COPiL CAMBRIDGE OCCASIONAL PAPERS *in* LINGUISTICS
Volume **17**, Article 2: pp. 22–36, 2025 | ISSN: 2050-5949

A Critical Review of Voice-Space Models*

Tianze Xu
University of Cambridge

1 Introduction

When humans listen to the voices of different speakers, some are perceived as sounding more similar than others. The phonetic and acoustic mechanisms underlying this perceived voice similarity remain an active area of research (Gerlach, McDougall, Kelly, Alexander & Nolan 2020). One prominent approach to understanding these mechanisms involves the development of voice-space models — representational spaces where perceptual voice identities are quantified using a few key acoustic dimensions (Baumann & Belin 2010, Latinus, McAleer, E.G., Bestelmeyer & Belin 2013). While these models have demonstrated great potential for applications, such as in developing voice perception assessments (Mühl, Sheil, Jarutytė & Bestelmeyer 2018, Humble, Schweinberger, Mayer, Jesgarzewsky, Dobel & Zäske 2023) and forensic analysis (Lee, Keating & Kreiman 2019), they also face methodological and conceptual limitations. This paper critically examines established voice-space models in the literature by first reviewing their development and underlying principles and then analysing the challenges posed by the variability observed both between and within speakers. It ultimately argues that the notion of *voice space* should be reconsidered in favour of *speaker space* (Nolan 1991, 1997) — an approach that more systematically incorporates the complexities of voice variability.

2 Voice-Space Models: Principles and Development

Human voices exhibit variability across a wide array of acoustic features, including pitch (fundamental frequency, f_0), formant frequencies (F_1–F_5) and formant dispersion (FD), loudness (amplitude), harmonic frequencies, and more (Kreiman, Lee, Garellek, Samlan & Gerratt 2021, Kreiman 2024, Lee et al. 2019). Among these features, are there a few prominent ones that can succinctly represent the voice identities perceived by listeners? This question is central to the development of voice-space models (Baumann & Belin 2010, Latinus et al. 2013). In these models, each voice is represented as a point in a low-dimensional space, with its coordinates defined by a small number of acoustic parameters (Kreiman 2024), and the similarity between two voices is then quantified by the Euclidean distance between their respective points. To achieve dimensionality reduction — from the seemingly

* I would like to thank Dr. Kirsty McDougall for her guidance on this essay. Any remaining errors are entirely my own. This work was supported by the Travel and Research (Postgraduate Study Aids) Grant from Christ's College, University of Cambridge.

infinite acoustic features that vary between speakers to a few key dimensions that effectively capture perceived identities — studies have examined the correlation between acoustic and perceptual measures. Perceptual measures are categorised into two types: behavioural ratings and neural responses.

2.1 Behavioural ratings as perceptual measures

Baumann & Belin (2010) developed a two-dimensional (2D) voice-space model, with f_0 and F_1 as the dimensions for female voices, and f_0 and FD between F_4 and F_5 for male voices, by correlating acoustic measures with behavioural ratings of voice similarity. In their study, 32 speakers (16 females) produced three sustained French vowels (/a/, /i/, and /u/), with each vowel lasting approximately one second. Ten listeners rated the similarity between pairs of vowel samples on a scale from 1 to 100, representing their confidence in whether the two samples were spoken by the same or different individuals. The researchers then applied multidimensional scaling (MDS) to analyse these similarity ratings. MDS works by identifying commonalities in the acoustic features listeners use to judge voice similarity, reducing the dimensionality of the perceptual space to a small set of shared dimensions (Carroll & Chang 1970). For their rating data, Baumann & Belin found that a 2D space was the most appropriate for interpretability, uniqueness, and explained variance.

To interpret the two dimensions (which are inherently abstract due to MDS calculations), the researchers correlated them with a wide range of acoustic measures. These included f_0, F_1–F_5, overall FD (related to vocal tract size), FD between F_4 and F_5 (relatively stable across vowels), jitter (reflecting local f_0 variation), shimmer (reflecting local amplitude variation), harmonics-to-noise ratio (HNR, reflecting periodicity), and utterance duration. All measures were averaged across the three vowels for each speaker. Correlation analyses revealed that the first dimension correlated exclusively with f_0, while the second dimension differed by gender: for female voices, it correlated most strongly with F_1, and for male voices, with FD between F_4 and F_5. These findings align with the source-filter theory of speech production, wherein f_0 reflects glottal source characteristics and F_1/FD reflects vocal tract filter characteristics, emphasising the importance of f_0 (Kreiman, Gerratt, Precoda & Berke 1992) and formant-related parameters (Bachorowski & Owren 1999) in distinguishing voices.

2.2 Neural responses as perceptual measures

Latinus et al. (2013) validated a 3D voice-space model using f_0, overall FD, and HNR by correlating each voice's distance to a prototypical voice defined in this space with neuroimaging data. They used the English syllable *had* uttered by 64 speakers (32 females) and represented each voice as a point in the f_0-FD-HNR space. The Euclidean distance between each voice and its gender-specific prototypical voice was then defined as the distance-to-mean. This distance-to-mean metric was correlated with both behavioural and neural measures. Behaviourally, listeners rated voice distinctiveness, and the ratings significantly correlated with acoustic

distance-to-mean: voices farther from the prototypical voice (i.e., acoustically more dissimilar) were perceived as more distinctive. Neurally, using functional magnetic resonance imaging, the researchers observed greater activity in the temporal voice areas — brain regions selective to human voices (Belin, Zatorre, Lafaille, Ahad & Pike 2000) — for voices that were farther from the prototypical voice.

