# Automatic pronunciation assessment for language learners with acoustic-phonetic features

*Vaishali Patil, Preeti Rao*
Department of Electrical Engineering,
Indian Institute of Technology Bombay,
Mumbai, India.
{vvpatil, prao}@ee.iitb.ac.in

ABSTRACT

Computer-aided spoken language learning has been an important area of research. The assessment of a learner's pronunciation with respect to native pronunciation lends itself to automation using speech recognition technology. However phone recognition accuracies achievable in state-of-the-art automatic speech recognition systems make their direct application challenging. In this work, linguistic knowledge and the knowledge of speech production are incorporated to obtain a system that discriminates clearly between native and non-native speech. Experimental results on aspirated consonants of Hindi by 10 speakers shows that acoustic-phonetic features outperform traditional cepstral features in a statistical likelihood based assessment of pronunciation.

KEYWORDS : acoustic-phonetic features, language learners, pronunciation, aspirated stops.

*Proceedings of the Workshop on Speech and Language Processing Tools in Education*, pages 17–24,
COLING 2012, Mumbai, December 2012.

## 1 Introduction

Fluency in spoken language by a language learner must be judged based on the achieved articulation and prosody in comparison with that of native speakers of the language. The articulation is influenced by how well the learner has mastered the pronunciation of the phone set of the new language as well as the usage of the phones in the context of the words. Language learning is an important activity and automating aspects of it via speech recognition technology has been an area of recent research globally (Strik et al., 2007). One aspect of spoken language learning that lends itself to such automation is the assessment of pronunciation. The phones uttered by the speaker can be compared to their native acoustic forms to provide corrective feedback about the extent and type of errors. Automatic speech recognition (ASR) technology would seem to provide the solution to automatic pronunciation error detection by its ability to decode speech into word and phone sequences and provide acoustic likelihood scores indicating the match with trained native speech models. However, state-of-the-art ASR systems fare poorly on phone recognition accuracy unless aided by powerful language models. In an application such as pronunciation assessment, language models would obscure genuine pronunciation errors by the non-native learner. Further, for better raw phone recognition accuracy, the acoustic models need to be trained on actual non-native speech. Such a speech database is unlikely to be available.

A way to deal with the problem of poor phone recognition accuracies from ASR is to exploit any available knowledge about the type of pronunciation errors. It is observed, for instance, that the errors made by a non-native speaker learning the language (L2) tend to be heavily influenced by his own native tongue (L1). These phone segment level errors arise from (1) the absence of certain L2 phones in the L1 leading to phone substitutions by available similar phones, and (2) phonotactic constraints of L1 leading to improper usage of phones in certain word-level contexts. A knowledge of the common phone-level errors in the non-native speech can help to reduce the search space in speech decoding thus improving the phone recognition accuracy. However, since the phone errors typically involve phone substitution by closely matching phones borrowed from the speaker's L1, the discrimination is more challenging. Proper feature design in the acoustic space can contribute to improved discrimination between different phones that otherwise share several articulatory attributes. In the present work, we investigate the design of acoustic features in the context of specific pronunciation errors made by learners of spoken Hindi. We restrict ourselves to speakers whose L1 is Tamil. This language-pair provides prominent examples of phone substitution errors arising from the differences in the phonetic inventories of languages of two distinct language groups viz. Indo-Aryan (Hindi) and Dravidian (Tamil). Although the reported work is restricted to a specific type of pronunciation error, namely that relating to aspirated stops, the methodology presented in this paper can be usefully generalised.

In the next section, we describe the task of pronunciation assessment in the context of the chosen languages and the design of databases for training and system evaluation. Acoustic-phonetic features are described next followed by an experimental evaluation involving traditional ASR features as well.

## 2 Database design

The pronunciation assessment is carried out by specially constructed word lists that are read out by the language learner. The recorded utterances are processed by a speech recognition system to determine the pronunciation quality of each of the phones of interest with respect to native speech phone models. With our focus on Tamil learners of spoken Hindi, we should ideally have access to labelled data of Hindi speech from native Hindi speakers as well as from non-native (Tamil) speakers in order to build phone models for the automatic phone recognition system. However, a sizeable database of non-native speech is an unavailable resource. Linguistic knowledge about the phonetic inventories of the two languages can help to overcome this problem partially with the substitute phone models trained on more easily available native speech databases (Bhat et al., 2010).

