Talley, J. (2016). What makes a Bostonian sound Bostonian and a Texan sound Texan? In J. Levis, H. Le, I. Lucic, E. Simpson, & S. Vo (Eds). *Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference*, ISSN 2380-9566, Dallas, TX, October 2015 (pp. 168-179). Ames, IA: Iowa State University.

WHAT MAKES A BOSTONIAN SOUND BOSTONIAN AND A TEXAN SOUND TEXAN?

Jim Talley, Linguistic Computing Systems, Austin, TX

This paper introduces a preliminary version of a new methodology for the automated, data-driven discovery of acoustic features of speech which potentially contribute to an accent's distinctiveness. The results discussed herein, while merely illustrative at this stage, provide reason to be optimistic about the prospects of evolving a truly useful and robust automated methodology for cataloging the characteristic acoustic aspects of accented speech. If this line of research were to fully fulfill its promise, the resulting comprehensive catalog of features would contribute to our explicit knowledge of the correlates of accent. The knowledge represented by such a catalog could potentially be directly applied by teachers of second language pronunciation, and it certainly would inform the development of the more capable and individualized computer-assisted pronunciation training (CAPT) tools of the future.

INTRODUCTION

At some level, all of us are aware of accents in speech in our native languages. We distinguish accents (acoustic characteristics due to the provenance or linguistic background of a speaker) from other acoustic idiosyncrasies of a speaker (such as those due to an individual's physical characteristics). With varying degrees of skill, some of us are able to identify the first language (L1) of a non-native speaker or the dialectal region of a native speaker (at least some of the time). This is a fairly difficult task, and harder yet is the task of identifying (especially in real-time) the acoustic features of speech which cause it to be perceived as accented. Yet this challenging task is an implicit requirement for teachers of second language (L2) pronunciation, since knowing what is making students' speech sound accented is a prerequisite for explaining to them how to sound less accented. The task is challenging enough when a teacher is intimately familiar with the L1(s) of his/her students, and still more challenging when the L1s of students are unfamiliar.

A comprehensive catalog of L2 pronunciation issues commonly exhibited by speakers of a specific L1 can be a useful tool for L2 pronunciation teachers, as long as teachers are cognizant of the fact that it merely provides an enumeration of *possible* pronunciation issues to watch for, rather than predicting exactly the pronunciation errors that all speakers with that L1 background will make. Neri, Cucchiarini, Strik & Boves (2002) also highlights the need for such knowledge in computer-assisted pronunciation training (CAPT). Discussing one of the most advanced

CAPT systems (the ISLE project), they say “this approach can only be adopted for specific L1-L2 pairs for which sufficient knowledge of typical pronunciation errors is available” (p.457).

Swan & Smith (2001) represents perhaps the most comprehensive attempt to manually catalog likely L2 pronunciation (and other) issues for a significant range of L1 languages. Derwing & Munro (2015, p.72) takes issue with Swan & Smith's “global prediction of difficulty,” but this criticism is less well founded if (as discussed above) such a catalog is viewed as an overly large set of *possible* L1-sourced issues. Derwing & Munro rightfully emphasize the individual variability with respect to actual pronunciation issues. Individualization is where a CAPT system, based on an extensive catalog of potential issues, could be well-equipped to shine. The (very preliminary) research reported on in this paper takes an approach quite unlike the human expert based cataloging of Swan & Smith. This complementary approach is a bottom-up method, starting from machine-detectable, acoustic features derived from corpora of recorded speech. It uses a newly elaborated, machine learning (ML) based methodology to attempt to automatically create a catalog of the characteristics which distinguish one speaker population from another (*e.g.*, Quechua speakers who are learning English vs. native English speakers). Since this knowledge is automatically derived directly from base recordings, it follows that it would be detectable (and actionable) in a CAPT framework.

The Speech Data

The method discussed below is generally applicable to characterizing accent differences between sub-populations of speakers given a representative corpus of speech upon which to train. The work described herein focuses on learning the distinguishing characteristics of regional dialects of American English, rather than, say, distinguishing characteristics for Malayalam-speaking learners of English, for no better reason than that the necessary type of training data was readily available in the form of the TIMIT database[1].

The TIMIT speech database (Garofolo, et al., 1993) consists of clean (laboratory) recordings from 630 speakers (70% male, 30% female, of varied ages). The speakers were categorized into 7 dialect regions (DRs) – New England, Northern, North Midland, South Midland, Southern, NYC, and Western – based on where they had grown up. It also defines an “Army Brat” pseudo-region for those who lived in multiple DRs during childhood, presumably becoming speakers of Standard American English (SAE). Figure 1, based upon a photo included in Garofolo, et al. (1993), illustrates the 7 geographical TIMIT dialect regions.

We do not necessarily endorse the choice of these DRs as being ideal, neither do we assume homogeneity within each DR – they are simply all that we have to work with. It is some consolation, however, that the TIMIT DRs correspond fairly well with the 6 DRs delineated by Labov, Ash, & Boberg (2006). We should also note that the TIMIT speakers were not selected for, nor evaluated on, the prototypicality of their regional accents, and many may have effectively been speakers of SAE rather than true DR representatives.

