Evaluating vowel pronunciation quality: Formant space matching versus ASR confidence scoring

Ashish Patil
Department of ECE,
NIT Tiruchirappalli,
Tamil Nadu, India
patilashishliv@gmail.com

Chitralekha Gupta, Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology Bombay,
Mumbai 400076, India
{chitralekha,prao}@ee.iitb.ac.in

***Abstract*— Quantitative evaluation of the quality of a speaker's pronunciation of the vowels of a language can contribute to the important task of speaker accent detection. Our aim is to qualitatively and quantitatively distinguish between native and non-native speakers of a language on the basis of a comparative study of two analysis methods. One deals with relative positions of their vowels in formant (F1-F2) space that conveys important articulatory information. The other method exploits the sensitivity of trained phone models to accent variations, as captured by the log likelihood scores, to distinguish between native and non-native speakers.**

Keywords - pronunciation; accent detection; formants; ASR confidence scoring

I. INTRODUCTION

Detecting the accent of a speaker is useful for automatic speech recognition (ASR) systems where acoustic models need to be matched to speaker characteristics for consistent performance. It is useful also in language learning tools where detection of improper accent can provide valuable feedback to the user. Accent refers to a pattern of pronunciation in the use of vowels or consonants, intonation, stress patterns and other prosodic features. It is usually characteristic of a regional or social grouping of people. In the modeling of accents the following four acoustic correlates are considered essential [1]

- Formants
- Intonation
- Duration and speaking rate
- Glottal pulse shape

A significant part of the acoustics of accents is normally due to the differences in the distributions of the formants of the vowels and diphthongs. Formants are the resonant frequencies of the vocal tract. Formants, in addition to conveying phonemic identity, are also affected by speaker and accent characteristics. This property can be utilized to distinguish speaker characteristics related to the speaker's accent.

In this work, we study the evaluation of vowel pronunciation, a part of the larger problem of accent detection. We propose a method based on comparing the formant space of the speaker's vowels with the formant space of vowels for the given language as obtained from native speakers with good pronunciation. The formant space based method is compared with a standard method of pronunciation scoring via ASR confidence values [2]. The confidence values are likelihood scores obtained by the recognition of the accented speech using acoustic models trained on native speakers of the language. Experimental results comparing the two methods are obtained on a dataset of continuous speech obtained from native and non-native speakers of Hindi. A part of the native speech database is used as training data. A mix of native and non-native sets of speaker utterances are then used to test the performance of the accent detection method in terms of its prediction of the goodness of pronunciation as expected from the known native/non-native character of the speaker.

II. DATABASE

The data sets used in the experiments are drawn from the TIFR Hindi speech database [3]. Designed on the lines of the TIMIT database, the continuous speech sentences spoken by 100 native speakers of Hindi are phonetically segmented and manually labeled. The speech data was recorded using a close-talking, directional microphone sampled at 16 kHz and stored in 16-bit PCM, mono format. The phonetically rich sentences have been designed at TIFR. Each sentence is typically 3-5 sec duration with each speaker contributing 10 sentences of which 2 sentences are common across all speakers.

The vowel set under consideration includes short and long vowels along with their nasalized counterparts as shown in Table I grouped under 6 classes that include short (lowercase), long (uppercase) and nasalized (M) vowels (phonetic labels as per TIFR Hindi Database [3]). Each of these classes is henceforth referred to as a "vowel". These vowels span the F1 – F2 formant space that, in this study, forms the basis of pronunciation quality assessment.

TABLE I. VOWEL CLASSES

Vowel class name	Vowels
a	a, aM
A	A, AM
e	e, E, eM EM
i	i, I, iM, IM
o	o, O, oM, OM
u	u, U, uM, UM

A. Training dataset

In the experiments performed, 25 male native speakers from the TIFR Hindi database were considered for the training set. The HMM models for the ASR confidence scoring method were trained on this set (described in section IV). The 6 vowel classes, viz. a, A, e, i, o, u were considered

from this continuous segmented data. About 90 vowel tokens were taken per speaker (i.e. 15 per vowel per speaker, except for a few cases of 'u' and 'o' in which it was less than 15 in number), a total of 25*90 tokens were used for the 25-speaker reference set across 6 vowels (i.e. 15*25 tokens per vowel).

