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;* [2]*Department of Computational Linguistics - Phonetics, University of Zurich, Switzerland.*

brucex.wang@polyu.edu.hk / lei.he@uzh.ch

'work in progress' poster

BACKGROUND

- Mel-frequency cepstral coefficients (MFCCs; Davis and Mermelstein, 1980)
- Capture the spectral characteristics of speech signal Spectral characteristics are a function of vocal tract (Fant, 1971)
- Claimed that MFCCs capture the shape and features of the human vocal tract
- Widely used as the input features in speech technology (Prakash & Gangashetty, 2012; Zhang et al., 2013; Pellegrini et al., 2014 Hughes et al., 2017)

HOWEVER,

- No studies have attempted to investigate how MFCCs and vocal tract features, if there are any, are related.

AIM

To investigated the interpretability of MFCCs from an articulatory perspective

HOW

- Three corner vowels: FLEECE, TRAP, FOOT
- Single words, i.e., not connected speech
- First 12 MFCCs
- 12 articulatory kinematics:
  - movement of tongue dorsum (TD), tongue lateral (TL), tongue blade (TB), upper lip (UL), lower lip (LL), lateral lip corner (LC)
  - Two dimensions, i.e., x: front and back, y:height,
- Speakers
  - 20 Midwestern standard American English speakers (10 male and 10 female; Ji et al. 2014)
- PCA performed on the first 12 MFCCs as well as 12 articulatory kinematics data (i.e., 6 sensors * x-axis * y-axis)

