ACOUSTIC MODELS FOR PRONUNCIATION ASSESSMENT OF VOWELS OF INDIAN ENGLISH

Shrikant Joshi, Preeti Rao
Department of Electrical Engineering, Indian Institute of Technology Bombay, India
{shrikant, prao}@ ee.iitb.ac.in

***Abstract*— We consider the pronunciation assessment of vowels of Indian English uttered by speakers with Gujarati L1 using confidence measures obtained by automatic speech recognition. The goodness-of-pronunciation measure as captured by the acoustic likelihood scores can be effective only when the acoustic models used are appropriate for the task i.e. detecting errors in the target language (Indian English) typical of speakers of Gujarati (the source language). Thus the speech data used for acoustic model training is expected to have a prominent influence on system performance. In the absence of labeled speech databases of either the source or target language, we investigate specific combinations of acoustic models trained on available databases of American English and Hindi. It is observed that Indian English speech is better represented by Hindi speech models for vowels common to the two languages rather than by American English models. Further, adaptation with a limited amount of Indian English speech improves the system performance.**

Keywords— Computer-assisted learning; pronunciation assessment; General Indian English; Bilingual models.

I. INTRODUCTION

India, as is well known, is linguistically very diverse with at least two major language groups, the Indo-Aryan and Dravidian. English is used across India both for official and social communication and often serves as the only common language among people from different regions of the country. Since Indian language phonologies differ from each other and from that of English, different L1 interferences lead to distinct colorations that give rise to specific regional varieties of spoken English. Since English is widely taught from primary school (even when the medium of instruction is a regional language), educated speakers of Indian English do not differ from British R.P. in grammar or vocabulary but fall short in pronunciation due to insufficient exposure to the spoken form. Faulty pronunciations arise both, from the mismatch of L1 and English phonologies, and from the unusual spelling-to-sound rules that contrast with the phonemic orthography of Indian languages. Fluency in English is widely viewed as enabling access to new opportunities in a growing economy, driving thousands to English coaching institutes across the country [1].

A desirable goal for an Indian English learner is the acquisition of the standard form of spoken English known as General Indian English (GIE) which is devoid of regional influences and intelligible across the country and outside it [2]. Although it owes its origin to the colonial legacy, GIE deviates from British R.P. in its phonology and quality of phones. Rather, it takes into account the common segmental features across Indian languages to bring in a more easily acquired version that is distinctly Indian, but without prominent regional influences. As for prosody, the other important dimension of speaking skill, there is no specific prescription in GIE. It is expected to follow British R.P. more or less [2].

Automatic assessment of segmental and prosodic aspects of a learner's speech with reference to native speech can be a valuable complement to the classroom teaching of spoken language. Such a tool requires the reliable segmentation of the learner's utterance at the phone level including detection of disfluencies and the extraction of prosodic attributes such as duration and pitch at the syllable level that can then be compared with the corresponding attributes of the target speech. At the segmental level, non-native speakers tend to mispronounce words by substituting phones from their native language (L1) and also make phone insertion and deletion errors, influenced by phonotactic constraints of their L1.

The computer-assisted learning of spoken language is closely tied to automatic speech recognition (ASR) technology. The challenges of automation are linked to the known deficiencies of state-of-the-art ASR systems where phone recognition accuracies are relatively low and an acceptable performance in practical tasks is achieved only through the constraints of a powerful language model. In an application such as pronunciation assessment, however, language models would obscure genuine pronunciation errors by the non-native learner. Witt and Young [3] and Franco et al. [4] have described pronunciation scoring systems focused on measuring pronunciation quality of a non-native speaker at phone level. They used acoustic likelihood-based confidence measures for automatic pronunciation assessment within the framework of a Hidden Markov Model speech recognition system. Franco et al. [4] presented a paradigm for automatic assessment of phone quality via probabilistic scores of selected time-aligned segments with respect to trained native acoustic models. Using non-native L2 speech trained models can help the phone recognizer cope better with systemic and realizational differences between the non-native and native phones [5]. However non-native speech datasets are not easily available. Possible solutions to get around this are to use acoustic models of L1 and L2 in parallel, or a combination of L1 and L2 models, and, optionally, also include intermediate phones [5, 6].