B.. Testing dataset

A native test dataset of 5 male native speakers from the same TIFR database but outside the reference set was taken.

A second test dataset comprised recordings from 5 male non-native speakers each reading out 10 sentences also drawn from the TIFR database but different from those used in the native set. The non-native speakers were from the southern states of India and spoke with accents characteristic of their individual native tongues (Telugu, Kannada, Tamil and Malayalam) apart from inserting occasional hesitation pauses in their speech. The data was recorded at IIT Bombay using a close-talking, directional microphone sampled at 16 kHz 16 bit PCM.

For the test set, 15 tokens per vowel per speaker (approx.) were considered for the set of 6 vowel classes, as was done for the reference set.

III. Formant Space Based Method

The first two formants (F1, F2) of a steady vowel utterance cue the phonemic identity of the vowel. In the case of a non-native speaker, the phonemic quality may deviate from its canonical form in the language, influenced possibly by the vowels in the speaker's own native language. This aspect is represented by a change in the relative position of the vowel in the continuous F1-F2 space. However, it is important to consider the fact that formant locations are influenced not only by the phonemic quality but also by the anatomical/physiological characteristics of the speaker such as vocal tract length and shape. Vowel normalization methods have been used in the past to reduce speaker differences in terms of anatomical/physiological variations while largely preserving phonemic and socio-linguistic variations [4].

There are different methods available for vowel normalization depending on the type of information employed. Vowel-intrinsic procedures use only the acoustic information within the single vowel token to normalize that vowel token. These include transformation of formants on the frequency scale with or without reference to other formants. Vowel-extrinsic normalization procedures use the knowledge of the formants of *all* the vowels of the speaker. In research investigating language variation [4], vowel-extrinsic formant normalization procedures have been found effective in reducing differences due to physiological variations.

In this section, we describe the method used for obtaining the formants of the vowel utterances in continuous speech and investigate the suitability of two different normalization procedures available in the literature. The methods are vowel-extrinsic in that they normalize the i^{th} formant of a given vowel token using the knowledge of the i^{th} formant values across all the vowel tokens of the language by the speaker in question. The normalization methods are compared with the baseline (i.e. no normalization) in terms of reducing the scattering in F1-F2 space. Finally a method is presented to quantify the extent of phonemic quality degradation by a distance measure in normalized F1-F2 space.

A. Formant detection and normalization

The first two formants were extracted from formant tracks generated automatically using the PRAAT speech interface. The LPC-Burg method was used for formant estimation. The available manual labeling was used to determine the mid-point of each marked vowel segment in the continuous utterances. The formant values at these instances were recorded to get a single (F1, F2) point in formant space per vowel utterance.

The entire set of training data vowels was used to derive a reference space for the 6 vowels. Before establishing the reference or test vowel space, a speaker intrinsic and vowel extrinsic normalization procedure was implemented on the formant data essentially to eliminate inter-speaker variations due to physiological differences and to preserve sociolinguistic/dialectal differences in vowel quality. Normalization helps in better clustering of similar native speakers and will help in better classification against non-native speakers. The two methods of normalization that we have explored are described below.

1) Lobanov Normalization: Lobanov's (1971) normalization procedure [4,5,6] standardizes the mean and the standard deviation for each speaker's vowels with the equation:

$$F_{inorm} = \frac{F_i - \overline{F_i}}{SD_i} \quad (1)$$

where F_i is a given formant, $\overline{F_i}$ is the average value of F_i across all vowels, and SD_i is the standard deviation of F_i about its mean for all vowels. The Lobanov method does an excellent job of factoring out physiologically-caused differences in formant values while retaining sociolinguistic differences [5,6].

