A comparison of voice similarity through acoustics, human perception and deep neural network (DNN) speaker verification systems

Suyuan Liu[1]*, Molly Babel*[1]*, Jian Zhu*[1]

[1]University of British Columbia, Canada

suyuan.liu@ubc.ca, Molly.Babel@ubc.ca, jian.zhu@ubc.ca

Abstract

Voice similarity can be assessed through acoustic analysis, perceptual judgments by human listeners, and the recent addition of automatic speaker verification systems. However, a comparison across the similarity judgments made from acoustics, listener perception, and deep neural network (DNN) based speaker verification systems has not yet been made. This project fills this gap by comparing acoustic similarity scores generated from 24 acoustic dimensions and verification scores generated by seven pretrained speaker verification models using the Wespeaker toolkit to perceptual similarity assessed by human listeners in an AX discrimination task and a (dis)similarity rating task. Results suggest verification similarities correlate with acoustic similarities, but not with human perceptual similarities when controlled for talker pair, indicating the correlation between listeners and speaker verification models happens at a gross-phonetic level rather than a fine phonetic level.

Index Terms: speaker recognition, human-computer interaction, computational paralinguistics

1. Introduction

What do we mean when we say two voices are similar? Voice similarity can be estimated through an assessment of acoustic similarity or judged perceptually by human listeners. Although perceptual similarity generally correlates with acoustic similarity [1, 2, 3, 4, 5], not all acoustic measurements can equally predict perceptual similarity of voices [6, 7]. For instance, listeners often rely more on f0 than other acoustic dimensions for voice similarity judgements [3, 4]. The imprecision in the alignment between acoustic and perceptual similarity is not surprising given that listeners do not have direct access to speech acoustics (for an overview see [8]). Moreover, perceptual similarity can also be affected by listener experience such as voice familiarity [9, 10] and speaking style [11, 12].

The ongoing discussion of how to accurately and efficiently assess voice similarity has been crucial for tasks such as voice parade and ear-witness identification in forensic linguistics [13] and voice actor casting [14]. Acoustic similarity itself does not address the similarities listeners perceive, but obtaining perceptual similarity judgements is often time-consuming and the results are prone to influence by the choice of stimuli. The development of automatic speaker recognition introduced another option for such assessment [15, 14]. It is unclear whether the similarity ratings generated by automatic speaker recognition systems merely reflect acoustic similarities or develop abstractions at a higher level that resemble human perception of voice similarity.

Automatic speaker recognition can be divided into speaker identification tasks and speaker verification tasks. Speaker verification tasks can be considered as a specific type of open-set speaker identification task, meaning that it involves novel speakers not present in the training data [16]. This study focuses on text-independent speaker verification tasks, which perform verification regardless of what was said or the phonological input and focus on features that are indicative of speakers' physiology and behavioural characteristics. These features are represented as speaker embeddings in automatic speaker recognition. Before the incorporation of deep neural network (DNN), i-vector was the common type of speaker embeddings used in speaker verification, which reduces the dimensionality of input acoustic signals to a constant length that captures speaker-specific characteristics. Currently, DNNs are used to learn other types of speaker embeddings in lieu of the i-vectors, but the benchmark DNN architecture has not been established. The common DNN representations include d-vectors, x-vectors and r-vectors (for a review, see [17]).

A few studies compared human perceptual similarity to automatically generated similarities by speaker verification systems (henceforth "verification similarity"). [18] showed that human perceptual similarities correlate with automatic similarities for male voices that are judged to be "similar" by an i-vector-based system, while female voices showed no correlation. [19] found positive correlations between perceptual similarity and verification similarities when judging speech in English for both English listeners and German listeners. In a more recent study, [20] further showed strong correlations between perceptual and verification similarities within and across dialects for male voices, especially when the speaker embeddings consisted of automatically extracted perceptually relevant phonetic features. [21] found that the correlations between verification and perceptual similarities can be affected by task designs (e.g. speaker clustering, lineup and binary-decision tasks). Specifically, [22] showed that perceptual similarities had a significant correlation with automatic similarities measured using Cosine Distance Scoring for an i-vector-based speaker verification model in a clustering task. [23] compared human performances to i-vector-based speaker verification systems in AX discrimination tasks of female voices with matched and mismatched speaking styles, showing a weak correlation between perceptual and verification scores. Humans always outperformed speaker verification systems except for the style-mismatched condition.