In the present work, we consider the question of training acoustic models for likelihood-based pronunciation detection of vowels of GIE. Thus GIE is the target language (L2). A study of the speech of several proficient speakers of Indian English revealed both phonetic and phonological influences of speakers' native language on their accent in Indian English [7]. This suggests that the GIE norm must account for a certain extent of L1 transfer effects. Our work on developing a pronunciation feedback system for GIE is therefore restricted to speakers of a specific L1, namely Gujarati. We hope the methodologies developed in this work can be extended to systems for other L1 speakers of Indian English. Since this is initial work, we restrict ourselves to vowels since vowels typically show more accent discriminability compared with other phone classes [8]. In the next section we present the phonology background with a view to understanding the challenges that arise in identifying appropriate speech data for the training of acoustic models for our task. The training and test datasets are described followed by a discussion of the pronunciation scoring method and experimental evaluation of performance.

TABLE I. MAPPING OF VOWELS IN BRITISH R.P. AND AMERICAN ENGLISH WITH IPA SYMBOLS

Sr. No.	British R.P.	American English	GIE
1	iː	iː	iː
2	ɪ	ɪ	ɪ
3	eɪ (diphthong)	eɪ (diphthong)	eː
4	e	ɛ	ɛ
5	æ	æ	æ
6	ə ɜː ʌ	ə, ɜː, ʌ	ə
7	ɑː	ɑ	aː
8	ɔː	ɔː	oː
9	ɒ	ɔ	ɒ
10	uː	uː	uː
11	ʊ	ʊ	ʊ

TABLE II. MAPPING OF AE, GIE, HINDI AND GUJARATI VOWELS WITH EXAMPLE WORDS

Sr. No	AE word	AE IPA	GIE word	GIE IPA	Hindi IPA	Gujarati IPA
1	beat	iː	beat	iː	iː	i
2	bit	ɪ	bit	ɪ	ɪ	
3	bait	eɪ	gate	eː	eː	e
4	bet	ɛ	get	ɛ	---	
5	bat	æ	bat	æ	æ	
6	about, bird, butter	ʌ, ə, ɜː	cut	ə	ə	ə
7	father	ɑ	past	aː	aː	aː
8	boat	ɔː	coat	oː	oː	ɔ
9	or, golf	ɔ	caught	ɒ	ɒ	
10	boot	uː	fool	uː	uː	u
11	book	ʊ	put	ʊ	ʊ	

II. TRAINING AND TESTING DATASETS

The appropriate training of acoustic models is critical to pronunciation scoring based on acoustic likelihoods. Ideally the training data should comprise speech that matches the achievable and desired target. In our case, this would be English utterances of proficient speakers of Indian English with Gujarati L1. A database of Indian English speech is unavailable, let alone that comprised of Gujarati L1 speakers of Indian English. We discuss different possibilities to overcome this with reference to the phonologies of the languages involved.

A. Language phonologies

General Indian English: There are 11 pure vowels in GIE [2]. These correspond to 12 pure vowels plus 1 diphthong of British R.P. as shown in Table I. The closely spaced central vowels are collapsed into one in GIE. Further, vowel quality differences for the same phoneme are expected between Indian English and British R.P., of course. In view of the observation that Indian English vowels are more similar to American English (AE) vowels [7], especially the front vowels, as well as the ready availability of a labeled American English database (TIMIT), we consider its use for training the target language (GIE) models. Table I shows the corresponding phones of AE.

Gujarati: has 6 pure vowels [9]. As seen from Table II, these correspond to the collapsed forms of each of 4 sets of GIE vowels, and the two remaining GIE vowels. The collapsed phonemes are of an "intermediate" quality and when used for English pronunciation give rise to the corresponding ambiguities, i.e. confusions between long and short vowels: /iː, I/, /u, uː/ and /eː, ɛ, æ /, /ɒ, oː/ [9]. For example, the English words "snack" and "snake" uttered by Gujarati L1 speakers are often indistinguishable, much as "coat" and "caught", "beat" and "bit", and "fool" and "full".