Additional analyses confirmed the robustness of these effects. The authors found similar results when they controlled for potential adaptation effects (caused by the repeated presentation of similar stimuli), used more natural speech materials (e.g., the English word *hello*), and directly manipulated the acoustic distance-to-mean (by creating artificially morphed stimuli). These findings led to the conclusion that the f_0-FD-HNR 3D space provides an adequate approximation of the space for representing voice identities.

3 Methodological Limitations: Beyond Vocalic Variability

Both Baumann & Belin's (2010) 2D voice-space model and Latinus et al.'s (2013) 3D model have proven highly informative in understanding perceived voice similarity, particularly in developing voice perception tests. These tests, which ask participants to judge the similarity between paired voice samples, rely on estimating perceived similarity from acoustic parameters, making voice-space models foundational to their development. For instance, the 2D model informed the Bangor Voice Matching Test (BVMT; Mühl et al. 2018), where a significant correlation was observed between the acoustic distance of voice sample pairs calculated in the 2D space and listeners' classification accuracy. Similarly, the f_0-FD-HNR 3D model underpinned the Jena Voice Learning and Memory Test (JVLMT; Humble et al. 2023), with acoustic distances derived from the model significantly correlating with behavioural judgments. Notably, the BVMT used syllables based on English (e.g., *had*, *ugu*), extending the 2D model's validity from vowels to syllables and from French to English. Meanwhile, the JVLMT, which employed meaningless pseudo-sentences based on English, broadened the 3D model's applicability from syllables to sentence-level materials.

However, when additional factors are introduced, the robustness of voice-space models diminishes. The Sisu Voice Matching Test (SVMT; Xu, Jiang, Zhang & Wang 2025), which relied on the f_0-FD-HNR 3D model and employed pseudo-words and pseudo-sentences constructed following Mandarin Chinese phonology, revealed a more nuanced relationship between acoustic distance and perceptual classification accuracy (Figure 1). For pseudo-word pairs (orange and blue lines), the correlation aligned well with the predictions of voice-space models: larger distances between samples (greater dissimilarity; e.g., the fifth vs. first quintiles) led listeners to classify them as different speakers, resulting in higher accuracy for different-identity pairs and lower accuracy for same-identity pairs. However, this correlation did not perfectly hold for same-identity pseudo-sentence pairs (green line; see the fifth vs. third quintiles). This raises critical questions: Are voice-space models valid across languages with different phonological systems? Can they reliably apply to longer linguistic units (e.g., sentences)? Examining these factors, among others, is essential for extending the validity of voice-space models.

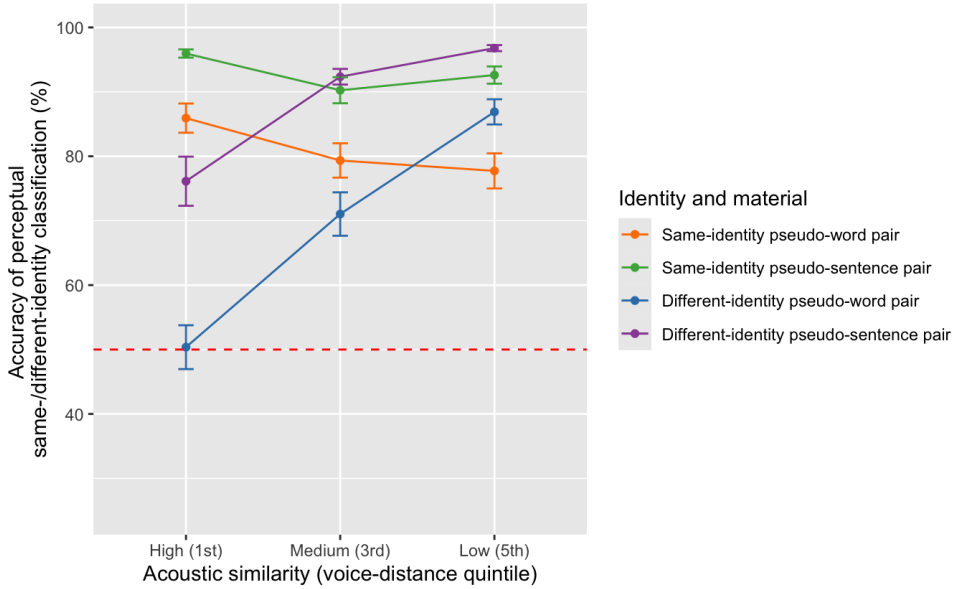

Figure 1 Accuracy of perceptual same-/different-identity classification for voice sample pairs grouped by varying levels of acoustic similarity (voice-distance quintile). The graph illustrates the relationship between acoustic similarity (high, medium, low), speaker identity (same or different), and material type (pseudo-word or pseudo-sentence) on classification accuracy. Quintiles: decreasing acoustic similarity from high similarity (1st) to low similarity (5th). Dotted line: the chance level (.5). Error bars: the standard error of the mean (SEM). Adapted from Xu et al.'s (2025) Figure 3.

