Development of Quantitative Measure for Speaker Distinction

Roberto Romeu
Cornell University
Spring 2019

Abstract

To empirically define qualitatively observable phenomena, it is necessary to have a quantitative reflection. In the context of speaker identification, this has a dual-necessity for both commercial applications and academic quantification of experimental phenomena and informing phonological theory. Previous work on quantifying speaker identity focuses on speaker classification and verification, using text-dependent methods, Gaussian Mixture Models, and Deep Neural Networks, among other methods. These methods, however, often prioritize non-speech (i.e. text) identifiers, treat identifying features in a black-box fashion, or involve scales of data and data processing that computational complexity becomes a limiting factor in analyses. In this paper, we construct and justify the construction of a profile of a speaker via a highly-multivariate bottle neck neural network algorithm, optimized by a low-level machine learning algorithm. Initially optimized results give correct classification at a rate of 65% using 1200-windows of data from 30 seconds of speech. By construction, this algorithm gives details as to focal points of distinction, runs in linear time with respect to both number of speakers and length of sample, and is only indirectly dependant of vocabulary. The weighting-mechanism optimized by machine learning on big data sets supports the idea of a formant-basis for addressing speaker variation.

Background

Classification and verification tasks can be broadly described as consisting of the following parts:

1. Define data and derived measures upon which to construct a profile
2. Construct a metric space using these measures, optionally maximizing distance between profiles
3. Construct a threshold for acceptance or rejection of a sample to a known profile
4. (Optional) Update the space, measures, profiles, or thresholds to optimize performance

The demands of the first step partially exclude text-dependent speaker classification for commercial contexts, and to a lesser extent academic contexts. While text-dependence invites available methods of authorship identification, which have relatively strong accuracy in both lexical-dependent and -independent methods[1], the learning data set must be relatively large (i.e. one sentence or phrase is usually insufficient), accuracy is lost substantially with more candidate authors, and the context of the text can shift the learned data from the testing data. Even experiments in using Big Data methods of analyzing short text samples[2] fail to have high accuracy in classification-verification tasks. the acoustic information and often operate in more constrained forms of speech given the computational complexity of processing a lexical or phrase datum. Given data under these constraints, TD analyses give high accuracy on short samples and training data, and are especially useful in verification tasks of a specific phrase[3]. The use of Deep or Recurrent Neural Networks (DNN and RNN, respectively) further give a means of feature generation

that potentially extrapolate to less-constrained speech[4], however, this has the constraint that these leaves these methods less robust as DNNs and RNNs in feature extraction can compound noise in deeper or recursive features. Further, NN-derived features are often sufficiently abstracted from high-level observables, that they effectively black-box information that academically informs the phenomenon of speaker diversity. In fact, due to the abstraction of NN-TD analyses is that they are highly susceptible to attack that replaces the text arbitrarily[5], rendering analyses informed by the lexical or phrasal datum invalid.

Text-independent (TI) analyses largely focus on profiling i-vectors of Mel-Frequency Cepstrum Coefficients MFCC, with Gaussian Mixture Models GMM being the preferred form for the computational efficiency and accuracy in profiling non-normal distributions, even with short utterances[6]. Within the context of this paper, the issues with GMMs are that in focusing on an MFCC, one ignores or reduces priority of high-level observables information, e.g. formants, and that using the GMM allows for multi-modal data to be treated as a singular variable, whereas the information on where academic focus should center on defining distinctions requires minor modes to be represented and ordered in priority.

The use of NN on TI verification-classification has improved the existing accuracy of the MFCC-GMM by extracting new features, however, these face similar critiques of NN-derived features in that they can cost robustness, but also black-box information that would lead to high-level phenomena. The use of bottle necks (BN) in DNN allows for potential corrections, but still faces similar constraints as the use has previously focused on BN-DNN on MFCC-GMM[7]. An approach that has similar applications to the aims of this project is the use of a convolutional NN for speaker verification-identification[8], as the strong inter-connectedness of derived features has the potential to give insights of the relationships of high-level observables for speaker distinction, however, this paper's focus lays more in the independent relationships, despite the likely weaker efficacy.