Lobanov has two main disadvantages. First, like other vowel-extrinsic formulas, it works optimally when all the vowels of speakers' vowel systems are included. When some vowels are excluded, vowel-extrinsic methods will yield skewed normalized values. That was the reason why equal numbers of tokens were taken for each vowel. The other disadvantage, also shared with other vowel-extrinsic methods, is that it may be impaired when different dialects or languages that show different vowel systems are compared. So, only Hindi language accents are compared here.

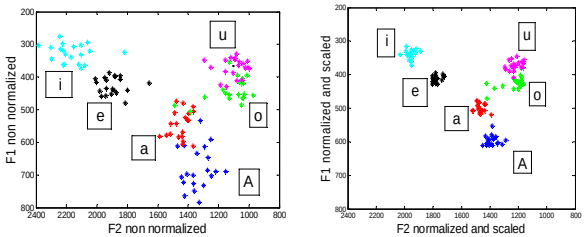

Figure 1. Non normalized (left) and Lobanov normalized (right) vowel clusters for training set speakers.

2) Nearey Normalization: To normalize with the Nearey method [4, 6], the following formula is used:

$$F*_{n[V]} = anti\log((\log(F_{n[V]})) - mean(\log(F_n))) \quad (2)$$

where $F*_{n[V]}$ is the normalized value for $F_{n[V]}$, formant n of vowel V, and $mean(\log(F_n))$ is the log-mean of all F_n s for the speaker in question.

Much of what was said about the Lobanov formula also applies to Nearey. It performed well in reducing physiological variation, and no worse than the other methods compared at preserving sociolinguistic variation. Disner (1980) [6] found that it reduced scatter the best of all the methods she compared. Nearey suffers from the same disadvantages as Lobanov.

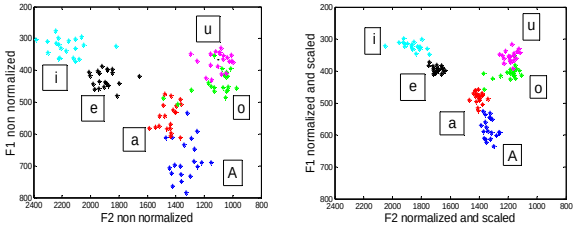

Figure 2. Non normalized (left) and Neary normalized (right) vowel clusters for training set speakers.

Normalization essentially reduces the scatter of the vowel clusters as is clearly seen in Fig.1 and 2. The parameter of squared coefficient of variation (SCV) was used for comparing scatter reduction between Lobanov and Nearey procedures. The percentage scatter reduction is the average of the reductions of all six vowels considered. SCV is defined as square of the ratio of deviation and mean for a distribution. As we can see from Table II, Nearey procedure reduces the scatter to a better extent compared to Lobanov.

TABLE II. PERCENTAGE OF SCATTER AREA AFTER NORMALIZATION

Normalization	**Percentage of scatter area remaining after normalization (with 100 percent indicating unnormalized total scatter)**
Unnormalized	100%
Lobanov	55%
Nearey	45%

B. Measuring vowel pronunciation quality

For a set of geographical units in the Cartesian coordinate system, the locus of the standard deviation of the x coordinates of the set forms a closed curve as the system is rotated about the origin. This curve is often referred to as 'standard deviational ellipse' (SDE) or 'standard deviation curve' (SDC) [7]. The SDE gives dispersion in two dimensions. The major and minor axes give standard deviation in X-Y direction. The orientation of the ellipse gives the direction of distribution of data. To describe the amount of scatter of geographical units, the following index is defined:

$$S_d = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\{(x_i-\bar{x})^2+(y_i-\bar{y})^2\}} = \sqrt{\sigma_{x_\alpha}^2+\sigma_{y_\alpha}^2} \quad (0<\alpha\le 2\pi) \quad (3)$$

The SDE helps in qualitative discriminant analysis of the native and non-native speakers.

