The Role of Voice Source and Filter in Automatic Speaker Recognition

Yuhan Huang, Chris Carignan, Josef Schlittenlacher

Automatic speaker recognition (ASR) gleans information about individual speaker characteristics from speech acoustics, including idiosyncratic details about unique source (i.e. larynx) and filter (i.e. vocal tract) traits. Previous studies have mostly used inverse filtering to acquire voice source information (Patel et al., 2011), ad-hoc selection of source and filter acoustic feature sets (Hughes et al., 2023), and/or opaque "black-box" models to increase performance (Kabir et al., 2021). However, these approaches raise questions regarding how the complex interaction of voice source and filter influence ASR in more naturalistic, ecologically valid ways.

In this study, we applied a complete feature set (88 features, eGeMAPS; Eyben et al., 2016) to two natural speech modes (modal speech and whispering), in order to broadly tease apart the effects of voice source + filter (modal) from the filter alone (whisper), using an interpretable machine learning model (XGBoost; Chen & Guestrin, 2016). Sixty professional actors were recruited to record 20 short sentences in both speech modalities in an anechoic chamber at University College London. We inferred the relative importance of voice and filter by observing the *decrease* in model accuracy from cross-mode model testing (e.g. training on modal, testing on whisper) compared to baseline within-mode model testing (e.g. training on modal, testing on modal). Model training/testing was carried out 400 times, in order to generate distributions of 100 accuracy values for each scenario.

We hypothesized that: (1) if the voice *source* is responsible for speaker-specific acoustic traits, we should observe a *larger decrease in accuracy* for models trained on modal and tested on whispered speech (MW models) compared to their baseline performance (MM models), since the *source-related* features present in model training would be absent in the test set; or (2) if the voice *filter* is responsible for speaker-specific acoustic traits, we should observed *relatively equal changes in accuracy* for both models trained on whispered/modal speech and tested on modal speech/whispered speech compared to their baseline performance (MM and WW models), since the *filter-related* features present in model training would also be present in both test sets. Our results reveal that MW models decreased by an average of 57.7% accuracy compared to their baseline (i.e. MM models), whereas WM models decreased by an average of 51.4% accuracy compared to their baseline (i.e. WW models); a t-test revealed the difference between the two cross-mode distributions to be significant, with a large effect size (Cohen's d: 6.117). Despite this significant effect, the difference is relatively small (6.4%) and the accuracies for both cross-mode scenarios are similar in comparison to chance (i.e. 5.16-5.27 times greater than chance accuracy). These results support hypothesis (2), suggesting that naturalistic filter-related acoustic information plays a crucial role in vocal identity.

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion*, 12(5), 1161–1179. https://doi.org/10.1037/a0025827

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing, 7*(2), 190–202. IEEE Transactions on Affective Computing. https://doi.org/10.1109/TAFFC.2015.2457417

Hughes, V., Cardoso, A., Foulkes, P., French, P., Gully, A., & Harrison, P. (2023). Speaker-specificity in speech production: The contribution of source and filter. *Journal of Phonetics, 97*, 101224. https://doi.org/10.1016/j.wocn.2023.101224

Kabir, M. M., Mridha, M. F., Shin, J., Jahan, I., & Ohi, A. Q. (2021). A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities. *IEEE Access, 9*, 79236–79263. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3084299

Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. *Biological Psychology, 87*(1), 93–98. https://doi.org/10.1016/j.biopsycho.2011.02.010