Principles of voice quality theory: A rtMRI study on labiodental production

Daniel Denian Lee[1,3], Scott Reid Moisik[1], Vimalan Vijayaragavan[2]

[1]Linguistics and Multilingual Studies, Nanyang Technological University, Singapore
[2]Cognitive and Neuroimaging Centre, Nanyang Technological University, Singapore
[3]Phonetics Laboratory, University of Cambridge, UK
ddl26@cam.ac.uk, scott.moisik@ntu.edu.sg, vimalan.vijay@ntu.edu.sg

ABSTRACT

This real-time magnetic resonance imaging (rtMRI) study investigates the theoretic relationship between voice quality settings and phonetic segments, through the lens of two Laverian principles of susceptibility and compatibility. The setting–segment interactions were examined using whole-vocal-tract data of native Singapore English speakers. Qualitative analyses in this paper aim to offer instructive insights into the underlying nature of voice quality, and how its influence on segmental articulation is hierarchically and asymmetrically organised. Specifically, the interaction between voice quality and labiodental articulation will be analysed using the Laryngeal Articulator Model (LAM), which builds upon Laver's seminal descriptions of voice quality as a phenomenon that merits robust phonetic treatment.

Taking stock of the realities of commingling speech patterns amidst 'superdiverse' migratory complexes, voice quality theory will be elucidated through real-world examples from natural speech. Conceptual, terminological, and metatheoretical issues will be discussed, concluding with suggestions for future voice quality research.

Keywords: *voice quality theory*, *supralaryngeal settings*, *segments*, *susceptibility*, *compatibility*

1. INTRODUCTION

VOICE QUALITY is a fundamental property of an individual's speech [1], thus qualifying it as a crucial area of study in phonetic sciences research, and linguistic scholarship more broadly. Voice quality is the phenomenon—and in this paper also a theoretical framework—that consists of visuo-auditory signals present in any given person's speech production. A key concept for elucidating the notion of voice quality is that of a SETTING, which is a shorthand for 'voice quality setting', referring to a specific articulator-wise component of the overall voice quality of a given speaker, language or group. For instance, in examining the voice quality of the 45th President of the United States of America Donald J. Trump, his speech may be described as exhibiting a PROTRUDED LIPS and OPEN JAW setting, and Trump's lips and mandible are the articulatory referents when discussing PROTRUDED LIPS and OPEN JAW setting respectively. These 'sub-phenomenal' strands of Trump's voice quality are manifested, and can be observed even by phonetically untrained laypersons, in the visual (articulation) and aural (audition and acoustics) domain. Therefore, to provide an efficacious examination of voice quality, it is important to investigate both its articulatory and acoustic correlates.

2. NOMENCLATURAL CONCERNS

2.1. Defining voice quality

A well-defined technical definition of voice quality can be traced back to [1]: 'A quasi-permanent quality running through all the sound that issues from his mouth.' This definition was invoked in [2] classic descriptions of voice quality, and later reinforced in the LAM. [1] was astute in in disambiguating between the well-established phonetic sense of 'voice', which refers to the phonatory vibration of the vocal folds, and the 'voice' in 'voice quality'. The latter carries a more generalist sense of 'voice', pointing to the overall notion of a person's speech, thereby including articulations beyond the true vocal folds and other structures in the lower vocal tract. In other words, 'voice quality' should not be treated as an interchangeable synonym for 'voicing' (as in vocal-fold oscillation) or phonatory quality.

For terminological clarity, it is useful to conceive of a 'broad' and 'narrow' sense of voice quality. The narrow sense is associated with phonatory quality (e.g. LOWERED LARYNX VOICE) as expounded in the paragraph above, whereas the broad, Laverian sense of voice quality encompasses articulatory activity across the whole vocal tract, beginning inferiorly from the laryngeal articulator and ending antero-superiorly at the lips.

2.2. Voice quality typology

Voice quality manifests at different levels in naturally occurring speech contexts. At the highest level is the level of the language. That is, each language possesses its own unique aggregate of semipermanent gross vocal tract postures, and unsurprisingly languages belonging to the same genealogical family may appear auditorily similar to the naïve listener, such as Indonesian Malay and Tagalog (Austronesian) and Thai and Lao (Austroasiatic). French is a language that is known to feature a preponderance of nasality, which activates the velopharyngeal port (i.e. frequently lowered velic posture), and lip rounding. In LAM terms, therefore, the 'French voice quality' enacts the velopharyngeal and labial settings to give rise to an overall impression of its visuo-auditory characteristics. Another prototypical language that involves salient setting activation is Russian, known for its distinctive LOWER LARYNX configuration.