1 TIMIT is distributed by the Linguistic Data Consortium (www.ldc.upenn.edu).

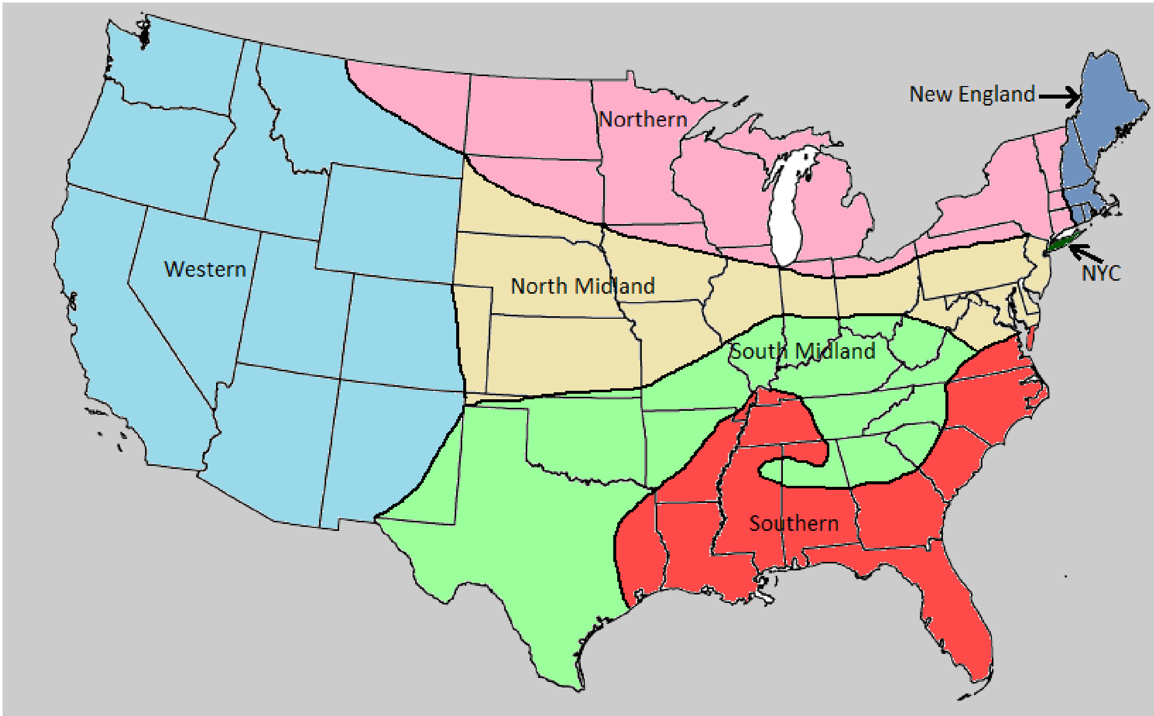

Figure 1. The TIMIT dialect regions (DRs).

There are three classes of recorded sentences in the TIMIT database – Dialect (SA), Compact (SX), and Diverse (SI). The two SA sentences (see Figure 2) were designed to expose differences in regional dialects and were recorded by each of the 630 TIMIT speakers. There were 450 distinct SX sentences, designed for phonetic diversity and coverage. Each of the speakers recorded 5 SX sentences, and each SX sentence was recorded by 7 speakers. The SI sentences were selected for phonetic context diversity from existing sources. Each TIMIT speaker recorded 3 unique SI sentences. For our purposes, we divided the corpus sentences into two groups: the 8 SI & SX sentences from each speaker vs. the 2 SA sentences from each speaker. We used the SI/SX subset for training of the front-end (FE) system, and the SA subset was used for back-end (BE) training and testing.

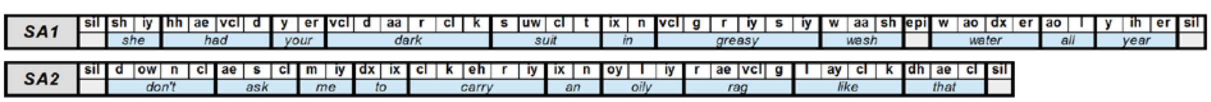

Figure 2. The (canonically transcribed) SA sentences from the TIMIT database.

Research Goals

The desired end goal for this line of research is to comprehensively catalog features of speech which distinguish accents of sub-populations. Such a catalog would serve to inform the development of CAPT tools and could serve as an additional resource for pronunciation training

practitioners. This paper provides an outline of a still underdeveloped methodology for achieving those goals and, as such, it has the modest objectives of 1) providing indications that it has the potential to be developed into an effective technique for meeting the ultimate goal, and 2) exposing the methodology to others in the field for vetting, feedback, and elaboration.

METHODS

This section describes how, starting from raw speech data plus transcriptions, we arrive at ranked lists of features for dialect identification. It has two primary blocks of ML (dubbed the Front-End [FE] and Back-End [BE]) which are connected by a number of data transformation steps (the "Glue").