This phenomenon of a reduced vowel set (from the original larger set of Sanskrit phonemes) is common to other Indo-Aryan languages such as Marathi, Oriya, Assamese and Bengali. Hindi (standard form), on the other hand, has a vowel system that is quite similar to English as depicted in Table II, although its consonants are very different [10]. This is attributed to the Persio-Arabic influences on the Sanskrit origin language. Considering the match between Indian English and Hindi vowel systems, as well as the availability of a labeled Hindi speech database, we also consider acoustic models for GIE that are trained on Hindi speech. Further, we investigate acoustic model adaptation with a small, specially recorded dataset of Indian English speakers as described next.

B. Training and adaptation databases

In this work we use the phonetically labeled TIMIT American English database [11] for training the AE vowel models of Table I to get our L2 (GIE) models. TIMIT comprises of 462 speakers across 8 dialect regions in the training set uttering 10 phonetically balanced sentences each. We also use a standard Hindi speech database [12] to train the 8 acoustic models selected from 10 vowels listed in Table II that are common to GIE and Hindi (the 8 Hindi vowels are from among the 10 listed in Table II, by omitting /ɒ, oː/ (bat, caught) since they have too few instances in the training data. The Hindi database is patterned on TIMIT but much smaller with 100 native Hindi speakers uttering 10 sentences each. Out of the 100 speakers, the train set of 76 speakers is used for training the acoustic models in this work.

GIE adaptation data: GIE adaptation data is collected from 12 "model" Indian English speakers (6M+6F) with 42 short sentences (3 to 5 second duration) uttered by each speaker. The model speakers of Indian English were identified by their absence of any recognizable L1 accent. These were college students proficient in spoken English living in Mumbai but of various native languages (L1) such as Marathi, Hindi, Kannada and Punjabi.

Sentence prompts from TIMIT core test set were used to elicit speech by the IE model speakers. This GIE adaptation data is manually labeled at phone level. All three databases, TIMIT, TIFR and GIE adaptation database are at 16 kHz sampling, 16-bit word length.

C. Test datasets

Our test data are word lists to elicit the 11 GIE vowel phones. Most words are monosyllabic but several have more syllables and hence contain more than one vowel. The use of words minimizes insertion and deletion errors allowing us to focus on the aspect of possible substitution of the target vowels by L1 phones. The word lists were read out by each of 20 model speakers (different from the speakers in part II B) of Indian English and 16 Gujarati-L1 speakers. Like the model speakers, the Gujarati-L1 speakers selected for testing the pronunciation assessment system were college students but who spoke English with various extents of proficiency and exhibited a preceptible L1 influence. Most of these students had been schooled in the Gujarati medium.

The English word lists each contained 11 words (one for each GIE vowel in Table II). The vowel example words are selected from [2]. Seventeen such word lists of distinct words were prepared. Each test

speaker read aloud the word lists with a reasonable pause between lists and a short pause between the words. The data was recorded using good quality microphone with 16 kHz sampling frequency and 16 bit mono wave file format in a quiet room. The test dataset details are given in Table III. The Gujarati-L1 data is manually annotated at

TABLE III. SUMMARY OF TEST DATASETS

	Database	Number of speakers	Minimum number of words / vowel
Testing database	Model-IE	20	340
	Guj-IE	16	272

TABLE IV. CONFUSION MATRIX OF GUJARATI TEST DATA: 'P' INDICATES PERCEIVED VOWEL AND 'A' INDICATE ACTUAL VOWEL

P / A	i:	ɪ	e:	ɛ	æ	ə	a:	o:	ɒ	u:	ʊ
i:	317	**17**									
ɪ	**41**	388		2		1					
e:			266	**1**	**1**						
ɛ	3	5	**24**	325	**35**						
æ			**9**	**9**	278						
ə			1	10	10	595					
a:							269				
o:								292	**29**		
ɒ							1	**18**	246		
u:										252	**17**
ʊ										**43**	221

phone level to obtain the surface transcription (that is, what is actually perceived) for each reference vowel instance. This helps us to identify the most common confusions, or mispronunciations by the Gujarati-L1 speakers of English as shown in Table IV. The major confusions observed are within the short-long vowels pairs like / iː , ɪ / and / uː , ʊ /. Equally important are confusions within the clusters /eː, ɛ, æ / and / oː, ɒ/. These observations are consistent with previous literature [9].