Voice quality is also analysable at the level of the individual. Speaker-specific voice qualities may be auditorily distinct to hearers owing to pathological bases (e.g. cleft lips, a short lingual frenulum) or simply stylistic factors. For instance, the American former professional boxer Mike Tyson is noted for his signature lisp, likely contributed by the gap in his central incisor space which impedes the canonical production of fricatives. For a non-pathological example of a speaker-specific voice quality, the English historian and television presenter Lucy Worsley features a distinctive CLOSE JAW VOICE that does not appear necessarily linked to pathological factors. Existing intermediately between the levels of the individual and language, is the group (or community). The group may be classified by age, sex, socioeconomic status, and even sexual orientation (e.g. see work done on the 'gay voice/persona').[1]

2.3. Articulatory setting

It is exceedingly crucial to avoid confusing 'voice quality' with another existent term in the literature: ARTICULATORY SETTING (AS). Coined by [3], AS carries a completely disparate meaning, to the extent that it may well qualify as a misnomer based on its prevalent but potentially facile usage among researchers, particularly non-linguists. Put briefly, AS is a term that solely refers to the physiological activity of speech, precluding the acoustic and perceptual correlates of voice quality. Compared against Laver's foundational framework, AS appears to be more like a synonym for the articulatory strand of voice quality (which is a tripartite phenomenon comprising articulation, acoustics and perception) than an analytic paradigm *stricto sensu*. Moreover, the AS nomenclature was founded on the basis of pronunciation pedagogy rather than analytic depth and explanatory power. Thus, the LAM is demonstrably more robust as a framework for describing the mechanisms of voice quality. Additionally, there seems to be a usual pairing of the AS nomenclature with the inter-speech posture (ISP) methodology featured in some works. The ISP technique faces a handful of theoretical and methodological concerns, with the most critical issue being that it discards speech data itself in favour of 'pause states', since only the between-utterance, pre-utterance and absolute rest postures are examined. Finally, the lexical-semantic transparency of 'voice quality' over 'articulatory setting', especially for laypersons (non-linguists), makes it difficult to recommend the continued usage of the AS nomenclature, as it may well propagate the muddying of terminological waters in the literature. With the aforementioned factors taken into consideration, there is a real impetus to promote scholarly concord in the field through the maintenance of the decidedly more appropriate term 'voice quality' in voice quality scholarship.

3. METHODOLOGY

3.1. Subjects

Subjects were recruited through word of mouth and publicity posters placed across several locations at Nanyang Technological University. A total of 53 subjects were recruited for this project, aged between 21 to 41 (inclusive) with 30 males and 23 females. Out of the total of 53, two (one male, one female) possessed atypical mandibular morphology (Class III malocclusion or negative overjet), and one (male) presented with an array of vocal tract abnormalities, including an elongated soft palate that ostensibly impeded his respiration and speech task performance. Hence, the remaining subjects (n = 50) form the usable sample for data analysis. All subjects are native or near-native Singapore English speakers. For the purpose of qualitative analysis in this paper, rtMRI data from five subjects will be presented for visual examination.

The five subjects are male, native speakers of Singapore English, ethnically Chinese (Singapore Mandarin as L2), under the age of 30, and either undergraduates or postgraduate professionals at the time of their participation. Controlling for these

[1] Refer to the author's thesis for a cited sources.

ethnolinguistic factors were aimed at mitigating sociolinguistically-grounded phonetic variation.

3.2. Data acquisition and rtMRI protocol

A 3.0 Tesla Siemens MAGNETOM Prisma scanner at the Cognitive Neuroimaging Centre (CoNiC) of the Lee Kong Chian School of Medicine was utilised for this study's experimental tasks. A real-time 2D and 3D GRE radial sequence (also known as radial FLASH)[2] was used for MR data acquisition, adapted from [4]. The steady-state free precession (SSFP) was attempted during pilot sessions, but a specific 'reconstruction' algorithm was necessary but absent in the system (V. Vimalan, personal communication, October 26, 2021) for the sequence to run successfully. Hence, [5]'s recommendation to utilise the SSFP was unfortunately not applicable for this study, despite the promise of yielding favourable rtMRI data from the 3.0 Tesla system.