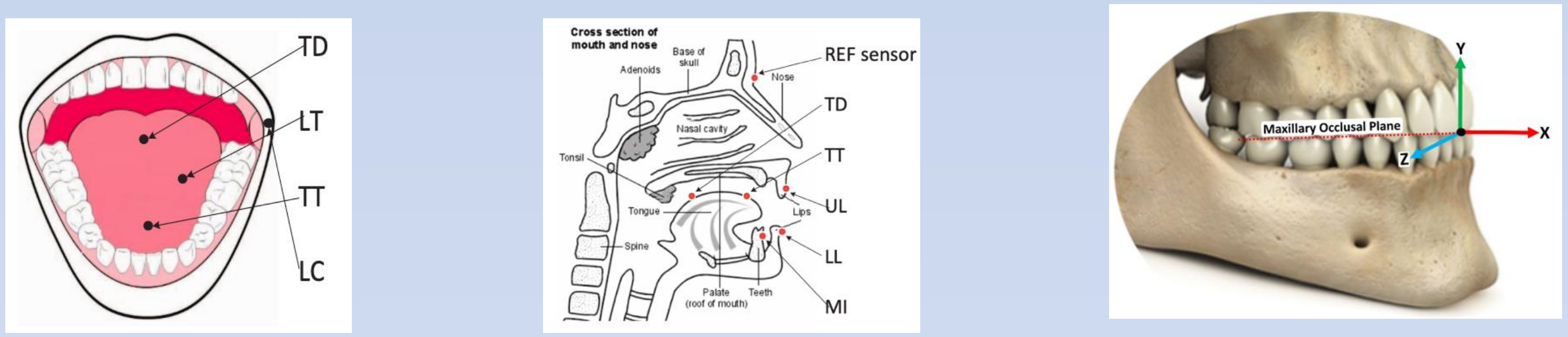