Construction of Model

To justify the model produced, we first characterize the data which we seek to model identification and classification. Drawing from the Buckeye Corpus, we have WAV files that are in turn analyzed with Praat with the To formant (burg) method to generate intensity and F1 to F5 estimates in 25ms non-overlapping time steps. The raw forms, then, of the data is a time-dependent 6-tuple of intensity and the first five formants. The method of formant extraction does not guarantee five formants, so not every time-step has six full measures. For mathematical representation, we have our data be represented by $X(s,t)$, for a speaker s, and a time t, returning the 6-tuple in the natural order following from intensity before formants:

$$X(s,t) : T \to \{x_0, x_1, ..., x_5\} \qquad (1)$$

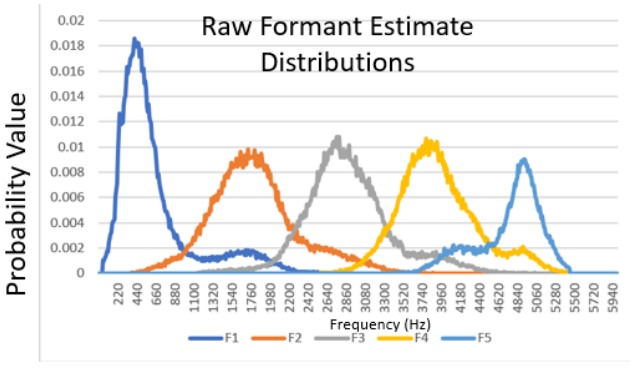


Figure 1: *Unweighted probability function for the formants' estimates' distribution along frequency using 10Hz windows of a speaker. This distribution is over 90,192 potential data points, with the areas of some formants not summing to 1 by the lack of an identifiable formant estimate at some time steps. The number of data points for the formant estimates for this speaker, in order, is 90,192—90,192—90,190—87,689—47,243*

Within Figure 1, we note three features in

particular: there is substantial overlap among formant estimates, and the distributions are partially normal, but have multi-modal characteristics, each formant estimate's distribution seems to be an approximate Euclidean transform of the other formants.

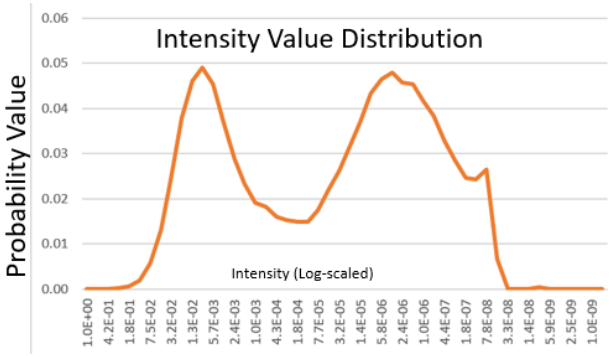


Figure 2: *Probability distribution of intensity along log-scaled A^2 for the same speaker as used in Figure 1*

Figure 2 shows us a clear bi-modal distribution of intensity, which can are likely interpretable as periods of vocalization (especially vowels) and periods of silence or unvoiced consonants. The 50^{th}-percentile falls approximately at 2.0×10^{-5}, meaning that the mode at lower intensity is three-to-four orders-of-magnitude less than the higher one, giving a natural means of identifying when speech is occurring.

The ideal form of this model is a computationally efficient means of creating a distance function on X that robustly minimizes distance for the same speaker. To account for the scale of data and to allow for variation, this model takes summary metrics on X and creates a distance function on these metrics. The summary metrics are addressed as a whole for a speaker by a Spkr class.