Notice that previous comparisons of perceptual similarity to verification similarity were made without consideration of acoustic similarity. This poses a question: how much can we attribute the correlation between human and speaker verification systems to acoustic similarity? As mentioned before, human perception of voice similarity does not correlate perfectly with acoustic similarity. How much would this mismatch be reflected

in verification scores, if at all? This project aims to explore this relationship between acoustic similarity, perceptual similarity, and verification similarity in DNN speaker verification models.

2. Materials: SpiCE corpus

The English portion of the Speech in Cantonese and English (SpiCE) corpus is used in this study [24]. SpiCE is an open-access corpus of conversational speech from 34 early Cantonese-English bilinguals (17 self-identified female, 17 self-identified male; Age range: 19 - 34). The recordings were made with a 44.1 kHz sampling rate and 16-bit resolution. Only female participants were used in this study.

3. Acoustic similarity measures

The acoustic similarity scores were taken from [25], where a detailed description of the measurement and calculation for each acoustic dimension can be found. To summarise, a total of 24 acoustic measurements including f0, F1, F2, F3, F4, H1$*$-H2$*$, H2$*$-H4$*$, H4$*$-H2kHz$*$, H2kHz$*$-H5kHz, CPP, Energy, SHR and their moving standard deviations (s.d.s), calculated for each measure to capture dynamic changes, were analyzed. These acoustic measures were selected based on [26], and are the perceptually validated measures of the psychoacoustic model of the voice [27]. Principle component analyses (PCA) on these 24 acoustic dimensions were performed to reduce the dimensionality of acoustic measurements. A pairwise numerical comparison of PCAs was then performed using the canonical correlation analysis (CCA). CCA provides redundancy indices that represent a metric of redundancy between the two PCAs. These redundancy indices are used as acoustic similarity scores in the analysis. Only within-English-cross-talker redundancy indices for female talkers were used for acoustic similarities in this study. Higher redundancy indices indicate higher similarity between voices.

4. Listener experiment

4.1. AX discrimination task

Participants: 530 participants with heterogeneous language backgrounds from the University of British Columbia (UBC) completed the task for partial course credit. The student population at UBC is comprised of a majority of early simultaneous bilinguals who have English as the most common dominant language [28].
Materials: Thirty-six 1-2 second (s) intervals of spontaneous interview speech were extracted for each talker. None of these items exhibited an observable disfluency or a codeswitch.
Procedure: The task was presented online via jsPsych. Participants were presented with two speech samples that were separated by a 1500 millisecond interstimulus interval. Participants' task was to determine whether the two speech samples came from the same or different speakers by pressing keys 'f' and 'j' on their keyboard; key assignment was counterbalanced across listeners. Voice samples were accompanied by a yellow circle (voice 1) and a blue circle (voice 2) to facilitate the parsing of the speech as two utterances. The trial advanced if no same/different response was entered after 5000 ms. Participants were exposed to a total of 186 trials. Listeners were randomly assigned to one of 55 different lists which contained exhaustive comparisons for 5 different female voices, in addition to the occasional presentation of other voices. In order to maximize the talker pairings, the number of same and different talker trials was unequal. Listeners were presented with the following trial types: different talker different language (n = 42), different talker same language (n = 82), same talker different language (n = 20), same talker same language (n = 42).

4.2. (Dis)similarity rating task

Participants: 64 participants from the same population as the AX task completed the task for partial course credit.
Materials: A selection of 1.5-2 second (s) intervals of spontaneous interview speech were extracted for each talker. From these items, the longest fluent interval without a codeswitch or disfluency was selected, resulting in materials that ranged from 1.64 - 2 s in duration, and all items had continuous speech during that interval. Because it is impractical to present all pairwise voice samples to listeners, separate lists (n = 174) containing 240 trials were created that each exhaustively compared 7 randomly selected voices in pairwise combinations within and across languages. Within-voice comparisons were not made and lists were not balanced by talker gender.
Procedure: Listeners were presented with two voice samples in succession with a 2 s interstimulus interval with the option of playing each stimulus a second time. Listeners' task was to compare the (dis)similarity of the voices on a visual-analogue scale with endpoints labelled as the same or different, which were counterbalanced across listeners. The cursor began at the midpoint landmark for each trial.

