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 for both 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 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>
<thead>
<tr>
<th>TABLE I. VOWEL CLASSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel class name</td>
</tr>
<tr>
<td>a</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>e</td>
</tr>
<tr>
<td>i</td>
</tr>
<tr>
<td>o</td>
</tr>
<tr>
<td>u</td>
</tr>
</tbody>
</table>

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
form 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 10 male native speakers from the same TIFR database but outside the reference set was taken.

A second test dataset comprised recordings from 10 male non-native speakers each reading out 10 sentences also drawn from the TIFR database same as those used in the native set. The non-native speakers were from southeastern and eastern states of India and spoke with accents characteristic of their individual native tongues (Telugu, Kannada, Tamil, Malayalam, Bengali and Oriya) 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. Some vowel normalization methods are effective in reducing speaker differences in terms of anatomical/physiological variations while largely preserving phonemic and socio-linguistic variations [4]. This property of the normalization procedures is instrumental in our native – non-native classification.

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-intrinsic normalization procedures have been found effective in reducing differences due to physiological variations while at the same time preserving phonemic and socio-linguistic variation.

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 and formant intrinsic, that is, they normalize the $i^{th}$ formant of a given vowel token using the knowledge of the $j^{th}$ 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 which clearly shows the variance reduction property.

Finally a method is presented to quantify the extent of phonemic quality degradation in the test speakers (native and non-native) 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 automatic labeling was used to determine the mid point of each marked vowel segment in the continuous utterances. The formant values at the mid point of the vowel segments 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, speaker intrinsic, vowel extrinsic and formant intrinsic normalization procedures were implemented on the formant data essentially to eliminate inter-speaker variations due to physiological differences and to preserve socio-linguistic/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:

$$F_{\text{norm}} = \frac{F_i - \overline{F}_i}{SD_i}$$

(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 socio-linguistic 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:

\[
F^*_{n[v]} = \text{anti log}((\text{log}(F_{n[v]})) - \text{mean}(\text{log}(F_n)))
\]  (2)

where \( F^*_{n[v]} \) is the normalized value for \( F_{n[v]} \), formant \( n \) of vowel \( V \), and mean(\( \text{log}(F_n) \)) 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 socio-linguistic 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.

<table>
<thead>
<tr>
<th>Normalization</th>
<th>Percentage of scatter area remaining after normalization (with 100 percent indicating unnormalized total scatter)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unnormalized</td>
<td>100%</td>
</tr>
<tr>
<td>Lobanov</td>
<td>55%</td>
</tr>
<tr>
<td>Nearey</td>
<td>45%</td>
</tr>
</tbody>
</table>

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:

\[
S_a = \frac{1}{n} \sum_{i=1}^{n} ((x_i - \bar{x})^2 + (y_i - \bar{y})^2) = \sqrt{\sigma_x^2 + \sigma_y^2} \quad (0 < \alpha \leq 2\pi)
\]  (3)

The SDE helps in discriminant analysis of the native and non-native speakers.
Fig. 5. Plots of the test tokens of a non-native speaker along with the SDE derived from the training set (Nearey normalization). Fig. 4 and 5 are the plots of a native and a non-native speaker respectively. We observe that several vowels of the non-native speaker lie outside the corresponding vowel SDEs while those of the native speaker 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 $\tau_i$ denotes the start time of the $i^{th}$ phonetic segment then the total log-likelihood of this segment $I_i$ can be computed, using an HMM, by

$$I_i = \sum_{t=\tau_i}^{t_{i+1}} \log(p(x_t | s_{-1}) p(x_t | s_t))$$

(5)

where $x_t$ and $s_t$ are the observed spectral vector and the HMM state at time $t$, respectively, $p(s_t | s_{-1})$ 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

$$L_v = \frac{1}{N} \sum_{i=1}^{N} l_i$$

(6)

where, the duration normalized log likelihood is given by $l_i = \log(l_i / d_i)$, $l_i$ 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 the non-native test sets 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 10 native (blue) and 10 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 reference 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 corresponding 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 one standard deviation of the same. If scores of more than two (out of six) 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>
<thead>
<tr>
<th>Method</th>
<th>Native (out of 10 speakers)</th>
<th>Non-native (out of 10 speakers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lobanov</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Nearey</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>ASR confidence</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

VI. DISCUSSION

It is observed that Nearey normalization method does better than Lobanov in classification of native and non-native speakers based on deviations in vowels formant space. 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 typically lower than that of the native speakers when the phone models are trained on native speakers. 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.