Front-End Machine Learning

The objective of the FE ML is to learn the models (neural network [NN] and hidden Markov model [HMM]) which enable us to convert from digital speech recordings into temporally segmented frames of descriptive features, with segmentation conforming to phone[2] boundaries (see Figure 3).

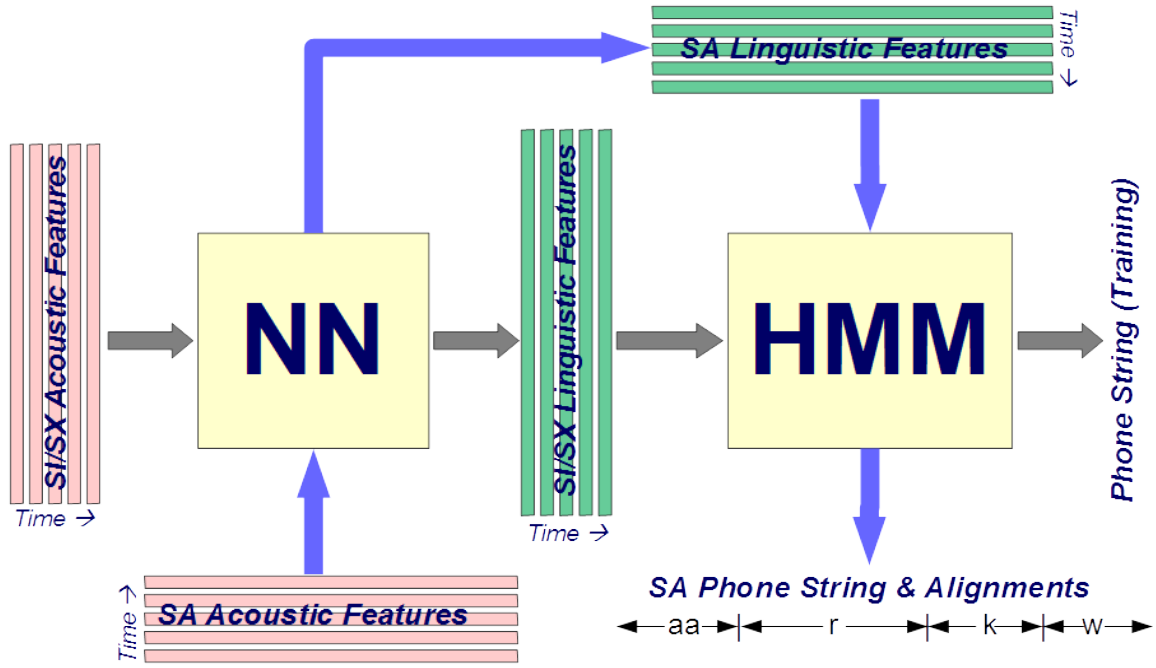

Figure 3. High-level depiction of the system's Front-End (FE) components.

2 The speech recognition community has a long history of using "phone" (or "phoneme") for what would be more accurately called a "phoneme-like, sub-word unit of modeling." Give the unwieldy nature of the more precise expression though, I'll just follow convention and use the "phones" misnomer in this paper.

Horizontally (and with gray arrows), Figure 3 depicts the model training processes, which use the TIMIT database's SI and SX sentences (with the standard train/test split). Input to the NN stage consists of sentential sequences of frames of MFCCs (Mel-frequency cepstral coefficients) – the predominant acoustic featurization used by the automatic speech recognition community. The output of the NN stage is comprised of sequences of frames of fairly standard acoustic/articulatory distinctive features (here labeled “linguistic features” and later referred to as “phonetic features”). The linguistic/phonetic feature sequences from the SI/SX sentences serve as training inputs to the HMM model which is tasked with learning to segment and classify the frame sequences into series of phones with associated start/stop times. Once this training is complete, we have no more use for the SI/SX recordings or their derived featural representations.

The next step in the FE is to put the trained NN and HMM to use, processing the two SA sentences as recorded by each of the TIMIT speakers. This is depicted vertically (and with blue arrows) in Figure 3. We retain the resulting sequences of frames of linguistic/phonetic features corresponding to the speakers' SA utterances to use in the post-FE stages of the method. Signal processing also extracts frame-by-frame pitch and loudness features (not illustrated in Figure 3). We utilize the trained HMM to force align each SA sentence's sequence of linguistic/phonetic frames to its canonical transcription (Figure 2). And then, the resulting phone alignments are used to segment the pitch, loudness, and phonetic feature frames (as illustrated in Figure 4) for the downstream processing.

