HINDI SPEECH DATABASE
Samudravijaya K1, P.V.S. Rao2, and S.S. Agrawal3

ABSTRACT
The design and development of an annotated and time-aligned speech database for Hindi language is described here. Although this continuous speech database is principally intended for training of a speech recognition system for Hindi, the design specifications of the database are general so that it can also be useful in tasks such as speaker recognition, study of acoustic-phonetic correlates of the language. The database consists of a total of 500 sentences spoken by 50 speakers. There are two sets of sentences. The first set of 2 sentences (containing most Hindi phonemes) was read by each and every speaker. The second sets of sentences (8 distinct sentences per speaker) were designed such that they collectively cover most phonemic contexts. The database is comprehensive enough to effectively capture phonetic, acoustic, intra-speaker and inter-speaker variabilities in Hindi speech. The speech data was simultaneously recorded using a close talking microphone and another desktop “far field” microphone. The former speech data was manually segmented and labeled in terms of sub-phonetic units by trained personnel.

The database was used to conduct a study of the prosodic characteristics of the Hindi vowels. There are five pairs of vowels in Indian languages; one member is longer in duration than the other. It was observed that native speakers of Hindi seem to give more importance to the duration attribute to contrast vowels in a vowel pair than non native speakers.

1 INTRODUCTION
Large vocabulary continuous speech recognition systems are designed to classify speech sounds in terms of selected subword units. Thanks to language specific effects, existing foreign language databases cannot be used ‘as is’ for Indian languages. Large, annotated, time aligned speech databases are required to train such systems. This paper describes the design and development of such a database, comprising of phonetically rich Hindi sentences, spoken continuously by multiple speakers, segmented and labeled in terms of acoustic-phonetic units.

The rest of the paper is organized as follows. In section 2, major characteristics of Hindi language are presented and the consequent choice of units of representation of speech are discussed. The desired characteristics of the sentence corpus, and the design strategies adopted to achieve the goal are dealt with in section 3. Data collection, segmentation and labeling processes are also briefly described here. The observations of a study of the prosodic characteristics of the Hindi vowels are presented in section 4. A summary is given in section 5.

2 SOUND UNITS
The choice of the units of representation should reflect the primary purpose of the database, viz., to support the development of Hindi speech recognition system. Normally spoken sentences are represented as sequences of phonemes, which are defined based on their linguistic relevance in a language. However, a phoneme (or even a phone) need not be an acoustically homogeneous unit. In the context of machine recognition of speech, acoustic compactness of the units of representation is important because it is the spectral similarity of different realizations of a unit which enables a recognition system to infer its identity. Thus the units should be acoustically homogeneous and fairly distinct from each other. These units need not have one-to-one correspondence with the phonemes; it is, however, desirable that there is a fair degree of association between these units and the phonemes of the language. Hence it is worthwhile to delve into the acoustic-phonetic features of Hindi.

2.1 Acoustic-phonetic Features of Hindi
The acoustic-phonetic profile of Hindi (and other Indian languages) differs considerably from that of European languages. The Hindi alphabet is shown in Table 2.1. It has three sections: the first section lists the vowels, the second section lists phonemes whose production involves complete closure of oral tract (plosives, affricates and nasals); the third section lists the semivowels and fricatives. Each cell in the figure represents a phoneme and has 3 rows: the first row is the Devanagari script, the the corresponding IPA symbol, and the third the roman script used to label the phoneme in a spoken Hindi sentence.
Aspiration is a phonemic feature in Hindi, unlike English, for example. There are 8 aspirated plosives and 2 aspirated fricatives in addition to their unaspirated counterparts. Retrusion is another feature which occupies a prominent place in the Hindi alphabet. Intentional nasalization of vowels carries linguistic information. Many intervocalic /r/ and retroflex plosives (in non-geminated context) manifest as taps or flaps. These factors was taken into account while choosing the units of segmentation.