We first classify each speaker by the weighted average, weighted standard deviation, and amount of measures on each formant estimate, adding these measures as a 3-tuple to the Spkr class. The weights are the intensity at t, that is $x_0(s,t)$. Recall that the distribution of intensities naturally minimizes influence of non-speech data. More specifically, we calculate the weighted mean and deviation by fixing a speaker s, letting $X_{s,i} \subset T$ correspond to times where $x_i(s,t) \neq \text{None}$ and $M \subset T$ where $x_0(s,t) \neq 0$:

$$\bar{x}_{s,i} := \frac{\sum_{t \in X_{s,i}} x_0(s,t) \cdot x_i(s,t)}{\sum_{t \in M \cap X_{s,i}} x_0(s,t)} \tag{2}$$

$$\sigma_{s,i} := \sqrt{\frac{\sum_{t \in X_{s,i}} x_0(s,t) \cdot (x_i(s,t) - \bar{x}_{s,i})^2}{\frac{|X_{s,i} \cap M| - 1}{|X_{s,i} \cap M|} \sum_{t \in X_{s,i} \cap M} x_0(s,t)}} \tag{3}$$

Combining this with the count of data points from $|X_i|$, we then combine this to the first attribute of Spkr, an array we will call whole:

$$\text{whole} = \begin{bmatrix} \bar{x}_{s,1} & \sigma_{s,1} & |X_{s,1}| \\ \vdots & \vdots & \vdots \\ \bar{x}_{s,5} & \sigma_{s,5} & |X_{s,5}| \end{bmatrix}$$

If we can assume that each speaker has a unique combination of means for each formant, then a distance measure between the means with a variance tolerance would serve well as a metric. On this assumption we introduce the basic intra-metric distance function: the Welch's t-test calculated on a metric j between two speakers $s, \hat{s}$:

$$d_j(s,\hat{s}) = \frac{\bar{x}_{s,j} - \bar{x}_{\hat{s},j}}{\sqrt{\frac{\sigma^2_{s,j}}{|X_{s,j}|} - \frac{\sigma^2_{\hat{s},j}}{|X_{\hat{s},j}|}}} \tag{4}$$

Note the benefits of this method:

- Periods of non-speech are systematically ignored by the weighting of intensity

  This is qualified that some points of speech are given undue weight for being more intense
- Speakers with more variability account for a larger range, and those with less variability account for a smaller range, but with more confidence
- Since number of samples, $|X_{s,j}|$, is accounted for, different size samples are comparable

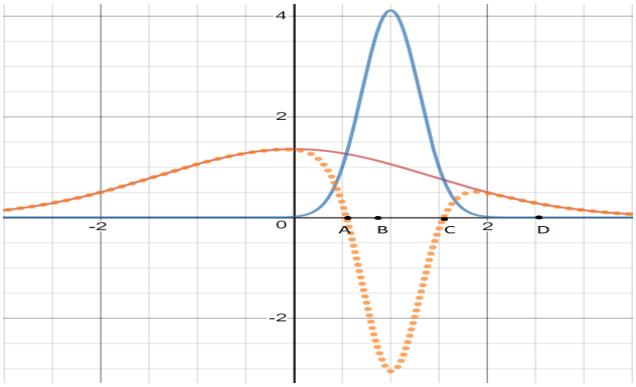


Figure 3: *One-dimensional example of t-test behavior: the solid red line* SRL *representing a distribution centered at* 0 *with a deviation of* 1*, the solid blue line* SBL *representing a distribution centered at* 1 *with a deviation* .5*, and the dotted orange line* DOL *representing the difference of the two. The range where the* DOL *is below the x-axis (range* $[A, C]$*) is the area where a sample, say* B*, would have a minimal t-value when compared to sbl, and above the x-axis, say* D*, being minimal to* SRL*. Note that this differs from intuitions about Euclidean distance, and is in stead a statement of approximate probability.*

To create a total-speaker distance, we must combine the distinct d_j into a single $d(s, \hat{s})$. We should avoid direct summation as we wish to avoid assumption of equal identifying capacity of all formants. In particular, we seek that this metric give some information that can be used to quantitatively differentiate speakers, as such the relationship of the formants to identification is itself a relevant problem. The model adopted is similar to a Gaussian Mixture Model, in that we create weights to each metric based on the metric's capacity to maximize differences. The intuition being that giving priority to where most pairwise differences can be seen will reflect global pairwise differences. We calculate the raw weights by the following, where S is a pool of at least two distinct speakers:

$$w_j(S) = \frac{1}{2} \sum_{s,\hat{s} \in S, s \neq \hat{s}} |d_j(s, \hat{s})| \tag{5}$$

