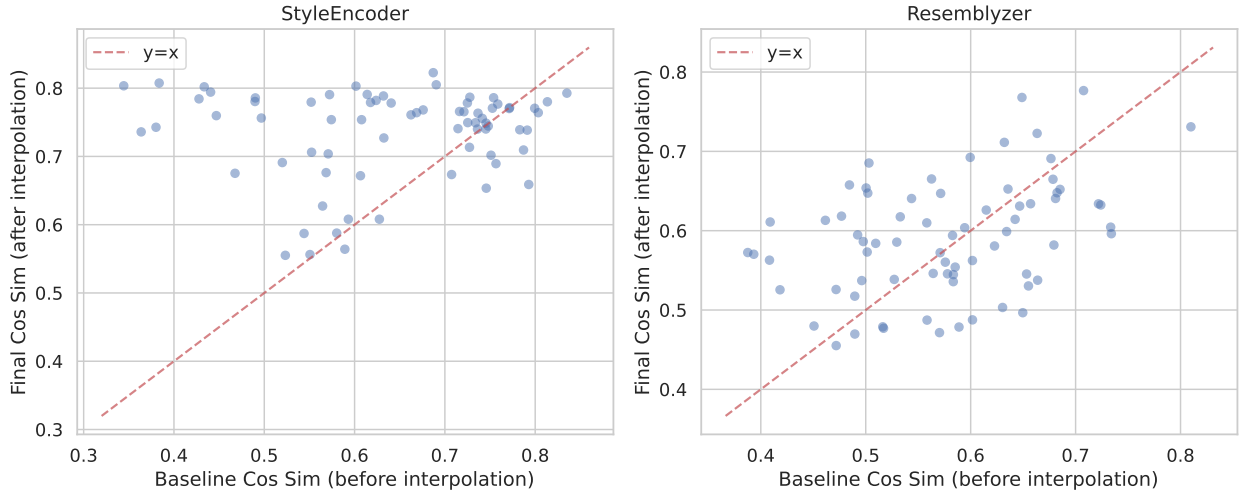


Figure 4.5: Baseline vs. final cosine similarity to the target speaker for all VCTK test pairs. Points above the $y = x$ diagonal indicate successful spoofing. Left: StyleEncoder. Right: Resemblyzer.

condition (post-interpolation). Across a range of thresholds, the spoofed curve lies well above the baseline, indicating that a substantially larger fraction of pairs would pass a fixed-threshold speaker verification system after modification. The gap between curves quantifies the practical spoofing risk introduced by StyleMod.

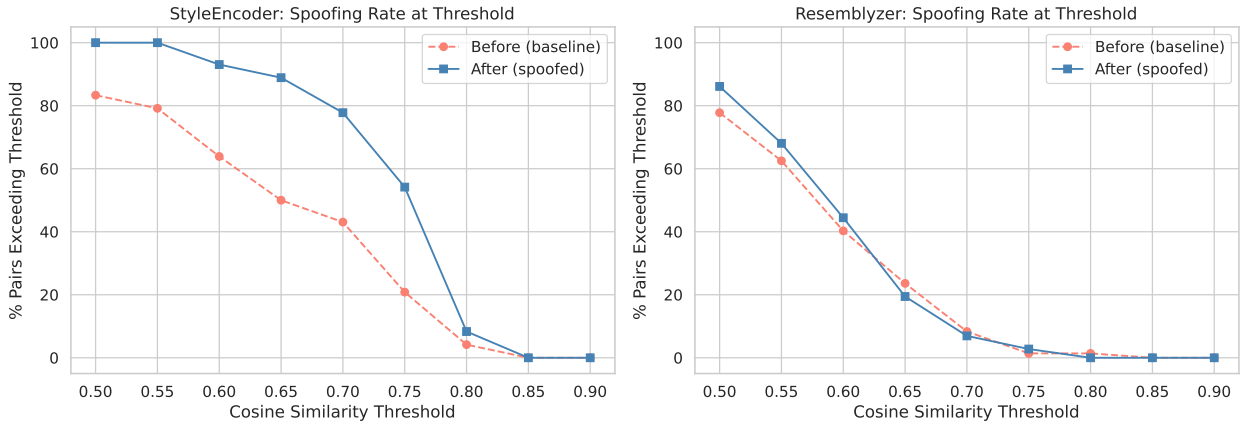


Figure 4.6: Percentage of VCTK test pairs exceeding a cosine similarity threshold before and after spoofing. Left: StyleEncoder. Right: Resemblyzer.

Table 4.5: Conditional pairwise ranking accuracy (≥ 20 yr gap). *Mean* averages per-speaker accuracies; *Overall* pools all pairs across speakers.

Model	Dataset	Pairs	Mean Acc.	Overall Acc.
StyleMod	CREMA	27	0.943	0.926
StyleMod	PVQD	27	0.681	0.704
StyleMod + Consistency	CREMA	30	0.971	0.967
StyleMod + Consistency	PVQD	30	0.767	0.800

Demographic (Age/Gender) Modification

Age Modification

To evaluate the perceptual validity of age modification, we conduct a within-speaker pairwise ranking experiment. For each source speaker, the model generates clips at several target ages; all pairwise comparisons among these clips form one *speaker group*. Listeners are presented with a pair of clips from the same speaker and asked: "From 5 years old to 80 years old, how old is the speaker in this voice clip?" We evaluate 20 speaker groups comprising 114 pairs where the target age gap is at least 20 years, ensuring that only pairs with a clearly distinguishable intended age difference are considered.