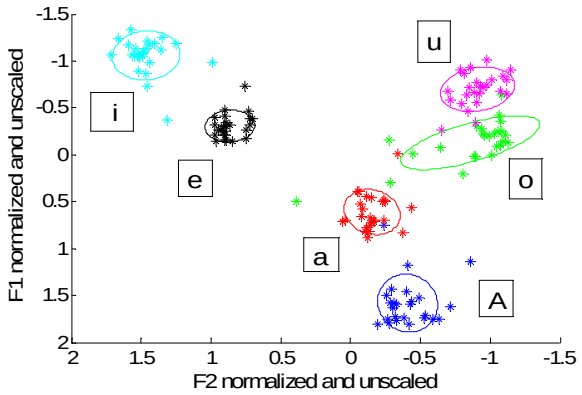

Figure 3. Lobanov-normalized vowel clusters of training set speakers enclosed in SDE. For each SDE, 76 percent of data is enclosed.

Once the reference clusters for each vowel class were prepared, a standard deviation ellipse was constructed for each of them which shows the distribution and orientation of the clusters in the formant space as shown in Fig 3. Now vowels from a test speaker were appropriately normalized and placed at their respective positions in the formant space and compared with the reference set. If more than 2 test vowel tokens lie outside their respective vowel SDEs then the test speaker is non-native, otherwise it can be classified as native. In this manner, a qualitative classification can be done as shown in Fig.4 and Fig.5. For quantitative classification, we use Mahalanobis distance rule that can determine the similarity of an unknown sample set to a known one. First of all, a threshold distance has to be fixed for each vowel class. If the test Mahalanobis distance is greater than the threshold then it can be classified as non-native. Formally, the Mahalanobis distance from a group of values with mean $\mu = (\mu_1, \mu_2, \mu_3, ..., \mu_N)^T$ and covariance matrix S for a multivariate vector $x = (x_1, x_2, x_3, ..., x_N)^T$ is defined as

$$D_M(x) = (x-\mu)^T S^{-1}(x-\mu) \quad (4)$$

The Mahalanobis distance is simply the distance of the test point from the center of mass divided by the width of the ellipsoid in the direction of the test point. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant.

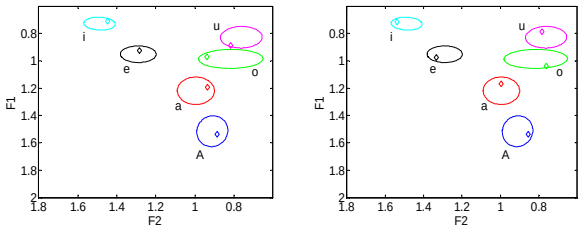

Figure 4. Plots of the test tokens of two of the native speakers along with the SDE derived from the training set.

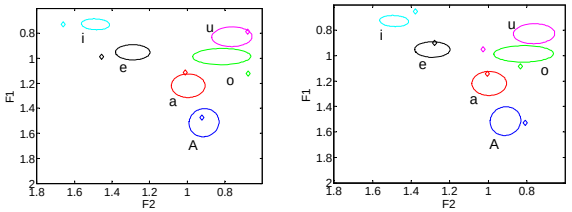

Figure 5. Plots of the test tokens of two of the non-native speakers along with the SDE derived from the training set.

Fig. 4 and 5 are the plots of two of the native and two of the non-native speakers respectively which give a qualitative method of classifying them. We observe that several vowels of the non-native speakers lie outside the corresponding vowel SDEs while those of the native speakers are within their SDEs. This suggests that non-native speakers may be distinguished from the native speakers by the higher number of outlying vowels.