That is we sum each pair's t-distance (the fraction is because the pairs are double counted) to get $w_j(S)$. To create a final weight, we 1-normalize the w_j with J being the set of metrics:

$$\omega_j(S) = \frac{w_j(S)}{\sum_{i \in J} w_i(S)} \tag{6}$$

This leads to the global distance metric of

$$D_S(s, \hat{s}) = \sum_{i \in J} \omega_i(S) |d_i(s, \hat{s})| \tag{7}$$

To classify an unknown sample $\hat{s}$ as a member of a pool of speakers S, we take $\min_{s \in S} D_S(s, \hat{s})$ to be the predicted classification, with a rejection threshold on the minimum value of D_S. That is, after calculating an appropriate rejection threshold r, we say that a sample $\hat{s} \notin S$ when $r < \min_{s \in S} D_S(s, \hat{s})$.

One means of evaluating the model's efficacy is to partition *temporally* the data and treat the time-partitions as speaker samples. This gives a natural means of empirically testing correct classification by creating a confusion matrix and calculating the proportion of guesses along the diagonal. We call this measure the P. The model as described to this point, only gives $P \leq .15$ even when only partitioning the global sample into halves.

The inaccuracy of the model using only the global formant estimate measures is due to the lack of corrective mechanism for overly variable speakers. Referencing Figure 3, we can see that our intuition from the Euclidean distance betrays what the metric is doing. The metric tells us an approximation of the probability that a specific point (more broadly a distribution) is of the same distribution as that to which it is compared. This creates a systemic preference to those speakers which have the greatest global variation, which is amplified by the lesser variation existing in smaller samples. This preference largely accounts for the ineffectiveness of this model due to over 90% of classifications falling to two speakers. p

To correct for the preference for the variability, we implement a partitioning on the

formant estimate values to give a finer picture of the distribution of the data. That is to say we create a new array of metrics for the Spkr class, also producing a mean, deviation, cardinality for the partitions. To define these partitions, we take a number of partitions, p, and create p windows on each x_i on the range $[\min_T x_i, \max_T x_i]$. We then calculate the above measures on each partition. We call this array splits, with notation:

$$\text{splits} = \begin{bmatrix} \bar{x}_{s,(1,1)} & \sigma_{s,(1,1)} & |X_{s,(1,1)}| \\ \vdots & \vdots & \vdots \\ \bar{x}_{s,(1,p)} & \sigma_{s,(1,p)} & |X_{s,(1,p)}| \\ \bar{x}_{s,(2,1)} & \sigma_{s,(2,1)} & |X_{s,(2,1)}| \\ \vdots & \vdots & \vdots \\ \bar{x}_{s,(5,p)} & \sigma_{s,(5,p)} & |X_{s,(5,p)}| \end{bmatrix}$$

Spkr Formants Partitions

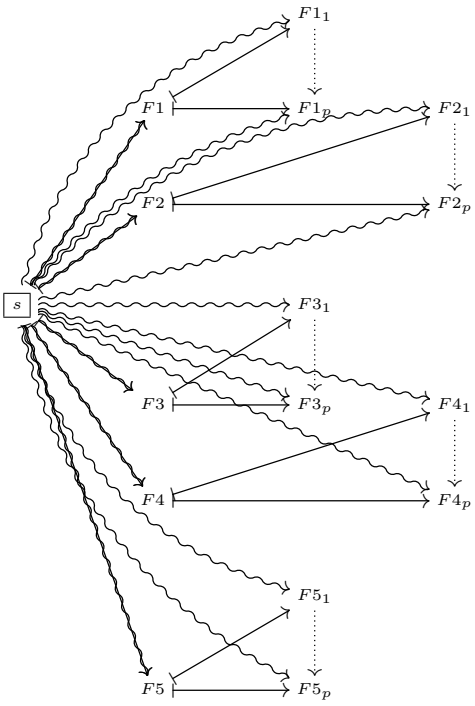


Figure 4: *Matrix representation of the network relationship. A $\mapsto$-arrow indicates that the measure is derived from the source, a $\rightsquigarrow$ indicates that there is a weight assignment to the end point of the arrow for the speaker ($s \rightarrow Fi$ has an overlap appearing as a bolded squiggle), and a $\cdots>$ indicates a continuation down the sequence.*

To implement this into our D_S, we treat each partition as a new metric for the weights, and weigh them just as we did when we only had five formants.

