Phoneme-Agnostic English Accent Classification

Alexander Huras
Systems Design Engineering
University of Waterloo
athuras@uwaterloo.ca

***Abstract*—Speech recognition is hard, partially because speaker-agnostic phoneme detection and classification is hard, which is exacerbated by dramatic individual variations in pronunciation typically attributed to unfamiliar 'accents'. In general, speech accents manifest themselves as particular patterns of phoneme *shifts* in spoken english. Direct approaches to classify a particular voice's accent typically attempt to apply sophisticated phonological models to unstructured spoken english, a relatively expensive proposition. We propose a novel unsupervised acoustic feature extraction technique, that leverages the constrained phone-spectra exhibited through a constructed english language elicitation. The feature's performance in separating the voices of native and non-native english speakers is assessed through application of a linear support vector machine—yielding a 8-fold cross-validated F1-score of 0.78 in an unbalanced data set—which is comparable to cepstral-backed hidden Markov models, indicating that the features perform well. Possible improvements to the technique are discussed, the most prominent of which is the use of more sophisticated feature selection/dimensionality-reduction techniques.**

I. Introduction

SPOKEN language is a highly organic, complex, and nuanced vehicle for communication. It is perhaps unsurprising then, that when we listen to someone speak, we don't cognizantly identify the individual articulated utterances that make up the underlying consonants, vowels, and syllables. Unfortunately, despite what we may perceive, we in fact communicate not in words or ideas, but in a much lower-level medium, that of vocalizations: frequency shifts, and tone patterns—these together form the canonical base units of speech. Each of which is subtly (or in some cases vastly) different, and yet our brains parse the ambiguity into phones, phonemes, and words.

This ambiguity is further compounded by variations in pronunciation among english speakers (the primary subject of this paper), the most consistent and marked of which are the phoneme-shifts we attribute to *accents*. The proper classification of accents is the subject of ongoing research within the computational linguistics, artificial intelligence, and signal processing communities, and has direct applications to fields such as speech recognition, telephony, speech mining, or natural language dialog systems [1]–[3].

The methodology proposed in this paper aims to build an acoustic (rather than phonological) accent detection framework for use either as a speech recognition model tuner, or novel classifier. In particular we explore the performance of a novel high-dimensional acoustic feature-extraction technique in the context of various classification tasks related to english accents.

II. Background

In the design of automatic speech recognition systems, the individual variance of the smallest units of our *vocabulary* (of sorts), either from physical characteristics such as gender or age [4], differing vocal tract traits [5], or higher-level characteristics such as dialect [6] leads to large amounts of ambiguity, leading to error [1].

A. The Problem with Phonemes

The standard approach for modern accent classification is based almost entirely on identifying pronunciation rules based on the phonemes and phonology of a recorded voice sample [2], [7]. This common starting point for most accent/speech recognition systems involves using acoustic information such as the standard Mel Frequency Cepstral Coefficients (MFCC) and occasionally Linear-Predictive Coefficients (LPC) [8], which have been shown to well-describe the tonal characteristics of spoken language [9], and are inexpensive to extract. Furthermore, tonal characteristics (MFCC alone) have been demonstrated to poorly separate speech attributes such as dialect (albeit in relatively small studies) [3].

In contrast, phonemes (and other temporally-associative audio higher-level features) are typically identified by using Hidden Markov Models (HMM) in conjunction with large Gaussian Mixture Models (GMM) [3], [7], [10]. Recent research has demonstrated the effectiveness of modern deep-learning techniques in combination with HMM, and Hierarchical-HMM (HHMM) [2]. In general, phoneme identification systems are quite good, but ultimately require a truly massive amount of computation for training with hilariously large data sets.