III. AUTOMATIC PRONUNCIATION SCORING

Phone level error detection algorithms in the literature are generally based on speech recognition confidence measures. Kim et al. [13] used HMM based acoustic log likelihood score, log posterior probability score and segment duration score (log probability of normalized duration) to compute the specific phoneme pronunciation score. Witt [14] used HMM based log likelihood score computed by taking difference of duration normalized log likelihood score of forced alignment and free decoding output for a specific phone segment. This "Goodness of Pronunciation" (GOP) measure exhibited a good correlation with human expert judges' ratings. In this study we have used Witt's GOP scoring algorithm.

A. GOP scoring algorithm

In GOP scoring method, the subject utters system-provided prompts. Thus it is assumed that the reference transcription of the test utterance is known and suitably trained HMM models are available. The alignment of the reference phone sequence with the given observation vector sequence is performed using trained HMM models to obtain the likelihood $p(O|p_j)$, where O is the observation vector and p_j is the underlying phone. With these assumptions the pronunciation score of a given phone p_j is the duration normalized log posterior probability $p(p_j|O)$. From the mathematical derivation presented in [14], this can be rewritten as follows.

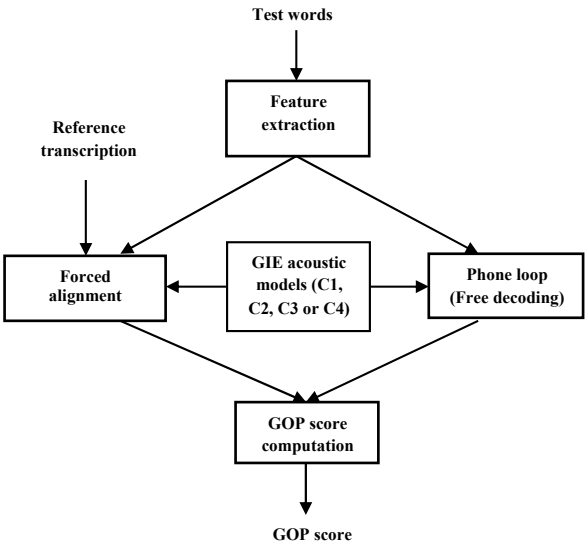

Fig.1. GOP scoring system [14]

$$GOP(p_j) = \frac{1}{NF(O)} \left| \log \left(\frac{p(O \mid p_j)}{\max_{i=1}^{I} p(O \mid p_i)} \right) \right| \quad (1)$$

Where I is the total number of phone models and $NF(O)$ is the number of frames corresponds to acoustic segment O. All the phones are assumed to be equal probable.

The implementation of the Equation (1) is as shown in the Fig. 1. The numerator is computed using forced alignment of the reference transcription using trained HMM models (also known as constrained decoding), while the denominator is computed using the phone loop (free decoding). GOP score for a given phone is the duration normalized difference between the log-likelihood score of forced alignment and phone loop. From Equation (1), it is expected that a correctly uttered phone would have the forced aligned and decoded phones matching and hence zero GOP score while badly articulated phones (with respect to the underlying acoustic model) would give rise to higher GOP scores.

B. Signal processing and feature extraction

The speech signal is pre-emphasized using filter ($1\text{-}0.97\ z^{-1}$). 12 MFCC coefficients and one normalized log energy coefficient is computed using 25 ms Hamming window and 10 ms hop size. Log energy normalization is done by subtracting maximum value of log energy in the utterance from log energy of every frame and then adding 1. Delta and acceleration coefficients are also computed using these 13 coefficients which form the 39 dimension feature vector.

C. HMM details

All acoustic models are context independent left to right 5-state HMMs with diagonal covariance Gaussian mixtures with 12 mixtures per state. First and last states are non-emitting. All the models except silence model are without skip state but silence model is ergodic with skips are allowed between state 2 and 4.

