

TOWARDS AN INTERPRETABLE REPRESENTATION OF SPEAKER IDENTITY VIA PERCEPTUAL VOICE QUALITIES

*Robin Netzorg*¹, *Bohan Yu*¹, *Andrea Guzman*¹,
*Peter Wu*¹, *Luna McNulty*², *Gopala Anumanchipalli*¹
¹University of California, Berkeley, ²Brown University

ABSTRACT

Unlike other data modalities such as text and vision, speech does not lend itself to easy interpretation. While lay people can understand how to describe an image or sentence via perception, non-expert descriptions of speech often end at high-level demographic information, such as gender or age. In this paper, we propose a possible interpretable representation of speaker identity based on perceptual voice qualities (PQs). By adding gendered PQs to the pathology-focused Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) protocol, our PQ-based approach provides a perceptual latent space of the character of adult voices that is an intermediary of abstraction between high-level demographics and low-level acoustic, physical, or learned representations. Contrary to prior belief, we demonstrate that these PQs are hearable by ensembles of non-experts, and further demonstrate that the information encoded in a PQ-based representation is predictable by various speech representations.

Index Terms— Speech Representation, Speaker Identity, Perceptual Qualities

1. INTRODUCTION

When a lay person hears a voice, they can quickly identify certain features, such as whether or not the voice is masculine or feminine, or old or young [1]. This level of abstraction, however, does not shed light on the building blocks of a voice. Intermediate representations of speech are seemingly locked behind a veil of expertise, time-consuming to understand and contained to sub-fields. While there have been certain combinations of various fields and speech processing, such as speech pathology and musical vocal training, these combinations have been disconnected and not applied towards creating a complete representation of speaker identity.

In this work, we holistically consider one such intermediate representation from speech pathology: perceptual voice quality. Prior work has explored the ability of crowdsourcing and machine learning methods to capture these individual qualities [2, 3], but, to the best of our knowledge, a study of perceptual qualities’ ability to represent speaker identity has not been conducted. Perceptual qualities related to voice atypicality alone lack the ability to provide a comprehensive

latent space of adult voices, since gender information is missing. As such, we take inspiration from musical and transgender vocal training, supplementing perceptual qualities from speech pathology with what we call gendered perceptual qualities. Combined with perceptual qualities from the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) protocol [4], we propose a 7-dimensional representation of spoken speaker identity based on perceptual qualities.

A perceptual quality-based representation of speaker identity provides two benefits currently lacking in other representations. Firstly, a perceptual quality-based representation of speaker identity is low-dimensional and interpretable, whereby any listener or ensemble of listeners, given minimal training, can hear the particular aspects of each perceptual quality. Secondly, the information encoded in subjective perceptual qualities has an objective basis, containing information present in varied representations of speech, from hand-crafted to self-supervised. Perceptual qualities have the potential to bring to speech processing what it has previously lacked: a perceivable and descriptive level of abstraction of the texture of a speaker’s voice.

2. RELATION TO PRIOR WORK

2.1. 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 [5, 6]. 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 [6].

Of particular interest to our work 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 [4, 7]. Non-surgical treatment of a voice involves a patient performing exercises to bring certain perceptual qualities into

healthy levels [4]. 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 [8, 9].

Prior work in automatic assessment of dysphonia [10] and emotion recognition [11] have explored the predictability of individual perceptual qualities. Deep neural network approaches to predicting perceptual qualities have been conducted [2, 12], 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 extend on prior work by demonstrating that automatic detection of perceptual qualities is possible at a human level across multiple representations of spoken sentences.

2.2. Representations of Speaker Identity

From speaker verification and identification [13, 14] to voice conversion [15], an informative representation of speaker identity is necessary for many problems in speech processing. Across nearly all modern methods, especially those based in deep learning [16], the highest performing methods are learned representations that, while containing information relevant to speaker identity, lack highly low-dimensional interpretability of the PQ-based representation proposed here.

3. PERCEPTUAL VOICE QUALITIES

In this section, we describe the collection and interpretability of perceptual voice qualities. While the space of possible PQs is vast, we limit our consideration to seven PQs, described below.

3.1. 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 [4]. SLPs undergo training to successfully identify the PQs: strain, loudness, roughness, breathiness, pitch, and severity [7]. For a complete description of the CAPE-V vocal qualities, please refer to the original protocol [4].

While this data is difficult to collect, the Perceptual Voice Qualities Database (PVQD) serves a publicly available ratings of PQs from the CAPE-V scale [7]. 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. In this work, we examine five of the six PQs, excluding severity, which is overall measure of vocal atypicality.

3.2. PVQD+: 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 of 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. Current pedagogies in transgender voice modification are particularly concerned with two primary PQs: vocal resonance and vocal weight. These correspond to two physiological differences [17] that distinguish male and female voices. 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 [18], 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) [19]. These ideas are also often used in singing lessons for vocalists, but the focus of transgender voice training on gender lines lends itself to a representation of spoken voice.