While sophisticated, these systems were designed for general speech processing and recognition as applied to arbitrary input under a wide range of recording conditions, an understandably hard problem—for which pure (single-layer) acoustic methods (such as MFCC extraction) are woefully insufficient. However that is not to say that acoustic methods are ineffective under severely constrained conditions (fixed vocabulary, known language, controlled recording environments etc.). While phonological features are the most obvious choice for accent detection (after all, accents are *defined* in terms of phonemes), they are decidedly difficult (computationally and conceptually) to obtain in the absence of massive amounts of data. Our premise is that intangible speech characteristics (such as accent) can be approximated by stateless acoustic information, an approach that in principal is shared by [8], [11] and verified through Cepstral-trained HMM by [3]. However in

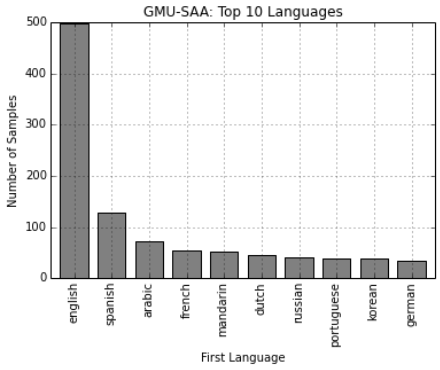

Fig. 1: The top 10 languages by population in the GMU-SAA data set

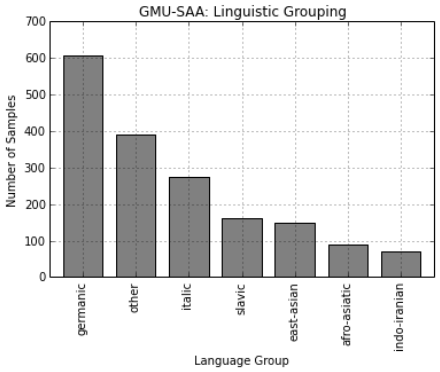

Fig. 2: The major linguistic groups in the GMU-SAA data set, including the catch-all 'other' group.

general, approaches tend to rely on state/memory, of hierarchical features derived from the underlying acoustic information, such as Stochastic Trajectory Modelling by [**?**]. We further propose that acoustic approximations to phonological features can be meaningful under constrained phonological scenario (such as a fixed vocabulary), and that this approach is novel in that it explicitly discards state/memory information in the signal. For context, HMMs tagged with speaker traits and evaluated on cepstral features were shown to differentiate similar speaker attributes (such as dialect) with accuracy of 44% [3] in unbalanced multi-class data sets, and roughly 70% for attributes with similar unbalance as we use within the Speech Accent Archive.

B. The Speech Accent Archive

The data set used for this paper is the GMU Speech Accent Archive (GMU-SAA) which was "established to uniformly exhibit a large set of speech accents from a variety of language backgrounds." [12]. It contains recordings of english speakers (both native and non-native) reading the same paragraph or *elicitation*—a complete transcription of which is available in [12].

Other than containing roughly two-thousand high-quality recordings of english speech, the primary allure of the GMU-SAA is that the subjects are all recorded speaking the same carefully constructed elicitation—which was designed to exhibit almost all of the tonal and phonological attributes found in standard American English. From a computational standpoint, this underlying phonological similarity (between recordings) reduces some of the perceived problems associated with context-less acoustic models, and thus serves to greatly constrain the accent-detection problem.

The general demographic representation of the GMU-SAA are displayed in Figs. 1 and 2, which serve to outline how each language (or linguistic proto-language) is represented within the data set, which ultimately contains recordings from 1200 non-native and 500 native english speakers. The extent of the class frequency imbalance is illustrated most evidently in Fig. 2, which illustrates that germanic languages dominate the class landscape.

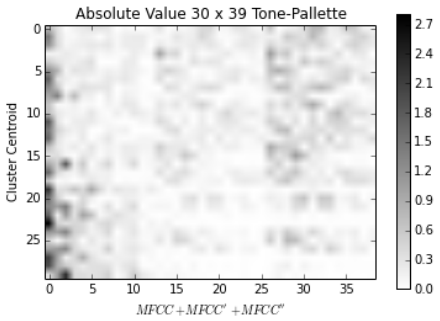

Fig. 3: Example tone palette with strictly positive elements (via. absolute value), k = 30. Note the vertical spectral banding associated with the derivative components

III. Phoneme-Agnostic Accent Classification

The novelty of our approach stems from the explicit avoidance of temporal relationships in the underlying audio stream. Through the use of unsupervised feature extraction, the phonological spectra is approximated through MFCC clustering, which is then used to train a support vector machine.