Table 4.5 reports per-speaker-group mean and overall pairwise ranking accuracy for pairs with a target age gap ≥ 20 years. Both models achieve strong accuracy on CREMA, with *StyleMod + Consistency* reaching 96.7% overall accuracy. Performance on PVQD is lower, likely reflecting the greater diversity of recording conditions and speaker characteristics in that dataset.

Table 4.6: Pairwise ranking accuracy: all pairs vs. conditional (≥ 20 yr gap).

Model	Dataset	All Pairs	Gap ≥ 20 yr
StyleMod	CREMA	0.786	0.926
StyleMod	PVQD	0.595	0.704
StyleMod + Consistency	CREMA	0.913	0.967
StyleMod + Consistency	PVQD	0.717	0.800

Table 4.6 compares pairwise accuracy computed over all pairs against the conditional subset (gap ≥ 20 years). Across both models and datasets, accuracy improves substantially when only large-gap pairs are considered, indicating that the age modifications produce perceptually consistent orderings for clearly separated target ages.

Gender Perception

We further evaluate whether StyleMod produces perceptually valid gender modifications. Listeners are presented with a single generated clip and asked to rate the perceived gender of the speaker on a continuous scale from 0.0 (Extremely Masculine) to 1.0 (Extremely Feminine). Each clip is generated with one of five intended gender directions: Extreme Masculine, Slight Masculine, Ambiguous, Slight Feminine, and Extreme Feminine.

Figure 4.7 reports the distribution of perceived femininity scores for both models on the CREMA and PVQD datasets. On both datasets, perceived femininity increases monotonically with the intended gender direction, confirming that the modifications shift perception in the expected manner. *StyleMod + Consistency* produces tighter distributions and a steeper gradient across categories compared to the base *StyleMod* model, suggesting that the consistency objective improves the controllability of gender modification. Horizontal reference lines indicate the ground-truth mean perceived femininity of the original dataset (0.54 for CREMA, 0.70 for PVQD), providing context for the absolute scale of the ratings.

Accent Stability

A desirable property of speaker identity modification is that attributes not targeted by the modification, such as accent, remain stable. To evaluate this, we classify the accent of both the original and modified speech using the Whisper-Large-v3 narrow accent classifier from Vox-Profile [95], which categorizes English speech into 16 regional accent groups. We measure the *accent preservation rate*: the fraction of pairs for which the predicted accent is unchanged after modification.

We evaluate accent stability across three modification sources: *demographic modifications* (age and gender), *PQ modifications* (single-axis PQ edits), and *speaker verification spoofing*

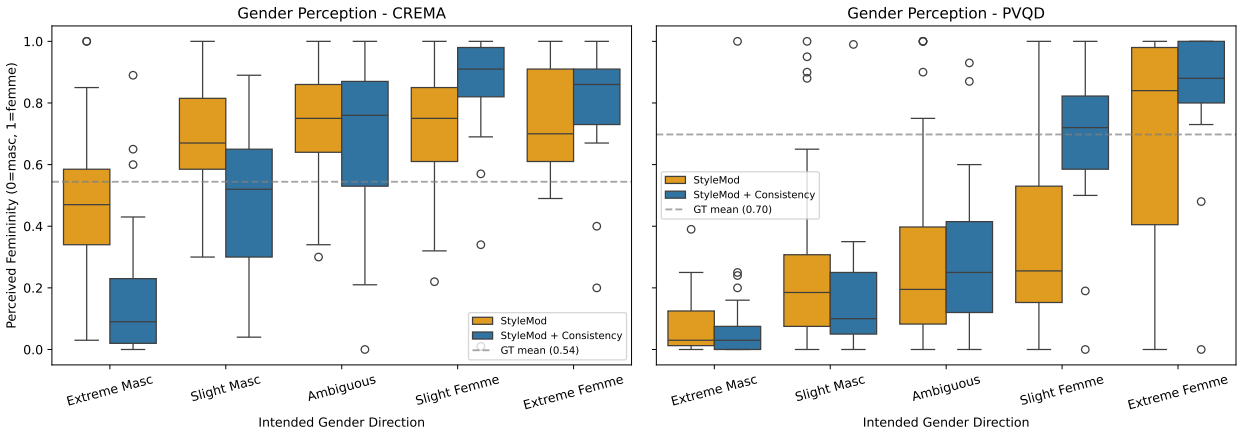


Figure 4.7: Perceived femininity (0 = masculine, 1 = feminine) as a function of intended gender direction on CREMA (left) and PVQD (right). Dashed lines indicate the ground-truth dataset mean.

Table 4.7: Accent preservation rate by model and modification source. Higher values indicate better accent stability.