As discussed above, three voice teachers listened to the 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 1. ICC is a common metric for measuring inter-rater reliability [7]. We see that the expert ICC of both resonance and weight are the maximum and minimum of the ICCs for all PQs, respectively. We note that the

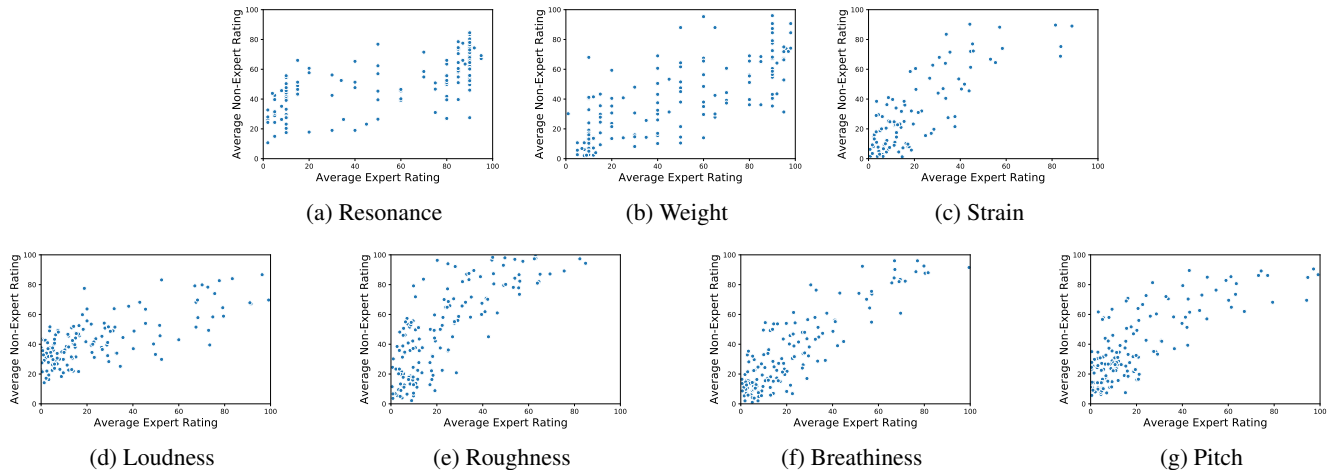


Fig. 1: Average Expert Rating (x-axis) vs. Average Non-Expert Rating (y-axis) across perceptual qualities.

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.

3.3. Can Non-Experts Hear Perceptual Qualities?

A common statement from voice teachers and speech language pathologists is that hearing perceptual qualities requires training [20, 21]. 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 [2]. Recent work has demonstrated that non-experts can accurately label the overall atypicality of a voice [3], 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.

In this section, we test the ability of non-experts to accurately rate the PQs of PVQD voice clips. Using the Amazon Mechanical Turk (AMT), we ask 6 workers with master’s qualifications to rate the clips using the CAPE-V protocol, and the resonance and weight as described in Section 3.2. Workers are 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 1 demonstrates, average non-expert ratings achieve a correlation 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, as reported in Figure 2, 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.

The RMSE between non-expert and expert ratings suggests a potential flaw in rating ability of non-experts, but visualizing the average non-expert ratings vs. average expert ratings, reported in Figure 1 reveals non-expert biases across PQs. 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, we see higher levels of deviation between non-expert and expert ratings, but similar trends hold. Non-experts appropriately label voices with high-values of a perceptual quality with a high-value, and vice-versa.

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.

Rater	Resonance	Weight	Strain	Loudness	Roughness	Breathiness	Pitch	Average
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 1: Correlation of Non-Expert Labels with Expert Labels. A measure of inter-rater agreement, Intra-class Correlation Coefficient, is reported for Expert Labels.

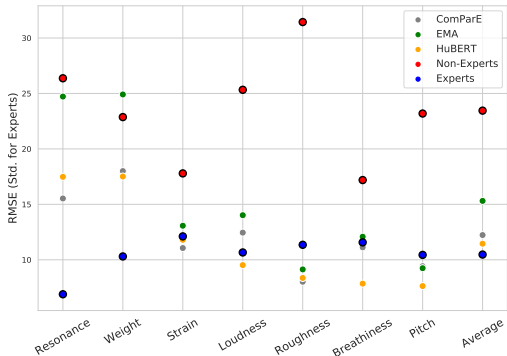


Fig. 2: Test RMSE of various rating methods when compared with the average expert rating for each perceptual quality. Standard Deviation reported for expert ratings.

4. CAN OBJECTIVE FEATURES PREDICT SUBJECTIVE QUALITIES?

While prior work has demonstrated the capability of predicting perceptual qualities directly from waveform and mel-spectrograms [2, 12], in this section, we explore the universality of perceptual qualities across all various representations of speech: acoustic, physical, and self-supervised.

4.1. 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 finetuned on the 20% validation set.