A. Unsupervised Feature Extraction

Firstly 13 (number 1 through 13, omitting the 0-th order) MFCCs associated with 4 ms audio frames are calculated yielding a time series of MFCC vectors. Then, the first, and second time derivatives of the MFCC stream are concatenated to it, yielding a stream of 39-dimensional vectors. It should be noted that this is accomplished by convolving the MFCC stream with a simple *delta* (edge-detector) filter, once for each derivative. Using a form of histogram *liftering*, the vectors exhibiting the lowest 10-percent (by euclidean norm of the vector) of magnitude are removed from the stream, which approximately removes 'dead air', from the readings—implicitly reducing the inter-speaker variability associated with reading speed, and the number of patterns to cluster. While artifacts such as pauses and reading speed may prove to be informationally dense (relative to the other acoustic features), they were explicitly omitted to maintain

The feature stream is then clustered using k-means into k clusters . The patterns within the feature-space are then

labelled according to their nearest centroid—contributing to the centroid's *population*. The cluster centroids are then sorted in ascending order in terms of their *populations*, and then concatenated together, forming a (in this case $k \times 39$-dimensional) pattern we refer to as a ''tone-palette'', an example of which (with k=40) is shown in Fig. 3.

The ordering of the centroids, represents the most significant aspect of the process, as it is ultimately this step which defines how different vectors within each tone-palette are compared (i.e. for classification). The placement of a given vector within the tone-palette thus indicates the extent that a particular speaker's voice exhibits a particular cepstral *moment*. Inter-palette comparisons are thus analogous to comparing the relative extent that a particular voice exhibits said *moment*, as well as the cepstral energy associated with that moment. In theory, this should simultaneously differentiate tonal variation (i.e. voices with different pitch), as well as (due to the derivative terms) cepstral shifts (that are likely analogous to phoneme boundaries).

B. Supervised Learning

Classification/training labels are provided (indirectly) by the GMU-SAA data set, which contains demographic information such as native language, years of experience speaking English, and which country the speaker is from. Unfortunately, the GMU-SAA labels do not identify the accent directly, this is particularly problematic for determining (for example) which accent is being spoken by any member of primarily english-speaking countries (i.e. USA, UK, Australia etc.), forcing us to use some proxy for determining the accent of the speaker. This ultimately leads to either severely unbalanced classes for multi-class classification, or very fairly superficial binary classification (i.e. those who's first language is english vs. the complement).

Within the context of this paper, we demonstrate the performance of high-dimensional *tone-palettes* as sparse acoustic features through the use of linear support vector classifiers.

C. Validation

For both experiments the classifiers were selected based on the highest weighted accuracy metric (F1-score) through stratified k-fold cross validation using exhaustive grid-search as described in [13], [14] using the Scikit-Learn python package ([15]), with $k = 8$ (12% of each class was withheld for testing), which draws training and testing *folds* such that the within-fold class distribution is representative of the entire training set's class distribution. The stratification was used as the underlying classes are relatively unbalanced, and the data set was not large enough to effectively subsample the larger (non-native english speaking) class.

Given that the classes are unbalanced, the standard accuracy score (total correct positives over total samples) doesn't effectively represent the performance of the classifier (optimizing hyperparameters with respect to overall accuracy heavily biases the classifier against less populous classes). It is for this reason that we use the F1-score (harmonic mean of model precision and recall) for evaluating the cross-validated model performance.

Accent Group	Precision	Recall	F1-Score	Support
non-native	0.80	0.79	0.80	184
native	0.52	0.53	0.52	78
avg / total	0.71	0.71	0.71	262

TABLE I: Classification report for Naive Linear SVM application to 1170-D tone-palettes

Accent Group	Precision	Recall	F1-Score	Support
non-native	0.89	0.77	0.83	184
native	0.59	0.77	0.67	78
avg / total	0.80	0.77	0.78	262

TABLE II: Classification report for application of Linear SVM to the first 50-principal components of the training tone-palettes.

IV. Classification Performance

When partitioned based on whether a speaker's first-language is english, both classes are well represented, but unbalanced in favor of the non-native (English as a non-primary language: colloquially ESL) english speakers. Throughout this section, the english speaker classes are referred to as 'Accent Groups', and refer to the differentiation between native, and non-native speakers.

