A Simple Feature Normalization Scheme for Non-native Vowel Assessment

Mitchell Peabody, Stephanie Seneff

Spoken Language Systems, CSAIL, MIT
{mizhi,seneff}@csail.mit.edu

Abstract

We introduce a set of speaker dependent features derived from the positions of vowels in Mel-Frequency Cepstral Coefficient (MFCC) space relative to a reference vowel. The MFCCs for a particular speaker are transformed using simple operations into features that can be used to classify vowels from a common reference point. Classification performance of vowels using Gaussian Mixture Models (GMMs) is significantly improved, regardless of which vowel is used as the target among /ɑ/, /i/, /u/, or /ə/. We discuss how this technique can be applied to assess pronunciation with respect to vowel structure rather than agreement with absolute position in MFCC space.

Index Terms: vowel assessment, classification, normalization

1. Introduction

Pronunciation assessment is an important component of Computer Aided Language Learning (CALL) systems. CALL systems frequently employ model scores to produce some measure of pronunciation quality. However, these scores can be very sensitive to intrinsic speaker differences that may not be the result of mispronunciation. Native and non-native speakers exhibit systematic differences in pronunciation. We explore here the possibility of exploiting these systematic differences to improve classification and measure pronunciation quality of non-native speech. We propose a simple normalization procedure that anchors the MFCC spaces of individual speakers to a common reference point. We justify this normalization for assessment by showing improved classification performance and improved correlation of statistical model distances to the rates of vowel substitutions provided by expert human labelers.

2. Background

Numerous approaches have been proposed to normalize speech to account for speaker dependent variation. Vocal tract normalization (VTLN) techniques model the length of the vocal tract and warp the acoustic signal to match a reference. In previous work, Nordstrom and Lindblom [1] scale the formants of the signal by a constant factor determined by an estimate of the vocal tract length from measurements of F_3. Fant [2] extended this by making the scale factor dependent on formant number and vowel class. These methods require knowledge of the formant number and frequencies. More recently, Umesh et al. [3, 4] introduced two automatic methods: one uses a frequency dependent scale factor that does not require knowledge of the formant number, and another based on fitting a model relating the frequencies of a reference speaker to frequencies of a subject speaker.

In contrast to operating on the acoustic signal, Maximum Likelihood Linear Regression (MLLR) [5] attempts to accomodate speaker to speaker variation by adapting the means and variances of existing acoustic models given a relatively small amount of adaptation data. It accomplishes this by estimating linear transformations of model parameters to maximize the likelihood of the adaptation data. Some normalization approaches work directly on the MFCCs extracted as features for speech recognition. Cox [6] implements speaker normalization in the MFCC domain utilizing a filterbank approach to shift MFCCs up and down in the spectrum. He shows that this is a form of vocal tract normalization, and has similarities to a constrained MLLR. Pitz and Ney [7] showed that frequency warping vocal tract normalization can be implemented as linear transformations of MFCCs.

Our approach is inspired by the work presented in [8, 9], which used the Bhattacharyya Distance [10] to compute the overall structure of speakers' phonetic spaces. This was conducted in the spirit of work by Jakobson [11] who argued that the study of the sounds of a language must consider the structure of the sound system as a whole. Thus, the structure created by Minematsu et al. modeled a phonetic space in a holistic fashion, as opposed to the typical method for modeling acoustic spaces using MFCCs or other localized features. They used this structure to measure the distortion between Japanese accented English and General American English and found a positive correlation with human assessments of pronunciation quality. One of the limitations of their technique was that it was unable to individually classify or assess sounds.

3. Approach

We hypothesize that vowels may be produced by humans via an internal relativistic model that attempts to maximize discriminability, akin to the principles in [12]. With this idea in mind, we decided to investigate a very simple normalization method based on relativizing the Cepstral coefficiants to those of a target reference vowel. We therefore propose a simple scheme that intuitively works by anchoring vowel spaces to a common reference point on a per speaker basis. Since speakers are using a common language, common phonetic inventory, and hence a similar vowel space shape, this anchoring should have the effect of shifting speaker vowel spaces into closer proximity.