5. Neural network experiment

5.1. Model description

Pretrained speaker verification models from Wespeaker [29] were used to rate voice similarities. Wespeaker is an open-source speaker-embedding learning toolkit designed for research and production purposes. To balance the diversity in model architectures, parameter counts and training methods, we evaluated seven pretrained speaker verification models: CAM++, CAM++LM, ResNet34, ResNet34LM, ResNet152LM, ResNet221LM, ResNet293LM. All were trained using the VoxCeleb corpus [30, 31], which contains around 2800 hours of speech from over 6000 celebrities with various backgrounds and accents extracted from YouTube. These pretrained models were trained on either one of two model architectures: ResNet and CAM++. ResNet models apply the original ResNet [32] to the speaker verification domain [33], with fully convolutional blocks and residue connections. Despite its simplicity, ResNet remains a strong baseline in all speaker verification benchmarks. CAM++ [34] is a state-of-the-art speaker verification model that incorporates context-aware masking (CAM) on top of the ResNet blocks to refine the speaker representations through the gating mechanism. The model variants with the LM suffixes further went through large-margin finetuning to enhance performance.

5.2. Materials for automatic experiment

All materials used for the neural network experiment were from the interview section of SpiCE corpus. Since pretrained Wespeaker speaker verification models recommend using audio stimuli longer than 5 seconds, sample audio files with both shorter duration (range: 1.64 - 2s) and longer duration (range: 4.75 - 5.77s) were used for the automatic experiment. The shortest stimuli were the same used in the listener experiment as described in 4.2. The longer stimuli were manually extracted to

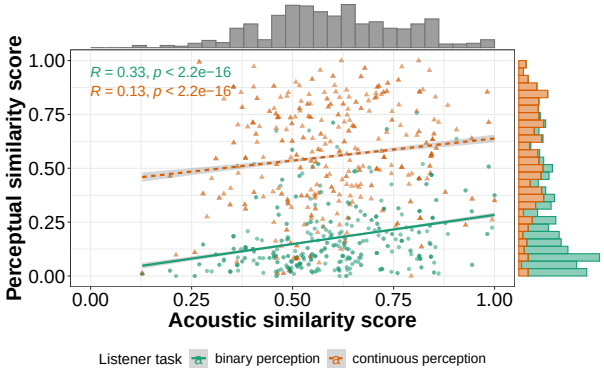

Figure 1: *Correlation and the Spearman's rank correlation coefficients between normalized perceptual similarity scores (x-axis, 1: most similar) and normalized acoustic similarity scores (y-axis, 1: most similar) grouped by listener task (green/solid line: AX discrimination task; orange/dotted line: (dis)similarity rating task). The distributions of scores are shown on the marginal histograms.*

avoid the accidental inclusion of long periods of silence. All speaker verification models convert speech stimuli into fixed-dimensional speaker embeddings and compute the cosine similarity between pairs of speaker embeddings as a proxy for voice similarity.

6. Results

All data manipulation and visualization were done using the {tidyverse} package [35] and all Bayesian models were fitted in Stan using the {brms} package [36] in R [37]. For all Bayesian models below, samples were drawn from the posterior distribution using a four-chain Hamiltonian Monte-Carlo sampling (3000 iterations, 1000 warm-ups). Weakly informative priors were used for all parameters. The mean of the posterior distribution, the 95% credible interval (CrI), and the probability of direction (PD) are reported. A CrI that does not encompass 0 suggests a meaningful effect and a PD greater than 95% suggests a probable direction of effect [38].

6.1. Data wrangling

All acoustic similarity scores (range: 0.636 - 0.951) were normalized on a 0 to 1 scale. For the AX discrimination task results, the percentage of "same" responses for each speaker pair was calculated for each listener and then averaged across all listeners. All perceptual dissimilarity ratings were converted to similarity ratings by subtracting from 100. The top and bottom 5% scores for each speaker pair were removed since listeners might not use the rating scale the same as one another (some might only rate the voices from 30 to 70 while others might use the endpoints). All converted perceptual similarity scores were then normalized by listener to a 0 to 1 scale, and averaged across all listeners for each speaker pair. All verification similarity scores were normalized by model type to a 0 to 1 scale.