In the present work, we investigate the pronunciation of consonants, in particular, the unvoiced stops of Hindi. The Tamil and Hindi phonetic inventories differ significantly in this category as seen in Table 1 (Thangarajan et al., 2008). Due to the absence of aspirated unvoiced stops in Tamil, learners of Hindi whose L1 is Tamil tend to substitute these phones with the closest available phone viz. the same place-of-articulation unvoiced unaspirated stop. Table 2 provides some example words with their observed pronunciations by native and non-native speakers. Assuming that the articulation of the unvoiced unaspirated stops is the same in both languages for a given place-of-articulation, we can use native Hindi speakers' speech for training the phone models for both native and non-native speakers. In the present work we use an available dataset of Marathi speech (another Indo-Aryan language that shares nearly all the phones of Hindi) as training data. The Marathi database comprises the word utterances from 20 speakers where the words cover all consonant-vowel (CV) occurrences in the language.

| Stop categories / PoA | Hindi/Marathi | | Tamil | |
|---|---|---|---|---|
| | Unaspirated | Aspirated | Unaspirated | Aspirated |
| Labial | p | $p^h$ | p | - |
| Dental | t̪ | t̪$^h$ | t̪ | - |
| Retroflex | ʈ | ʈ$^h$ | ʈ | - |
| Velar | k | $k^h$ | k | - |

TABLE 1 – IPA chart showing difference in unvoiced stops of Hindi/Marathi and Tamil.

| Word | Meaning | Articulation by native speaker | Articulation by non-native speaker |
|---|---|---|---|
| खामोशी | Silence | k$^h$AmoshI | kAmoshI |
| ठिकाना | Location | ʈ$^h$ikAnA | ʈikAnA |
| तारिका | Starlet | t̪ArikA | t̪ArikA |
| पेशा | Profession | peshA | peshA |

TABLE 2 – Examples of unvoiced stops in word initial as articulated by native and non-native speakers

The test data comprises of recordings of Hindi words uttered by 5 native (3M and 2F) and 5 non-native (3M and 2F) Hindi speakers. The words list comprises each unvoiced stop consonant 'C' in word initial position with 8 words per consonant. These words are embedded in two Hindi sentences (one statement and one question form) acting as carrier phrases. The native speakers had good fluency in Hindi. In case of the non-native speakers (familiar with Hindi but not using it regularly), we focussed on Tamil L1speakers. They could read the Devnagiri script of Hindi having studied the language formally in school. The speakers were asked to go through the word list (Devnagiri spelling with English meaning) mentally before recording to ensure that they have no difficulty in reading. All recordings were made at 16 kHz sampling rate.

## 3 Proposed pronunciation assessment system

Given the recorded words from a test speaker, the system should quantify the “distance” of a realised phone from the canonical model of the intended phone.  An overall measure of pronunciation quality can then be obtained by summarising the distances over the entire word list along with feedback on specific pronunciation errors. A widely used distance measure in pronunciation assessment is a normalized acoustic likelihood score obtained within a statistical speech recognition framework (Strik et al., 2007). The effectiveness of this measure depends on the acoustic features used to derive the observation vector. Standard HMM-based systems use MFCC features representing the spectral envelope of speech sounds. Such systems are known to confuse phones within the same broad manner classes and depend heavily on language modelling for acceptable speech recognition accuracy. On the other hand, research by speech scientists over the years has suggested that acoustic-phonetic cues obtained by the understanding of speech production can be usefully applied to discriminate phones that differ in a single attribute (Niyogi and Ramesh, 2003; Truong et al., 2004). In the next section, we present acoustic-phonetic features for discriminating aspiration in stop phones. These features are incorporated in a statistical phone recognition system. For comparison, a standard MFCC-based system is also implemented and evaluated on the same task. Both systems share a common first stage of broad-class segmentation (by aligning with the known transcription from the word list) with a 3-state diagonal covariance HMM-based system using the 39-dim MFCC vector (Young et al., 2006). The segmented unvoiced stops so obtained are used in pronunciation assessment of the aspiration attribute separately with each of two feature sets, viz. acoustic-phonetic and MFCC.