We consider anchoring points at the vowels /ɑ/, /i/, and /u/, as these quantal vowels [13] exist at relative extremes in the Universal Vowel Space [12], are found in nearly all languages, and should provide relatively stable points of reference. We also considered the use of /ə/ as an anchor, as Puppel and Jahr argue that one of the forces acting on the location of /ɑ/, /i/, and /u/ is a thrust away from the neutral /ə/ in order to maximize discriminability and Diehl [14] notes that in some respects, /ə/ is slightly more stable.

Anchoring the vowel space entails computing the difference between the mean MFCC values for each anchoring vowel and the MFCCs for a sample under consideration. Mathematically,

$$\begin{aligned}\Delta MFCC_{i,\mathrm{ɑ}} &= S_i - \overline{MFCC_{\mathrm{ɑ}}}\\ \Delta MFCC_{i,\mathrm{i}} &= S_i - \overline{MFCC_{\mathrm{i}}}\\ \Delta MFCC_{i,\mathrm{u}} &= S_i - \overline{MFCC_{\mathrm{u}}}\\ \Delta MFCC_{i,\mathrm{ə}} &= S_i - \overline{MFCC_{\mathrm{ə}}}\end{aligned} \tag{1}$$

where S_i is the MFCC sample at segment i and $\overline{MFCC_{\mathrm{ɑ}}}$, $\overline{MFCC_{\mathrm{i}}}$, $\overline{MFCC_{\mathrm{u}}}$, $\overline{MFCC_{\mathrm{ə}}}$ are the mean MFCCs for a speaker's productions of /ɑ/, /i/, /u/, and /ə/, respectively.

Our data come from two corpora. The first corpus is the TIMIT corpus [15], consisting of 4,620 (3,260 male, 1,360 female) training utterances and 1,180 (800 male, 380 female) test utterances from native English speakers. The second corpus is the Chinese University Chinese Learners of English (CU-CHLOE) corpus [16], consisting of 33,026 (16,511 male, 16,515 female) training utterances and 3,760 (1,835 male, 1,835 female) test utterances. Speakers in training sets did not appear in the test sets. Recordings were sampled at 16kHz using close-talking microphones.

The data were force-aligned using a standard SUMMIT [17] recognizer with native English landmark models to obtain a segmentation and assigned reference label for each target vowel. We averaged the MFCCs (14 dimensions) at five regions relative to the vowel endpoints for each segment: 30ms-0ms before the segment (pre), at 0%-30% (start), 30%-70% (middle), and 70%-100% (end) through the segment, and to 30ms after the segment (post). We computed the mean MFCC value for the anchor vowels, /ɑ/, /i/, /u/, and /ə/ of each speaker. For each measurement, we computed the difference between the measured MFCCs and the mean of a speaker's anchor vowel as shown in Equation 1 at the corresponding part of the segment. The measurements from all five regions plus the log duration of the segment were combined into a 71-dimension feature vector.

Some speakers did not have enough instances of an anchor vowel. In the cases where there were fewer than 5 samples of the anchor vowel, we used a fallback model consisting of the mean of all the training data for the anchor vowel. This was critical in the TIMIT data where each speaker spoke only 10 utterances and some of the anchor vowels suffered from data sparseness issues on a per speaker basis.

We created a number of different feature sets based on these measurements for use in our experiments. The MFCCs (baseline), /ɑ/-anchor, /i/, /u/, and /ə/ anchor features (Table 1) were used to train Gaussian Mixture Model (GMM) classifiers using k-means clustering.

To evaluate the effect of the anchoring, three classification experiments were then performed for each feature: native test data with native-trained models, non-native test data with non-native-trained models, and non-native test data with native trained models. We also qualitatively and quantitatively evaluated the effect of the transformation on the anchor feature distribution at the middle of the sound events. Finally, we measured correlation of native and non-native model distances for each vowel with human error assessment.