Model	Source	Preserved/Total	Rate
Baseline	F_0 Shift	269/300	0.897
Baseline	Formant Shift	193/300	0.643
Baseline	F_0 + Formant Shift	202/300	0.673
Baseline	All	664/900	0.738
StyleMod	Demographic	1820/3830	0.475
StyleMod	PQ Mod	1036/1809	0.573
StyleMod	All	2856/5639	0.507
StyleMod + Consistency	Demographic	1686/3830	0.440
StyleMod + Consistency	PQ Mod	912/1809	0.504
StyleMod + Consistency	SV Spoofing	275/665	0.414
StyleMod + Consistency	All	2873/6304	0.456

(identity interpolation toward a target speaker). Table 4.7 reports overall accent preservation rates by model and modification source.

Both models preserve accent in roughly 45–57% of cases, with PQ modifications exhibiting higher preservation rates than demographic or spoofing modifications. This is expected: demographic modifications (particularly gender) alter pitch, resonance, and weight simultaneously, which are among the acoustic cues that accent classifiers rely on, while single-axis

Table 4.8: Accent preservation rate by PQ dimension and modification direction for single-axis PQ edits ($n = 67$ per cell).

	StyleMod				StyleMod + Consistency			
PQ	↓	Mid	↑	Avg	↓	Mid	↑	Avg
Pitch	.612	.463	.522	.532	.358	.433	.328	.373
Resonance	.597	.612	.493	.567	.478	.537	.433	.483
Weight	.582	.537	.612	.577	.328	.493	.567	.463
Breathiness	.493	.582	.567	.547	.478	.552	.597	.542
Roughness	.687	.627	.567	.627	.567	.627	.552	.582
Strain	.582	.537	.582	.567	.552	.478	.463	.498
Loudness	.582	.552	.567	.567	.567	.522	.463	.517
Creak	.522	.612	.582	.572	.582	.552	.522	.552
Nasality	.567	.597	.627	.597	.537	.522	.522	.527

PQ edits produce smaller perturbations. The base *StyleMod* achieves slightly higher preservation rates than *StyleMod + Consistency* across all sources, suggesting that the consistency loss, by enforcing tighter adherence to PQ targets, may induce slightly larger acoustic shifts.

Within demographic modifications, age-targeted edits preserve accent at a higher rate than gender-targeted edits (StyleMod: 52.6% vs. 42.8%; StyleMod + Consistency: 47.9% vs. 40.5%), consistent with the observation that gender modifications involve larger changes to the pitch–resonance–weight subspace.

Table 4.8 reports accent preservation rates broken down by PQ dimension and modification direction for PQ modifications. Across both models, roughness modifications are the most accent-stable (StyleMod: 62.7%; + Consistency: 58.2%), while pitch modifications are the least stable (StyleMod: 53.2%; + Consistency: 37.3%). The low accent stability of pitch modifications is expected, as fundamental frequency is a primary cue for both accent and gender perception. The health-related PQs (breathiness, roughness, strain) generally exhibit higher accent stability than the gender-related PQs (pitch, resonance, weight), consistent with the hypothesis that gender-correlated acoustic features overlap more heavily with accent-discriminative features.

To contextualize these results, we compare against a signal-processing baseline that applies pitch shifting and formant shifting directly to the waveform, without any learned generative model. The baseline achieves an overall accent preservation rate of 73.8%, substantially higher than either StyleMod variant on the same dimensions. The higher overall stability of the signal-processing baseline reflects its more constrained modification: it adjusts only pitch and formant frequencies, whereas StyleMod's learned modifications may induce correlated changes across a broader set of acoustic features.

4.5 Discussion

The results above demonstrate that grounding speaker identity modification in perceptual voice qualities yields a system capable of interpretable, continuous, and targeted manipulation of voice. We organize the discussion around three key findings.

StyleMod enables targeted PQ edits with minimal collateral perturbation. The single-axis modification results (Tables 4.3 and 4.4) confirm that StyleMod + Consistency can modify individual perceptual qualities while largely preserving the remaining dimensions. With an average on-axis MAE of 2.04 and off-axis MAE of 1.39, modifications are both faithful to the intended target and contained to the targeted axis. This disentanglement mirrors the acoustic-physical-perceptual correlations established in Chapter 2: just as $\bar{F}_0$ and ΔF exhibit largely disentangled correlations with pitch and resonance at the production level (Table 3.3), StyleMod achieves largely disentangled modifications at the generation level. The consistency loss is critical for this property—without it, on-axis MAE increases by 38% and the model fails to reliably rank adjacent PQ levels (Table 4.4).

StyleMod produces perceptually valid demographic modifications, but with asymmetric difficulty. Gender modification benefits from the well-characterized relationship between pitch, resonance, and weight established in Chapter 2 and Section 3.4: the hard-coded offset strategy succeeds because these three PQ axes capture the primary perceptual correlates of gendered voice, as validated by crowdsourced perception studies (Section 2.7). Age modification, by contrast, is inherently more diffuse. The grid-search heuristic achieves strong pairwise ranking accuracy for large age gaps (96.7% on CREMA-D), but lower performance for fine-grained distinctions (Table 4.6). This mirrors the perceptual instability of age estimation found in Section 2.7, where listeners demonstrated reliable ordinal but poor interval-scale age perception. StyleMod's age modification inherits the same asymmetry: coarse aging effects are achievable through simultaneous manipulation of strain, roughness, and breathiness, but precise age targeting remains limited by the diffuse and overlapping nature of age-related vocal changes.