D. Training and adaptation of acoustic models

Acoustic models are separately trained from the AE and Hindi speech databases. HTK 3.4 [15] is used for all the experiments. The AE dataset is clustered into 22 broad phone classes. That is, the 61 phones of TIMIT are mapped to 11 GIE vowels, 6 diphthongs, semivowel, nasal, obstruent, silence and voice-bar which form 22 phone classes. Similarly Hindi phones are mapped into 15 classes as 8 GIE vowels, 2 diphthongs plus semivowel, nasal, obstruent, silence and voice-bar. Other than for vowels and diphthongs, broad classes are used rather than fine phone classes to reduce the mismatch between the native speech models and non-native speech test data to the extent possible and provide for better time alignment with the canonical phone sequence [16]. So each acoustic model combination consists of 22 acoustic classes.

Table V shows a count of the instances of the 11 GIE vowels present in each of the AE and Hindi databases. We note that there were too few instances of / æ and ɒ / (bat and caught**)**, although these vowels belong to the Hindi phonology. Both the model sets (AE and Hindi) are tested on the corresponding test datasets: TIMIT (168 speakers) and Hindi test (24 speakers), giving correct phone recognition of 71% and 74% respectively. With the above two distinct sets of models (AE and Hindi), we investigate four combinations as follows:

C1. AE-models only (the 11 GIE vowels are treated as a subset of the 12 AE monothongs as seen in Tables I and II)
C2. AE models adapted with GIE model speakers' data
C3. Bilingual models (i.e. 8 Hindi vowels + 3 AE vowels)
C4. Bilingual models adapted with GIE model speakers' data

The above trained acoustic models are used to perform GOP scoring on the test words data of model-IE and Gujarati-English speakers. All the four model combinations include 11 GIE vowels, 6 diphthongs and 5 broad classes as semivowel, nasal, obstruent, voice-bar and silence. The 5 broad classes are common across all four model combinations and for uniformity selected from TIFR Hindi models. The model combination C1 is 11 GIE vowels plus 6 diphthongs and 6 common broad classes mentioned above. C3 consists of 11 GIE vowels (8 TIFR Hindi vowel models plus 3 from TIMIT) plus 2 diphthongs from TIFR and 4 diphthongs from TIMIT and 5 common broad classes. Model combinations C2 and C4 are formed by adaptation of C1 and C3 respectively by GIE adaptation data. The adaptation is performed using Maximum a posteriori (MAP) approach using HTK framework [15]. A GOP is computed for each of the 11 GIE vowels using all the four model combinations C1, C2, C3 and C4 for Model-IE and Guj-IE test data set.

E. GOP score validation methodology

Qualitative assessment on the performance of the four acoustic models is carried out using scatter plot of GOP score. GOP scatter plot for Model-IE data is plotted and the spread of GOP score is used to assess the quality of the model sets. Ideally the GOP score should be concentrated near zero for the set of vowels uttered by the model IE speakers. The dispersion around zero would serve to indicate the quality of the match of the underlying acoustic models with IE speech.

The GOP scores distribution for each IE vowel in the model speakers' data also serves to normalize thresholds applied on GOP scores for the test data in order to detect pronunciation errors with respect to that vowel. That is, if there is a large spread in the model GOP scores, the threshold of acceptance of a pronunciation should be higher. We compute pronunciation error detection accuracy at various thresholds in order to obtain precision-recall curves for each acoustic model set. For a given point on a curve, the thresholds for each vowel are given by a fixed scaling applied to the zero-mean standard deviation of the GOP score distribution of the vowel.

Each test vowel instance is categorized as shown in Table VI based on its ground-truth (i.e. subjectively judged pronunciation as summarized in Table IV) and its pronunciation detection system output. In the confusion matrix, TP indicates true positive where correctly pronounced vowel/token classified correctly and TN indicates true negative where mispronounced vowel/token classified correctly as mispronounced. FN and FP are false negative and false positive respectively.

TABLE V. TRAINING TOKEN COUNTS IN AE (TIMIT) AND HINDI DATABASE CORESPONDING TO EACH GIE VOWEL

Vowel IPA symbol	Example word	Training Token count	
		TIMIT	TIFR
i:	beat	4595	1353
ɪ	big	11479	1331
e:	gate	2266	2380
ɛ	get	2919	----
æ	bat	2292	----
ə	cut	6102	4026
a:	past	2256	2875
o:	coat	1653	1037
ɒ	caught	1865	----
u:	fool	1933	244
ʊ	put	495	788