6.2. Acoustic similarity and perceptual similarity

Figure 1 shows the correlation and Spearman's rank correlation coefficients between normalized perceptual similarity scores (x-axis, 1: most similar) and normalized acoustic similarity scores (y-axis, 1: most similar) grouped by listener task. The perceptual similarity scores for the AX discrimination task have a Spearman's rho of 0.33, suggesting a moderate correlation with the acoustic similarity scores. The perceptual similarity scores for the (dis)similarity rating task have a Spearman's rho of 0.13, suggesting a negligible correlation with the acoustic similarity scores.

Two Bayesian mixed-effect linear regression models were fitted for each listener task. One was fitted with the acoustic similarity scores as the outcome variable and the by-listener normalized similarity ratings as the predictor variable; the other was fitted with the dummy-coded AX similarity decisions as the predictor variable (same vs. different; reference level: same). Both models included by-talker and by-listener random intercepts. The Bayesian model outputs showed little evidence for the correlation between similarity ratings and acoustic similarity scores (β = 0.01, CrI = [-0.01, 0.04], PD = 87.14%) but strong evidence for a positive correlation between AX similarity decision and the acoustic similarity scores (β = 0.03, CrI = [0.02, 0.03], PD = 100%).

6.3. Verification score and acoustic similarity

A Bayesian mixed-effect linear regression model was fitted with the verification score as the outcome variable and predictor variables of acoustic similarity scores, dummy-coded duration of stimuli (short vs. long, reference level: short) and their interaction. The Bayesian model provides strong evidence that verification similarity has a positive correlation with acoustic similarity (β = 0.09, CrI = [0.05, 0.13], PD = 99.99%). Longer audio stimuli, in general, receive higher verification similarity scores (β = 0.10, CrI = [0.07, 0.12], PD = 100%). There is also weak evidence for the interaction effect between stimuli duration and acoustic similarity scores: longer stimuli exhibited a weaker correlation between acoustic and verification scores (β = -0.04, CrI = [-0.09, 0.01], PD = 95.64%). The green lines in Figure 2 showed the predicted verification (y-axis, 1: most similar) score and its correlation with acoustics similarity (x-axis, 1: most similar).

6.4. Verification score and perceptual similarity

6.4.1. Verification score and (dis)similarity rating

Two Bayesian mixed-effect linear regression models were fitted with the verification similarity as the outcome variable and predictor variables of perceptual similarity ratings, dummy-coded duration of stimuli (short vs. long, reference level: short) and their interaction. Both Bayesian models included by-listener and by-model random intercepts. The difference lies in the inclusion of by-talker random intercept. The Bayesian model *with* by-talker random intercept showed little evidence for the correlation between verification scores and similarity ratings (β = 0.00, CrI = [-0.00, 0.01], PD = 80.62%) regardless of interaction with stimuli duration (β = -0.01, CrI = [-0.02, 0.00], PD = 89.84%). The Bayesian model *without* by-talker random intercept showed strong evidence for a positive correlation between verification scores and similarity ratings (β = 0.05, CrI = [0.04, 0.06], PD = 100%) despite interaction with stimuli duration (β = 0.01, CrI = [-0.01, 0.02], PD = 72.72%). The purple lines in Figure 2 showed the predicted verification scores (y-axis, 1: most similar) and their correlation with perceptual similarity ratings (x-axis, 1: most similar). The dotted lines are posterior draws from the Bayesian model without by-talker random intercept.