Of additional note is the extensive appearance of the *linear* support vector machine. While it is of course beneficial to use a simple classifier to illustrate the performance of a particular feature extraction technique, we were surprised by the extent to which the linear-kernel SVM outperformed the more traditional/contemporary radial-basis-function, or polynomial kernel machines. Throughout tangential experiments, non-linear SVMs were the most biased towards the dominant class, and thus performed poorly under the F1-score for this task.

A. Naive binary classification of accent groups

The first pass at classification involved training a linear support vector machine on the tone-palettes directly. Through a small-scale grid-search, the parameter $C = 1.2$ was found to yield an F1-score of 0.71—not outstanding. The naive classification results are shown in Fig. I, and illustrate not only that the model didn't perform exceptionally well, but that the unbalanced nature of the classes drove the accuracy (across the board) in favor of the more populous class. This is demonstrated unequivocally by the near 50% scores for the native speaking group (which also has dramatically less support in the data set). Furthermore, due to the high-dimensionality of tone-palettes, it is likely that there was insufficient data to properly train the SVM.

B. Dimensionality Reduction

Given the relatively poor native-english classification, principal component analysis (of the training data) was conducted to achieve higher performance—both in terms of the overall F1-score, but particularly to increase the F1-score for the native english class. The resulting classifier (still linear, with $C = 1.2$), achieved incremental improvement in total F1-performance (up to 78%), but more importantly, drew those

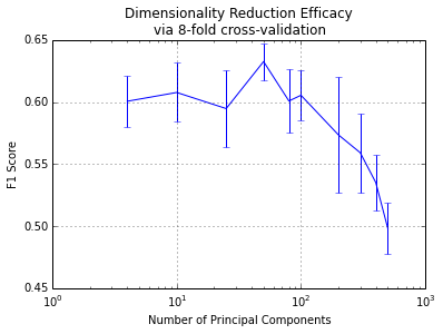

Fig. 4: F1-Score of Linear SVM in binary classification of native/non-native english as a function of the number of Principle Components.

improvements most significantly from better differentiating the native class (naturally at the expense of non-native recall performance).

This performance improvement indicates that the tone-palette (at least for this particular classification task) is rather sparse. The extent to which the selectivity of a linear SVM (with $C = 1.2$, and sample weights inversely proportional to class frequency), varies with the number of principal components selected from the tone-pallettes is shown in Fig. 4.

V. Conclusions and Future Work

Ultimately, a lot relies on effectively reducing the sparsity of the underlying tone-palette, while still retaining nuanced information from the underlying acoustic stream. However given the relatively small effort placed on optimizing the SVM classifier (using raw features, or naively dimensionally reduced features) we believe the tone-palette may still have utility for classification tasks involving audio–particularly those involving speech.

A. Tone-palettes as acoustic features

Through the proxy application of relatively simple (linear kernel, suboptimal) support vector classification, the tone-palette exhibits mediocre performance. That said, as a stateless, purely acoustic feature generated via truly unsupervised methods, it performs surprisingly well, achieving raw-classification performance (in a small data set relative to the dimensionality) of $F1 = 0.71$. Through the use of principle component analysis, the sparsity of the features can be reduced dramatically (where 50-components represents 94% compression), which dramatically increases the classification performance to $F1 = 0.78$, while also yielding a relatively balanced classifier from a dramatically *unbalanced* data set.

B. On the validity of SVM tuning and model selection

Throughout this paper we have discussed the various merits and downsides associated with using the tone-palette as a high-dimensional acoustic feature; specifically within the context of aiding support vector classification. It is highly probable that given more sophisticated feature selection/search techniques, such as FS/SFS discussed in [16] or SVM suboptimization as discussed in [17], the resulting F1-score could be dramatically improved. Additionally, more sophisticated techniques could have been used to determine the optimal regularization, or gain parameters within the SVM model itself [13].

Additionally, while we chose to pursue PCA to reduce the dimensionality of the tone-palette (see Fig. 4, and Table II), it has been suggested that similar (or better) performance could be achieved through slightly different (still algebraic) means, such as LDA [18]. Along a similar course, we never performed feature selection (outside of the original unsupervised feature extraction—which could be interpreted as feature selection), such methods (i.e. through application of ensemble methods) are likely to further increase the performance, and the use of accent-specific trees (admittedly for supervised feature-extraction and model-selection) in [19] show promise.

