The interpretability of Mel-Frequency Cepstral Coefficients: A pilot study using articulatory phonetics

WANG Xiao, HE Lei

Abstract In recent years, Mel-frequency cepstral coefficients (MFCCs; Davis and Mermelstein, 1980) have been widely used as the input features in speech technology (Prakash & Gangashetty, 2012), e.g., automatic speech recognition, forensic speaker recognition (Hughes et al., 2017; Zhang et al., 2013) and automatic disordered speech recognition (Pellegrini et al., 2014), i.e., speakers with unilateral facial palsy (UFP). Although some studies showed that using MFCCs as the input features for automatic speech/speaker recognition yielded better performance when comparing to the use of other acoustic features (e.g., formants), the interpretability of MFCCs is never properly discussed and the number of MFCCs (e.g., 12 coefficients) extracted in automatic systems is often performance oriented (i.e., lower error rate) (Hughes et al., 2023).

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 the 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; nevertheless, the interpretability of MFCCs has rarely been properly investigated as well as the question of *why* MFCCs, despite a higher dimensionality, outperform traditional acoustic phonetic features in automatic systems.

In the current pilot study, we investigated the interpretability of MFCCs from an articulatory perspective. We extracted the first 12 MFCCs and articulatory kinematics from three corner vowels (i.e., FLEECE, TRAP, FOOT) in single words (i.e., not connected speech). 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, i.e., x: front and back, y: height, z: left and right (Figure 1). We performed 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 different components (e.g., PC1, PC2). Only the front and back (x-axis) and height (y-axis) dimensions from the EMA kinematics data were used as the left and right (z-axis) dimension seems to be less relevant to the production of corner vowels investigated here. The implementation of PCA was carried out using the `nsprcomp` (Sigg & Buhmann, 2008) package, and the plots were generated using the `factoextra` (Kassambara & Mundt, 2020) package in R (R Core Team, 2022).

Figure 2 shows the PCA representation in the first 10 dimensions of the FOOT vowel of male and female speakers respectively. Using the first 12 MFCCs as well as 12 kinematics data (i.e., 24 dimensions in total) as the input for PCA, Figure 2 shows that the first six dimensions account for over 60% of variance explained for both male (64.13%) and female (60.85%) speakers. Further, we accessed the squared cosine values to investigate which variables (e.g., MFCCs or EMA kinematics or both) are well represented in which principal component. Figure 3 shows the squared cosine values (an estimate of the quality of the representation of MFCCs and EMA kinematics data across principal components, marked as *Dim* (dimension) in the plot) of all variables on the first 10 dimensions. The labels on top indicate the EMA kinematics and MFCCs (e.g., c1 represents the first coefficient), while labels on the left indicate the *n*th dimension. The colour scheme indicates the value of the squared cosine, and a higher squared cosine value indicates a good representation of the variable on the principal component. Taking dimension one for example, Figure 3 shows that the tongue dorsum at the vertical movement (i.e., TDy) is well represented for both male and female speakers; meanwhile in dimension two, the lower lip movement at the front and back direction (i.e., LLx) is well represented for both genders as well. These are sensible similarities between male and female speakers from an articulatory perspective. This is because the production of the FOOT vowel involves a lowering tongue dorsum and a constriction at the lower lip. However, there are some discrepancies in terms of the wellness of the representation of MFCCs in the principal components for different genders. For example, the second, fourth, fifth and seventh coefficients (i.e., c2, c4, c5, c7) are relatively well represented at the first dimension for male speakers, while it is the fourth, sixth and seventh (c4, c6, c7) for female

speakers. This discrepancy of the MFCC representations in principal components between male and female speakers can be explained by the fact that MFCCs capture vocal track characteristics as well as articulatory movement, while EMA kinematics only captures articulatory movement. In general, male speakers have longer vocal track and larger larynxes and thicker vocal folds than female speakers (Yule 2010:275) biologically. Further, from a sociolinguistic perspective, male and female speakers are likely to use different variants, e.g., female speakers tend to use the standard variant more than male speakers (see Labov 1990 for more details), which in a way would make a difference in the articulatory gestures as well as the acoustic features (i.e., MFCCs here).

