On featural bias and consistency in MFCC representations of acoustic distance

Charles Redmon
Department of Language and Linguistics, University of Essex

While much of the research in acoustic phonetics has focused on the study of discrete parameters with a clear articulatory or auditory link to phonological features such as voicing, sibilance, or nasalisation, the use of a more theoretically agnostic parameterisation, such as via Mel-frequency cepstral coefficients (MFCCs) has played a similarly critical role in some areas of speech analysis, such as in automatic speech recognition [2, 6]. This latter use has largely been motivated by practical concerns with an eye toward prediction; however, more recently researchers have begun to use MFCCs in an inferential way as an approximate measure of phonetic distance [see, for example, 1, 3, 4]. However, despite their wider practical use and increasing application in a range of speech science research, there remain critical gaps between our use of the MFCC parameterisation and our understanding of the degree to which it reflects different underlying phonetic or phonological distinctions used in the encoding of the speech signal.

This study uses two large speech databases (1F, 1M) of single word productions from the MALD project [5] to test how different contrasts are reflected in measures of acoustic distance based on MFCC representations of the signal. In particular, 6500 words were selected from each database, which comprise over 95% of the cumulative frequency of the set (based on measurements from several different written and spoken corpora; see [5] for details). Among these items 7542 minimal pair distinctions were identified (here we restrict contrasts to those with the same phonological length) and compared for acoustic distance along several different parameterisations of MFCCs. The HTK parameterisation [6] was utilised as a baseline upon which further testing of the effects of frequency range (upper-limit: 8, 10, 12 kHz), step size (5, 10, 15 ms), and window size (15, 25, 35 ms) was done, as well as the effect of inclusion of *delta* and *delta-delta* coefficients.

All models had the mean log amplitude of the power spectrum included in the coefficient set. MFCC matrices were then aligned using dynamic time warping (DTW), and then compared via the aggregate normalised distance from the DTW alignment. For simplicity, the default parameterisation used in the *dtw* R package (Euclidean distance, fixed edges, symmetric alignment, no windowing) was used for the alignment of signals, though in a future study the impact of different parameterisations of the alignment algorithm will also be tested (alongside wider analysis of how temporal information is encoded; e.g., in hidden Markov model, HMM, or long short-term memory, LSTM, states). In total 50 unique MFCC-DTW alignments were run and analysed.

Two main analyses of the acoustic distance measures according to different signal parameterisations were done: (1) a comparison of the distances measured for different featural distinctions; and (2) a comparison of the relative stability of distance measures by contrast for different parameterisations.

Figure 1 shows the mean and variance (+/- 1σ) of acoustic distance measures for different featural distinctions as derived from the baseline, delta, and delta-delta models. Figure 1 illustrates the expected result that those contrasts that have greater spectral information are better reflected in the MFCC parameterisation than are contrasts that rely to a greater extent on temporal information (e.g., noise duration) or dynamic spectral information (e.g., formant transitions). In terms of general effects of parameterisation, the frequency range has little apparent effect, and even declines slightly at 12 kHz (*norm. dist.* = 37.9) as compared to 8 (*d* = 38.4) and 10 kHz (*d* = 38.3). Window size has a greater effect, predictably showing reduced discriminability at larger sizes (from 39 to 37.5 from 5 to 20 ms), while on the other

hand discriminability increases with larger step sizes (from 37.5 to 39 from 15 to 35 ms). These results are then discussed in the wider context of encoding and acoustic modeling, including relating to alternative approaches, such as self-supervised learning, which have different spectro-temporal properties and constraints.