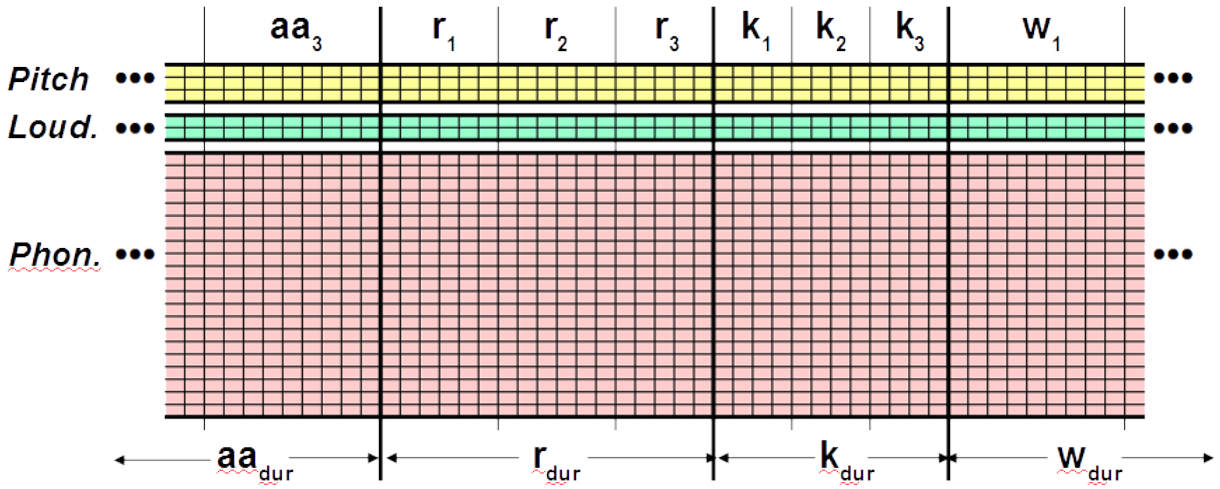

Figure 4. A partial sequence of pitch, loudness, and phonetic feature frames of one utterance, segmented into phones.

Super-Vector Selection, Projection, and Normalization – FE to BE “Glue”

Our first step in getting ready for the BE ML is to transform each SA utterance's segmented sequence of feature vectors (Figure 4) into a single super-vector (SV) for that utterance (one of the rows illustrated in Figure 5). In order to make phone internal (*e.g.*, vowel inherent spectral change (Nearey & Assmann, 1986)) and edge co-articulations available to the subsequent ML,

we temporally split each phone into thirds. Each third of each distinct phone (*e.g.*, the 6 frames corresponding to the center third of the /r/ phone, r_2, in Figure 4 above) is summarized. We, then, form an utterance SV (*Utt_SV*) by concatenating the resulting summarized representations for each phone third's pitch, loudness, and phonetic features plus duration (along with the sentence's global rate of speech, mean loudness, and mean f_0).

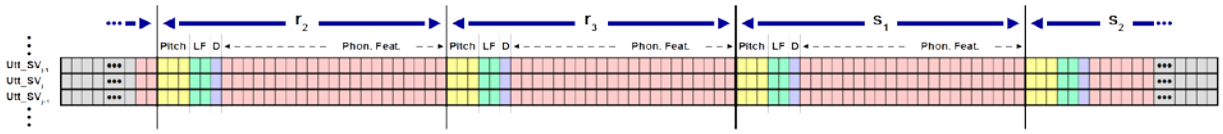

Figure 5. Partial representation of a few SA utterance super-vectors (*Utt_SV*s) derived from the segmented sequences of feature vector frames for those SA utterances.

We could, in theory, go straight to the analysis of the features of these *Utt_SV*s to determine which feature combinations are most useful in distinguishing dialects. However, there is a practical problem which must be addressed – the *Utt_SV*s are very long (on the order of 2000 elements) which makes exploration of the combinations not computationally feasible. The adopted solution to this problem was to define *partitionings* of the *Utt_SV*s into meaningful groups of features. Figure 6 illustrates such a *partitioning*, splitting the phonetic features of an *Utt_SV* into 4 groups – the thirds of the phone /r/ (*i.e.*, r_1, r_2, & r_3) plus a background group of all of the phonetic features which are not from the phone (*i.e.*, ¬ r). We can then define a *partitioning series* which consists of a similar partitioning for each of the phones of a particular SA sentence. Comparisons can be made across the elements of the partitioning series since each phone-specific partitioning partitions the same global set of features from the super-vectors. Such comparisons are the basis of the graphs that we will look at in the *Results* section below.

Within a partitioning (such as that of Figure 6) we exhaustively explore each combination of feature groups, where a particular combination is conveniently represented as a group inclusion bit vector (*e.g.*, 1010 in Figure 6 represents the "not /r/" features plus the features from the middle third of /r/). For each such combination of the partitioning's groups, we select the features from each full Utt_SV_j to create a corresponding selected utterance SV ($Utt\_Sel\_SV_j$) as illustrated in Figure 7. The resulting collection of *Utt_Sel_SV*s represents the particular subset of the features from the original super-vectors which the BE ML will have available to it for training and testing for the specified combination of feature groups (*e.g.*, 1010) within the current partitioning.

Phonetic Features				
¬r	r_3	r_2	r_1	
0	0	0	0	*Baseline1*
0	0	0	1	
0	0	1	0	
0	0	1	1	
0	1	0	0	
0	1	0	1	
0	1	1	0	
0	1	1	1	*Topline2*
1	0	0	0	*Baseline2*
1	0	0	1	
1	0	1	0	
1	0	1	1	
1	1	0	0	
1	1	0	1	
1	1	1	0	
1	1	1	1	*Topline1*

Figure 6. Partitioning.