IV. ASR Confidence Scoring

It can be assumed that with HMM models trained on native speech data, the log of the likelihood of any input speech data, as computed by Viterbi decoding, would provide a measure of the similarity of the test data to native speech. This is the basis for the local average log likelihood scoring for pronunciation quality [2].

In this section, we describe the details of the trained acoustic models and the method employed to compute the likelihood scores. HMM Toolkit (HTK 3.4) [8] was used to train the acoustic models for the 36 phonetic classes on the training dataset. All the models were context independent, 5-state HMM (first and fifth states were non-emitting) left to right without skip state, except the 5-state back-forth silence model (forward and backward transitions between first and third emitting states), all with 8 Gaussian mixtures (diagonal covariance) trained with flat-start initialization. The standard 39 dimensional pre-emphasized and energy normalized MFCC, delta and acceleration feature vector was computed for the 16 kHz sampled signals at 10 ms intervals. A null grammar network of monophones is used to preserve language independence.

For each sentence, the phone segment boundaries were obtained, along with the corresponding log-likelihood scores of each segment by operating the recognition engine in the forced alignment mode. In this mode, the recognition network is constructed from the orthographic phone level transcription and the duration normalized log likelihood scores are obtained.

If τ_i denotes the start time of the i^{th} phonetic segment then the total log-likelihood of this segment l_i can be computed, using an HMM, by

$$l_i = \sum_{t=\tau_i}^{\tau_{i+1}-1} \log(p(s_t \mid s_{t-1}) p(x_t \mid s_t)) \qquad (5)$$

where x_t and s_t are the observed spectral vector and the HMM state at time t, respectively, $p(s_t|s_{t-1})$ is the HMM transition probability and $p(x_t|s_t)$ is the so-called output distribution of state s_t [2].

To compensate the effect of duration of phones (longer phone score dominating over that of shorter phone), 'local average log likelihood' is computed for every vowel type v, given by

$$L_v = \frac{1}{N} \sum_{i=1}^{N} \frac{l_i}{d_i} \qquad (6)$$

where, the duration normalized log likelihood is given by $\frac{l_i}{d_i}$ where l_i is the log likelihood of the i^{th} phonetic segment and $d_i = \tau_{i+1} - \tau_i$ is the duration in frames of the i^{th} phonetic segment (as obtained after forced alignment of the utterance). Rather than computing the average over all phone segments of the utterance [2], we have computed it over all the segments (N) of a particular vowel type.

Utterances from the native and non-native test set were forced aligned with their orthographic transcriptions, i.e. transcriptions as would have been if spoken by a native speaker. The local average log likelihood scores for each of the six vowel classes were computed and an average score per vowel per speaker was obtained. As it is clear from Fig.6, the scores for the non-native speakers is significantly less than that for the native speakers. So an appropriate threshold score for each vowel class will distinguish between native and non-native speakers (described in the next section).

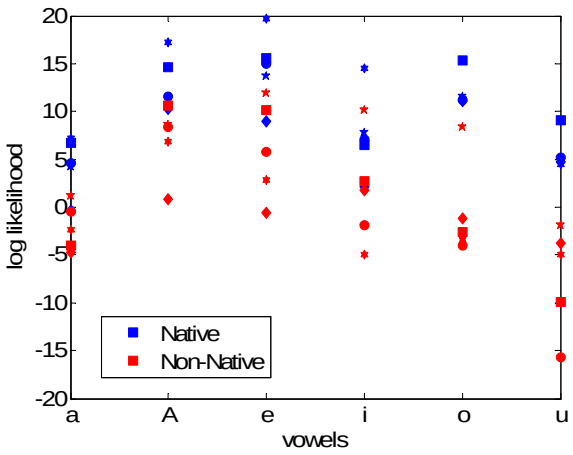

Figure 6. Local average log likelihood scores for the six vowel classes for the 5 native (blue) and 5 non-native (red) speakers