4. Results

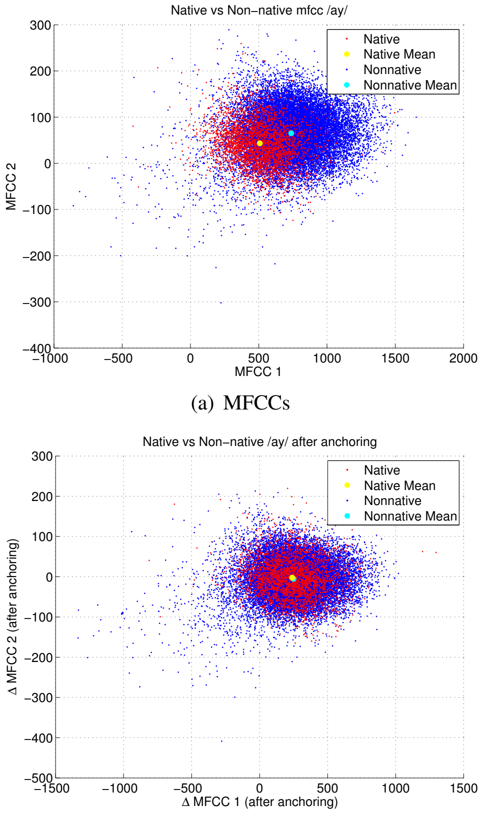
The results for our classification experiments are presented in Table 1. Our baselines for comparison are features from Table 1 row (a). These are standard sets of MFCCs used for segment models in our classifier. The poor performance for CHLOE, particularly when TIMIT is used for training, reflects the difficulting in pronouncing a non-native vowel.

Table 1 presents the error rates when the means of the anchor vowel MFCCs are computed from the labeled test data. The relative performance increases range from 11.2% to 13.0% for the native classifier with native speech, 10.7% to 11.5% for non-native speech with non-native classifier, and 8.6% to 9.8% for non-native speech with the native classifier.

		Training Data	TIMIT	CHLOE	TIMIT
		Test Data	TIMIT	CHLOE	CHLOE
Features	a	MFCCs	33.0%	38.3%	48.8%
	b	/ɑ/-anchor	28.7%	33.9%	44.6%
	c	/i/-anchor	29.0%	34.2%	44.4%
	d	/u/-anchor	29.3%	34.1%	45.2%
	e	/ə/-anchor	29.0%	34.2%	44.0%

Table 1: Percent error vowel classification. The classification error decreases significantly with normalization with respect to any vowel.

Diehl [14] points out that some studies have found that /ə/ actually has smaller F_2 variance than some of the quantal vowels. The reason for this is that the cross-sectional area of the vocal cavity is more uniform when /ə/ is produced. In contrast, /ɑ/, /i/, and /u/ all have non-uniform cross-sectional areas, caused by tongue and jaw position. The classification results in Table 1 rows b-e confirm that /ə/ as an anchor performs comparably to the other vowels. When these facts are considered, along with its high usage frequency, a strong case can be made for using /ə/ as an anchor vowel.



(a) MFCCs

(b) /ə/-normalized

Figure 1: Distributions of the first two dimensions of the feature vectors for /ɑʸ/ spoken by native and non-native speakers.

To qualitatively understand why we see these performance improvements and why this scheme may be beneficial for assessment, it is helpful to visualize the transformation. Figure 1 depicts the effect of the transformation on the native and non-native data for MFCCs 1 and 2 for the dipthong /ɑ^{y}/. As can be seen from the figure, the mean of the non-native distribution is shifted closer to the native mean. This effect was seen for almost all pairs of vowel distributions. Note that MFCC 1 captures the total energy of the MFCC spectrum, so this normalization effectively corrects for differences in microphone gain as well. By using only one vowel as the reference point, we are essentially shifting the entirety of the speaker's vowel space without affecting its shape. This creates a feature space in which the samples still exist in the same relative proximity to each other. This would be important for pronunciation assessment of individual vowels.