The current pilot study shows that certain MFCCs and EMA kinematics can be well represented on the *same* principal component, e.g., TDy and c2, c4, c5 and c7 of the FOOT vowel produced by male speakers, and TDy and c4, c6, c7 of the FOOT vowel produced by female speakers. Further studies are scheduled using more controlled data to explore the strength of association between MFCCs and articulatory data and how does the change of articulatory gestures affect the MFCCs values.

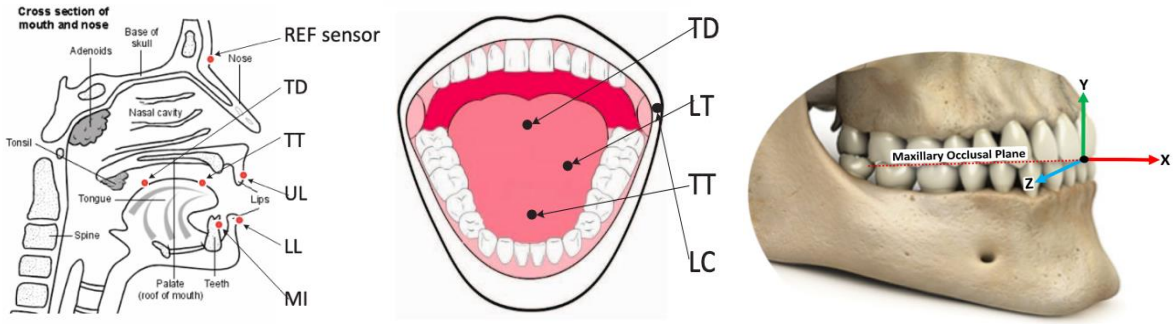

Fig. 1 Leftmost and central panel: sensor placement (Figure 1 from Ji et al. 2014). Rightmost panel: Target anatomically-referenced coordinate system. Positive increases in sensor values denote forward, upward, and rightward movement along x, y, z, respectively (Figure 2 from Berry et al., 2016 EMA-MAE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

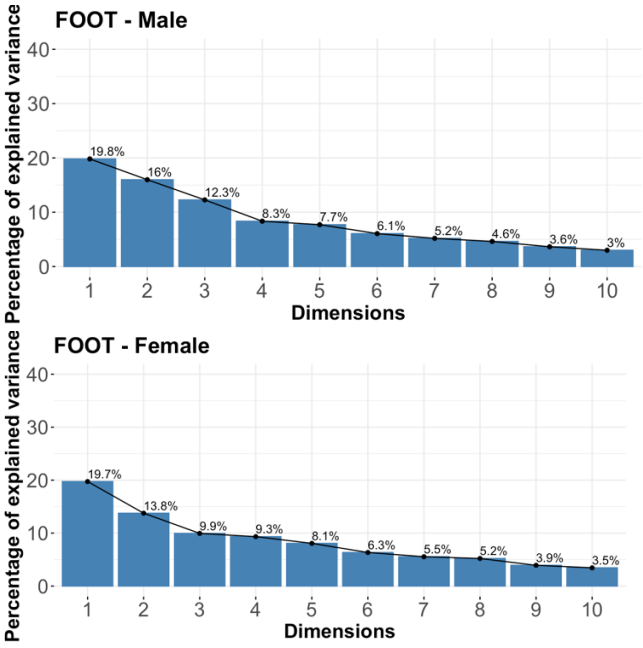

Fig. 2 Scree plots of the FOOT vowel (upper panel: male speakers; bottom panel: female speakers) showing the variance of percentage explained in the first 10 dimensions (i.e., PC1 to PC10).