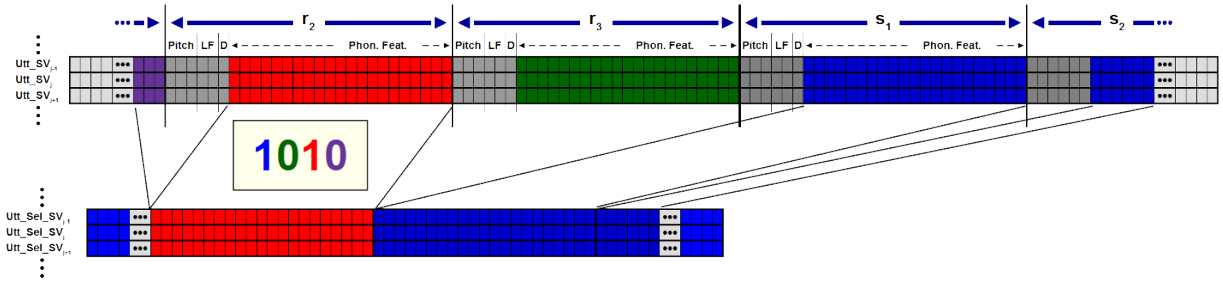

Figure 7. Selection of feature groups for training corpus creation (resulting in the *Utt_Sel_SV*s).

The variation in the number of features selected (the vector dimension of the *Utt_Sel_SV*s) with different group selection specifications (*e.g.*, 0001 vs. 1111) is still problematic with respect to comparison of results across training conditions. For comparability, we would like the models in the BE ML to have the same number of parameters. This final "glue" issue was resolved by calculating an affine transformation matrix, via LDA (linear discriminate analysis), and applying that matrix to project each *Utt_Sel_SV*$_j$, into a very short normalized feature vector, *Utt_Norm_SV*$_j$. By producing such a collection of standard length feature vectors for each group combination within each partitioning across the partitioning series, we create conditions for fair comparisons since each NN training/evaluation in the BE ML employs the same (small) NN architecture (with equivalent modeling power). This is crucial to the feature ranking process.

Back-End Machine Learning and Evaluation

The BE ML consists of training hundreds of multi-layer perceptron (MLP) NNs – one for each feature group combination within each partitioning across the entire partitioning series – and evaluating the classification accuracy for each resulting NN for each of the DRs. For each such training, the original, full super-vectors (the *Utt_SV*s) are selected, projected, and normalized as described above to create corresponding *Utt_Norm_SV*s specific to the desired information subset.

Within each partitioning, the DR-specific accuracies are used to calculate a ranking metric score for each (non-background) feature group of the partitioning. This is done for each dialect region (DR). The process is repeated for all partitionings in the partitioning series, collecting the ranking metric scores into DR-specific tables of scores representing all of the (non-background) feature groups from across the entire partitioning series. Then, because the method has been designed to allow fair comparison across partitionings, we simply take the highest ranked feature groups within each DR table as the salient aspects of that DR's accent. We'll take a look at some examples in the results section below, but first, in order to understand those graphs, we need to briefly examine the ranking metric calculation.

Ranking Metric

The ranking metric is designed to enable fair comparison between the different (non-background) feature groups of a partitioning (*e.g.*, r_1, r_2, and r_3 of Figure 6) and, also, fair comparison of groups from the various partitionings of a partitioning series. It is a weighted average of five indicators of a feature group's importance with respect to identifying a DR. The

graphs of the *Results* section below show each of those five indicators (*db1%+*, *db2%+*, *dt1%-*, *dt2%-*, and *dm%+*) as bars in addition to the final ranking metric (*RMetric*) for each graphed feature group. Each indicator represents a normalized change in accuracy versus distinct references. Though not previously discussed, four of those references are marked on the right side of Figure 6 above. Those references represent feature selections where no information (0000) is used in training (*Baseline1*), only the background ("non-X") features (1000) are used (*Baseline2*), all of the partitioning's information (1111) is available for training (*Topline1*), and all of the features except the background features (0111) are used (*Topline2*).

Indicator calculations are for a given feature group of a partitioning – for example, the first third of the /r/ phone (r_1) of the partitioning of Figure 6 (*i.e.*, the 2nd bit in the feature group combination bit vector, x**1**xx). The *db1%+* and *db2%+* indicators represent deltas which we expect to be positive as we add the (r_1) group's features to a reference which does not include them. The indicator *db1%+* is the normalized increase in classification accuracy obtained when the group's features are added to the *Baseline1* features (0000→0100). The indicator *db2%+* is similar except it is the increase versus the *Baseline2* features (1000→1100).