TABLE VI. CONFUSION MATRIX, WHERE CORRECT REFERS TO CORRECT PRONUNCIATION AND WRONG REFERS TO WRONG PRONUNCIATION

		Classified output	
		Correct	Wrong
Input	Correct	TP	FN
	Wrong	FP	TN

Recall and precision values for different thresholds values for each of the model sets are computed as follows. Recall and precision of true negative tokens is given by

$$Recall\_TN = \frac{TN}{(TN+FP)} \quad (2)$$

$$Precision\_TN = \frac{TN}{(TN+FN)} \quad (3)$$

IV. RESULTS AND DISCUSSION

Qualitative comments on the performance of the underlying acoustic models are possible by observations on the spread of the GOP measure. The spread is expected to be limited and close to the zero value for properly articulated vowels (i.e. those of the model-IE speakers) if the acoustic models are a good match to GIE realizations. The GOP values and spread for Gujarati-English vowels should ideally correspond with the known pronunciation errors of Gujarati L1 speakers. We make some vowel-specific observations in following sections IV-A. Objective evaluation of the underlying acoustic models on Guj-IE test data is performed using precision-

recall measures given in the section III-E and the results are discussed here.

A. Observations on model-IE speakers

From Fig. 2, the model set C2 shows less scatter as compared to model set C1 where as C4 is slightly better than C3. Overall it indicates that there is better match of GIE data with Hindi models (C3) than with C1 (American English). Further, upon adaptation by GIE data, model set C2 shows significant improvement in the GOP scatter over C1, while C4 shows only slight improvement over C3. This shows that while the AE models (C1) were not quite suitable for GIE speech, a limited amount of adaptation data achieves significant improvement. On the other hand Hindi model set C3 are a better match for GIE data and with adaptation improved only slightly. A major point emerging from these observations is the Indian English speech is better represented by Hindi speech models for the vowel phones common to the two languages rather than by AE speech models.

In particular, we note from Fig. 2-A and B that the acoustic models set C2 (AE adapted with GIE) shows the least dispersion or spread for the vowels / i:, u: and ʊ / (as in beat, fool and put). For vowel / e: / (as in gate), model set C3 (Hindi models) shows the least dispersion in GOP while / ɪ , o: / (as in bit and coat) shows best performance in adapted Hindi model set C4. Though the models for the vowels / ɛ, æ and ɒ / (as in get, bat and caught) are borrowed from AE models and used without adaptation in C3 and with adaptation in C4 model set, these vowels shows slightly better performance in adapted model set C2 than C4. This might be because of the effect of other vowel class models affects the GOP score slightly in these vowel classes. Overall, when Hindi speech data is available, GIE speech is modeled reasonably well by Hindi-trained models. Further, acoustic model adaptation with even limited amounts of IE speech improves the performance. For some vowel classes, AE models adapted by GIE speech data works well.

B. Objective evaluation of Gujarati-IE vowels

GOP scores are computed for the Guj-IE speakers' vowels using each of the acoustic model sets C1, C2, C3 and C4. Considering the task to be the detection of mispronunciations, Fig. 3 shows the precision-recall plot obtained by varying the score threshold as discussed in Sec. III-E. Since, as indicated by the subjective judgements of Table IV, Guj-IE vowel confusions are restricted to 8 vowels (i.e. the vowels /e:, ə and a: / are realized correctly). Hence the precision-recall curve is plotted in Fig. 3 considering the 8 vowels only. From Fig. 3, we note that model sets C2 and C4 shows better performance than model sets C1 and C3, confirming that adaptation with matched data improves performance. Model sets C2 and C4 perform similarly for the recall value range of 0.65 to 0.8. For the recall values between 0.35 to 0.65, the bilingual adapted models (C4) provide better precision. The precision values are low overall due the relatively low proportion (10% only) of subjectively labeled mispronunciations of the total utterances. This happened because only gross mispronunciations (i.e. clear vowel substitutions) were labeled by the judge whereas there actually were several milder mis-articulations present in the data as evident in casual listening.

