Exploring the Articulatory Perspective of Mel-Frequency Cepstral Coefficients: Unravelling the Link between MFCCs and Vocal Tract Features

Bruce Xiao Wang[1], *Lei He*[2]
[1]*Department of Chinese and Bilingual studies, Hong Kong Polytechnic University, HK*
brucex.wang@polyu.edu.hk
[2]*Department of Computational Linguistics - Phonetics, University of Zurich, Switzerland.*
lei.he@uzh.ch

'work in progress' poster

In recent years, Mel-frequency cepstral coefficients (MFCCs; Davis and Mermelstein, 1980) have been widely used as the input features of semi-automatic (Nolan & Grigoras, 2005) forensic voice comparison (FVC) systems, and some studies have shown that MFCC features yield better speaker discriminatory performance (e.g., lower EER and/or C_{llr}) than traditional acoustic phonetic features (e.g., vowel formants).

MFCCs capture the spectral characteristics of speech signal, and spectral characteristics are a function of vocal tract (Fant, 1971), e.g., oral cavity. It is claimed that MFCCs capture shape and features of the human vocal tract; however, no studies have attempted to investigate how MFCCs and vocal tract features, if there are any, are related. A recent study (Hughes et al., 2023) has partially discussed the correlation between MFCCs and formant values; however, the interpretability of MFCCs has rarely been properly discussed or investigated as well as the question of *why* MFCCs, despite a higher dimensionality, outperform traditional acoustic phonetic features.

In the current work-in-progress paper, we aim to investigate MFCCs from an articulatory perspective. We extracted the first 12 MFCCs and articulatory kinematics from three vowels (i.e., FLEECE, TRAP, FOOT) in single words. The data, obtained from Ji et al. (2014), contained the raw recordings of single words produced by 20 Midwestern standard American English speakers (10 male and 10 female) as well as the articulatory kinematics. The articulatory kinematics data was measured using electromagnetic articulography (EMA containing the movement of tongue dorsum (TD), tongue lateral (TL), tongue blade (TB), upper lip (UL), lower lip (LL), lateral lip corner (LC) measured in three dimensions (Figure 1, i.e., x: front and back, y: height, z: left and right). We will perform principal component analysis (PCA) on the first 12 MFCCs as well as 12 articulatory kinematics data (i.e., 6 sensors * x-axis * y-axis) aiming to explore to what extent the MFFCs and articulatory kinematics contribute to the first three components.

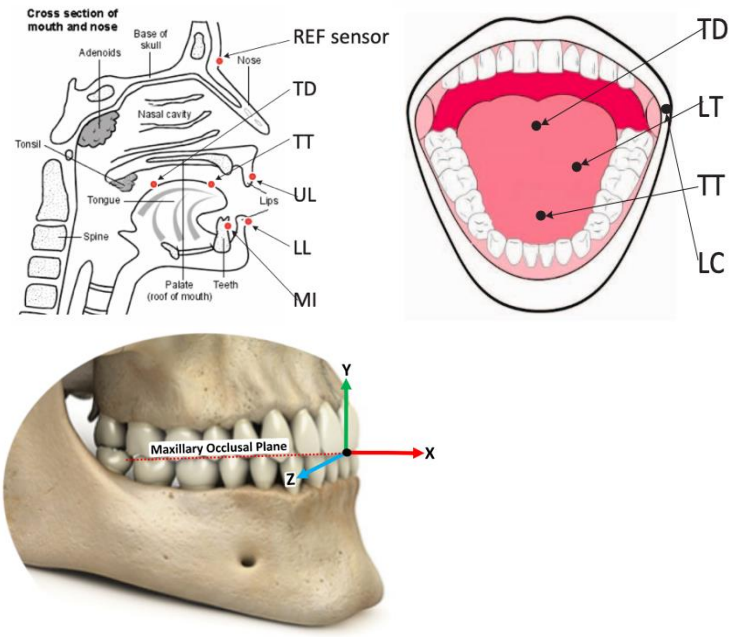

Figure 1. Leftmost and central panel: sensor placement (Figure 1 from Ji et al. 2014). Rightmost panel: Target anatomically-referenced coordinate system, Positive increases in sensor values denote forward, upward, and rightward movement along x, y, z, respectively (Figure 2 from Berry et al., 2016 EMA-MAE corpus User's Handbook).

References

Berry, J., Ji, A. & Johnson, T. (2016). *EMA-MAE Corpus User's Handbook* (Version 2.0). Marquette University, Milwaukee, WI, USA.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing, 28*(4), 357-366.

Fant, G. (1971). *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations* (No. 2). Walter de Gruyter

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.

Ji, A., Berry, J. J., & Johnson, M. T. (2014, May). The Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data. In 2014 *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 7719-7723). IEEE.

Nolan, F., & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law, 12*(2), 143-173.