3.1 Across languages, dialects, and accents

Previous research has shown that listeners generally identify voices more accurately in their native language than in unfamiliar ones — a phenomenon known as the language familiarity effect (Gerlach et al. 2020, Perrachione 2019). For example, Perrachione, Tufo & Gabrieli (2011) found that English-speaking participants with dyslexia (a disorder undermining phonological awareness) performed 40% worse than non-dyslexic participants when identifying voices in English. However, the two groups performed similarly poorly when identifying voices in Mandarin, which were unfamiliar to both groups. These findings underscore the critical role of phonological differences in shaping voice perception across languages (Kuhl 2011).

Acoustic studies across different languages have identified variations in the contribution of acoustic measures to voice variability. Lee & Kreiman (2022b) used principal component analysis to determine which acoustic measures most strongly explain voice differences in languages with varying phonation and tone contrasts, including American English, Seoul Korean, and White Hmong. Their findings revealed that the principal components differed based on the phonological structure of each language. For instance, variability in f_0 accounted for significant voice variability in Korean and Hmong, which use tonal contrasts, but not in non-tonal

English. Similarly, the amplitudes of lower harmonics, related to phonation contrasts, explained significant variance in Hmong voices but not in English or Korean. This reflects Hmong's use of phonation contrasts (e.g., breathy or creaky vowels vs. modal vowels), a feature absent in English and Korean. These findings further highlight the role of cross-linguistic phonological differences in shaping voice variability.

Therefore, the differences observed in the fit of voice-space models in different voice tests may be attributed to the language of the materials. Both the JVLMT and the SVMT were based on the f_0-FD-HNR 3D model and used pseudo-sentences as stimuli. While the JVLMT followed English phonology, consistent with the English syllable/word materials used when developing the f_0-FD-HNR model, the SVMT was based on Mandarin phonology. Given the significant phonological differences between the two languages (e.g., Mandarin's tonal contrasts vs. English's lack thereof), the voice-space model might encounter limitations when applied to sentence-length materials in a different phonological system.

Moreover, differences in dialects and accents may also introduce variability. Although Baumann & Belin (2010) and Latinus et al. (2013) developed their models with speakers of the same language, these studies showed limited control over speakers' demographic factors, such as age, place of birth/residence, dialect, and other personal characteristics. Previous research has shown that dialect and accent differences can significantly add to between-speaker acoustic variability (Jacewicz, Fox & Wei 2010, McDougall 2011, McDougall, Duckworth & Hudson 2015). This lack of control in design and analysis may limit the generalisability of voice-space models in contexts involving diverse dialects or accents.

Future research applying these models should consider the differences across languages, dialects, and accents. To enhance the robustness of the voice-space models across diverse languages and contexts, studies could investigate effective methods for integrating linguistic, dialectal/accentual, cultural, and other variations.

3.2 Higher levels of linguistic units

A discrepancy exists between the levels of linguistic units used in developing voice-space models and those employed in practical applications. Baumann & Belin's (2010) 2D model was developed using French sustained vowels (/a/, /i/, and /u/), while Latinus et al.'s (2013) 3D model used English syllables and words (*had* and *hello*). In contrast, voice-space models are often applied to materials at the word (e.g., the BVMT and the SVMT) or sentence level (e.g., the JVLMT and the SVMT) for establishing voice tests. This mismatch in linguistic materials between development and application may undermine the validity of these models.

Indeed, variability in voice source characteristics is evident between sustained vowels and continuous speech (Gerratt, Kreiman & Garellek 2016). For instance, Moon, Chung, Park & Kim (2012) recorded 202 Korean speakers producing a sustained vowel /a/ for three seconds and reading sentences from a text lasting approximately 50 seconds and analysed acoustic differences between these two speech types. The results showed that for male speakers, both f_0 and contact quotient (the

ratio of vocal fold contact duration to one vocal fold period) were higher in sentences than in sustained vowels. Similarly, differences in f_0 between continuous speech and isolated sustained vowels have been observed in Danish speakers (Iwarsson, Nielsen & Næs 2020). Importantly, Iwarsson et al. (2020) also identified differences in between-speaker contrasts across speaking conditions (Figure 2). For example, while Speakers S8 and S12 both exhibited higher f_0 in the isolated vowel than in continuous reading, S8 had a higher f_0 than S12 in continuous reading but a lower f_0 in the isolated vowel. Such differences could alter how listeners perceive these two speakers. Additionally, Lederle, Barkmeier-Kraemer & Finnegan (2012) reported differences in vocal tremor (relatively periodic modulations in f_0 and amplitude) between vowels and sentences. These findings highlight the acoustic variations of voices across levels of linguistic units.

Continuous speech, such as sentences, is considered more ecologically valid because it better reflects the dynamic attributes of voice in natural communication (Maryn, Corthals, Cauwenberge, Roy & Bodt 2010). For example, vocal fluctuations during voicing onset and termination, as well as variations in amplitude and f_0, are more evident in continuous speech (Awan, Roya, Jetté, Meltzner & Hillman 2010). These phonetic cues may aid in voice differentiation. Given the current focus of voice-space models on vowels or syllables, further research is needed to expand these models to accommodate longer linguistic units, thereby enhancing their practical applications.