## 4 Acoustic-phonetic (AP) features for aspiration detection

The production of aspirated stops is similar to that of the same-place unaspirated stops but for the additional glottal aspiration that accompanies the onset of the following vowel.  The main differentiating acoustic cues are either durational in terms of the longer voicing onset time (VOT) in the case of aspirated CV, or spectral, by way of the presence of cues associated with breathy voice quality. The features are presented next (Patil and Rao, 2011).

Frication duration: is the duration between the burst release and voicing onset. These landmarks are determined by refinement of the segment boundaries obtained in the HMM based broad class segmentation (Patil et al., 2009).

Difference between the first and second harmonic (H1-H2): is an indicator of breathiness since amplitude of the first harmonic relative to that of the second is proportional to the open quotient of glottis. It is computed from the short-time spectrum of a 25 ms window and averaged over 5 frames in the region 6 to 10 ms from the vowel onset.

Spectral tilt A1-A3: is the difference between the strongest spectral component in [100, 1000 Hz] and the one in [1800, 4000 Hz] thus capturing the difference of the energy between the first and third formant regions. Breathy voices are characterised by greater spectral tilt and hence greater A1-A3. It is computed by the same short-time spectrum averaging as used for H1-H2.

Signal-to-noise ratio: The superimposed aspiration leads to a lower harmonics-to-noise ratio in the vowel region immediately following an aspirated stop. This can be accurately estimated by a cepstrum based SNR computed over duration 25 ms starting 6 ms from the detected vowel onset.

B1-band energy: It is computed in the band of 400 Hz to 2000 Hz from the average power spectrum of vowel region using 25ms window for 5 frames every 1ms from vowel onset point. Lower values of this parameter characterize the presence of breathiness due to aspiration.

## 5 Experiments and results

The two systems using two different acoustic feature sets are used to obtain the statistical likelihood for the presence or absence of aspiration for each unvoiced stop across the entire list of word utterances by a test speaker. The MFCC-based system uses the HMM phone recogniser in a 2-class (aspirated/unaspirated) forced alignment mode on the unvoiced stops using 39 MFCCs. The AP system uses a 5-dimensional feature vector in a GMM framework.

A likelihood ratio distance measure is computed using equation (1) (Niyogi and Ramesh, 2003).

$$d(x) = \log\left(\frac{L(x \mid \wedge 1}{L(x \mid \wedge 2}\right) \qquad (1)$$

where $L(x| \wedge 1)$ is the likelihood of a test point x in the observation space for model of class 1 (likewise $L(x| \wedge 2)$ for class 2). Here class1 refers to unaspirated stops and class2 to aspirated stops. In case of proper articulation, d(x) is expected to be greater than zero for unaspirated stops and less than zero for aspirated stops.

For each test speaker, we compute the distribution of the likelihood ratios computed across the speaker's set of *intended* unaspirated stops and also across the set of *intended* aspirated stops. If the stops are all properly articulated, we expect a good separation of the two distributions. Fig. 1 show the distributions obtained for each of the 10 native and non-native speakers using the AP features system. We note the prominent difference in the extent of overlap between the likelihood ratios in the case of native speakers with respect to that of non-native speakers. Fig. 2 shows the corresponding results for the MFCC feature system. While there is a difference in the overlap observed for the non-native speakers, the distinction between native and non-native speakers is much more clear across speakers with the AP features.

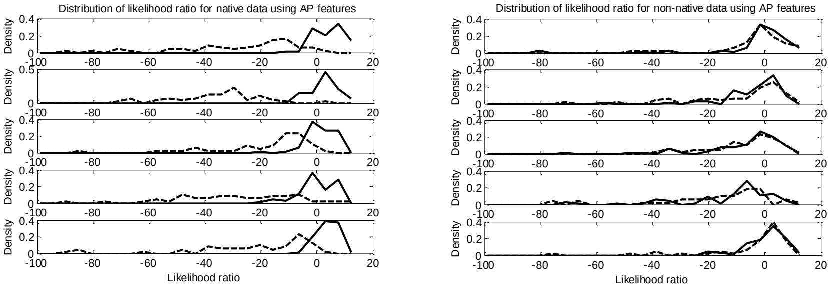

FIGURE 1 – Speaker wise distribution of likelihood ratio for native and non-native data using AP cues (solid line: intended unaspirated; dashed line: intended aspirated)