V. Non-Native Accent Detection Experiments

For quantitative classification between native and non-native speakers, a threshold 'distance' and 'score' were computed for each vowel class for the formant space method and the ASR confidence method respectively. In the formant space method, each vowel cluster comprises of 25 points (one value per speaker obtained by averaging over around 15 tokens of that vowel of that speaker, for the 25 speakers in the reference set). A Mahalanobis distance of each point from its own cluster was calculated. These 25 distance values for each vowel were sorted in an ascending order. It was observed that after 20th speaker, the distance values increase steeply. So the 19th Mahalanobis distance for each vowel was fixed as the threshold for classification. We applied both the normalization procedures for a test speaker and if there were more than two vowel classes whose distance was greater than the threshold distance, then the speaker was classified to be non-native.

In the ASR confidence method, utterances of the 25 speakers from the reference set were forced aligned with their orthographic transcription, and the local average log likelihood scores for each of the vowel class was computed. An average value of the scores (averaged over around 15 tokens) per vowel per speaker was obtained which gave 25 scores per vowel. The mean and the standard deviation of the scores for each vowel were computed. A threshold for each vowel was defined as the mean of the scores over all the speakers for that vowel minus the standard deviation of

the same. If scores of more than two vowel classes for a test speaker are less than the threshold of the respective classes, then the speaker was classified to be non-native. This is in agreement with the fact that higher the value of log likelihood score, more likely it is that the uttered phone has been pronounced similar to a native speaker. The test results of the three methods are given in Table III.

TABLE III. NUMBER OF CORRECT DETECTIONS

Method	Native (out of 5 speakers)	Non-native (out of 5 speakers)
Lobanov	2	2
Neary	4	4
ASR confidence	4	5

VI. DISCUSSION

It is observed that Neary normalization method does better than Lobanov in classification of native and non-native speakers. The ASR confidence method does a good job in classifying most of the speakers appropriately because of the general trend of the average log likelihood score for a particular vowel class of the non-native speakers being lower than the threshold values and that of the native speakers being higher. An interesting observation was that one of the non-native speaker that was classified as native by the formant space methods and also had scores nearing native (close to threshold) by the ASR confidence method. This shows a consistency in the observations of the two methods.

The ASR confidence score, however, is merely a number that at best quantifies the amount of deviation in the pronunciation but provides no information on the nature of the deviation; on the other hand, the formant space based method can provide useful articulatory feedback from the knowledge of the precise location in formant space with respect to the reference positions of all the vowels. Formant extraction, however, is a challenging task especially for high-pitched voices whereas ASR confidence scoring relies only on easily obtained broad spectral envelope parameters such as MFCCs.

ACKNOWLEDGEMENTS

This work was supported by the TTSL IIT Bombay Center of Excellence in Telecommunication (TICET) at IIT Bombay.

REFERENCES

[1] Q. Yan, S. Vaseghi, D. Rentzos, C.H. Ho, *Analysis and Synthesis of formant spaces of British, Australian and American Accents*, IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No 2, February 2007.

[2] Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, Mitchel Weintraub, *Automatic scoring of pronunciation quality*, Speech Communication 30 (2000) 83 - 93

[3] Samudravijaya K, P.V.S.Rao, and S.S.Agrawal, "Hindi Speech Database", Proceedings of International Conference on Spoken Language Processing, 2000, China.

[4] P. Adank, R. Smits, R. van Hout, A comparison of vowel normalization *procedures for language variation research*, Journal of the Acoustical Society of America, 2004, 116:3099-107.

[5] B.M. Lobanov, Classification of Russian Vowels spoken by different *speakers*, Journal of the Acoustical Society of America 49:606-08, 1971

[6] S.F. Disner, Evaluation of vowel normalization procedures, Journal of the Acoustical Society of America 67:253-61, Jan.1980.

[7] J. Gong, Clarifying the standard deviational ellipse, Geographical Analysis, Vol 34, April 2002.

[8] Young S. et al., "The HTK Book v3.4", Cambridge University, 2006