The differences the in long-short vowel pairs in / i:, ɪ / and /u:, ʊ / are durational and often cannot be clearly labeled since the non-native realization tends to have an intermediate duration. On the other hand, the vowels / ɛ - æ and o: - ɒ / (as in get, bat and coat, caught) provide more prominent phonemic differences and make mispronunciation detection easier. The objective performance of the system for these four vowel classes is shown separately by the precision-recall curves

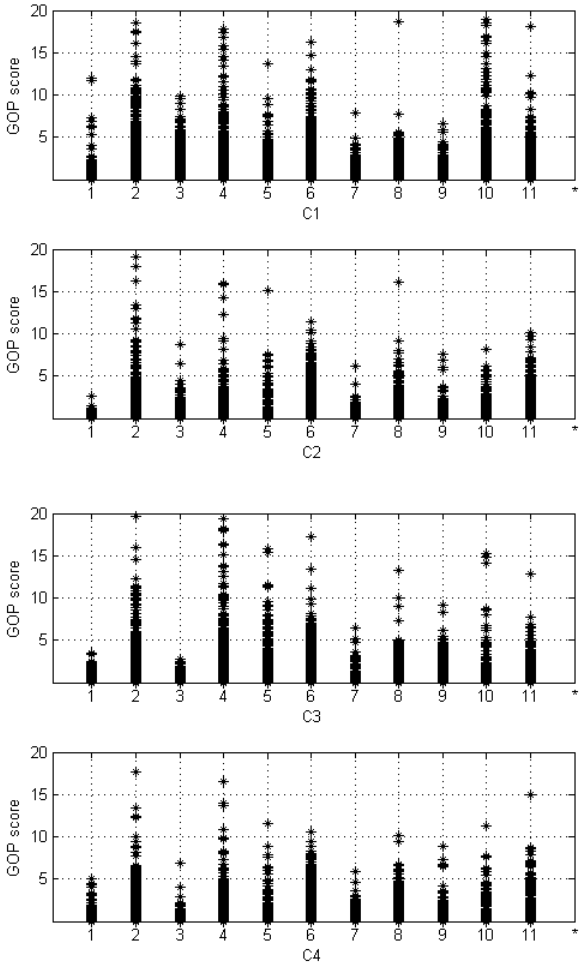

Fig. 2 The scatter of GOP values for the model-IE speakers for each of the 11 vowels arranged from front to back as in Table II using: C1- AE models, C2- Adapted AE models, C3- Adapted Bilingual models, C4- Adapted Bilingual models.

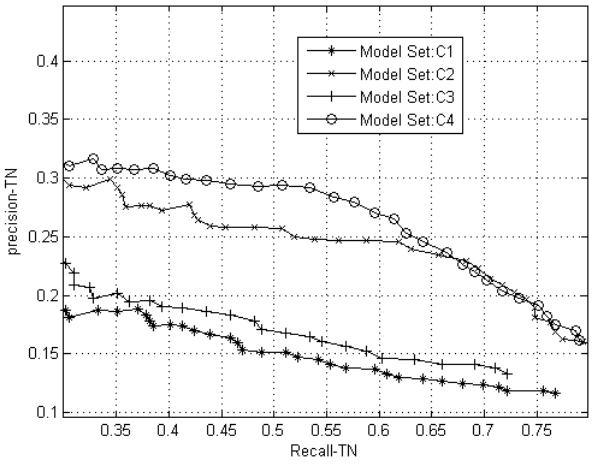

Fig. 3. Precision-recall for mispronunciation detection across 8 GIE vowels.

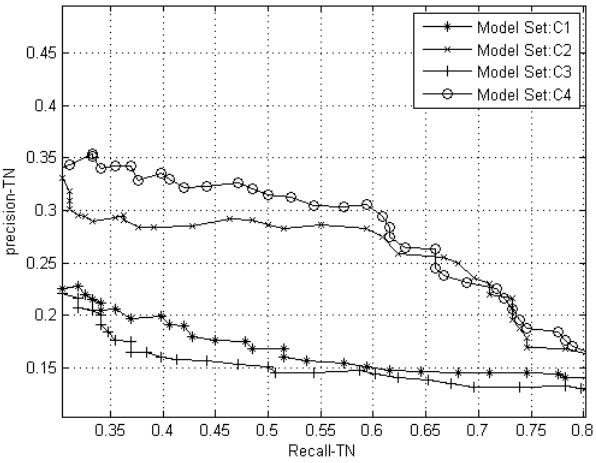

Fig. 4. Precision-recall for mispronunciation detection across 4 GIE vowels.

in Fig. 4. We see that the adapted model sets C2 and C4 shows a trend similar to Fig. 3 but the precision-recall curves for model set C1 and C3 are interchanged. This is the interesting scenario and justifies using selection of individual vowel models based on the GOP scatter plot for model-IE speaker data as shown in Fig. 2. Comparing GOP scatter plots for vowel / ɛ, æ / and / oː, ɒ /, model set C1 shows less spread than C3, which is reflected in the 4-vowel set performance.