C. Improving the Tone-Palette

Given that the tone-palettes are high-dimensional and sparse, in general a lot of samples are required for sufficient training (not only for SVMs but for artificial neural network techniques as well). On a fundamental level, this is particularly evident in the high-frequency regions, by excluding these the number of dimensions could be notably reduced with minimal loss of fidelity. The tone-palette is a first pass approach at selecting the *best* acoustic *moments* (centroids) to represent the underlying vocal aspect of a recording, and as such there are many possible improvements and extensions to the feature generation itself. The most obvious of which is the introduction of an explicit signal preprocessing phase to remove noise—while the recordings used for feature extraction were high-quality, there was no process applied to remove microphone noise, or handle dynamic signal compression (i.e. for volume normalization), doubtless this could improve the feature fidelity slightly.

References

[1] M. Yusnita, M. Paulraj, S. Yaacob, and A. Shahriman, "Classification of speaker accent using hybrid dwt-lpc features and k-nearest neighbors in ethnically diverse malaysian english," in *Computer Applications and Industrial Electronics (ISCAIE), 2012 IEEE Symposium on*, Dec 2012, pp. 179–184.

[2] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[3] I. Shafran, M. Riley, and M. Mohri, "Voice signatures," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. IEEE, pp. 31–36.

[4] S. Deshpande, S. Chikkerur, and V. Govindaraju, "Accent classification in speech," in *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on*, Oct 2005, pp. 139–143.

[5] P. Cook, "Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing," Master's thesis, Stanford University, Stanford, California, 12/1990 1990. [Online]. Available: https://ccrma.stanford.edu/files/papers/stanm68.pdf

[6] E. Shriberg, "Higher-level features in speaker recognition." in *Speaker Classification (1)*, ser. Lecture Notes in Computer Science, C. Mller, Ed., vol. 4343. Springer, 2007, pp. 241–259. [Online]. Available: http://dblp.uni-trier.de/db/conf/speakerc/speakerc2007-1.html#Shriberg07

[7] F. Biadsy, "Automatic dialect and accent recognition and its application to speech recognition," Ph.D. dissertation, Columbia University, 2011, ph.D., Columbia University.

[8] U. Shrawankar and V. Thakare, "Techniques for feature extraction in speech recognition system : A comparative study," *ArXiv e-prints*, May 2013.

[9] G. Doddington, "Speaker recognition; identifying people by their voices," *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1651–1664, Nov 1985.

[10] M. Marolt, "Gaussian mixture models for extraction of melodic lines from audio recordings." in *ISMIR*, 2004. [Online]. Available: http://dblp.uni-trier.de/db/conf/ismir/ismir2004.html#Marolt04

[11] X. Lin and S. Simske, "Phoneme-less hierarchical accent classification," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Eighth Asilomar Conference on*, vol. 2, Nov 2004, pp. 1801–1804 Vol.2.

[12] S. Weinberger, "Gmu speech accent archive," http://accent.gmu.edu, 2014.

[13] K. Baumann, "Cross-validation as the objective function for variable-selection techniques," *TrAC Trends in Analytical Chemistry*, vol. 22, no. 6, pp. 395–406, 2003.

[14] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143. [Online]. Available: http://dl.acm.org/citation.cfm?id=1643031.1643047

[15] F. e. a. Pedregosa, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] Y. Liu and Y. F. Zheng, "Fs_sfs: A novel feature selection method for support vector machines," *Pattern Recognition*, vol. 39, no. 7, pp. 1333–1345, 2006.

[17] M. H. Nguyen and F. de la Torre, "Optimal feature selection for support vector machines," *Pattern Recognition*, vol. 43, no. 3, pp. 584 – 591, 2010.

[18] A. M. Martinez and A. Kak, "Pca versus lda," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 228–233, Feb 2001.

[19] Y. Z. et al., "Accent detection and speech recognition for shanghai-accented mandarin," in *Proc. Interspeech*, 2005.