The indicators *dt1%-* and *dt2%-* are analogous to *db1%+* and *db2%+*, but they represent deltas which we expect to be negative as we remove the (r_1) group's features from a reference which includes them. The indicator *dt1%-* is the normalized decrease in classification accuracy obtained when the group's features are removed from the *Topline1* features (1111→1011). The indicator *dt2%-* is similar except it is the decrease versus the *Topline2* features (0111→0011).

And, the final indicator *dm%+* (the mean normalized delta increase) is the average increase in classification accuracy obtained by adding the (r_1) feature group into each combination of features which do not already include it. In order not to double count the other indicators, combinations which involve *Baseline1/2* or *Topline1/2* are excluded. In our example using the first third of /r/, the delta accuracy increases from 0001→0101, 0010→0110, 1001→1101, and 1010→1110 would be included in the *dm%+* average.

RESULTS

Returning to the titular question regarding what characteristics occurring in a Bostonian's speech make it readily identifiable as being from Boston, and likewise what aspects of a Texan's speech make it identifiably Texan, we'll take a brief look at a couple of example results from applying the above procedure. It should be noted that, though these are real results, they should be regarded merely as selected illustrations of the kinds of results that one might obtain across the board as the methodology is further refined.

The first example ranking (Figure 8) is with respect to speech "from Boston," where we're generously letting TIMIT's New England DR stand in for Boston. This example is drawn from a partitioning series over the phones of the SA1 sentence using only the phonetic features. Each partitioning was into feature groups X_1, X_2, & X_3 (temporal thirds of X) plus ¬X (as discussed above), where X represents an SA1 phone. Furthermore, in this case, the components of the ranking metric for each third of X were averaged (*e.g.*, *db2%+* for /r/ is the mean of the *db2%+* values for r_1, r_2, and r_3), so that each phone is considered as a whole.

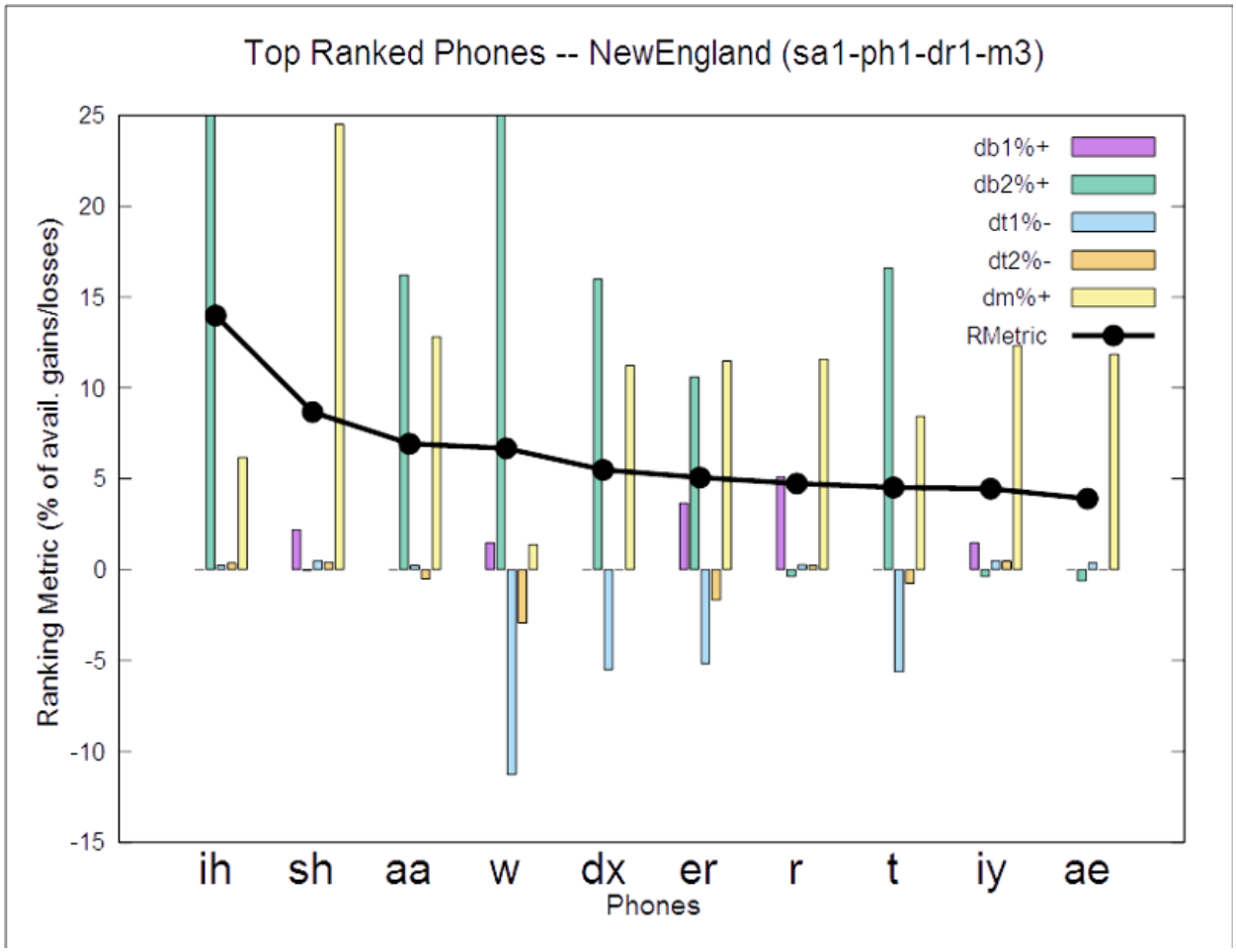

Figure 8. Example "Bostonian" results.