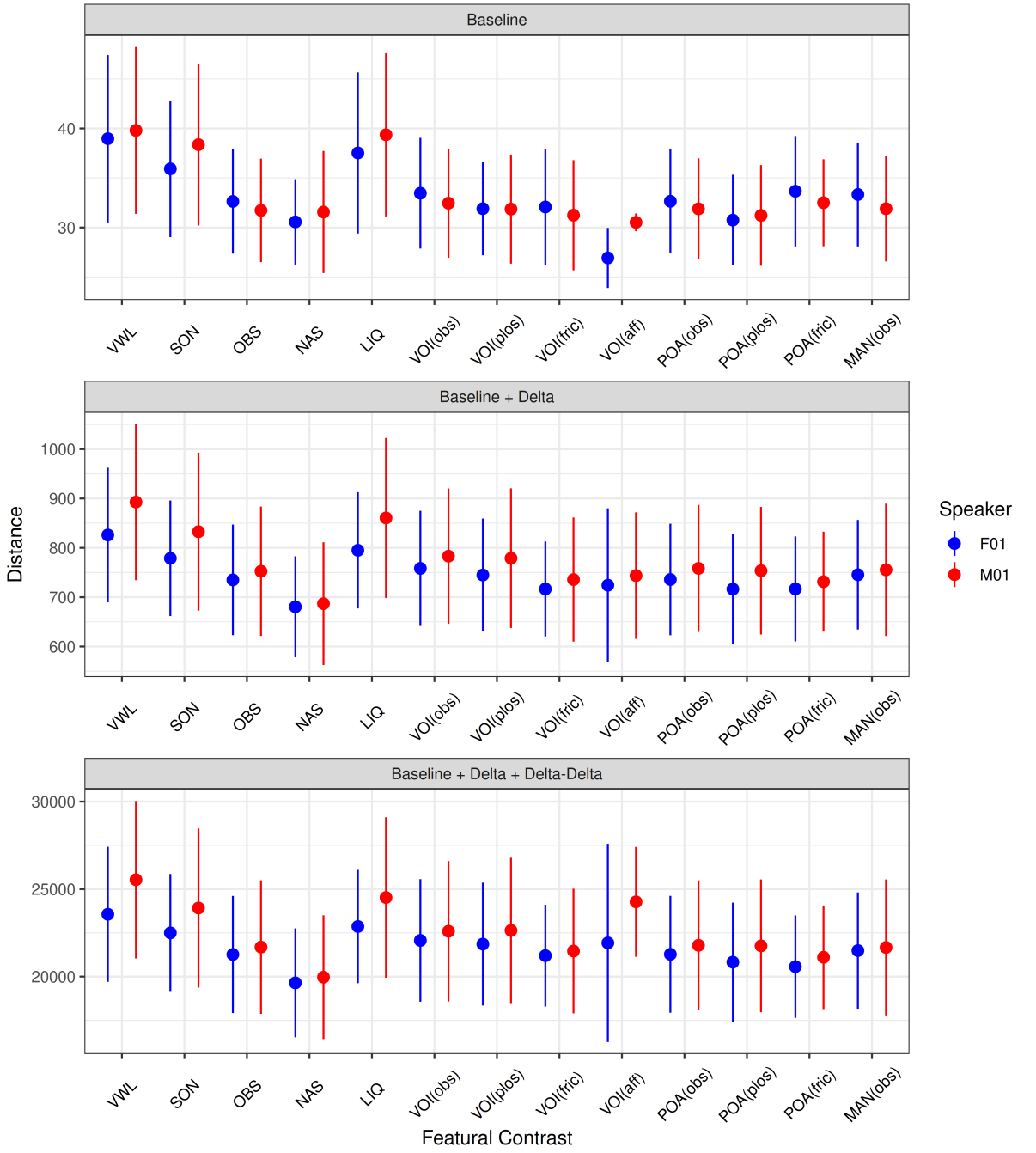

Figure 1. *Normalised distance between minimal pairs of different featural distinctions in the female (F01) and male (M01) lexical databases (point ranges represent +/- 1σ).*

[1] Bartelds, M., Richter, C., Liberman, M., and Wieling, M. (2020). A new acoustic-based pronunciation distance measure. *Front. Artif. Intell.*, *3*(39).

[2] Kelly, A.C., & Gobl, C. (2011). A comparison of mel-frequency cepstral coefficient (MFCC) calculation techniques. *Journal of Computing*, *3*(10), 62-66.

[3] Mielke, J. (2012). A phonetically based metric of sound similarity. *Lingua, 122*, 145-163.

[4] Miodonska, Z., Marcin D. Bugdol, M.D., & Krecichwost, M. Dynamic time warping in phoneme modeling for fast pronunciation error detec*tion. Computers in Biology and Medicine, 69*, 277–285.
[5] Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The massive auditory lexical decision (MALD) database. *Behavior research methods*, *51*, 1187-1204.
[6] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., ... & Woodland, P. (2002). The HTK book. *Cambridge University Engineering Department*, *3*(175), 12.