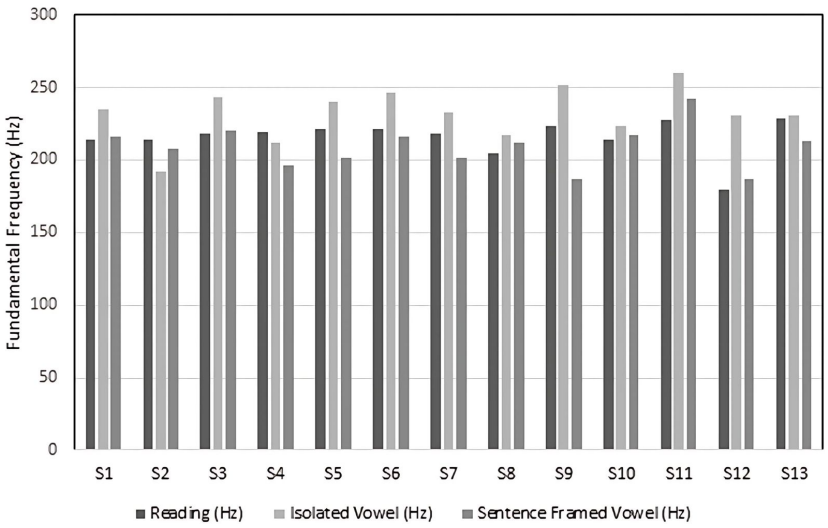

Figure 2 Mean fundamental frequency for Danish speakers under three speaking conditions: continuous speech, isolated vowel, and sentence-framed vowel [1]. Adapted from Iwarsson et al.'s (2020) Figure 3.

[1] Continuous speech: reading the standard text *Vinden og solen* ('The Wind and the Sun'; 108 words / 11 sentences). Isolated vowel: sustained /a/ (3 seconds) at a self-chosen pitch. Sentence-framed vowel: prolonged /a/ (3 seconds) in the word *gartneren* within the sentence *Var gartneren syg?* ('Was the gardener sick?').

3.3 Consonantal, suprasegmental, and other features

The focus on vowels during the development of voice-space models introduces another issue: insufficient consideration of consonantal, suprasegmental, and other features. For instance, the neglect of consonants has led to two problematic outcomes in applying the models: either including consonants with unsuitable acoustic parameters (as in the JVLMT) or excluding them entirely from acoustic analyses (as in the BVMT and the SVMT). Both approaches present significant limitations.

Including consonants for analysing features primarily suited for vowels can undermine acoustic analyses and affect downstream tasks (Gerratt et al. 2016). For example, in extracting the three acoustic parameters — f_0, FD, and HNR — the JVLMT analysed entire pseudo-sentences (e.g., *ble sulpty debepts thek henbly stopapt*; Humble et al. 2023). However, some consonants, such as plosives, fricatives, and affricates, typically lack formant features. This makes automatic parameter extraction prone to inaccuracies. Additionally, pauses and silences within a sentence could further distort the parameter estimates.

Meanwhile, excluding consonants from acoustic analysis is also non-ideal, since it overlooks important sources of between-speaker variability. Studies have shown that consonantal features also contribute to between-speaker acoustic variability (see also Xu et al. 2025). For example, voice onset time (VOT), which reflects timing in speech production, exhibits individual differences for plosives and affricates even when the speaking rate is controlled (Alle, Miller & DeSteno 2003, Chodroff & Wilson 2017). Importantly, Chodroff & Wilson further observed strong, systematic linear relationships among VOTs for plosives (e.g., /p^{h}/ and /k^{h}/) across speakers in American English, indicating consistent speaker-specific variation patterns in this feature. Similarly, the centre of gravity (COG), a measure of spectral energy distribution that is influenced by turbulence frequency in fricatives and affricates, varies between speakers (Kavanagh 2012, Smorenburg & Heeren 2020). For instance, Smorenburg & Heeren identified through multinomial logistic regression that COG was the most critical acoustic feature for accurately predicting speaker identities in Dutch fricatives /s/ and /x/. Furthermore, individual variability in f_0 and FD has been observed for nasals and lateral approximants due to their shared acoustic properties with vowels (Kavanagh 2012, Smorenburg & Heeren 2021).

Beyond consonants, suprasegmental and other features such as intonation (Loakes, Fraser & McDougall 2023, Moon et al. 2012, Nolan 2002), speaking rate (Jacewicz et al. 2010, Morrill, Baese-Berk & Bradlow 2016), voice quality (Jessen 1997, Kreiman et al. 2021, Vaňková & Skarnitzl 2014), and disfluencies (McDougall & Duckworth 2017, 2018, McDougall et al. 2015, McDougall, Rhodes, Duckworth, French & Kirchhübel 2019) also contribute to between-speaker variability. For example, in terms of speaking rate, Wisconsin speakers (northern American English) speak significantly faster than North Carolina speakers (southern American English). Regarding disfluencies, based on British English speech data, McDougall & Duckworth (2017: 18) proposed a 'taxonomy of fluency features' that revealed notable differences between speakers, including unfilled pauses (e.g., *I came … and he left*), filled pauses (e.g., *he was, er, running*), repetitions (e.g., *but it- it did*), prolongations (> 200 ms), and interruptions

(e.g., *I prob- I don't know*). Differences between speakers in these disfluency features were evident in both their types and their frequencies, with each speaker's disfluency profile showing consistency across telephone and interview styles (McDougall & Duckworth 2018).