In a prototypical heavy Boston accent, speakers delete /r/s (*e.g.*, "Pahk duh cah in duh yahd."). So a word like "harbor" (with a canonical [ARPAbet] pronunciation of /hh aa r b er/) has a very distinctive Boston pronunciation. Looking at the Top 10 list of significant phones for identifying the New England pronunciation, we see that it includes all of those distinctive (vocalic & rhotic) phones of "harbor" (/aa/, /r/, & /er/). Also, the only occurrence of /ih/ in SA1 is juxtaposed with /er/ in the word "year" (canonically /y ih er/). We see, again, that the expected Boston (non-SAE) pronunciations of /ih/ and /er/ were automatically flagged.

Our second example ranking (Figure 9) is for speech "from Texas," where we're letting TIMIT's Southern DR represent Texas. This example is drawn from a partitioning series over the phones of the SA2 sentence using only phonetic and duration features. Each partitioning was into feature groups X_1, X_2, & X_3 (temporal thirds of X), plus X_{dur} and ¬X, where X represents an SA2 phone. Unlike in the first example, here, the phone's duration and the phone thirds were ranked independently. Figure 9 shows the Top 15 feature groups (from this larger set).

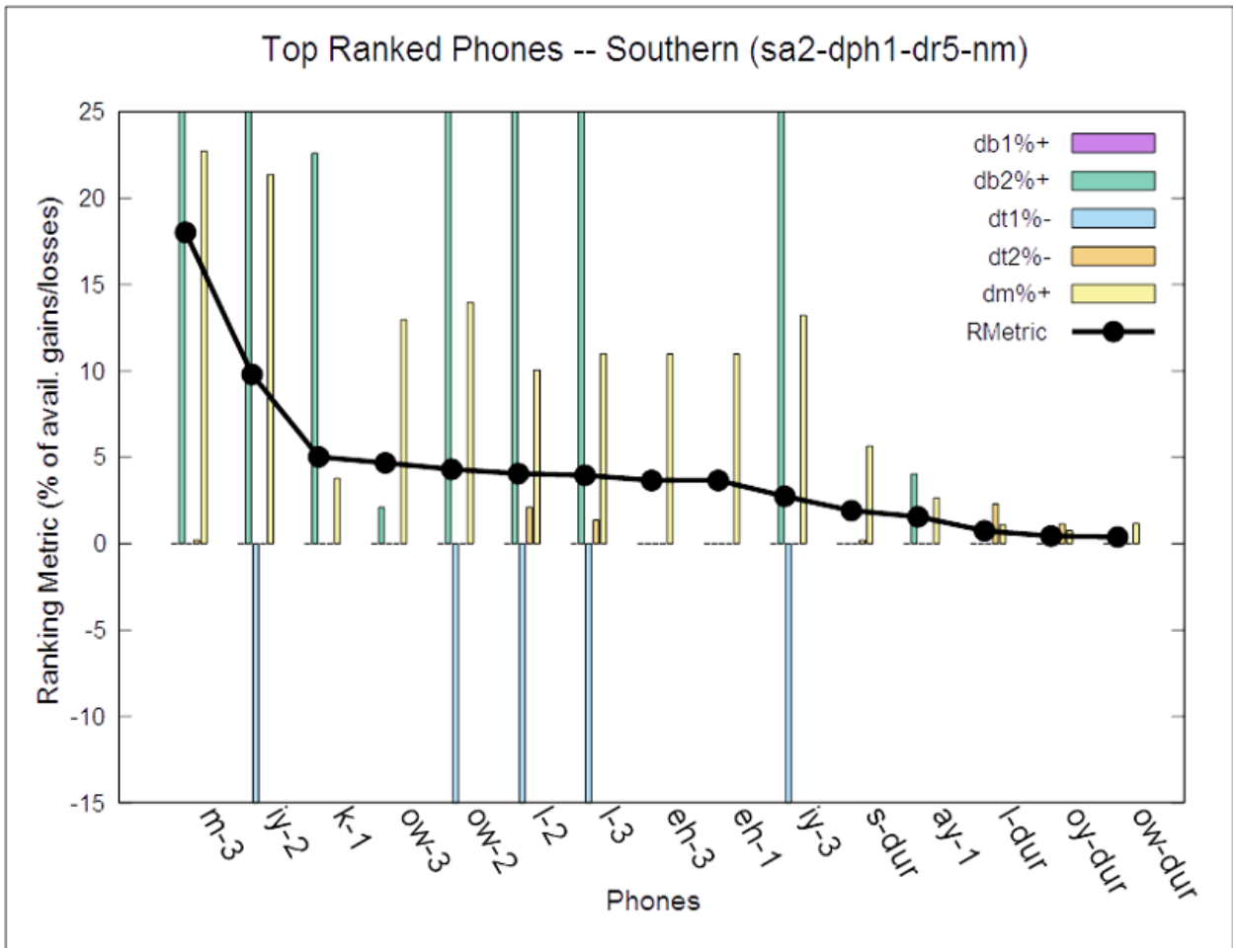

Figure 9. Example "Texan" results.