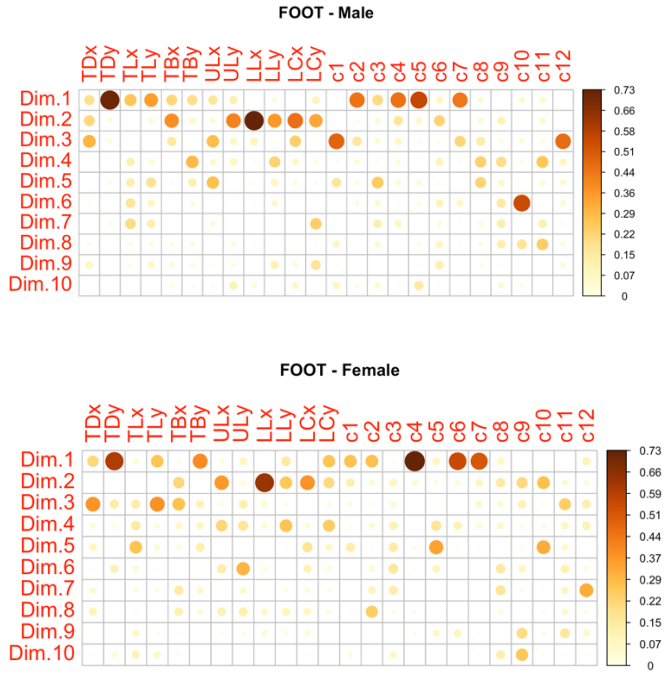

Fig. 3 Squared cosine values shows the relative importance of EMA kinematics and MFCCs of the FOOT vowel represented in the first 10 dimensions.

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., Harrison, P., Foulkes, P., French, P., Kavanagh, C., & Segundo, E. S. (2017). Mapping Across Feature Spaces in Forensic Voice Comparison: The Contribution of Auditory-Based Voice Quality to (Semi-)Automatic System Testing. *Interspeech* 2017, 3892–3896. https://doi.org/10.21437/Interspeech.2017-1508

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.

Kassambara, A. and Mundt, F. (2020) Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R Package Version 1.0.7. https://CRAN.R-project.org/package=factoextra.

Labov, W. (1990). 'The intersection of sex and social class in the course of linguistic change'. Language Variation and Change 2: 205-254.

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.

Prakash, C., & Gangashetty, S. V. (2012). Fourier-Bessel cepstral coefficients for robust speech recognition. 2012 *International Conference on Signal Processing and Communications* (SPCOM), 1–5. https://doi.org/10.1109/SPCOM.2012.6290031

Pellegrini, T., Fontan, L., Mauclair, J., Farinas, J., & Robert, M. (2014). The goodness of pronunciation algorithm applied to disordered speech. *Interspeech* 2014, 1463–1467. https://doi.org/10.21437/Interspeech.2014-357

R Core Team (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Sigg, C. D. & Buhmann, J. M. (2008). "Expectation-Maximization for Sparse and Non-Negative PCA.". In *Proc. 25th International Conference on Machine Learning*. doi:10.1145/1390156.1390277.

Yule, G. (2010). *The Study of Language*. 4th ed. Cambridge: Cambridge University Press.

Zhang, C., Morrison, G. S., Enzinger, E., & Ochoa, F. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison – Female voices. *Speech Communication*, 55(6), 796–813. https://doi.org/10.1016/j.specom.2013.01.011.

WANG Xiao (Bruce)
Ph.D, Postdoctoral researcher at the Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University. His research interests lie in forensic speech science, sociophonetics, acoustic phonetics and Bayesian statistics. He is the corresponding author of this paper.
Email: brucex.wang@polyu.edu.hk

HE Lei
Ph.D, Group Leader at the Department of Computational Linguistics - Phonetics, University of Zurich. His research interests lie in speech acoustics, articulatory kinematics and phonatory and articulatory processes.
Email: lei.he@uzh.ch