V. Summary and Future Work

In summary, the speech data used for acoustic model training has a prominent influence on the performance of an acoustic-likelihood measure for goodness of pronunciation of non-native speech. In the scenario considered in this paper, suitable training and adaptation of acoustic models with available databases of languages (American English and Hindi) other than the source (Gujarati) or target (Indian English) language were used to achieve a pronunciation scoring system that could predict typical error patterns from acoustic likelihoods of IE vowels uttered by Gujarati L1 speakers. Performance improvements were obtained by adapting the AE and Hindi models using a relatively small dataset of model IE speech. Considering the prominent phonological differences between Gujarati and GIE, future work will address the added use of Gujarati phone models to obtain more accurate decoding of Gujarati-English speech for GOP score computations. The well known confusions in Gujarati-English are within the groups, /iː, ɪ /, /eː, ɛ, æ/, /oː, ɒ/ and /uː, ʊ/ which suggests that it would be interesting to investigate the discriminative training of acoustic models. Efforts are also underway to collect larger and more balanced datasets for the study of pronunciation error detection.

Acknowledgement

The authors would like to thank Prof. Peri Bhaskararao for valuable discussions on the language phonologies.

VI. References

[1] D. Nagpal-D'souza and I. Akbar, "Spoken in English Vinglish", Eye Magazine, Indian Express, Oct.7, 2012.

[2] R. K. Bansal and J. B. Harrison., "Spoken English", Orient Blackswan Private Limited, Mumbai, 2009.

[3] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and accessment for interactive language learning", Speech Communication, 30(2-3), pp. 95-108, Feb. 2000.

[4] H. Franco et al. 2000. Automatic scoring of pronunciation quality. Speech Communication, Vol. 30, pp. 83-93, 2000.

[5] G. Kawai and K. Hirose, "A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training", In: Proc. ICSLP 1998, Sydney, Australia,Nov.30-Dec.4,1998.

[6] C. Bhat, K. L. Srinivas and P. Rao, "Pronunciation scoring for Indian English learners using a phone recognition system", In: Proc. IITM'10,pp. 135-139,2010.

[7] C. R. Wiltshire and J. D. Harnsberger, "The influence of Gujarati and Tamil L1s on Indian English: a preliminary study", World Englishes, 25(1), pp. 91-104, 2006.

[8] P. Angkititrakul and J. H. L. Hansen, "Use of Trajectory Models for Automatic Accent Classification", In: Proc. Interspeech 2003, Geneva, Switzerland, Sept. 1-4,2003.

[9] K. B. Vyas, " A Comparative study of English and Gujarati phonological systems", Ph.D. thesis, Saurashtra Uni. English Dept, Rajkot, 2010.

[10] http://en.wikipedia.org/wiki/Hindustani_phonology

[11] J. S. Garofolo et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus", Linguistic Data Consortium, Philadelphia, 1993.

[12] K. Samudravijaya, P. V. S. Rao, S. S. Agrawal, "Hindi speech database", In: Proc. ICSLP 2000, Beijing, China, October 16-20, 2000.

[13] Y. Kim, H. Franco and L. Neumeyer, "Automatic pronunciation scoring of specific phone segment for language instructions", In. Proc. Eurospeech, Rhodes, Greece, 1997

[14] S. M. Witt, "Use of speech recognition in computer-assisted language learning", unpublished thesis, Cambrige Uni. Eng. Dept, 1999.

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

[16] V. Patil, S. Joshi and P. Rao, "Improving the robustness of phonetic segmentation of to accent and style variation with a two-staged approach", In: Proc. Interspeech 2009, Brighton, UK, Sept. 6-10,2009.