2.2 Label Inventory

The third row in each cell of figure 2.1 shows the ASCII script used to represent segments corresponding to the phoneme in the cell. Some phonemes are represented by more than one symbol, each symbol corresponding to a distinct acoustic segment of the phoneme. For example, a prevoice is represented as a closure followed by a segment comprising of burst, friction, and aspiration, if any. This convention is similar to that of the TIMIT database [1]. In addition, a scheme of suffixes were employed to denote the nasalization and flapping. Special symbols were used to represent devoicing of a voiced closure, voicing of an unvoiced closure, voiced glottal fricative (which is predominant in Hindi), tongue clicks, glottal stops, voiced segments (without formant structure) which sometimes occur between a vowel and silence, and foreign vowels used while speaking words of English. The closure of a geminated stop is indicated by repeating the closure label. Table 2 lists such special symbols along with a brief explanation of their usage.

<table>
<thead>
<tr>
<th>symbol</th>
<th>explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>glottal stop</td>
</tr>
<tr>
<td>lv</td>
<td>voiced /l/</td>
</tr>
<tr>
<td>vb{g,j,D,d,b}</td>
<td>voice bar</td>
</tr>
<tr>
<td>cl{k,c,T,t,p}</td>
<td>unvoiced closure</td>
</tr>
<tr>
<td>cl{g,j,D,d,b}</td>
<td>devoiced closure of a voiced stop</td>
</tr>
<tr>
<td>vb{k,c,T,t,p}</td>
<td>voicing in closure of unvoiced stop</td>
</tr>
<tr>
<td>cls</td>
<td>epiphonic closure</td>
</tr>
<tr>
<td>clp</td>
<td>closure of geminated /p/</td>
</tr>
<tr>
<td>sil</td>
<td>silence, pause</td>
</tr>
<tr>
<td>vbl</td>
<td>leading or trailing voicebar</td>
</tr>
</tbody>
</table>

Table 2: Symbols used to represent special characteristics of acoustic segments in Hindi utterances. Some of these symbols are used as suffixes to the symbols in Figure 2.1.

3 DESIGN OF SENTENCES AND DATA COLLECTION

Training of a speech recognition system needs a large number of labeled segments to account for large variations in the acoustic manifestations of phonemes due to factors such as variations of phonemic context, and differences between speakers and dialects. Hence, the sentences to be spoken by the subjects should (a) preferably contain all the phonemes, (b) be rich in phonemic context, and (c) aid the identification of dialect/accent. The sentences should be syntactically valid, meaningful, natural, simple and short. The design of phonetically rich sentences is summarized below. A detailed account can be found in [2].

It was decided that each subject will speak 10 sentences consisting of two parts. The first part consists of two 'dialect' sentences which preferably contain all
the phonemes of the language. These two sentences will be spoken by each and every speaker. The second part consists of eight sentences which cover as much phonetic context as possible. Thus, the second part is different for each speaker. Then, the speech database as a whole is likely to cover a large set of diverse phoneme contexts.

3.1 Design strategy

The 100 sets of Hindi sentences to be spoken by 100 speakers were chosen from a corpus of machine readable Hindi text from diverse sources. Only short sentences representable by less than 80 ASCII characters were retained. A text-to-phoneme(TTP) program was used to derive the phonetic sequence associated with the Hindi sentences after taking into account the morpho-phonemic rules of Hindi. Several design strategies were examined and the most successful one is used as described below.

**Dialect sentences:** A computer program was written to select a pair of sentences which contains maximum number of distinct Hindi phonemes. These dialect sentences should minimally contain speech sounds of all manners and places of articulation. Also, (a) Phonemes /陂/ and /陂/ should be present as retroflexion is a special feature of Indian languages. (b) The acoustic-phonetic attributes of the missing phonemes, (i.e., voicing, aspiration etc.) should be uniformly distributed. The selected pair of sentences contain all phonemes except /陂/, /陂/, /陂/, and /陂/. Also, the set has two nasalized sounds and a flap as well. The two dialect sentences are as follows:

भोजन जब आँकर उठती तो देखती कि चौका साफ पड़ा है और बर्तन में रहे हुए हैं।

यहाँ से लगभग पांच मील दक्षिण पश्चिम में कट्टर गांव है।

**Sentences with rich phonetic context:** Since a set of eight sentences cannot cover all phonetic contexts, it was decided to cover as many pairs of broad acoustic classes (BAC) of phonemes as possible. These BACs were {front, middle, back) vowels, {velar, retroflex, dental, labial) plosives, affricates, nasals, glides, liquids, fricatives and silence. It is desirable that each set of eight sentences contains as many distinct BAC pairs as possible. The most successful design algorithm was to form a set of 7 sentences containing rare BAC pairs, first. Then, choose a sentence which contains as many BAC pairs as possible which are missing in the first 7 sentences. Each of the 100 sets of 8 sentences generated by this algorithm [3] covers about 77% of the broad acoustic contexts despite the fact that certain contexts are not permitted by Hindi language.