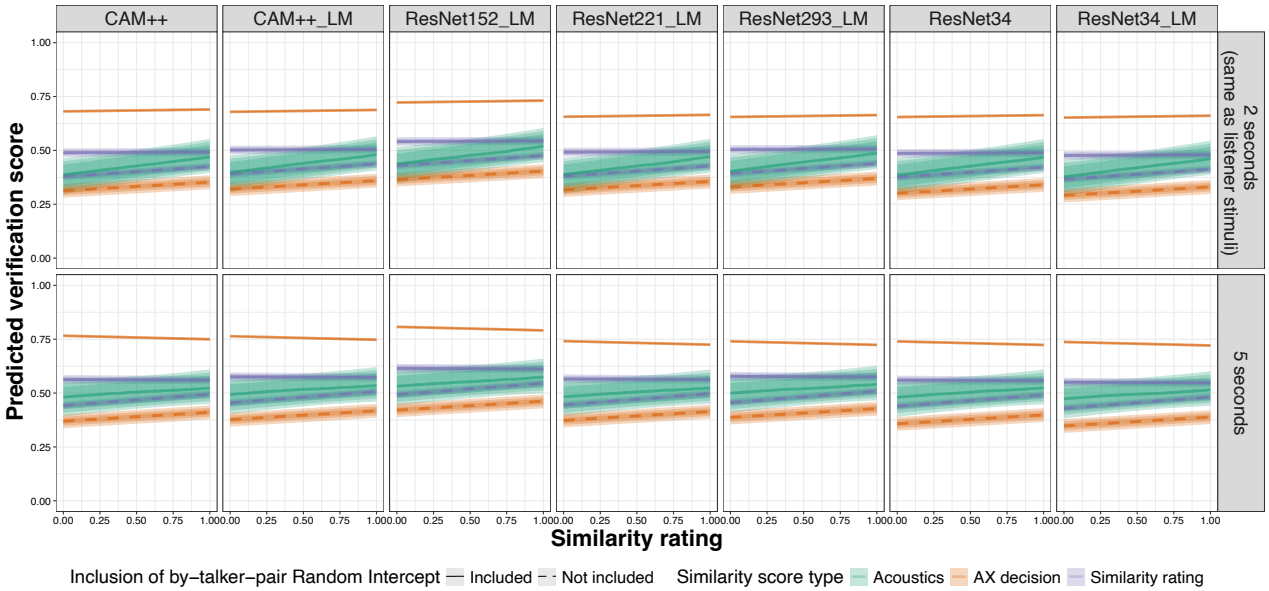

Figure 2: *Predicted verification score (y-axis, 1: most similar) drawn from the posterior distribution in relation to acoustic similarity (x-axis, green line, 1: most similar), AX decision (x-axis, orange line, 1: most similar) and similarity rating (x-axis, purple line, 1: most similar) grouped by stimuli duration (rows) and speaker verification model type (columns). The dotted lines are posterior draws from the mixed-effect model with by-talker random intercept. The solid lines are posterior draws from the mixed-effect model without by-talker random intercept.*

6.4.2. Verification score and AX similarity decisions

Two Bayesian mixed-effect linear regression models were fitted with the verification score as the outcome variable and predictor variables of dummy-coded AX similarity decision (same vs. different, reference level: same), dummy-coded duration of stimuli (short vs. long, reference level: short) and their interaction. Both Bayesian models included by-listener and by-model random intercepts but only one included a by-talker random intercept. The Bayesian model *with* by-talker random intercept showed no evidence for the correlation between AX similarity decision and verification similarity (β = 0.00, CrI = [-0.00, 0.00], PD = 51.78%) regardless of interaction with stimuli duration (β = -0.00, CrI = [-0.00, 0.00], PD = 52.58%). The Bayesian model *without* by-talker random intercept showed strong evidence for a positive correlation between AX similarity decision and verification similarity (β = 0.02, CrI = [0.02, 0.03], PD = 100%). Longer stimuli samples elicit a stronger positive correlation as seen from the interaction effect (β = 0.02, CrI = [0.02, 0.03], PD = 100%).

7. Discussion and Conclusion

This project again showed that human perception of voice similarity does not always correlate with acoustic similarity and can be task-dependent. The correlation between perceptual similarity and acoustic similarity only surfaces in the results of the AX discrimination task but not to the same extent in similarity rating data.