StyleMod shifts voice type but does not fully spoof speaker identity. The speaker verification spoofing results (Section 4.4) reveal that while StyleMod can shift cosine similarity toward a target speaker within the StyleStream embedding space, the effect is substantially weaker in the independent Resemblyzer space. This is expected given the findings of Section 3.7: a 9-dimensional PQ vector captures a meaningful but incomplete subspace of speaker identity (EER of $\sim$19% vs. Resemblyzer's 3%). The "non-PQ" information remains encoded in the estimated noise vector $\tilde{z}_0$ during inversion, preventing full identity transfer. From a safety perspective, this is a desirable property: StyleMod enables interpretable modification of *how* a voice sounds without enabling wholesale impersonation of *who* a voice belongs to. The gap between in-distribution spoofing (StyleStream) and cross-model spoof-

ing (Resemblyzer) quantifies the degree to which learned modifications are system-specific rather than perceptually general.

Broader implications. Taken together, these results validate the central thesis: that perceptual voice qualities provide a scientifically grounded and practically useful intermediate representation for speaker identity modification. The PQ space defined in Chapter 2, validated as perceivable by non-experts and correlated with acoustic measurements, and shown in Section 3.7 to encode meaningful demographic and identity information, here serves as the control interface for a generative model that produces perceptually valid modifications. The approach trades the fine-grained speaker discrimination of high-dimensional embeddings for interpretability and controllability. We believe this trade-off is appropriate for applications in gender-affirming voice training, speaker anonymization, and expressive speech synthesis where understanding *what* is being modified matters as much as the modification itself.

4.6 Limitations

The PQ space provides a low-dimensional, interpretable intermediate representation of speaker identity, but several limitations of the current approach merit discussion.

First, StyleMod's performance is dependent on the underlying speaker identity representation. The results presented here are primarily obtained using StyleStream embeddings, and the lower reconstruction quality observed with DDSP-SPARC (Table 4.1) suggests that the approach does not transfer equally well to all synthesis backbones. The geometry of the speaker embedding space—how speaker identity is distributed and how PQ-relevant information is encoded—varies across architectures, and StyleMod's conditional flow matching must learn a different manifold structure for each. There is no fundamental reason why this schema could not generalize to additional architectures; in our prior work [96], we generated mel-spectrum features in addition to speaker embeddings.

Second, accent stability remains a challenge. StyleMod preserves accent in only 45-57% of modifications, compared to 73.8% for a simple signal processing baseline. Entanglement between voice quality and accent has been studied in phonetics and speech forensics [97, 1], but these insights have not been integrated into speech models. While we believe that PQs can be modelled in an accent-independent manner (which the VVD does demonstrate for resonance and weight), larger and more diverse datasets of voice quality labels are needed to properly generalize across accents.

Finally, the static, utterance-level PQ-vector misses temporal dynamics that are necessary for modeling emotion and speaking style. As shown in Section 3.7, the PQ-vector achieves only 40.6% accuracy on emotion prediction, well below learned embeddings that capture prosodic variation. Extending PQs to time-varying representations and better standardization of the gendered PQs (particularly weight) are promising directions for future work.

Chapter 5

Conclusion

This thesis has taken a step toward understanding the limits of the human voice, and toward building systems that operate at the level of human perception rather than below it.

5.1 Summary of Contributions

The central claim of this work is that perceptual voice qualities provide the right level of abstraction for understanding and modifying speaker identity: low-dimensional enough to be interpretable, perceptually grounded enough to be meaningful, and acoustically substantive enough to drive generation. We have supported this claim through three stages of research.

In Chapter 2, we established the empirical foundations. The Versatile Voice Dataset and Disentangled Source-Filter Dataset demonstrated that single speakers can traverse the perceptual space of gender through deliberate modification of pitch, resonance, and weight. This degree of vocal flexibility challenges assumptions underlying most speech systems. Crowdsourced perception studies demonstrated that non-experts can reliably hear perceptual qualities (average correlation of 0.77 with experts), and that gender perception is stable across task types while age perception is ordinally reliable under pairwise comparison. These findings validated PQs as a perceivable and meaningful representation of voice.

In Chapter 3, we demonstrated the computational consequences of vocal flexibility: state-of-the-art speaker verification systems degrade from sub-1% EER to 21–29% EER on the VVD, failing in proportion to the magnitude of vocal modification. We established disentangled acoustic correlates of pitch ($\bar{F}_0$, 99.3% ranking accuracy), resonance (ΔF, 93.8%), and weight (contact quotient, 93.3%) on a single speaker. Scaling to multiple speakers, we presented the PQ-Encoder, which predicts a 9-dimensional PQ-vector from speech and captures meaningful information about speaker identity, gender, age, and affect, occupying an interpretable subspace of what high-dimensional embeddings encode opaquely.

In Chapter 4, we demonstrated that perceptual grounding enables not only analysis but modification. StyleMod, a conditional flow matching model with PQ consistency regularization, achieves targeted single-axis modifications (on-axis MAE of 2.04, off-axis MAE of 1.39), perceptually valid gender modification across the masculinity-femininity spectrum,

and age modification with 96.7% pairwise ranking accuracy for large age gaps. Crucially, these modifications preserve aspects of identity that PQs do not capture, as evidenced by the gap between in-distribution and cross-model spoofing.

