The phonetic meaning of Mel-Frequency Cepstral Coefficients (MFCC) for vowels

Alessandro Vietti, Free University of Bozen-Bolzano

Mel-frequency Cepstral Coefficients (MFCCs) are parameterizations of the speech signal that have been used for many decades in the area of Automatic Speech Recognition (Davis and Mermelstein 1980), and they continue to be used in many modern AI systems, e.g. HuBERT (Hsu et al. 2021) and Whisper (Radford et al. 2022), as their quantification of the speech signal tends to be less-speaker specific. It is commonly believed, however, that these parameters are not interpretable (Tracey et al., 2023; Burridge and Vaux, 2023) in the way that formants are, for instance. In this talk, I will show that MFCCs are indeed interpretable for vowel spectra, using simple Fourier-Theoretic arguments, effect-size statistics, and measures from modern information theory (Partial Information Decomposition, Timme and Lapish, 2018; Willams and Beer 2010). Our arguments are based on analyses of all vowel data from the TIMIT database, with large amounts of speaker, context, prosodic, and dialectal variability. We also ask the question of why certain MFCCs are responsible for different phonetic class distinctions.

I will discuss how this work can have relevance to the phonetic sciences and speech technologies, and could help bridge research in both fields. For the phonetic sciences, use of more speaker-independent parameters could advance acoustic-phonetic research into the phonetic differences amongst communities, especially in big data research. For speech technologies, research on explainability and interpretability of the intrinsic working of modern ASR systems would benefit from a phonetic interpretation of the inputs to the nonlinear parallel transformations computations in the neural networks.

This is joint work with Khalil Iskarous, University of Southern California.

Burridge, J. & Vaux, B. (2023). Low dimensional measurement of vowels using machine perception. *Journal of the Acoustical Society of America,* 153, 1.

Davis, S.B. & and Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences“. *IEEE Transactions of Acoustics, Speech, and Signal Processing*, 28, 4, 357-366.

Hsu, W-N, Bolte, B., Tsai, Y-H, Lakhotia, K., & Salakhutdinov, R. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29,* 3451-3460.

Radford, A., Kim, J-W., Tao, Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. arXiv: https://arxiv.org/abs/2212.04356.

Timme, N. & Lapish, C. (2018). A Tutorial for Information Theory in Neuroscience*, eNeuro,* 5.

Tracey, B., Volfson, D., Glass, J., R'mani, H., Kostrzebski, M., Adams, J., Kangarloo, T., Brodtmann, A., Dorsey, E.R., and Vogel, A. (2023). Towards interpretable speech biomarkers: exploring MFCCs. *Scientific Reports*, 13, 22787.

Williams, P. & Beer, R. (2010). Nonnegative Decomposition of Multivariate Information. arXiv: 1004. 2515. https://arxiv.org/abs/1004.2515.