The Southern accent's vowels are known to differ from SAE in a variety of ways (Allbritten, 2011). It famously reduces the diphthong /ay/ to the monophthong /aa/, and its "drawl" embellishes other vowels with additional movement (*e.g.*, /iy/→/iy ah/ and /ow/→/ow ah/). We would, therefore, expect Southern speech to be distinguishable by its diphthongs, and indeed, all of the diphthongs which actually occur in SA2 (/ow/, /iy/, /oy/, & /ay/) are in our Top 15 list. Southern speech also shifts /eh/ into the space occupied by /ih/ in SAE (*e.g.*, "get"→"git"). We see that the Southern /eh/ is flagged. And, finally, the word "oily" (in SAE, /oy l iy/), which occurs in the SA2 sentence, has a non-standard Southern pronunciation (realized with a dark /l/ and modified diphthongs). All three phones of "oily" show up in the Top 15 distinguishing feature groups list.

DISCUSSION

The work presented in this paper represents a first cut at creating a methodology which ultimately aspires to automatically comprehensively catalog features of speech which distinguish accents of specific sub-populations. Given its provisional nature, we outlined a modest objective for the research at this stage – to provide indications that the method can be further developed into an effective technique to realize those aspirations. The example results, discussed in the

prior section, demonstrate that the method has potential. That said, those results (and others) also show evidence of spurious findings, likely due to overly powerful ML latching onto insignificant statistical regularities within the limited data. While future methodological improvements should reduce the counter-intuitive findings, there will always be some of those with a method such as this – since human and machine learners are, inherently, working on different problems with different tools at their disposal. ML merely extracts statistical regularities, exclusively based upon the limited data made available to it, while humans can't help but bring deep, interconnected knowledge to any task.

A top item on the future elaborations/improvements list is to reduce the method's modeling power. It is apparent that the combination of LDA data projection followed by multi-layer perceptron NNs in the BE allowed spurious idiosyncrasies of the data to produce better than justified accuracies. Replacing LDA (supervised) with principal components analysis (unsupervised) data projection should go a long way towards rectifying this. It would also be germane to experiment with a less powerful BE ML method such as kNN (k-Nearest Neighbor) classifiers, which would have the additional benefits of facilitating an efficient jackknifed evaluation design (also opening the door to elimination of the data projection step altogether). Fuller utilization of the available features (the current results only used the phonetic and duration features) is a priority as well.

This preliminary form of a methodology for the automated, data-driven discovery of accent discriminating acoustic features shows initial promise. Especially with the elaborations suggested above (plus numerous other improvement possibilities), there is reason to be optimistic about the prospects of evolving a viable methodology for creating useful catalogs of the characteristic acoustic aspects of sub-populations' accents. A comprehensive catalog of such (automatically-derivable) features would contribute to our explicit knowledge of the correlates of accent, but perhaps more significant would be its potential to enable the more capable and individualized CAPT tools of the future.

ACKNOWLEDGMENTS

The (open source) Kaldi speech recognition toolkit (Povey, et al., 2011) was invaluable in carrying out this work.

ABOUT THE AUTHOR

Jim Talley is the founder and CTO of Linguistic Computing Systems, an early stage start-up focusing on data-driven approaches to interesting linguistic applications. Prior to starting LingCosms, Jim worked for decades as a research scientist in industry research labs (mostly Motorola Labs [in its various incarnations] and MCC [a pre-competitive research consortium]) on speech and handwriting recognition, predictive analytics, machine learning, and human interface. Jim's graduate education was at UT Austin in Linguistics (Acoustic Phonetics), following Computer Science (and Latin American Area Studies) degrees at the University of Kansas. He was an ESL instructor in the ancient past.

REFERENCES

Derwing, T. M. & Munro, M. J. (2015). *Pronunciation Fundamentals: Evidence-based perspectives for L2 teaching and research* (Vol. 42). John Benjamins Publishing Company.

Allbritten, R. M. (2011). *Sounding Southern: Phonetic features and dialect perceptions*. Unpublished Ph.D. thesis, Georgetown University.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon Technical Report N, 93, 27403.

Labov, W., Ash, S., & Boberg, C. (2005). *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter.

Nearey, T. M. & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *JASA*, 80, 1297-1308.

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. *Computer assisted language learning*, *15*(5), 441-467.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., ... & Vesely, K.(2011). The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1-4.

Swan, B. & Smith, M. (Eds.) (2001). *Learner English: A teacher's guide to interference and other problems*. (2nd ed.) Cambridge Univ. Press.