5.2 What This Work Means

This thesis argues for a particular way of doing speech science and technology: grounding computational representations in human perception, centering diverse voices rather than treating them as invisible edge cases, and valuing interpretability alongside performance. The PQ framework is not merely a technical contribution; it reflects a belief that the right abstractions for understanding the voice already exist in the communities that study and modify voices daily: speech-language pathologists, voice teachers, and trans individuals navigating the world through sound.

The results show both the promise and the boundaries of this approach. Nine perceptual dimensions capture enough information to predict gender, estimate relative age, and drive targeted voice modification, but not enough to uniquely identify a speaker or model the temporal dynamics of emotion. This is not a failure; it is a feature. The gap between what PQs capture and what full speaker embeddings encode is precisely the space in which non-targeted identity lives (accent, prosodic habits, articulatory idiosyncrasies), and preserving that space is what makes PQ-based modification interpretable rather than destructive.

5.3 Future Work

The limitations of this thesis point directly toward its continuation.

Loudness-controlled weight modification. Weight remains the most difficult PQ to model, with the lowest expert ICC (0.77), the lowest ranking accuracy among gendered PQs, and persistent entanglement with loudness. One explanation for the gap between human experts and the PQ-Encoder on weight is an over-reliance on loudness differences in the training signal. A targeted data collection effort, asking voice teachers to modify weight while explicitly controlling loudness, could disentangle these two dimensions, if such isolated modification is physiologically possible. This would also address the broader question of whether weight is a single perceptual construct or a composite of multiple laryngeal properties.

Time-varying voice quality. This thesis considers voice qualities at the utterance or speaker level, but much of what makes a voice unique lies in temporal dynamics: how qualities shift across an utterance, a conversation, or a lifetime. The 40.6% emotion prediction accuracy of the static PQ-vector reflects this limitation directly. Extending PQs to time-varying representations, perhaps as trajectories through PQ space rather than static points,

would enable modeling of prosodic style, emotional expression, and the micro-variations that distinguish speakers with similar average PQ profiles.

Accent-independent PQ modeling. StyleMod preserves accent in only 45–57% of modifications, revealing entanglement between PQ-relevant acoustic features and accent-discriminative cues. The VVD demonstrates that resonance and weight *can* be modified independently of accent by skilled voice teachers, suggesting that the entanglement is a property of the model rather than of the PQ space itself. Larger and more diverse datasets of PQ labels, spanning accents, languages, and speaking styles, are needed to learn accent-invariant PQ representations.

Standardization of perceptual qualities. The disagreements among voice teachers on the precise definition of weight ("buzzy" vs. "rumbly"), and the varying interpretations of "low," "medium," and "high" configurations in the VVD, reflect the absence of formalized research on gendered PQs. The pairwise comparison framework adopted in this thesis sidesteps calibration differences, but a long-term goal should be developing standardized protocols, analogous to CAPE-V for clinical PQs, that enable consistent labeling of resonance, weight, and creak across institutions and communities.

Beyond English, beyond gender. The datasets and models in this thesis are limited to English speakers and to the gendered and health-related PQ axes most relevant to the GAVT community. Voice qualities vary across languages, tonal systems, and cultural contexts in ways that the current framework does not address. Extending PQ-based representations to multilingual settings, and exploring additional perceptual dimensions relevant to other communities and applications, is a natural next step.

The voice is not a fixed property of a speaker but a flexible instrument. This thesis has shown that perceptual voice qualities are the language for describing how it is played, and that this language can ground computational systems in human understanding. We have so much more to learn about how voices work, how they are heard, and how they can change. But the foundation is here: a way of listening that is both human and computational, both interpretable and generative.

Bibliography

[1] J. H. Esling, S. R. Moisik, A. Benner, and L. Crevier-Buchman, *Voice quality: The laryngeal articulator model*. Cambridge University Press, 2019, vol. 162.

[2] G. B. Kempster, B. R. Gerratt, K. V. Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol," *American Journal of Speech-Language Pathology*, vol. 18, no. 2, pp. 124–132, 2009.

[3] L. Zimman, "Gender as stylistic bricolage: Transmasculine voices and the relationship between fundamental frequency and/s," *Language in Society*, vol. 46, no. 3, pp. 339–370, 2017.

[4] R. K. Adler, S. Hirsch, and J. Pickering, *Voice and communication therapy for the transgender/gender diverse client: A comprehensive clinical guide*. Plural Publishing, 2018.

[5] A. Z. Huff, "Modern responses to traditional pitfalls in gender affirming behavioral voice modification," *Otolaryngologic Clinics of North America*, vol. 55, no. 4, pp. 727–738, 2022.

[6] R. Netzorg, B. Yu, A. Guzman, P. Wu, L. McNulty, and G. K. Anumanchipalli, "Towards an interpretable representation of speaker identity via perceptual voice qualities," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 391–12 395.