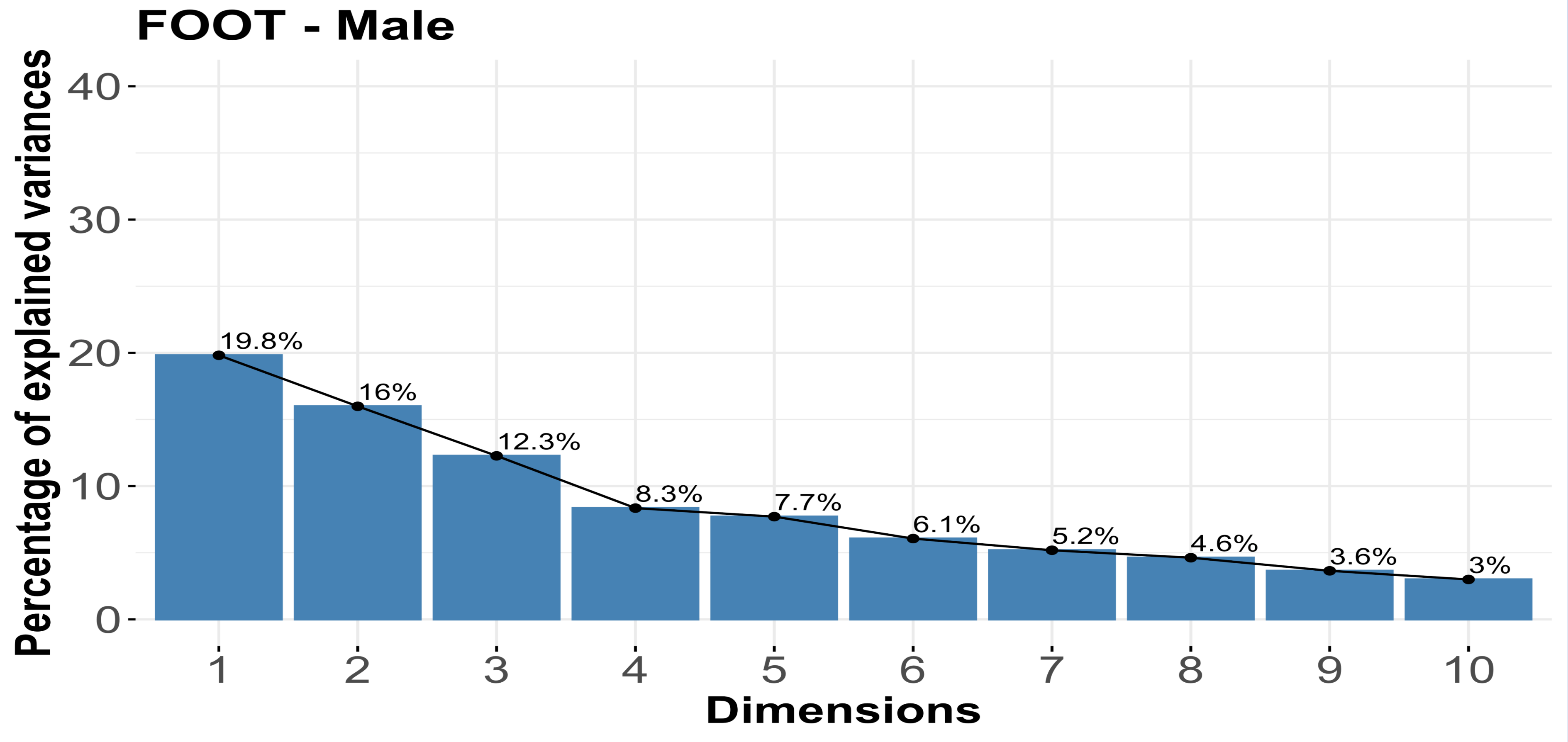
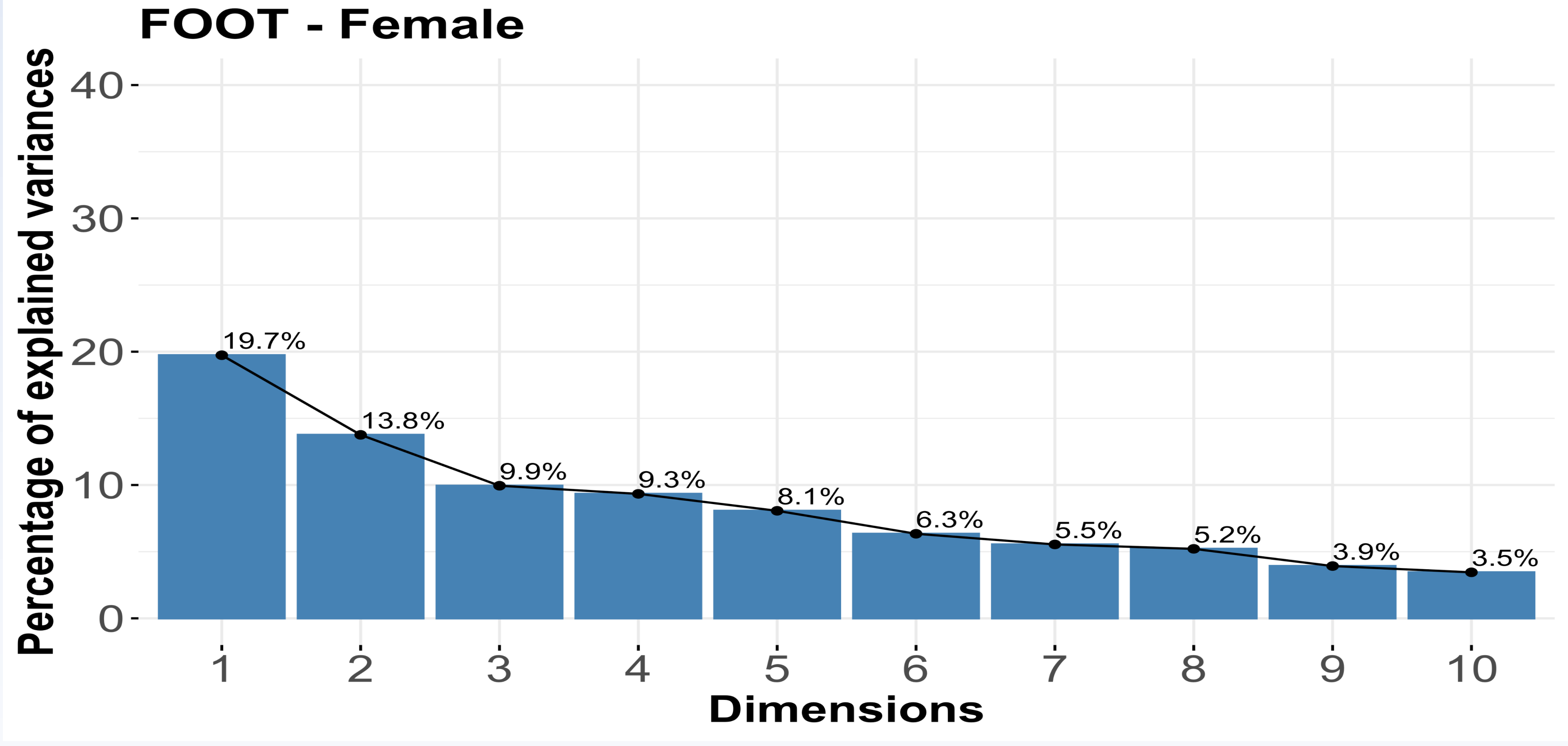