This substantially corrects for the inaccuracy with low temporal partitions, giving $P = .91$ at a 5-time-partitioned analysis on 20 speakers. This correction is attributable to the existence of more measures diminishing the relative power of one more variable measure dictating the D_S, and to the greater resolution to the distributions of speakers by allowing for a multi-modal normal distribution, better reflecting the nature of the data.

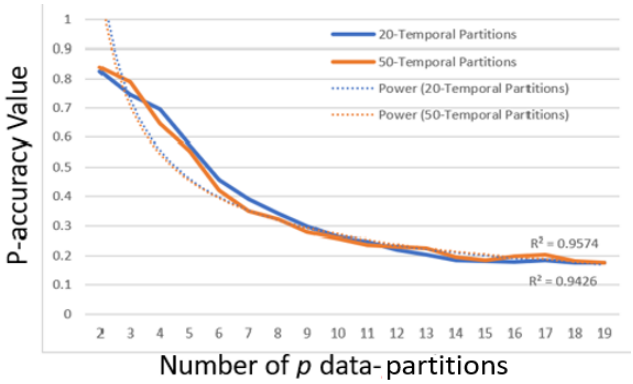


Figure 5: *P-value of correct classification on twenty-speakers, shifting the number of temporal partitions linearly. For scale, 20-temporal partition corresponds to an approximate average of 80 seconds of speech, 50-partitions: 30 seconds. Of note is the near perfect R^2 values when fit to a x^{-1} distribution.*

Despite the increase in accuracy the method only works for low temporal partitions, decreasing in an inverse proportional rate to increases in number of partitions. The 50-measure partition case on a 20-temporal partition (not shown in Figure 5), in particular has a $P = .12$, which is effectively identical to the no-partitions model with samples of approximately 30-45s.

To correct for this, we create a scaling on the partitions such that the minimum of the partition is 0, and the maximum is 1. That

is, where previously $\bar{x}_{(s,(1,1))}$ may have been 75 (corresponding to 100Hz) on the partition from [0,100], we now take $\bar{x}_{(s,(1,1))}$ to be .75 and adjust the deviation accordingly. We adopt the same notation for this modification.

This corrects substantially for the inverse-proportional decrease in P as it places the partitions on more reasonable comparatives. Previously, we saw the decline largely due to the global sample often having outlier maxima and minima that smaller temporal partitions are less likely to have, creating distorted weights towards the maxima and minima. In 1-scaling the data, we account for this better as a relative, rather than absolute, comparative. We can consequentially see a greater preservation of accuracy with $P = .78$ at the 5-measure partition case on the 20-temporal partitions, and $P = .74$ at a 20-measure partition.

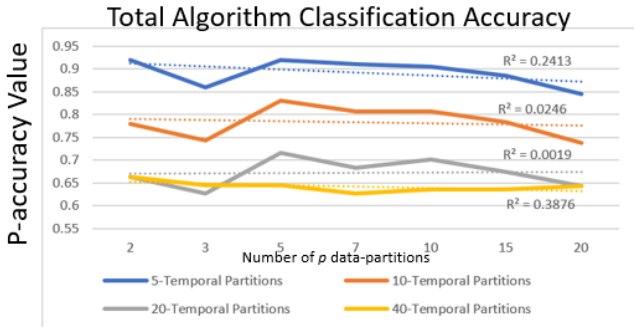
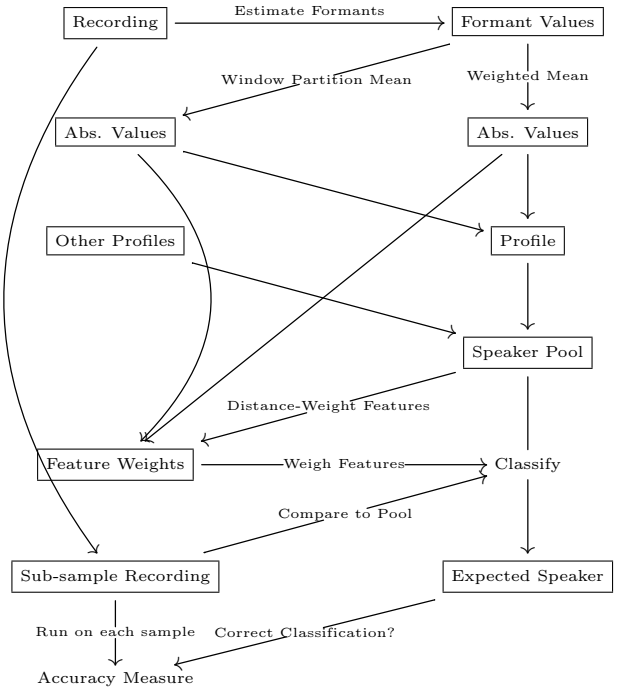