[7] F. Rautenberg, M. Kuhlmann, F. Seebauer, J. Wiechmann, P. Wagner, and R. Haeb-Umbach, "Speech synthesis along perceptual voice quality dimensions," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[8] J. Narain, V. Kowtha, C. Lea, L. Tooley, D. Yee, V. Mitra, Z. Huang, M. Espi Marques, J. Huang, C. Avendano, and S. Ren, "Voice quality dimensions as interpretable primitives for speaking style for atypical speech and affect," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2025)*, Rotterdam, The Netherlands, Aug. 2025, pp. 4628–4632.

[9] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, p. 292, 2013.

[10] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[11] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[12] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech 2020*. ISCA, Oct. 2020. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2650

[13] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang, Z. Wu, T. Qin, X.-Y. Li, W. Ye, S. Zhang, J. Bian, L. He, J. Li, and S. Zhao, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," 2024.

[14] R. D. Kent, "Vocal tract acoustics," *Journal of Voice*, vol. 7, no. 2, pp. 97–117, 1993.

[15] N. Lavan, "The time course of person perception from voices: A behavioral study," *Psychological Science*, vol. 34, no. 7, pp. 771–783, 2023.

[16] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality: review, tutorial, and a framework for future research," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 1, pp. 21–40, 1993.

[17] E. Keller, *The Analysis of Voice Quality in Speech Processing*. Berlin, Heidelberg: Springer-Verlag, 2005, p. 54–73.

[18] P. Walden, "Perceptual voice qualities database (pvqd)," 2020.

[19] L. Carew, G. Dacakis, and J. Oates, "The effectiveness of oral resonance therapy on the perception of femininity of voice in male-to-female transsexuals," *Journal of voice*, vol. 21, no. 5, pp. 591–603, 2007.

[20] M. A. García and A. L. Rosset, "Deep neural network for automatic assessment of dysphonia," 2022.

[21] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, "Emotion intensity and its control for emotional voice conversion," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 31–48, 2022.

[22] S. Hidaka, Y. Lee, M. Nakanishi, K. Wakamiya, T. Nakagawa, and T. Kaburagi, "Automatic grbas scoring of pathological voices using deep learning and a small set of labeled voice data," *Journal of Voice*, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199722003472

[23] S. Fujimura, T. Kojima, Y. Okanoue, K. Shoji, M. Inoue, K. Omori, and R. Hori, "Classification of voice disorders using a one-dimensional convolutional neural network," *Journal of Voice*, vol. 36, no. 1, pp. 15–20, 2022.

[24] N. Lavan, "Person perception from voices: Impressions of physical, trait, and social characteristics emerge with 800ms of exposure," 2022.

[25] K. Johnson and M. J. Sjerps, "Speaker normalization in speech perception," *The handbook of speech perception*, pp. 145–176, 2021.

[26] C. Brennan, A. C. Hannah, J. Romick, J. W. Lewon, and C. Meyers, "Differences and similarities in the perception of voice gender for individuals who are or are not members of the lgbt+ community," *Journal of Voice*, 2022.

[27] M. Hope and J. Lilley, "Gender expansive listeners utilize a non-binary, multidimensional conception of gender to inform voice gender perception," *Brain and language*, vol. 224, p. 105049, 2022.

[28] B. M. Merritt, *Perceptual representation of speaker gender*. Indiana University, 2022.

[29] B. Merritt and T. Bent, "Revisiting the acoustics of speaker gender perception: A gender expansive perspective," *The Journal of the Acoustical Society of America*, vol. 151, no. 1, pp. 484–499, 2022.

[30] C. E. Howerton, D. P. Buckley, K. L. Dahl, and C. E. Stepp, "The effects of speaker head posture on auditory perception of vocal masculinity," *Journal of Voice*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199724002492

[31] R. Netzorg, A. Cote, S. Koshin, K. V. Garoute, and G. K. Anumanchipalli, "Speech after gender: A trans-feminine perspective on next steps for speech science and technology," in *Proc. Interspeech 2024*, 2024, pp. 3075–3079.

[32] E. Székely and M. Hope, "An inclusive approach to creating a palette of synthetic voices for gender diversity."

[33] E. J. Bush, B. I. Krueger, M. Cody, J. D. Clapp, and V. D. Novak, "Considerations for voice and communication training software for transgender and nonbinary people," *Journal of Voice*, vol. 38, no. 5, pp. 1251–e1, 2024.

[34] Z. K. Li, M. M. Chen, Y. Zhong, P. Liu, and Z. Duan, "Gtr-voice: Articulatory phonetics informed controllable expressive speech synthesis," in *Interspeech*, vol. 2024, 2024, pp. 1775–1779.

[35] F. Rautenberg, M. Kuhlmann, F. Seebauer, J. Wiechmann, P. Wagner, and R. Haeb-Umbach, "Speech synthesis along perceptual voice quality dimensions," 2025. [Online]. Available: https://arxiv.org/abs/2501.08791

[36] T. L. Hardy, J. M. Rieger, K. Wells, and C. A. Boliek, "Acoustic predictors of gender attribution, masculinity–femininity, and vocal naturalness ratings amongst transgender and cisgender speakers," *Journal of Voice*, vol. 34, no. 2, pp. 300–e11, 2020.