Despite their importance, voice-space models have yet to quantify consonantal, suprasegmental, and other features (e.g., disfluencies). Further research is essential to investigate their contributions to voice variability and how they can be integrated into the models (as some features may not apply uniformly across different utterances).

4 Conceptual Limitations: Speaker Space versus Voice Space

Thus far, the focus has been on between-speaker variability. However, there is not only between-speaker variability but also significant within-speaker variability in voices. For instance, an individual's voice may change depending on specific words or phrases, speaking styles, physical conditions, and other factors. Despite this, within-speaker variability is not accounted for in existing voice-space models. In Latinus et al.'s (2013) 3D voice-space model, only one syllable (i.e., *had* or *hello*) was analysed at a time. Similarly, in Baumann & Belin's (2010) 2D model, the only within-speaker differences were those among the three vowels. However, even these differences were excluded during averaging, with each speaker represented as a single point in the space. While averaging over the same vowels for all speakers may arguably approximate within-speaker variations, it fails to explicitly encode these differences. By representing each speaker as a single point, voice-space models overlook the reality that a speaker can have various voices depending on numerous factors. This omission represents the core conceptual limitation of voice-space models.

4.1 Within-speaker variability

Perceptual and acoustic studies have revealed significant within-speaker variations. For instance, Loakes et al. (2023) demonstrated that in a six-speaker conversation with greater within- than between-speaker variability in f_0 and intonation, even trained phoneticians struggled to attribute utterances to the correct speaker. This underscores the influence of within-speaker acoustic variability on voice identity perception.

Lee and colleagues (Lee et al. 2019, Lee & Kreiman 2022a,b, 2023) applied principal component analysis to acoustic measures (e.g., f_0, F_1–F_4, FD, harmonic spectral shapes, and noise metrics) across languages including American English, Seoul Korean, White Hmong, and Thai. They identified a hierarchy of factors driving voice variability: biological, linguistic, and individual. Universally, the first principal component reflected variations in the balance of harmonic and inharmonic energy, associated with voice quality perception (e.g., strained to breathy) and arousal signalling across species (Anikin 2020, Kreiman 2024, Lee & Kreiman 2022b). FD (reflecting physiological vocal size) also consistently emerged early (Fitch 1997,

Lee & Kreiman 2022a). Together, these universal biological factors accounted for significant variability across languages. Language-specific factors followed. For example, variability in f_0 significantly explained voice differences in tonal languages (e.g., Korean and Hmong) but not in non-tonal ones (see also Section 3.1).

Despite this, shared acoustic structures explained only about half of the variability in group data, with the remainder stemming from within-speaker variability. This variability may arise from factors such as speaking style, expressiveness, and specific words or phrases (Lavan, Burston & Garrido 2019a, Lavan, Burston, Ladwa, Merriman, Knight & McGettigan 2019b, Lee et al. 2019). These details are essential for telling voices together (i.e., correctly determining that different samples come from the same person) despite misleading differences. For example, a speaker's voice may vary in f_0 between reading a given sentence and speaking spontaneously (Lee & Kreiman 2022a), resulting in at least two distinct voices belonging to the same underlying speaker. Combined with universal and language-specific between-speaker factors, they are also important for telling voices apart, where subtle acoustic distinctions in the acoustic space must be assessed. However, the current voice-space models, which represent each speaker as a single point, cannot accommodate this complexity.

4.2 Speaker space as an alternative framework

Given the conceptual limitations of voice space, the notion of speaker space (Hudson, McDougall & Hughes 2021, Nolan 1991, 1997) should be used to account for within-speaker variability in the acoustic-perceptual space. As illustrated in Figure 3, voice space (left panel) represents each speaker as a single point, calculated based on either one utterance (Latinus et al. 2013) or an average across multiple utterances (Baumann & Belin 2010). In contrast, speaker space (right panel) allows each speaker to be represented by a 'cloud' of points (Hudson et al. 2021: 639), reflecting the diversity of their voices across different utterances, speaking styles, emotions, or other factors. The speaker-space framework thus offers a more comprehensive conceptualisation of acoustic space, with the goal of approximating the perceptual space as accurately as possible.

In real-life scenarios, comparisons often involve two voices containing different linguistic content, whether from the same speaker or from different speakers. For example, in voice discrimination/recognition tests (Humble et al. 2023, Mühl et al. 2018, Xu et al. 2025), voice samples in each pair are intentionally selected to differ in words or sentences. This design both simulates real-life communication and controls for task difficulty. In this context, the perceived similarity between two speakers reflects the similarity between the voices as they utter different materials from their voice repositories. Additionally, studies have suggested that humans may form a norm or prototypical voice as a baseline during voice perception (Fontaine, Love & Latinus 2017, Lavan, Knight & McGettigan 2019c). This norm or prototype is based on exposure to a variety of voices from multiple speakers, rather than a single processed point representing each speaker. Therefore, in both respects, speaker space offers a more realistic representation of voice perception than voice space.