Figure 6: *P-value of correct classification on twenty-speakers, shifting the number of p-partitions quasi-quadratically. Of note is the low R^2 values, suggesting the number of partitions is irrelevant to accuracy.*

We can summarize the process with the following flow-chart:



Discussion

Computational Efficiency

This algorithm utilizes a structure of processing in limited size constraints the most computationally involved processes. With respect to length of input file (ℓ)(WAV-file), converting to an interpretable array runs in $O(\ell)$, linear, time. This is by having fixed-window length FFT produced MFCC and formant estimate generation, meaning that while this process is itself runs at $O((mn + m^2)n \log n)$ (m being the number of desired formant estimates), because both m and n are fixed, this can be treated as having an upper-bound per unit time, thus only making a linear increase in both time and space demands for initial processing.

The calculation of mean, standard deviation, and cardinality are all linear processes, acting on $O(p)$ time and space per speaker (recall p to be our determined quantity of

partitions). Because each speaker is pair-wise compared on $\propto p$-measures, this requires $O(p)$ space for each metric's distance-sum and $O(ps^2)$ time (where s is the number of speakers). This, however, can be constant if one assumes that the weights are universal and generalize the weights from a standard pool.

Classification, being a comparison against each speaker, requires $O(s)$ space, and $O(sp)$ time. This gives a total time for processing each speaker, extracting measures, generating weights and comparisons is $O(s\ell + sp + s^2p + sp) \implies O(s(\ell + sp))$ time, or $O(s(\ell + p))$ time if weights are assumed generalizable. Space constraints in both cases is $O(s(p + \ell))$.

Informing High-Level Phenomena

The most informative details of this algorithm are derived not from the accuracy of the method but from what the optimization tells us. The weighting algorithm, in establishing a relationship of prominence to both absolute measures and relative areas of significance tells us where in a speaker's distribution their most unique information can be found. That is, measures that have the greatest relative weight, by construction, inform the most about where speakers most uniquely deviate from each other. We can see a brief summary of the weights in the 20-speaker pool with varying numbers of p-partitions to the right.

The following are the comparative probability graphs of weights. Maxima reflect areas of higher priority for distinction. The areas don't sum to 1 as each formant's partition weight does not sum to 1. Since the ranges are entirely relativistic, the x-axis is left unmarked as it does not correspond to direct frequencies consistently. Figure 13 is an overlain set of formant weights for the 10-partition weights, for relative scale.

Note that varying relative weights per formant mean that the increasing number of partitions is not a sequentially better approximation of the same continuous limit. That is, these data do not conclusively support the idea of greater accuracy with increasing p. It is also noteworthy the variance in the location of the mode when comparing the formants; it suggests that there is not simply a preference for the center where noise is excluded, but rather some high-level phenomena shifting the modes.