We can quantitatively confirm that the distributions between native and non-native speakers have moved closer together by measuring the Bhattacharyya distance between native and non-native single gaussian distributions of MFCC values taken from the middle of the segments and the distance between native and non-native single gaussian distributions normalized with /ə/-anchors (see Table 2). We are specifically interested in Bhattacharyya distance, because this was the major normalizing component in Minematsu's work [8] on sound structure.

Table 2 also shows the rate of substitution (the number of times another vowel was substituted for the correct vowel) for each of the vowels, as judged by experts. To obtain this information, we compared the results of the forced alignment from the recognizer, which we considered the baseline truth of the sounds the speakers should have produced, with the human labeled data. It is interesting to note that the two vowels with highest error as judged by humans, /ɚ/ and /ɪ/ are also the only two vowels whose Bhattacharyya distance increased after normalization. These two vowels are both missing from the Cantonese vowel inventory.

The sample normalized correlations between human judged substitution rate and the Bhattacharyya distances, and the distance of the normalized distributions were computed using $C(X,Y) = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$. Normalized distributions have significantly higher correlation (0.916) than MFCC distributions (0.824). What these results show is that, by normalizing the data, we are better able to correlate the distances between native speakers and non-native speakers at a phonemic level with how frequently the phones are mispronounced.

The vowel /ɚ/, for example, is most frequently mispronounced by a significant margin and correspondingly has a very large distance (0.736) distance. We hypothesize that anchoring each of the speaker's vowel spaces to a common reference point may have some relation to a normalization process that occurs when humans perceive non-native speech.

If we compute the overall human error rate, vowels are substituted at a rate of 0.215. If we ignore the common case of vowels reduced to /ə/, which are not necessarily pronunciation errors, then the human error rate is 0.139. If we consider the substitution of /ə/ for /ɚ/ a pronunciation error, then the overall human error rate is 0.189. The difference in classification error rate between native and non-native data using TIMIT trained models with MFCCs as a feature is 15.8% (33.0% vs 48.8%). This difference is very close to the human error rates. We can interpret this to mean that the difference in classifier performance is largely explained by mispronunciation of vowels by the non-native speakers.

One might imagine that the distance between the distributions of native and non-native models would be a good measure of the degree of difficulty non-native speakers have with that particular vowel. Figure 2a depicts a representation of the MFCC vowel spaces of native and non-native speakers. The points represent the means of a subset of the vowel distributions for both sets of speakers. Figure 2b depicts the vowel spaces after they have been anchored by /ə/.

Vowel	Human Error Rate	MFCC	/ə/-anchor
/ɑ/ [aa]	0.076	0.407	0.247
/æ/ [ae]	0.229	0.324	0.155
/ʌ/ [ah]	0.135	0.294	0.072
/ɔ/ [ao]	0.074	0.282	0.223
/ɑ^{w}/ [aw]	0.134	0.468	0.269
/ɑ^{y}/ [ay]	0.111	0.378	0.204
/ɛ/ [eh]	0.289	0.332	0.225
/ɚ/ [er]	0.674	0.678	0.736
/e/ [ey]	0.156	0.419	0.319
/ɪ/ [ih]	0.353	0.299	0.336
/i/ [iy]	0.140	0.404	0.279
/o/ [ow]	0.134	0.256	0.138
/ɔ^{y}/ [oy]	0.045	0.542	0.298
/ʊ/ [uh]	0.149	0.233	0.163
/u/ [uw]	0.063	0.307	0.210
Correlation		0.824	0.916

Table 2: Bhattacharyya distances between native and non-native models trained on different feature sets and their correlations with vowel substitution rate provided by human expert labelers.