A practical challenge with implementing speaker space lies in calculating the acoustic similarity between speakers. In a voice-space model, the acoustic (dis)-similarity between two speakers is calculated as the Euclidean distance between their representative points, with each speaker being associated with only one point. This acoustic similarity is then assumed to approximate the perceptual similarity between speakers. However, in a speaker-space model, where each speaker is represented by multiple voice points, simply calculating Euclidean distance to represent overall similarity between speakers is not feasible. A potential solution could be mathematically comparing the trajectory between pairs of voice points within a speaker to the corresponding trajectory in another speaker. For instance, in Figure 4, each type of line connects the same pair of voices (when uttering the same pair of words) across speakers (e.g., the dash-dotted lines link the voices uttering *dasou* and *gehao* for each speaker). These trajectories form vectors with both magnitude and direction, and all trajectories for a speaker can collectively function as a set of vectors (i.e., a matrix). Mathematical comparisons of these sets could then be used to quantify acoustic similarity between speakers.

The validity of this representation should be investigated empirically. Possible applications include the development of future voice tests and forensic phonetic tasks where disguised speech is often involved (Hudson et al. 2021, Lee et al. 2019).

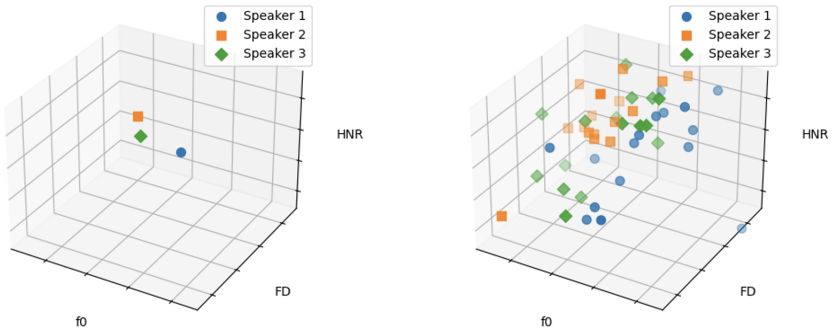

Figure 3 Illustration of voice space (left) versus speaker space (right) using the same three dimensions: f_0, FD, and HNR. Data adapted from Xu et al. (2025).

5 Conclusion

Representing perceived voice identities acoustically faces two major challenges: the vast number of acoustic dimensions that can differentiate speakers and the dynamic, ever-changing nature of voices (Latinus et al. 2013). Voice-space models, such as those proposed by Baumann & Belin (2010) and Latinus et al. (2013), have made significant strides by identifying two or three key acoustic parameters that adequately differentiate speakers in many cases. These models have found impactful applications in fields such as the development of voice perception tests. However, they still face notable methodological and conceptual limitations. Methodologically, the models fail to account for between-speaker variability arising from factors beyond the defined dimensions, such as language, dialect, and accent differences, higher

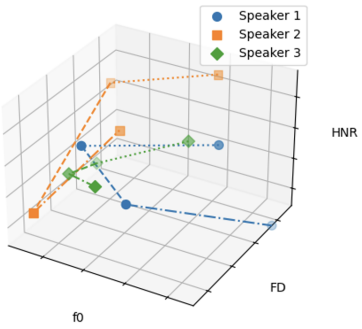

Figure 4 Illustration of between-voice variations for individual speakers in speaker space defined by three dimensions: f_0, FD, and HNR. Each line type (dotted, dashed, dash-dotted) connects the same pair of voices (e.g., uttering the same pair of words) across speakers. Data adapted from Xu et al. (2025).

levels of linguistic units (e.g., sentences), and consonantal, suprasegmental, and other non-vocalic features. Conceptually, they overlook within-speaker variability, which cannot be accommodated within their current design.

Accordingly, future studies could focus on two main directions. First, they could aim to capture more sources of between-speaker variability — not necessarily by increasing dimensionality, but by integrating new dimensions into existing frameworks in a succinct and efficient manner (Gerratt et al. 2016). Second, research could reconsider the conceptual foundation of voice-space models to systematically account for within-speaker acoustic and phonetic dynamics. A shift towards a speaker-space framework is needed to acknowledge that a single speaker can exhibit multiple voices (Hudson et al. 2021), enabling the comparison of between-voice variations across speakers and providing deeper insights. What truly matters is not the term *speaker space* but rather the capacity to accommodate the dynamic variability inherent within speakers.

References

Alle, J. S., J. L. Miller & D. DeSteno. 2003. Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America* 113(1). 544–552. doi:10.1121/1.1528172.

Anikin, A. 2020. A moan of pleasure should be breathy: The effect of voice quality on the meaning of human nonverbal vocalizations. *Phonetica* 77(5). 327–349. doi:10.1159/000504855.

Awan, S. N., N. Roya, M. E. Jetté, G. S. Meltzner & R. E. Hillman. 2010. Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: Comparisons with auditory-perceptual judgements from the CAPE-V. *Clinical Linguistics & Phonetics* 24(9). 742–758. doi:10.3109/02699206.2010.492446.

Bachorowski, J.-A. & M. J. Owren. 1999. Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in

running speech. *The Journal of the Acoustical Society of America* 106(2). 1054–1063. doi:10.1121/1.427115.