For speaker verification systems, verification similarities correlate with acoustic similarity with or without control for talker pairs. However, when controlled for talker pairs, the correlation between verification similarities and perceptual similarities disappears, regardless of the listener's tasks. This suggests that the positive correlation observed between verification similarities and perceptual similarities could be dragged by the agreements between speaker verification systems and human listeners when the talkers in comparison are either very similar or very different. Controlling for talker pairs indirectly controls for the *gross-phonetic details* between talkers. When these gross-phonetic details are controlled, we can evaluate whether speaker verification systems and human listeners react to talker pairs that have acoustic overlap in a similar manner. In this case, the similarity between talkers requires judgements based on *fine phonetic details*. Since our results suggest the correlation between verification similarities and perceptual similarities disappear when controlled for talker pairs, we can infer that the verification-perception correlation only exists at a gross-phonetic level. Relating to previous comparisons of verification similarities and perceptual similarities, this project adds to the interpretation of their correlation: speaker verification systems and human listeners make similar judgements of voice similarity at the gross-phonetic level, but not at the fine phonetic level.

To summarise, this study underscores the potential discrepancies introduced by different evaluation metrics in assessing voice similarity. Perceptual similarities determined by human listeners can be task-dependent and do not always correlate with acoustic similarity. Verification similarities, while correlating with acoustic similarity, do not capture human judgments at the fine-phonetic level. Therefore, it is crucial to consider the specific aspects of voice similarity that each assessment method reflects. For instance, when evaluating the feasibility of developing automatic speaker recognition systems for forensic applications, it is important to recognize that verification similarities might not accurately reflect human judgments at a fine-phonetic level. To minimize the influence of varying evaluation metrics, we advocate for transparent and detailed descriptions of the assessment methods used when reporting evaluations of voice similarity.

8. References

[1] Y. Adachi, S. Kawamoto, S. Morishima, and S. Nakamura, "Perceptual similarity measurement of speech by combination of acoustic features," in *2008 IEEE ICASSP*, 2008, pp. 4861–4864.

[2] M. Weirich and L. Lancia, "Perceived auditory similarity and its acoustic correlates in twins and unrelated speakers." in *ICPhS*, 2011, pp. 2118–2121.

[3] F. Nolan, K. McDougall, and T. Hudson, "Some acoustic correlates of perceived (dis) similarity between same-accent voices." in *ICPhS*, vol. 17, 2011, pp. 1506–1509.

[4] T. K. Perrachione, K. T. Furbeck, and E. J. Thurston, "Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices," *The Journal of the Acoustical Society of America*, vol. 146, no. 5, pp. 3384–3399, 2019.

[5] N. Lavan, J. Kreitewolf, J. Obleser, and C. McGettigan, "Familiarity and task context shape the use of acoustic information in voice identity perception," *Cognition*, vol. 215, p. 104780, 2021.

[6] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, and P. Corthals, "Acoustic measurement of overall voice quality: A meta-analysis)," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2619–2634, 11 2009. [Online]. Available: https://doi.org/10.1121/1.3224706

[7] J. Kreiman and D. Sidtis, *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons, 2011.

[8] M. B. Winn and C. E. Stilp, "Phonetics and the auditory system," in *The Routledge handbook of phonetics*, W. A. Katz and P. Assmann, Eds. Routledge, 2019, pp. 164–192.

[9] N. Lavan, S. K. Scott, and C. McGettigan, "Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices." *Journal of Experimental Psychology: General*, vol. 145, no. 12, p. 1604, 2016.

[10] S. V. Stevenage, A. E. Symons, A. Fletcher, and C. Coen, "Sorting through the impact of familiarity when processing vocal identity: Results from a voice sorting task," *Quarterly Journal of Experimental Psychology*, vol. 73, no. 4, pp. 519–536, 2020.

[11] H. M. Smith, T. S. Baguley, J. Robson, A. K. Dunn, and P. C. Stacey, "Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance," *Applied Cognitive Psychology*, vol. 33, no. 2, pp. 272–287, 2019.

[12] S. V. Stevenage, R. Tomlin, G. J. Neil, and A. E. Symons, "May i speak freely? the difficulty in vocal identity processing across free and scripted speech," *Journal of Nonverbal Behavior*, vol. 45, no. 1, pp. 149–163, 2021.

[13] N. Atkinson, "Variable factors affecting voice identification in forensic contexts," Ph.D. dissertation, University of York, 2015.