Along with the MR data obtained, noise-suppressed audio was concurrently retrieved using an MR-compatible optical microphone (Dual Channel-FOMRI, Optoacoustics, Or Yehuda, Israel). The noise cancellation software of the optical microphone suppressed the operating noise from the scanner, allowing the subjects' speech audio to come through more audibly for auditory and acoustic analyses. A 64-channel head-neck radiofrequency coil was utilised to amplify the MR signal in the vocal tract region. The following acquisition parameters were applied: FOV = 198 × 198 mm; ST = 7.0 mm, with 50 slices/slab; TR/TE = 476.89/2.25 ms; flip angle = 12 °; resolution = 1.5 × 1.5 × 7.0 mm^3 ; TA = 0.43 s. The region of interest (ROI) is approximately bounded infero-superiorly by the base of the 6th cervical vertebra (C6) and nasopharynx roof, and antero-posteriorly by the nose tip and spinal cord. Occasionally, the subcutaneous fat (which displays as bright white in T1 and T2-weighted imaging) posterior to the cervical spinal column can show unwanted visual artefacts in the form of bright streaks, as was the case in pilot sessions conducted for this study. Therefore, it may be recommended to exclude the rear neck subcutaneous fat from the ROI.

3.3. Research questions

To understand the two Laverian principles of SUSCEPTIBILITY and COMPATIBILITY, the following Research Questions (RQ1 and RQ2) are posed: (RQ1) How do voice quality settings impact labiodental articulation?; (RQ2) Are voice quality settings hierarchically ordered, and how do they interact with each other?

To briefly recapitulate Laver's descriptions, susceptibility refers to the openness of segments or settings to external influence, typically a voice quality setting. Compatibility is associated with the articulatory resources necessary and available for the production of speech sounds. It is anticipated that answering RQ1 and RQ2 will illuminate the hitherto less-understood aspects of voice quality.

4. RESULTS

The midsagittal MR images of five select subjects are presented given in the Appendix, with the montage depicting the articulation of the consonantal segment [v] using three voice quality settings (i.e. NEUTRAL, PROTRUDED LIPS, and PROTRUDED JAW). The rtMRI data of five subjects are used for these montages, and to protect their anonymity they are referred to as Sub-A (Subject A, designated with subject ID #20 in the DICOM database), Sub-B (#26), Sub-C (#32), Sub-D (#43), and Sub-E (#51) in this paper. Qualitative analyses are provided based on visual examination of the MRI montage, and categorised by articulatory markers of analysis (AMAs)[3] per subsection. Comparing the inline MR images vertically reveals the cross-setting differences in how voice quality influences segmental articulation, allowing RQ1 and RQ2 to be addressed from an articulatory perspective.

4.1. Labial

The articulation of the voiced labiodental fricative [v] appears to be canonical across all five subjects in the NEUTRAL row of the MRI montage; the upper incisors come into contact with the lower lip. When considering the PROTRUDED LIPS voice quality and its effect on [v] articulation, an interesting pattern emerges unanimously: All subjects realised the labiodental using the inner surface of where the mentalis muscle is located in place of the lower lip. Both upper and lower lips are advanced anteriorly, which prohibits canonical productions of [v] because

[2] Fast low angle shot (FLASH) magnetic resonance imaging.
[3] The four AMAs include LABIAL, LINGUAL, VELOPHARYNGEAL, and LARYNX HEIGHT parameters. AMAs are distinct from settings to distinguish between independent (settings) and dependent (AMAs) variables.

the lower lip is now displaced and unable to meet the upper incisors. Hence, as a natural compensatory action, all subjects made contact between the upper incisors and the inner mentalis muscle wall to achieve an auditorily similar [v] realisation.

Continuing with the analysis of [v] realisation, in the bottom row of the montage, it appears that Sub-A and Sub-E (first and last subjects) made the least use of the dental articulators under the PROTRUDED JAW condition. Corroborating this observation with their speech audio, the non-naturalness of their productions can be detected auditorily. With the physiologic constraint imposed by the PROTRUDED JAW configuration, it appears as though both subjects utilised their upper and lower lips to create a [v]-like sound, in place of the canonical labiodental configuration. Interestingly, the audio data for Sub-A reveal that he managed to achieve a continuous fricative-like sound, but Sub-E clearly struggled to sustain the sound with his lips and only made instantaneous realisations of [v]-like sounds in his utterance. In contrast, the intermediate Subjects B, C, and D appear to adopt a dentolabial strategy to achieve a fricative sound auditorily similar to [v]. This articulation is in theory the more natural outcome as a result of the lower muscular effort required to produce a dentolabial with a negative overjet configuration [6]. This difference in strategy between Subjects A and E on the one hand, and B, C and D on the other is a worthwhile area of further inquiry, using quantitative methods.