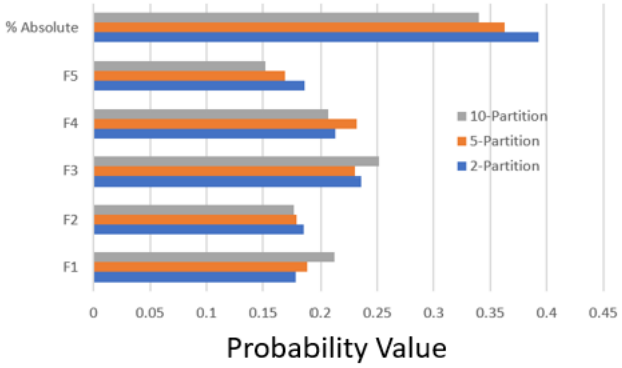


Figure 7: *Graph of weights direct value after running the algorithm on* 2, 5, *and* 10 *partitions. The top trio refers to the sum of weights of the formant estimate values, i.e. the weight to the absolute measures. The lower five trios refer to the sum of weights for all measures corresponding to that formant, that is the summative weight of* FX *and* $FX_{[1,p]}$*. Of note is the gradual decline, but overall preference for the absolute measures, and the preference for the* 3^{rd} *and* 4^{th} *formants.*