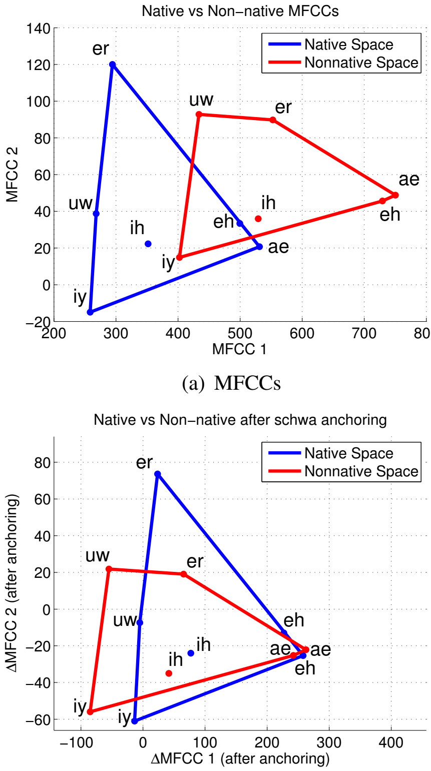
The overall shapes of the spaces have not been affected by the anchoring, but the spaces now directly overlap each other. The anchoring provides a direct comparison of the vowel spaces when the relative positions of the vowels are considered. For example, we can clearly see /ɚ/, the sound that is most often mispronounced by the non-native speakers, is located in very different relative positions between the native and non-native populations. Additionally, /æ/ and /ɛ/ are all clustered together and the non-native /ɛ/ exists in a different position relative to the non-native /æ/ when compared with the relative positions of the native equivalents.

5. Discussion and Future work

This work introduced a simple feature normalization scheme for vowel classification and subsequent vowel assessment of non-native speakers. The MFCC features for particular speakers were transformed using simple operations into features anchored at a common reference point. We showed that this results in increased classifier performance. We examined the effect of the transformation on the distributions and the shape of the vowel space, and showed that it resulted in better correlation with human assessment.

Our fallback models until this point have been based on the overall anchor vowel MFCC means for the entire training set. For the purposes of evaluation this may prove inadequate: we could try separating speakers by gender or performing a more refined clustering. An intermediate step would determine the best fallback model to apply to a given speaker.

The transformation performed is similar to the MLLR technique developed in [18], with the transformation matrix set to the identity matrix. The attraction of transforming the MFCCs using our technique is that it is very simple to implement and only requires instances of a speaker's common anchor vowels in order to be applied. Future work includes comparing the per-



(a) MFCCs

(b) /ə/-normalized

Figure 2: Comparison of feature space for the first two dimensions. The large points represent the means of the features measured at the mid-point for the corresponding vowel. The outlined shapes (red and blue) form the convex hull of the space.

formance of our transformation with the MLLR technique and exploring simple methods that account for variance in our technique. We should also compare our technique with VTLN, although, because VTLN shows the most significant gains when normalizing for child speech and between genders, we are not sure how it will perform when moving between native and non-native speakers.

We anticipate the transformation to be useful for assessing the pronunciation quality of non-native speakers. Because the transformation effectively positions the vowel spaces of non-native speakers with native speakers, we should be able to use this to help pinpoint and assess pronunciation errors in non-native speech using the log-likelihood scores from the classifier.

6. Acknowledgements

This research was supported by Industrial Technology Research Institute (ITRI) in Taiwan. We would like to thank Prof. Helen Meng at the Chinese University of Hong Kong for providing the non-native data used for this research.

7. References

[1] P. NORDSTROM and B. LINDBLOM, “A Normalization Procedure For Vowel Formant Data,” in *The International Congress Of Phonetic Sciences*, Leeds, Aug. 1975.