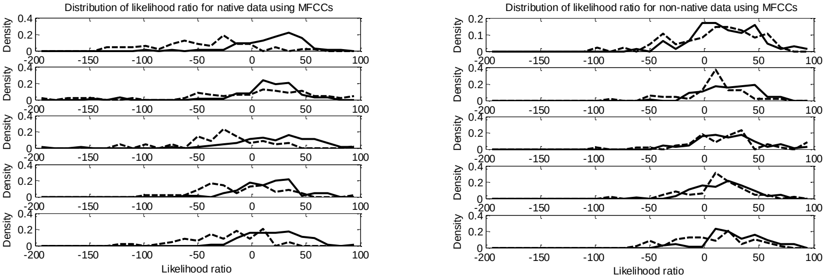

FIGURE 2 – Speaker wise distribution of likelihood ratio for native and non-native data using MFCCs (solid line: intended unaspirated; dashed line: intended aspirated)

| Native test set | | | Non-native test set | | |
|---|---|---|---|---|---|
| Speaker no. | AP | MFCCs | Speaker no. | AP | MFCCs |
| 1 | 132.79 | 79.66 | 1 | 0.01 | 2.11 |
| 2 | 373.42 | 2.12 | 2 | 3.3 | 11.38 |
| 3 | 76.57 | 12.89 | 3 | 0.3 | 0.29 |
| 4 | 113.87 | 23.09 | 4 | 0.56 | 1.08 |
| 5 | 74.72 | 67.88 | 5 | 6.91 | 14.41 |

TABLE 3 – Speaker wise F-ratio of unaspirated-aspirated likelihood ratio for native and non-native test sets.

The difference between the performances of MFCC and AP features in the task of detecting non-native pronunciation can be understood from the values of F-ratios across the 10 speakers in Table 3. The F-ratio is computed for the pair of corresponding of unaspirated-aspirated likelihood ratio distributions for each speaker and each feature set. A larger value of F-ratio indicates a better separation of the particular speaker's aspirated and unaspirated utterances in the corresponding feature space, which may be interpreted as higher intelligibility. We see from Table 3 that this intelligibility measure takes on distinctly different values in the case of the AP feature based system, and consequently an accurate detection of non-nativeness is possible. In the case of the MFCC features, however, there is no clear threshold separating the F-ratios of non-native from native speakers.

To summarise, we have proposed a methodology for evaluating pronunciation quality in the context of a selected phonemic attribute. It was demonstrated that acoustic-phonetic features provide better discriminability between correctly and incorrectly uttered aspirated stops of Hindi compared with the more generic MFCC features. Future work will address other phonemic attributes while also expanding the dataset of test speakers.

## References

Bhat, C., Srinivas, K. L. and Rao, P. (2010). Pronunciation scoring for language learners using a phone recognition system. *Proc. of the First International Conference on Intelligent Interactive Technologies and Multimedia 2010*, pages 135-139, Allahabad, India.

Niyogi, P. and Ramesh, P. (2003). The voicing feature for stop consonants: recognition experiments with continuously spoken alphabets. *Speech Communication*, 41: 349-367.

Patil, V., Joshi, S. and Rao, P. (2009). Improving the robustness of phonetic segmentation to accent and style variation with a two-staged approach. *Proc. of Interspeech 2009*, pages 2543-2546, Brighton, U.K.

Patil, V. and Rao, P. (2011). Acoustic features for detection of aspirated stops. *Proc. of National Conf. on Communication 2011*, pages 1-5, Bangalore, India.

Strik, H., Troung, K., Wet F. and Cucchiarini, C. (2007). Comparing classifiers for pronunciation error detection. *Proc. of Interspeech 2007*, pages 1837-1840, Antwerp, Belgium.

Thangarajan, R., Natarajan, A. and Selvam, M. (2008). Word and Triphone Based Approaches in Continuous Speech Recognition for Tamil Language. *WSEAS Transactions on Signal Processing*, 4(3): 76-86.

Truong, K., Neri, A., Cuchiarini, C. and Strik, H. (2004). Automatic pronunciation error detection: an acoustic-phonetic approach. *Proc. of the InSTIL/ICALL Symposium*, 2004, pages 135–138, Venice, Italy.

Young, S. et al. (2006). The HTK Book v3.4. Cambridge University.