[14] A. Gresse, M. Quillot, R. Dufour, V. Labatut, and J.-F. Bonastre, "Similarity metric based on siamese neural networks for voice casting," in *ICASSP 2019-2019 IEEE ICASSP (ICASSP)*. IEEE, 2019, pp. 6585–6589.

[15] N. Obin and A. Roebel, "Similarity search of acted voices for automatic voice casting," *IEEE/ACM Transactions on Audio, Speech, & Lang. Processing*, vol. 24, no. 9, pp. 1642–1651, 2016.

[16] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[17] M. Jakubec, R. Jarina, E. Lieskovska, and P. Kasak, "Deep speaker embeddings for speaker verification: Review and experimental comparison," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107232, 2024.

[18] F. Kelly, A. Alexander, O. Forth, S. Kent, J. Lindh, and J. Åkesson, "Identifying perceptually similar voices with a speaker recognition system using auto-phonetic features." in *Interspeech*, 2016, pp. 1567–1568.

[19] L. Gerlach, K. McDougall, F. Kelly, A. Alexander, and F. Nolan, "Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features," *Speech Communication*, vol. 124, pp. 85–95, 2020.

[20] F. K. A. A. Linda Gerlach, Kirsty McDougall, "Automatic assessment of voice similarity within and across speaker groups with different accents," in *Proceedings of the 20th International Congress of Phonetic Sciences*, 2023, pp. 3785–3789.

[21] B. O'Brien, C. Meunier, A. Ghio, C. Fredouille, and J.-F. Bonastre, "Discriminating speakers using perceptual clustering interface," in *Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications*, vol. 8, Zurich, Switzerland, Feb. 2021, pp. 97–111.

[22] B. O'Brien, C. Meunier, A. Ghio, C. Fredouille, J.-F. Bonastre, and C. Guarino, "Discriminating speakers using perceptual clustering interface," in *Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications*, vol. 8, Zurich, Switzerland, Feb. 2021, pp. 97–111.

[23] S. J. Park, G. Yeung, N. Vesselinova, J. Kreiman, P. A. Keating, and A. Alwan, "Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles," *JASA*, vol. 144, no. 1, pp. 375–386, 2018.

[24] K. A. Johnson, "SpiCE: Speech in Cantonese and English," 2021. [Online]. Available: https://doi.org/10.5683/SP2/MJOXP3

[25] K. A. Johnson and M. Babel, "The structure of acoustic voice variation in bilingual speech," *The Journal of the Acoustical Society of America*, vol. 153, no. 6, pp. 3221–3221, 2023.

[26] Y. Lee and J. E. Kreiman, "Within-and between-speaker acoustic variability: Spontaneous versus read speech," *JASA*, vol. 146, no. 4, pp. 3011–3011, 2019.

[27] J. Kreiman, Y. Lee, M. Garellek, R. Samlan, and B. R. Gerratt, "Validating a psychoacoustic model of voice quality," *JASA*, vol. 149, no. 1, pp. 457–465, 2021.

[28] M. Babel, L. Lloy, Z. K. Pirbaluti, L. Suite, and R. Soo, "Celebrating and quantifying the linguistic diversity of the ubc student community," *UBC Occasional Papers in Linguistics*, vol. Papers in Honour of Hotze Rullmann, no. 1, pp. 49–60, 2023.

[29] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *ICASSP 2023-2023 IEEE ICASSP (ICASSP)*. IEEE, 2023, pp. 1–5.

[30] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Science and Language*, 2019.

[31] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[34] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "Cam++: A fast and efficient network for speaker verification using context-aware masking," *arXiv preprint arXiv:2303.00332*, 2023.

[35] H. Wickham *et al.*, "Welcome to the tidyverse," *Journal of Open Source Software*, vol. 4, no. 43, p. 1686, 2019.

[36] P.-C. Bürkner, "Bayesian item response modeling in R with brms and Stan," *Journal of Statistical Software*, vol. 100, no. 5, pp. 1–54, 2021.

[37] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[38] B. Nicenboim and S. Vasishth, "Statistical methods for linguistic research: Foundational ideas—part ii," *Language and Linguistics Compass*, vol. 10, no. 11, pp. 591–613, 2016.