[2] G. Fant, “Non-uniform vowel normalization,” *Speech Trans. Lab. Q. Prog. Stat. Rep*, pp. 2–3, 1975. [Online]. Available: http://www.speech.kth.se/prod/publications/files/qpsr/1975/1975_16_2-3_001-019.pdf

[3] S. Umesh, S. Kumar, M. Vinay, R. Sharma, and R. Sinha, “A SIMPLE APPROACH TO NON-UNIFORM VOWEL NORMALIZATION,” in *IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING*. Citeseer, 2002. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.5.8157\&amp;rep=rep1\&amp;type=pdf

[4] S. V. B. Kumar and S. Umesh, “NON-UNIFORM SPEAKER NORMALIZATION USING FREQUENCY-DEPENDENT SCALING FUNCTION,” in *Proc. of International Conference on Signal Processing and Communications (SPCOM)*, 2004.

[5] M. Gales, D. Pye, and P. Woodland, “Variance Compensation within the MLLR Framework for Robust Speech Recognition and Speaker Adaptation,” in *Proc. ICSLP ’96*, vol. 3, Philadelphia, PA, USA, Oct. 1996, pp. 1832–1835.

[6] S. Cox, “SPEAKER NORMALIZATION IN THE MFCC DOMAIN,” in *Sixth International Conference on Spoken Language Processing*, 2000, pp. 4–7. [Online]. Available: http://fizz.cmp.uea.ac.uk/Research/speechgroup/cox-pub-archive/cox-vocal-icslp00.pdf

[7] M. Pitz and H. Ney, “Vocal tract normalization as linear transformation of MFCC,” in *Eighth European Conference on Speech Communication and Technology*, no. Cc. Citeseer, 2003. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.72.3962\&amp;rep=rep1\&amp;type=pdf

[8] N. Minematsu, “Yet another acoustic representation of speech sounds,” in *Proc. ICASSP*, 2004, pp. 585–588. [Online]. Available: http://www.ece.umassd.edu/Faculty/acosta/ICASSP/Icassp_2004/pdfs/0100585.pdf

[9] M. Suzuki, L. Dean, N. Minematsu, and K. Hirose, “Improved Structure-based Automatic Estimation of Pronunciation Proficiency,” in *Proc. SLaTE*, vol. 5, 2009. [Online]. Available: http://www.eee.bham.ac.uk/SLaTE2009/papers/SLaTE2009-21-v2.pdf

[10] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions,” *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.

[11] R. Jakobson and L. Waugh, *The sound shape of language*. Mouton de Gruyter, 1987. [Online]. Available: http://scholar.google.com/scholar?hl=en\&btnG=Search\&q=intitle:The+Sound+Shape+of+Language\#0

[12] S. Puppel and E. H. Jahr, *The theory of universal vowel space and the Norwegian and Polish vowel systems*. Berlin: Mouton de Gruyter, 1997, vol. 2, pp. 1301—-1324.

[13] K. N. Stevens, “The Quantal Nature of Speech: {E}vidence from Articulatory-Acoustic Data,” in *Human Communication: A Unified View*, E. E. David Jr. and P. B. Denes, Eds. New York: McGraw-Hill, 1972.

[14] R. L. Diehl, “Acoustic and auditory phonetics: the adaptive design of speech sound systems.” *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 363, no. 1493, pp. 965–78, 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17827108

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “{DARPA TIMIT} Acoustic-Phonetic Continuous Speech Corpus {CD-ROM},” Gaithersburg, MD, 1993.

[16] H. Meng, Y. Y. Lo, L. Wang, and W. Y. Lau, “Deriving salient learners’ mispronunciations from cross-language phonological comparisons,” in *Proc. of ASRU*, 2007.

[17] V. Zue, J. R. Glass, D. Goodine, M. Phillips, and S. Seneff, “The {SUMMIT} Speech Recognition System: Phonological Modeling and Lexical Access,” in *Proc. ICASSP*, 1990, pp. 49–52.

[18] D. Giuliani, M. Gerosa, and F. Brugnara, “Speaker normalization through constrained MLLR based transforms,” in *Proc. ICSLP*, vol. 1, no. 2, p. 3. [Online]. Available: http://pfstar.itc.it/public/publications/itc-Icslp-2004-1.pdf