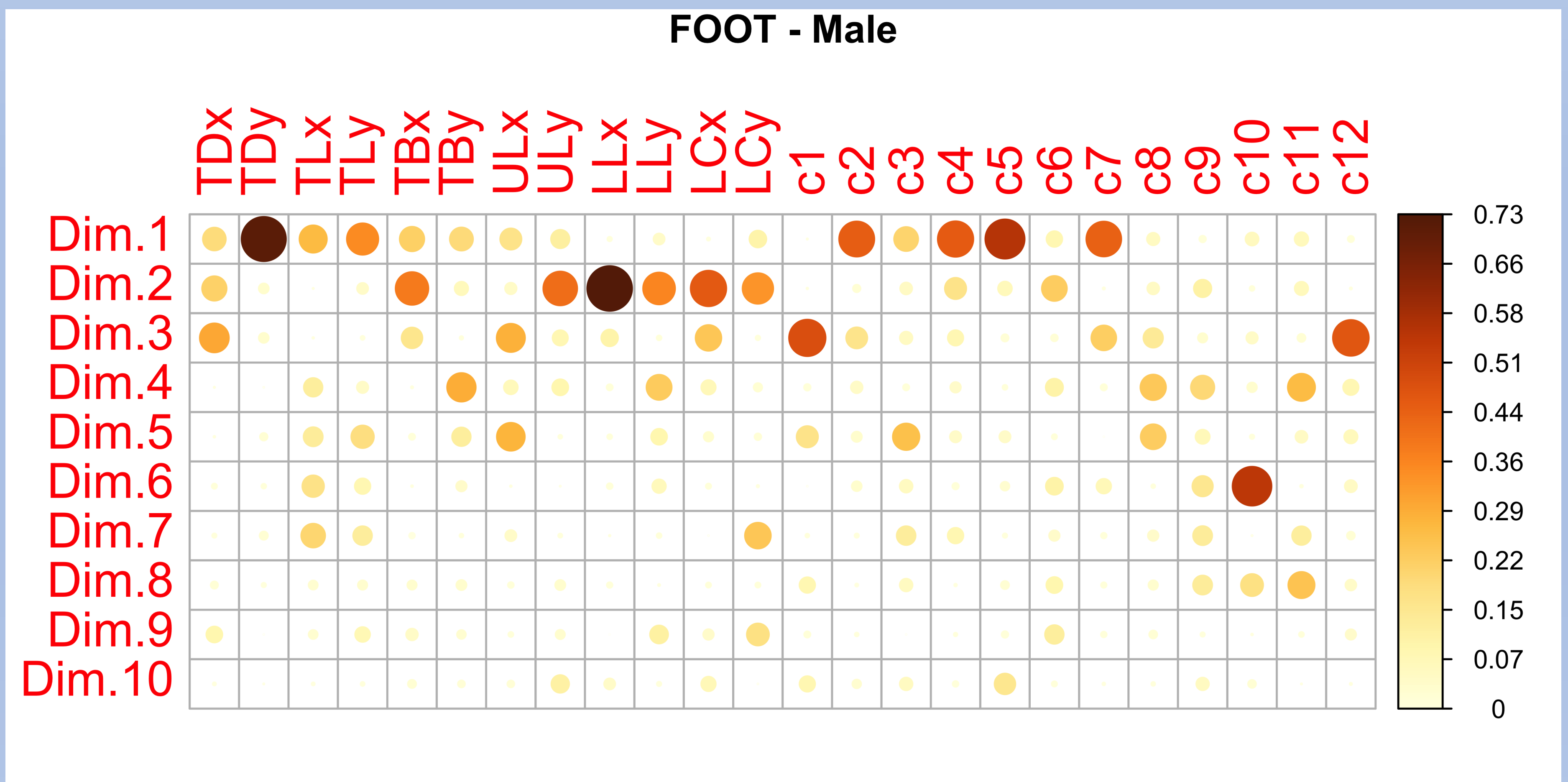