3.2 Data Collection

100 speakers read the sets of 10 phonetically rich sentences. The speech data was digitally recorded using two microphones in a quiet room. A close talking microphone was kept at a distance of about 3 cm from the mouth. Another directional microphone was mounted on a desk at a distance of 1 metre; this data would be influenced by the room acoustics. The rationale for the simultaneous recording was that only the speech data collected with the close talking microphone would be segmented and labeled; the same information for the far field microphone data can be easily obtained by a simple shift of the time indices. The data was sampled (at 16kHz) and digitized (with 16 bits). In the first phase, half the data were segmented and labeled. The speech data, recorded using close microphone, was hand segmented in terms of units described in section 2 and labeled by trained personnel using visual displays of speech such as waveform, spectrogram.

4 VOWEL DURATIONS

Acoustic-phonetic studies of the sounds of a language are useful for understanding and capturing systematic variations occurring in natural speech. This should lead to better modeling of recognition units for speech recognition as well as synthesis of high quality speech. In general vowels occur more frequently than consonants. Hence a study of vowels using a moderate size database is likely to be statistically more meaningful than that of consonants. Here we report the preliminary results of a study of the prosodic characteristics of the Hindi vowels. The 10 vowels of Indian languages occur in pairs as seen in Table 2.1; in addition to minor spectral differences, one member is longer in duration than the other. The mean and standard deviation (SD) of durations of the 10 Hindi vowels are shown in Table 3. The number of occurrences of each vowel is also shown in the last column.

We were curious to examine whether the durational characteristics of Hindi vowels spoken by native Hindi speakers exhibit any systematic behaviour distinct from those of vowels spoken by non-native speakers. Hindi is the mother-tongue of 38 speakers and the rest spoke Hindi fluently. The heights of the white and black bars in Figure 1 show the mean durations of vowels spoken by native and non-native speakers of Hindi respectively. It may be noted that the difference in durations of O and o is small in case of vowels spoken...
by non-native speakers while the difference is significant in case of native speakers. Similar behaviour is observed in case of vowels pairs [E, e]. This behaviour is better demonstrated in Figure 2 where the difference in mean durations of short and long vowels normalized by the sum of the standard deviations is shown as a bar graph. For example, the height of the right most black bar is the quantity $(\mu_1 - \mu_2)/\{(\sigma_1 + \sigma_2)\}$ where $\mu_1$, $\mu_2$, $\sigma_1$, $\sigma_2$ are the means and standard deviations of O and o spoken by non-native speakers respectively. Similar behaviour was observed even in case of vowels excised exclusively from dialect sentences where the phonemic contexts of vowels are identical for all speakers. Thus, when duration of the shorter vowel of an vowel pair is intrinsically high (i.e., in case of vowels o and e), native speakers appear to emphasize the duration attribute of vowel for maintaining the phonemic distinction between members of an vowel pair more than non-native Hindi speakers.

5 SUMMARY

A general purpose, multi-speaker, continuous speech database was developed for Hindi language. The sentences have been segmented and labeled in terms of sub-phonetic units which take into account the special acoustic-phonetic features of Hindi language. A preliminary study indicates that native Hindi speakers emphasize the duration attribute to contrast vowels in a vowel pair more than non-native speakers.

Acknowledgements

The contributions of A.K.Dhamija and S.S.Sarma in data collection, and of Anand, Preetam, Vijaya, Shoeb, Dinesh, Prashant, and Girish in labeling effort are acknowledged. This project was sponsored by the Department of Electronics, Government of India.

References