Baumann, O. & P. Belin. 2010. Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research* 74(1). 110–120. doi:10.1007/s00426-008-0185-z.

Belin, P., R. J. Zatorre, P. Lafaille, P. Ahad & B. Pike. 2000. Voice-selective areas in human auditory cortex. *Nature* 403(6767). 309–312. doi:10.1038/35002078.

Carroll, J. D. & J.-J. Chang. 1970. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika* 35(3). 283–319. doi:10.1007/bf02310791.

Chodroff, E. & C. Wilson. 2017. Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics* 61. 30–47. doi:10.1016/j.wocn.2017.01.001.

Fitch, W. T. 1997. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *The Journal of the Acoustical Society of America* 102(2). 1213–1222. doi:10.1121/1.421048.

Fontaine, M., S. A. Love & M. Latinus. 2017. Familiarity and voice representation: From acoustic-based representation to voice averages. *Frontiers in Psychology* 8. Article 1180. doi:10.3389/fpsyg.2017.01180.

Gerlach, L., K. McDougall, F. Kelly, A. Alexander & F. Nolan. 2020. Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features. *Speech Communication* 124. 85–95. doi:10.1016/j.specom.2020.08.003.

Gerratt, B. R., J. Kreiman & M. Garellek. 2016. Comparing measures of voice quality from sustained phonation and continuous speech. *Journal of Speech, Language, and Hearing Research* 59(5). 994–1001. doi:10.1044/2016_JSLHR-S-15-0307.

Hudson, T., K. McDougall & V. Hughes. 2021. Forensic phonetics. In R.-A. Knight & J. Setter (eds.), *The Cambridge Handbook of Phonetics*, 631–656. Cambridge University Press. doi:10.1017/9781108644198.026.

Humble, D., S. R. Schweinberger, A. Mayer, T. L. Jesgarzewsky, C. Dobel & R. Zäske. 2023. The Jena Voice Learning and Memory Test (JVLMT): A standardized tool for assessing the ability to learn and recognize voices. *Behavior Research Methods* 55(3). 1352–1371. doi:10.3758/s13428-022-01818-3.

Iwarsson, J., R. H. Nielsen & J. Næs. 2020. Mean fundamental frequency in connected speech and sustained vowel with and without a sentence-frame. *Logopedics Phoniatrics Vocology* 45(2). 91–96. doi:10.1080/14015439.2019.1637455.

Jacewicz, E., R. A. Fox & L. Wei. 2010. Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America* 128(2). 839–850. doi:10.1121/1.3459842.

Jessen, M. 1997. Speaker-specific information in voice quality parameters. *The International Journal of Speech, Language and the Law* 4(1). 84–103. doi:10.1558/ijsll.v4i1.84.

Kavanagh, C. M. 2012. *New consonantal acoustic parameters for forensic speaker comparison*: University of York dissertation. https://etheses.whiterose.ac.uk/3980/. Doctoral dissertation.

Kreiman, J. 2024. Information conveyed by voice quality. *The Journal of the Acoustical Society of America* 155(2). 1264–1271. doi:10.1121/10.0024609.

Kreiman, J., B. R. Gerratt, K. Precoda & G. S. Berke. 1992. Individual differences in voice quality perception. *Journal of Speech, Language, and Hearing Research* 35(3). 512–520. doi:10.1044/jshr.3503.512.

Kreiman, J., Y. Lee, M. Garellek, R. Samlan & B. R. Gerratt. 2021. Validating a psychoacoustic model of voice quality. *The Journal of the Acoustical Society of America* 149(1). 457–465. doi:10.1121/10.0003331.

Kuhl, P. K. 2011. Who's talking? *Science* 333(6042). 529–530. doi:10.1126/science.1210277.

Latinus, M., P. McAleer, P. E.G., Bestelmeyer & P. Belin. 2013. Norm-based coding of voice identity in human auditory cortex. *Current Biology* 23(12). 1075–1080. doi:10.1016/j.cub.2013.04.055.

Lavan, N., L. F. K. Burston & L. Garrido. 2019a. How many voices did you hear? natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology* 110(3). 576–593. doi:10.1111/bjop.12348.

Lavan, N., L. F. K. Burston, P. Ladwa, S. E. Merriman, S. Knight & C. McGettigan. 2019b. Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology* 72(9). 2240–2248. doi:10.1177/1747021819836890.

Lavan, N., S. Knight & C. McGettigan. 2019c. Listeners form average-based representations of individual voice identities. *Nature Communications* 10. Article 2404. doi:10.1038/s41467-019-10295-w.

Lederle, A., J. Barkmeier-Kraemer & E. Finnegan. 2012. Perception of vocal tremor during sustained phonation compared with sentence context. *Journal of Voice* 26(5). 668.e661–688.e669. doi:10.1016/j.jvoice.2011.11.001.

Lee, Y., P. Keating & J. Kreiman. 2019. Acoustic voice variation within and between speakers. *The Journal of the Acoustical Society of America* 146(3). 1568–1579. doi:10.1121/1.5125134.

Lee, Y. & J. Kreiman. 2022a. Acoustic voice variation in spontaneous speech. *The Journal of the Acoustical Society of America* 151(5). 3462–3472. doi:10.1121/10.0011471.