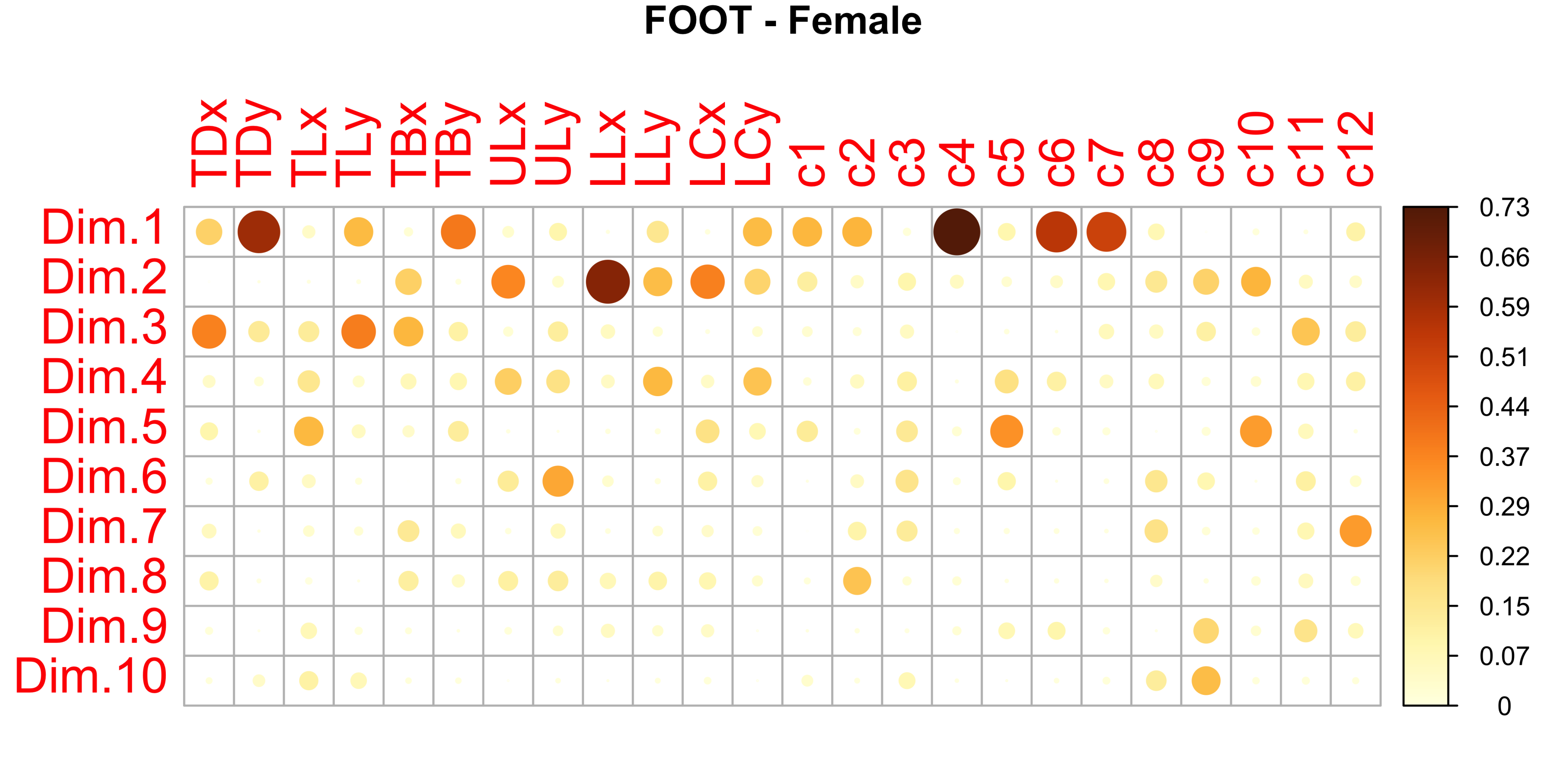
Fig. 1 Leftmost and central panel: sensor placement (Figure 1 from Ji et al. 2014). Rightmost panel: Target anatomically-referenced coordinate system. Positive increases in sensor values denote forward, upward, and rightward movement along x, y, z, respectively (Figure 2 from Berry et al., 2016 EMA-MAE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

SOME RESULTS









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).

Discussion

Figures 1 &2

First six dimensions account for over 60% of variance explained for male (64.13%) and female (60.85%)

Figures 3 & 4

Dim. 1 TDy (vertical movement) well represented for both genders

Dim. 2 LLx (front and back) well represented for both genders

- **These are sensible similarities between male and female speakers from an articulatory perspective.**

Discrepancies in the wellness of the representation of MFCCs

Dim. 1 c2, c4, c5, c7 well represented for male speakers
c4, c6, c7 for female speakers

Sensible discrepancies in MFCC representation between male and female?

- MFCCs capture vocal track characteristics & articulatory movement
- Articulatory kinematics only captures articulatory movement
- Biological factor - male generally have longer vocal track and larger larynxes and thicker vocal folds than females (Yule 2010:275).

Future studies

- How can we take biological factors into consideration?
- More controlled data in terms of participants' height and weight?
- Explore the strength of association between MFCCs and articulatory data
- How does the change of articulatory gestures affect the MFCCs values?

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

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.

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.