[37] R. Singh, *Profiling humans from their voice*. Springer, 2019, vol. 41.

[38] K. Sebastian and L. Mary, "Fasr: Effect of voice disguise," in *2016 International Conference on Emerging Technological Trends (ICETT)*, 2016, pp. 1–4.

[39] P. Staroniewicz, "Effect of deliberate and nondeliberate natural voice disguise on speaker recognition performance," *Acoustics, Acoustoelectronics and Electrical Engineering*, vol. 888, pp. 312–325, 2021.

[40] B. Merritt and T. Bent, "Perceptual evaluation of speech naturalness in speakers of varying gender identities," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 7, pp. 2054–2069, 2020.

[41] K. Povinelli, H. Zhu, Y. Zhao *et al.*, "Beyond the" industry standard": Focusing gender-affirming voice training technologies on individualized goal exploration," *arXiv preprint arXiv:2410.09958*, 2024.

[42] J. Gill-Peterson, *A Short History of Trans Misogyny*. Verso Books, 2024.

[43] D. V. Dolquist and B. Munson, "Clinical focus: The development and description of a palette of transmasculine voices," *American journal of speech-language pathology*, vol. 33, no. 3, pp. 1113–1126, 2024.

[44] M. Hope, C. Ward, and J. Lilley, "Nonbinary american english speakers encode gender in vowel acoustics," 2023.

[45] Z. Zhang, "Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model," *The Journal of the Acoustical Society of America*, vol. 139, no. 4, pp. 1493–1507, 2016.

[46] R. Netzorg, N. Carvalho, A. Guzman, L. Wang, J. Francis, K. V. Garoute, K. Johnson, and G. Anumanchipalli, "On the production and perception of a single speaker's gender," in *26th Interspeech Conference 2025, Rotterdam, Netherlands, Kingdom of the, August 17-21, 2025*. International Speech Communication Association, 2025, pp. 669–673.

[47] H. Lameris, S. H. B. Satish, J. Gustafson, and Éva Székely, "Lost in phonation: Voice quality variation as an evaluation dimension for speech foundation models," 2025. [Online]. Available: https://arxiv.org/abs/2510.25577

[48] V. C. Mathad, N. J. Scherer, K. L. Chapman, J. M. Liss, and V. Berisha, "A deep learning algorithm for objective assessment of hypernasality in children with cleft palate," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 10, pp. 2994–3003, 2021.

[49] T. C. Stone and M. L. Erickson, "Experienced listeners' perception of timbre dissimilarity within and between voice categories," *Journal of Voice*, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199722004271

[50] Z. Zhang, J. Zhang, and J. Kreiman, "Effects of laryngeal manipulations on voice gender perception," *Interspeech Proceedings*, 2022.

[51] D. Markova, L. Richer, M. Pangelinan, D. H. Schwartz, G. Leonard, M. Perron, G. B. Pike, S. Veillette, M. M. Chakravarty, Z. Pausova *et al.*, "Age-and sex-related variations in vocal-tract morphology and voice acoustics during adolescence," *Hormones and behavior*, vol. 81, pp. 84–96, 2016.

[52] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein *et al.*, "Usc-timit: A database of multimodal speech production data," USC, Tech. Rep., 2013.[Online] http://sail. usc. edu/span/usc-timit . . . , Tech. Rep., 2013.

[53] A. Wrench, "A multichannel articulatory speech database and its application for automatic speech recognition," in *Proc. 5th seminar on speech production: models and data, 2000*, 2000.

[54] K. F. Nagle, "Clinical use of the cape-v scales: Agreement, reliability and notes on voice quality," *Journal of Voice*, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199722003666

[55] T. McAllister, C. Nightingale, G. Moya-Gale, A. Kawamura, and L. Ramig, "Crowdsourced perceptual ratings of voice quality in people with parkinson's disease before and after intensive voice and articulation therapies: Secondary outcome of a randomized controlled trial," *Journal of Speech, Language, and Hearing Research*, vol. 66, pp. 1–22, 04 2023.

[56] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[57] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)." University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019.

[58] J. Lian, C. Zhang, and D. Yu, "Robust disentangled variational speech representation learning for zero-shot voice conversion," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6572–6576.

[59] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[60] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, "Ecapa-tdnn embeddings for speaker diarization," in *Interspeech 2021*. ISCA, Aug. 2021. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2021-941

[61] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker anonymization using neural audio codec language models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4725–4729.

[62] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," *arXiv preprint arXiv:1905.13561*, 2019.

[63] V. A. Trinh, P. Ghahremani, B. King, J. Droppo, A. Stolcke, and R. Maas, "Reducing geographic disparities in automatic speech recognition via elastic weight consolidation," in *Interspeech 2022*. ISCA, Sep. 2022. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2022-11063

[64] T. Feng, R. Hebbar, N. Mehlman, X. Shi, A. Kommineni, and S. Narayanan, "A review of speech-centric trustworthy machine learning: Privacy, safety, and fairness," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 3, 2023. [Online]. Available: http://dx.doi.org/10.1561/116.00000084

[65] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan, J. Wang, I. Cruz, B. Akula, A. Akinyemi, B. Ellis, R. Moritz, Y. Yungster, A. Rakotoarison, L. Tan, C. Summers, C. Wood, J. Lane, M. Williamson, and W.-N. Hsu, "Audiobox: Unified audio generation with natural language prompts," 2023.