Lee, Y. & J. Kreiman. 2022b. Linguistic versus biological factors governing acoustic voice variation. In *Proceedings of Interspeech 2022*, 640–643. doi:10.21437/Interspeech.2022-10847.

Lee, Y. & J. Kreiman. 2023. Within- versus between-speaker acoustic variability in Thai. *The Journal of the Acoustical Society of America* 153(3_supplement). A295. doi:10.1121/10.0018911.

Loakes, D., H. Fraser & K. McDougall. 2023. A preliminary investigation of the acoustic factors impacting decision making in speaker attribution. *The Journal of the Acoustical Society of America* 154(4_supplement). A159. doi:10.1121/10.0023122.

Maryn, Y., P. Corthals, P. V. Cauwenberge, N. Roy & M. D. Bodt. 2010. Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels. *Journal of Voice* 24(5). 540–555. doi:10.1016/j.jvoice.2008.12.014.

McDougall, K. 2011. Acoustic correlates of perceived voice similarity - a comparison of two accents of English. In *20th International Association for Forensic Phonetics and Acoustics Conference*, Vienna, Austria. https://projects.ari.oeaw.ac.at/publications/iafpa_abstracts/nr19_revised_mcdougall.pdf. Conference session.

McDougall, K. & M. Duckworth. 2017. Profiling fluency: An analysis of individual variation in disfluencies in adult males. *Speech Communication* 95. 16–27. doi:10.1016/j.specom.2017.10.001.

McDougall, K. & M. Duckworth. 2018. Individual patterns of disfluency across speaking styles: A forensic phonetic investigation of Standard Southern British English. *The International Journal of Speech, Language and the Law* 25(2). 205–230. doi:10.1558/ijsll.37241.

McDougall, K., M. Duckworth & T. Hudson. 2015. Individual and group variation in disfluency features: A cross-accent investigation. In T. S. C. for ICPhS 2015 (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK: the University of Glasgow. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0308.pdf.

McDougall, K., R. Rhodes, M. Duckworth, P. French & C. Kirchhübel. 2019. Application of the 'TOFFA' framework to the analysis of disfluencies in forensic phonetic casework. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (eds.), *Proceedings of the 19th International Congress of Phonetic Sciences*, 731–735. Australasian Speech Science and Technology Association and International Phonetic Association. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_780.pdf.

Moon, K. R. M., S. M. Chung, H. S. Park & H. S. Kim. 2012. Materials of acoustic analysis: Sustained vowel versus sentence. *Journal of Voice* 26(5). 563–565. doi:10.1016/j.jvoice.2011.09.007.

Morrill, T., M. Baese-Berk & A. Bradlow. 2016. Speaking rate consistency and variability in spontaneous speech by native and non-native speakers of english. In *Proceedings of Speech Prosody 2016*, 1119–1123. doi:10.21437/SpeechProsody.2016-230.

Mühl, C., O. Sheil, L. Jarutytė & P. E. G. Bestelmeyer. 2018. The Bangor Voice Matching Test: A standardized test for the assessment of voice perception ability. *Behavior Research Methods* 50(6). 2184–2192. doi:10.3758/s13428-017-0985-4.

Nolan, F. 1991. Forensic phonetics. *Journal of Linguistics* 27(2). 483–493. doi:10.1017/s0022226700012755.

Nolan, F. 1997. Speaker recognition and forensic phonetics. In W. J. Hardcastle & J. Laver (eds.), *The Handbook of Phonetic Sciences*, 744–767. Blackwell Publishers.

Nolan, F. 2002. Intonation in speaker identification: an experiment on pitch alignment features. *The International Journal of Speech, Language and the Law* 9(1). 1–21. doi:10.1558/sll.2002.9.1.1.

Perrachione, T. K. 2019. Recognizing speakers across languages. In S. Frühholz & P. Belin (eds.), *The Oxford Handbook of Voice Perception*, 515–538. Oxford

University Press. doi:10.1093/oxfordhb/9780198743187.013.23.
Perrachione, T. K., S. N. D. Tufo & J. D. E. Gabrieli. 2011. Human voice recognition depends on language ability. *Science* 333(6042). 595. doi:10.1126/science.1207327.
Smorenburg, L. & W. Heeren. 2020. The distribution of speaker information in dutch fricatives /s/ and /x/ from telephone dialogues. *The Journal of the Acoustical Society of America* 147(2). 949–960. doi:10.1121/10.0000674.
Smorenburg, L. & W. Heeren. 2021. Acoustic and speaker variation in dutch /n/ and /m/ as a function of phonetic context and syllabic position. *The Journal of the Acoustical Society of America* 150(2). 979–989. doi:10.1121/10.0005845.
Vaňková, J. & R. Skarnitzl. 2014. Within- and between-speaker variability of parameters expressing short-term voice quality. In *Proceedings of Speech Prosody 2014*, 1081–1085. doi:10.21437/SpeechProsody.2014-206.
Xu, T., X. Jiang, P. Zhang & A. Wang. 2025. Introducing the Sisu Voice Matching Test (SVMT): A novel tool for assessing voice discrimination in Chinese. *Behavior Research Methods* 57(3). Article 86. doi:10.3758/s13428-025-02608-3.

Tianze Xu
University of Cambridge
tx243@cam.ac.uk