Discussion*of The statistical analysis of acoustic phonetic data: exploring differences between spoken Romance languages

Shahin Tavakoli
University of Warwick, Coventry, UK
s.tavakoli@warwick.ac.uk

August 14, 2018

I congratulate the authors for their inspiring paper. I have a couple of comments:

Within and between word sound covariance The paper's finding is that the within word sound covariance seems to be equal for all the word available, and that the within word "marginal covariance operators are promising features to represent phonetic structure at the level of a language". From personal reflection, it seems that the within word sound covariance (or more generally patterns of variations) is an important feature that helps one *distinguish* between the sounds of different words in the same language. However when it comes to what a language *sounds like*, it seems that the between word covariance (or pattern of variation) is an important feature.

Representation of sound signals The 2-dimensional representation of sound signals used by the authors is the spectrogram, which is the log of the squared modulus of a local Fourier transformation of the sound wave $(s(t))_{t=1,2,\ldots,T}$. While spectrograms are valid representations of the sound wave for doing statistics (in particular taking averages of spectrograms, with the view of transforming the resulting spectrograms into sound waves), they suffer from the fact that energy peaks at high frequencies of two sounds of the same word are usually misaligned: taking averages smears out these peaks, and from personal experiments, the resulting sound waves becomes bland. An alternative 2-dimensional representation of the sound waves are given by Mel-frequency cepstral coefficients, or MFCC (see Figure 1 for a definition and Figure 2 for an example). The advantages of working with MFCCs is that they are less prone to smearing the high-frequency energies, while still allowing the MFCC to be transformed back into a sound wave. A drawback of working with MFCCs is that they are harder to interpret. Various implementation of the MFCC exist, Erro et al. (2011) being one of them that allows for high-fidelity speech sound resynthesis.

Registration The registration used in the paper is based on the discrepancy

$$D_\lambda(W_1, W_2, g) = \int_0^\infty \int_0^1 \left[W_1(\omega, g(t)) - W_2(\omega, t)\right]^2 dt d\omega + \lambda \,\text{Penalty}(g),$$

*published in 'The statistical analysis of acoustic phonetic data: exploring differences between spoken Romance languages', Pigoli, D., Hadjipantelis, P., Coleman, J. & Aston, J., *Journal of the Royal Statistical Society (Series C)* 2018, 67(5), pp. 1132-1134

between the spectrograms W_1 and W_2. While the L^2 metric is a natural measure of dissimilarity for registration of curves, the time registration of spectrograms could be done using other choices of dissimilarity measure (e.g. weighted L^2 metric, cross-correlation (Somervuo 2018) or the distance between the relative volume of the two sound waves (Tavakoli et al. 2019)).

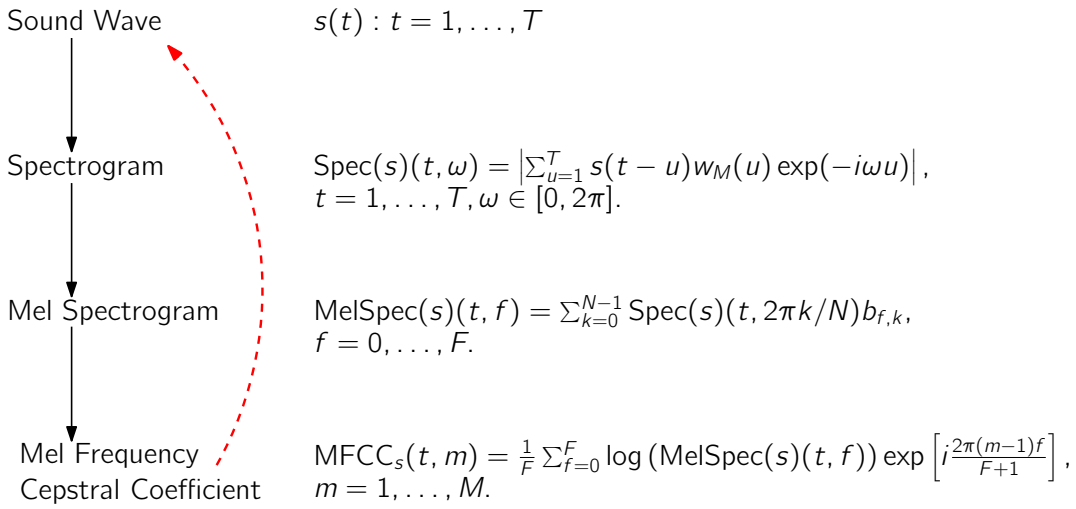

Figure 1: Basic implementation of the MFCC of a sound wave. The $b_{f,k}$s are filter banks.

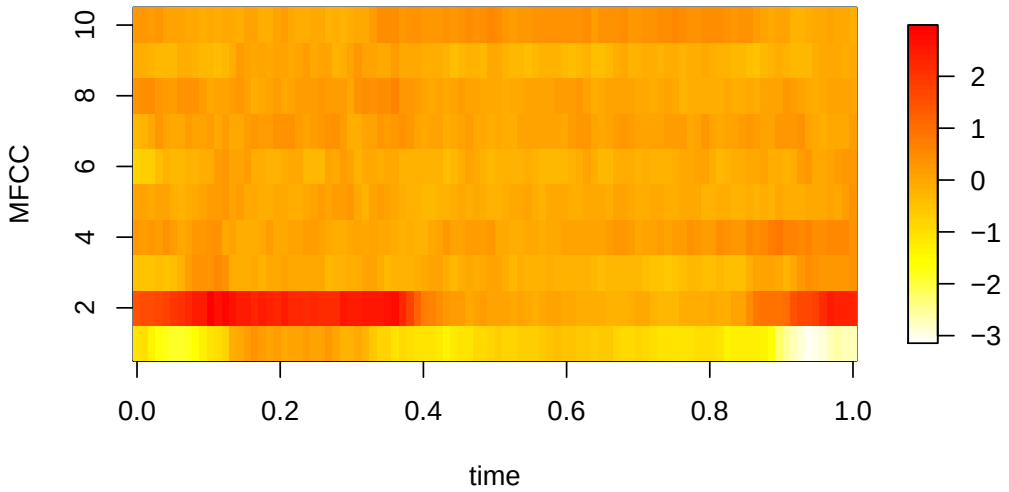

Figure 2: MFCC of a sound of the word "last".

References

Erro, D., Sainz, I., Navas, E. & Hernáez, I. (2011), HNM-based MFCC+F0 extractor applied to statistical speech synthesis, *in* 'ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings', pp. 4728–4731.

Somervuo, P. (2018), 'Time–frequency warping of spectrograms applied to bird sound analyses', *Bioacoustics* **0**(0), 1–12.
URL: *https://doi.org/10.1080/09524622.2018.1431958*

Tavakoli, S., Pigoli, D., Aston, J. A. & Coleman, J. (2019), 'A spatial modeling approach for linguistic object data: Analysing dialect sound variations across great britain (with discussions)', *Journal of the American Statistical Association* **114**(527), 1081–1096.
URL: *https://doi.org/10.1080/01621459.2019.1607357*