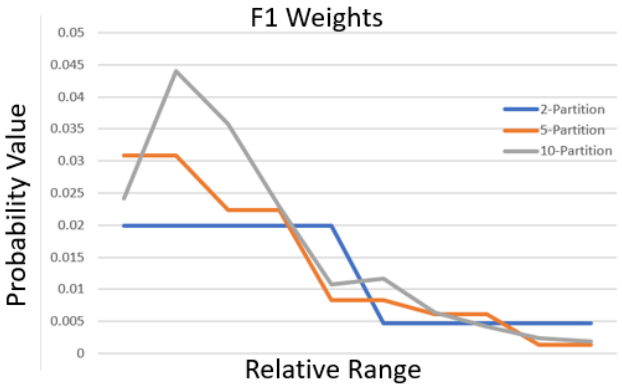


Figure 8: *F1 distributions*

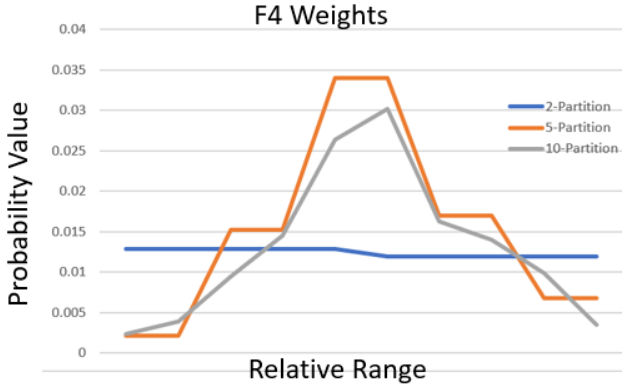


Figure 11: *F4 distributions*

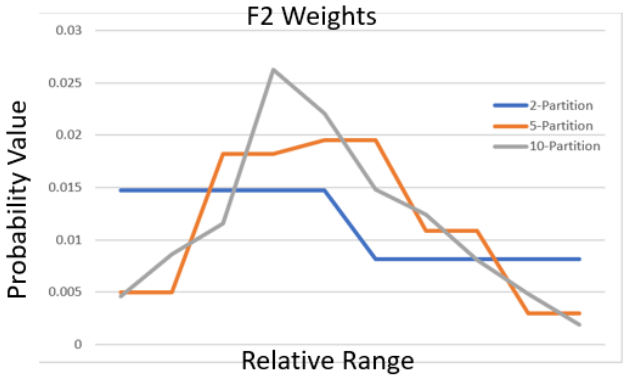


Figure 9: *F2 distributions*

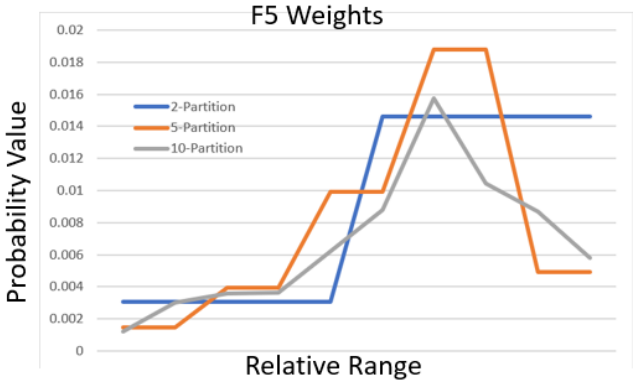


Figure 12: *F5 distributions*

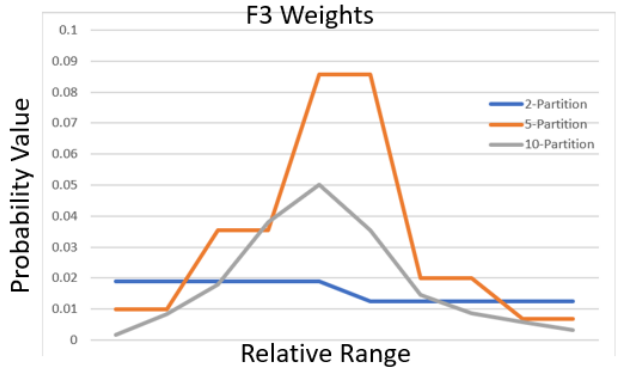


Figure 10: *F3 distributions*

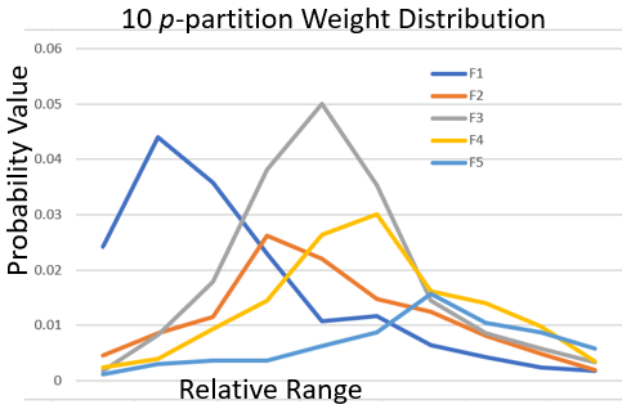


Figure 13: *Overlain distribution weights for 10-partition*

There are some broad claims that these data inform, perhaps best organized by those from F1 & and F2, then those from F3-f5. The marked feature of the first two formants weight-distribution is that it prefers the lower half of the formant estimates' relative range. Referencing Figure 1, the areas of

focus correspond to the modes of F1 and F2, potentially reflecting the capacity to identify vowels using F1-F2. That is, by identifying the vowel distribution of a time segment, one is running an indirect lexical analysis of the speech sample. This, however, is qualified by the comparative lack of weight (less than a total 40%), meaning that the possibility of this vowel-distribution analysis is at best secondary.

The noteworthy aspect is the weight given to the higher formants, especially F3 and F4 over the lower formants. Particularly the near-normal weight-distribution of these formants tell us there is broad information stored in the higher formants. An explanation suggested by the data of Fujisaki and Kawashima (1968)[9] is that higher formants serve as a reference point to understand the speaker variation in the F1 and F2 relationship for distinguishing vowels. That is, the position of the higher formants (and pitch) serve as a basis for understanding the vectors of F1 and F2 for a given vowel. This would have that the higher formants are, in fact, sources of identity in combination with the lower formants. Given that this algorithm's data supports this idea, this begets an area for extended research.

References

[1] E. Stamatatos, N. Fakotakis, and G. Kokkinakis (2001). Computer-Based Authorship Attribution Without Lexical Measures. Computers and the Humanities

[2] Al-Badarmeh, J. et al. (2015) Using Big Data Analytics For Authorship Authentication of Arabic Tweets

[3] Bhattacharya, G. et al. (2016) Deep Neural Network based Text-Dependent Speaker Recognition: Preliminary Results.

[4] Variani, E. et al. (2014) Deep Neural Networks for Small Footprint Text-Dependent Speaker Verification.

[5] Carlini, N. & Wagner, D. (2018) Audio Adversarial Examples: Targeted Attacks on Speech-to-Text

[6] Reynolds, D. & Rose, R. (1995) Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models

[7] Matejka, P. et al. (2016) Analysis of DNN Approaches to Speaker Identification.

[8] Nagrani, A. et al. (2018) VoxCeleb: a large-scale speaker identification dataset.

[9] Fujisaki, H. & Kawashima, T. (1968) The roles of pitch and higher formants in the perception of vowels.

The code for the algorithm can be found by clicking this line.