[66] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "Prompttts: Controllable text-to-speech with text descriptions," 2022.

[67] E. B. Kang, "On the praxes and politics of ai speech emotion recognition," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 455–466. [Online]. Available: https://doi.org/10.1145/3593013.3594011

[68] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[69] E. Harper, S. Majumdar, O. Kuchaiev, L. Jason, Y. Zhang, E. Bakhturina, V. Noroozi, S. Subramanian, K. Nithin, H. Jocelyn, F. Jia, J. Balam, X. Yang, M. Livne, Y. Dong, S. Naren, and B. Ginsburg, "Nemo: a toolkit for conversational ai and large language models," 2024. [Online]. Available: https://nvidia.github.io/NeMo/

[70] S. Ellis, S. Goetze, and H. Christensen, "Moving towards non-binary gender identification via analysis of system errors in binary gender classification," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[71] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1652–1656.

[72] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech 2017*. ISCA, Aug. 2017. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-950

[73] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 161–165.

[74] O. Perrotin and I. McLoughlin, "A spectral glottal flow model for source-filter separation of speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7160–7164.

[75] H. Ghasemzadeh, R. E. Hillman, and D. D. Mehta, "Toward generalizable machine learning models in speech, language, and hearing sciences: Estimating sample size and reducing overfitting," *Journal of Speech, Language, and Hearing Research*, vol. 67, no. 3, pp. 753–781, 2024.

[76] B. Yu and K. Kumbier, "Veridical data science," *Proceedings of the National Academy of Sciences*, vol. 117, no. 8, p. 3920–3929, Feb. 2020. [Online]. Available: http://dx.doi.org/10.1073/pnas.1901326117

[77] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," *arXiv preprint arXiv:2204.02152*, 2022.

[78] P. Wu, L.-W. Chen, C. J. Cho, S. Watanabe, L. Goldstein, A. W. Black, and G. K. Anumanchipalli, "Speaker-independent acoustic-to-articulatory speech inversion," 2023.

[79] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[80] M. Paierl, T. Röck, S. Wepner, A. Kelterer, and B. Schuppler, "Creapy: A python-based tool for the detection of creak in conversational speech," in *20th International Congress on Phonetic Sciences*, 2023.

[81] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "Libritts-r: A restored multi-speaker text-to-speech corpus," *arXiv preprint arXiv:2305.18802*, 2023.

[82] Y. Liu, N. Lee, and G. Anumanchipalli, "Stylestream: Real-time zero-shot voice style conversion," 2026. [Online]. Available: https://arxiv.org/abs/2602.20113

[83] T. Li and K. He, "Back to basics: Let denoising generative models denoise," *arXiv preprint arXiv:2511.13720*, 2025.

[84] Y. Hu, C. Zhang, J. Shi, J. Lian, M. Ostendorf, and D. Yu, "Prosodybert: Self-supervised prosody representation for style-controllable tts," 2022.

[85] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[86] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles *et al.*, "De novo design of protein structure and function with rfdiffusion," *Nature*, pp. 1–3, 2023.

[87] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[88] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[89] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[90] Z. Du, C. Gao, Y. Wang, F. Yu, T. Zhao, H. Wang, X. Lv, H. Wang, C. Ni, X. Shi *et al.*, "Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training," *arXiv preprint arXiv:2505.17589*, 2025.

[91] Y. Wang, Z. Ju, X. Tan, L. He, Z. Wu, J. Bian, and S. Zhao, "Audit: Audio editing by following instructions with latent diffusion models," *arXiv preprint arXiv:2304.00830*, 2023.

[92] C. J. Cho, P. Wu, T. S. Prabhune, D. Agarwal, and G. K. Anumanchipalli, "Coding speech through vocal tract kinematics," *IEEE Journal of Selected Topics in Signal Processing*, vol. 18, no. 8, pp. 1427–1440, 2024.

[93] Y. Liu, B. Yu, D. Lin, P. Wu, C. J. Cho, and G. K. Anumanchipalli, "Fast, high-quality and parameter-efficient articulatory synthesis using differentiable dsp," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 711–718.

[94] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[95] T. Feng, J. Lee, A. Xu, Y. Lee, T. Lertpetchpun, X. Shi, H. Wang, T. Thebaud, L. Moro-Velazquez, D. Byrd *et al.*, "Vox-profile: A speech foundation model benchmark for characterizing diverse speaker and speech traits," *arXiv preprint arXiv:2505.14648*, 2025.

[96] R. Netzorg, B. Yu, A. Guzman, P. Wu, L. McNulty, and G. K. Anumanchipalli, "Towards an interpretable representation of speaker identity via perceptual voice qualities," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 391–12 395.

[97] E. Gold, C. Kirchhübel, K. Earnshaw, and S. Ross, "Regional variation in british english voice quality," *English World-Wide*, vol. 43, no. 1, pp. 96–123, 2022. [Online]. Available: https://www.jbe-platform.com/content/journals/10.1075/eww.20007.gol