# Interpreting the Dimensions of Speaker Embedding Space

*Mark Huckvale*
*Speech, Hearing and Phonetic Sciences, University College London, U.K.*

Speaker embeddings are widely used in speaker verification, forensic phonetics and other applications where it is useful to characterise the voice of a speaker with a fixed-length vector [1]. These embeddings tend to be treated as "black box" encodings, and how they relate to conventional acoustic and phonetic dimensions of voices has not been widely studied. In this paper we investigate how a state-of-the-art speaker embedding system represents the acoustic characteristics of speakers as described by conventional acoustic descriptors, age, and gender.

We use a large, gender-balanced dataset of 10,000 speakers from the GLOBE corpus [2], and a state-of-the-art speaker embedding system [3] to create vectors of length 192 to represent each speaker. We also analyze the speakers using a small set of 9 acoustic parameters describing average voice source and filter characteristics, chosen to be "interpretable". See Table 1

*Table 1 - Acoustic parameters and interpretation*

| Interpretable Voice Property | Physical Parameter | Units |
|---|---|---|
| **Pitch height** | FXMEDIAN | st |
| **Pitch range** | FXIQR | st |
| **Irregularity** | PPQ | % |
| **Breathiness** | GNE | 0-1 |
| **Brightness** | SLOPE | dB/kHz |
| **Size** | VTLEN | cm |
| **Loudness** | LEVEL | dB |
| **Intelligibility** | STOI | 0-1 |
| **Signal Quality** | PESQ | 1-5 |

We show first that the embedding space is not well represented by a linear decomposition into basis vectors as afforded by principal components analysis. The lower principal components do not capture much variance in the data and are seen to operate differently for male and female speakers. Instead, we use a non-linear clustering algorithm, UMAP, which seeks to preserve nearest neighbours in the embeddings within fewer dimensions. We find that the quality of the UMAP clustering reaches a plateau after 6 dimensions.

We next correlate our acoustic parameters, together with speaker gender and age to understand how these six dimensions relate to conventional phonetic properties. We fit a linear model that predicts each dimension from the parameters, using a greedy algorithm and cross-validation to find the best combination of parameters, see Table 2. Two of the six dimensions are clearly influenced by speaker gender and well predicted by FXMEDIAN, VTLEN and PPQ parameters.

*Table 2 - Correlations of acoustic parameters with dimensions*

| Dim | Linear model | | Non-Linear | Non-Linear+ Accent |
|---|---|---|---|---|
| # | Best formula | Corr. | Corr. | Corr. |
| **1** | FXMEDIAN+LEVEL+GENDER | 0.499 | 0.579 | 0.660 |
| **2** | GNE+STOI+SLOPE | 0.331 | 0.557 | 0.558 |
| **3** | VTLEN+AGE+FXIQR | 0.209 | 0.275 | 0.616 |
| **4** | FXMEDIAN+PPQ+VTLEN | 0.906 | 0.939 | 0.941 |
| **5** | GNE+LEVEL+SLOPE | 0.314 | 0.466 | 0.555 |
| **6** | VTLEN+FXMEDIAN+GENDER | 0.944 | 0.982 | 0.982 |

Since we should not expect the UMAP dimensions to be linearly related to acoustic parameters, we also perform a non-linear regression for the dimensions from all the parameters (plus gender and age) using an MLP with non-linear hidden units. The cross-validated correlations show small increases over the linear model and are also shown in Table 2. Since the GLOBE corpus also has some accent labels, we can incorporate a one-hot embedding of the accents (put into 8 broad groups) alongside the parameters in the MLP regression, which further improves correlations, particularly for dimension 3.

In summary, we show that the set of average voice characteristics listed in Table 1 do at best a partial job of explaining the most significant dimensions of speaker embedding space. Considerable amounts of variability remain unexplained, of which some appears to be related to accent. Further work might explore how much more of the variance might be explained with features related to spectral dynamics and prosody.

[1] S. Wang, Z. Chen, B. Han, H. Wang, et al. "Advancing speaker embedding learning: Wespeaker toolkit for research and production", *Speech Communication*, 162, 2024.

[2] W. Wang, Y. Song, S. Jha, "GLOBE: A High-quality English Corpus with Global Accents for Zero-shot Speaker Adaptive Text-to-Speech", downloaded from https://arxiv.org/abs/2406.14875.

[3] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in Proc. Interspeech, 2020, pp. 3830–3834.