4.2. Lingual

All subjects show radically different overall tongue shapes in their production of [v] (top row, NEUTRAL setting), but there is a slight concavity to all subjects' tongues (albeit to varying degrees). This lingual concavity is interesting, because the lingual profile should not act as a significant contributor to shaping the auditory and acoustic properties of the fricative [v]; the most salient articulatory manipulation is to bring together the upper central incisors and the lower lip to create turbulent egressive airflow. It should be reiterated that in the experiment, subjects were tasked to articulate the nonce word <AVA>, and the inter-vocalic [a] context may have a role in biasing this tongue shape. It is interesting to note that Sub-C and Sub-D's NEUTRAL production of [v] appear to involve relatively less tongue retraction (ascertained by the larger linguo-PPW[4] distance) compared to the other three subjects.

4.3. Velopharyngeal

In the top row of the montage, the velopharyngeal port is closed in NEUTRAL productions of [v]. Since the outermost part of the vocal tract (lips, incisors) are the most salient structures for canonical [v] production, it might be natural to conclude that the velum has no relation to the production of labiodentals/dentolabials. This seems to be the case for the PROTRUDED LIPS series data, where the velum remains raised for all subjects (Sub-B may in fact even be showing a tighter velo-PPW constriction, as the appears to be less of a gap compared to its NEUTRAL counterpart). However, in the bottom row of the montage, the PROTRUDED JAW series clearly shows obvious lowering of the velum in Sub-B and Sub-C, compared to the sealed velopharyngeal port seen in Sub-B and Sub-C's NEUTRAL correlates for [v] production.

4.4. Larynx height

A horizontal white line is overlaid across each setting row in the montage, positioned at the base of the 6th cervical spine (C6) belonging to Sub-A, matched against the cricoid lamina. On the whole, it does not seem that the settings analysed in this study have an apparent effect on larynx height in [v] production. There is a possible exception in the case of Sub-A, where his [v] articulation appears to utilise a slight RAISED LARYNX posture in the non-neutral settings. The connection between larynx height and labial/mandibular settings is not explicit, and the current results do not provide enough information to reveal any interesting articulatory patterning in terms of voice quality.

5. CONCLUSION

In answering the research questions: (RQ1) Voice quality settings impact labiodental in asymmetrical and non-linear ways not easily categorisable through preliminary analysis; (RQ2) voice quality settings are hierarchically ordered and take precedence over one another depending on the phone in question. To improve the generalisability of the findings presented, other segmental articulations would need to be examined, and a quantitative analysis using techniques from geometric morphometrics may help yield deeper insights into the hidden structures of voice quality theory.

[4] Posterior pharyngeal wall.

6. REFERENCES

[1] D. Abercrombie, *Elements of general phonetics*. Edinburgh: Edinburgh University Press, 1967.

[2] J. Laver, *The phonetic description of voice quality*, vol. 31. Cambridge: Cambridge University Press, 1980.

[3] B. Honikman, ‘Articulatory settings’, in *In honour of Daniel Jones*, D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott, and J. L. M. Trim, Eds. London: Longman, 1964, pp. 73–84.

[4] B. P. Sutton, C. A. Conway, Y. Bae, R. Seethamraju, and D. P. Kuehn, ‘Faster dynamic imaging of speech with field inhomogeneity corrected spiral fast low angle shot (FLASH) at 3 T’, *J. Magn. Reson. Imaging*, vol. 32, no. 5, pp. 1228–1237, 2010, doi: 10.1002/jmri.22369.

[5] S. G. Lingala, B. P. Sutton, M. E. Miquel, and K. S. Nayak, ‘Recommendations for real-time speech MRI’, *J. Magn. Reson. Imaging*, vol. 43, no. 1, pp. 28–44, 2016.

[6] D. E. Blasi, S. Moran, S. R. Moisik, P. Widmer, D. Dediu, and B. Bickel, ‘Human sound systems are shaped by post-Neolithic changes in bite configuration’, *Science*, vol. 363, no. 6432, p. eaav3218, 2019.

APPENDIX

MRI montage