4.2. Feature Sets

Three feature sets are used in the random forest regression models. For the acoustic features, we use ComParE 2016 feature set, which consists of a combination of functionals computed over prosodic, spectral, and sound quality-based features [22]. For the physical features, we use the Electromagnetic articulography (EMA) representation of speech, which tracks movement of articulators with midsagittal x, y coordinates of jaw, lips, and tongue positions [23]. For the self-supervised features, we use the 7th layer of a pre-trained

HuBERT model, which distills speech into a 1024-dimension learned representation via training on masked audio samples [24].

4.3. Results

The test RMSE for the above representations are reported in Figure 2. Due to the lack of data, RMSE for resonance and weight is reported against only one voice teacher’s ratings.

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 (Section 3.2), 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.

5. DISCUSSION AND FUTURE WORK

Perceptual qualities are a perceivable and predictable representation of speaker identity. With minimal examples, non-experts can label perceptual qualities with remarkable correlation to expert labels. Across multiple speech representations, perceptual qualities are predictable, and, for CAPE-V PQs, at an error that is lower than inter-expert variation.

While the current work points at the ability of perceptual qualities to capture information about speaker identity, future work is needed to explore the limits of perceptual qualities. To what extent can PQ-based representations uniquely identify a voice? Can perceptual qualities be used to guide speech synthesis or voice conversion? Many questions on perceptual quality’s applications remain, but the perceivability and ubiquity of perceptual qualities promises to bring unseen interpretability and flexibility to speech processing systems.

6. REFERENCES

- [1] Stefan R. Schweinberger, Hideki Kawahara, Adrian P. Simpson, Verena G. Skuk, and Romi Zäske, “Speaker perception,” *WIREs Cognitive Science*, vol. 5, no. 1, pp. 15–25, 2014.
- [2] Shunsuke Hidaka, Yogaku Lee, Moe Nakanishi, Kohei Wakamiya, Takashi Nakagawa, and Tokihiko Kaburagi, “Automatic grbas scoring of pathological voices using deep learning and a small set of labeled voice data,” *Journal of Voice*, 2022.
- [3] Tara McAllister, Christopher Nightingale, Gemma Moya-Gale, Ava Kawamura, and Lorraine 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.
- [4] Gail B. Kempster, Bruce R. Gerratt, Katherine Verdolini Abbott, Julie Barkmeier-Kraemer, and Robert 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.
- [5] Jody Kreiman, Bruce R Gerratt, Gail B Kempster, Andrew Erman, and Gerald 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.
- [6] Eric Keller, *The Analysis of Voice Quality in Speech Processing*, p. 54–73, Springer-Verlag, Berlin, Heidelberg, 2005.
- [7] Patrick Walden, “Perceptual voice qualities database (pvqd),” 2020.
- [8] Noa Diamant and Ofer Amir, “Examining the voice of israeli transgender women: Acoustic measures, voice femininity and voice-related quality-of-life,” *International journal of transgender health*, vol. 22, no. 3, pp. 281–293, 2021.
- [9] Lisa Carew, Georgia Dacakis, and Jennifer 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.
- [10] Mario Alejandro García and Ana Lorena Rosset, “Deep neural network for automatic assessment of dysphonia,” 2022.
- [11] Kun Zhou, Berrak Sisman, Rajib Rana, Björn W Schuller, and Haizhou Li, “Emotion intensity and its control for emotional voice conversion,” *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 31–48, 2022.
- [12] Shintaro Fujimura, Tsuyoshi Kojima, Yusuke Okanou, Kazuhiko Shoji, Masato Inoue, Koichi Omori, and Ryusuke Hori, “Classification of voice disorders using a one-dimensional convolutional neural network,” *Journal of Voice*, vol. 36, no. 1, pp. 15–20, 2022.
- [13] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
- [14] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev 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.
- [15] Jiachen Lian, Chunlei Zhang, and Dong Yu, “Robust disentangled variational speech representation learning for zero-shot voice conversion,” 2022.
- [16] Zhongxin Bai and Xiao-Lei Zhang, “Speaker recognition based on deep learning: An overview,” 2021.
- [17] Diana Markova, Louis Richer, Melissa Pangelinan, Deborah H Schwartz, Gabriel Leonard, Michel Perron, G Bruce Pike, Suzanne Veillette, M Mallar Chakravarty, Zdenka 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.
- [18] Raymond D Kent, “Vocal tract acoustics,” *Journal of Voice*, vol. 7, no. 2, pp. 97–117, 1993.
- [19] Zhaoyan 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.
- [20] Kathleen F. Nagle, “Clinical use of the cape-v scales: Agreement, reliability and notes on voice quality,” *Journal of Voice*, 2022.
- [21] Amelia Zheanna 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.

- [22] Felix Weninger, Florian Eyben, Björn W Schuller, Marcello Mortillaro, and Klaus R Scherer, “On the acoustics of emotion in audio: what speech, music, and sound have in common,” *Frontiers in psychology*, vol. 4, pp. 292, 2013.
- [23] Peter Wu, Li-Wei Chen, Cheol Jun Cho, Shinji Watanabe, Louis Goldstein, Alan W Black, and Gopala K. Anumanchipalli, “Speaker-independent acoustic-to-articulatory speech inversion,” 2023.
- [24] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” 2021.