DELIVERABLE IDENTIFICATION

Identification number	King-ASR-389
Type	Technical Report
Title	Definition of Corpus, scripts, standards and Specifications of environment/speaker coverage for Hindi languages
Status	
Date	2016-02-14
Version	1.0
Number of pages	11
Author	Hui WANG, Qiong SONG, Ke LI, Dr. Yufeng HAO
Business contact point:	SpeechOcean China T/A: Beijing Haitian Ruisheng Science Technology Ltd Tel: +86-10-62660928; +86-10-62660053 ext.8080 Email: contact@speechocean.com Website: www.speechocean.com Add: D-801, U-center Building, No.28 Chengfu Road, Haidian District, Beijing, China (100083)
Supplementary notes	
Key words	Desktop Speech, Conversational Speech, ASR database, contents, design, description, speaker coverage.
Abstract	This document provides a specification of the contents, speaker and environment coverage of the conversational speech database collected over the headset microphone for the Hindi language.
Status of the abstract	Public

Contents

List of Tables

List of Figures

1. Introduction

This is a Hindi conversational speech database, which is collected over the headset microphone. This database is owned by Beijing Haitian Ruisheng Science Technology Ltd (SpeechOcean, www.speechocean.com)

The corpus contains recordings of daily spontaneous conversational speech which were from 1581 speakers, and performed in diverse environments, such as office, home, restaurant and street.

Speakers were recruited in batches of 4, and paired up as per the following scheme (for example, speakers MaleA, MaleB, FemaleC and FemaleD); recordings were scheduled as follows, giving 60-clock-minutes conversation per speaker:

- Two 15-clock-minute conversations between MaleA and MaleB;
- Two 15-clock-minute conversations between FemaleC and FemaleD;
- Two 15-clock-minute conversations between MaleA and FemaleC;
- Two 15-clock-minute conversations between MaleB and FemaleD.

The total recording time is around 739 hours (1-channel) including reasonable short pause during the conversation. The total size of the data base is about 168G.

1.1 Speech File Format

The utterance waves of each channel are stored as 16 KHz, 16 bit, Mono channel; Windows uncompressed PCM format. All the wave files are stored in \DATA\WAVE directory.

1.2 Directory Structure

The wave files are under the WAVE folder, and their names are defined as “HI<speakerID>_HI<speakerID>_<sessionID>_<roleID>.wav”, where each part is defined in Table 1-1.

E.g. HI0001_HI0002_1_A.wav

<speakerID>	Defined as <nnnn>; Where <nnnn> is a number from 0001 to 2651 (numbers were randomly used), and there are two speakers for each conversation.
<sessionID>	Defined as the number 1 or 2.
<roleID>	Headset<A/B> Each speaker wear a headset during the conversation, so there are two recording audio files for each conversation.

	HeadsetA: the volume of the speaker A is loud and clear, the volume of speaker B is very low and sometime inaudible. HeadsetB: the volume of the speaker B is loud and clear, the volume of speaker A is very low and sometime inaudible.

Table 1-1 Naming Description of Audio Files

1.3 Transcription Files

The speech audio files were transcribed manually after recording. The transcription files are under \DATA\SCRIPT folder, named as “HI<speakerID>_HI<speakerID>_<sessionID>_<roleID>.wav.TextGrid”. All of the transcription files were transcribed by native speakers, which also contain the noise labels and reflect the conversation content.

An example is shown as below:

```
File type = "ooTextFile"
Object class = "TextGrid"

xmin = 0
xmax = 969.24
tiers? <exists>
size = 1
item []:
    item [1]:
            class = "IntervalTier"
            name = "HI0001_HI0002_1_A"
            xmin = 0
            xmax = 969.24
            intervals: size = 153
            intervals [1]:
                    xmin = 0
                    xmax = 2.78028066547375

                    text = "<S>"

            intervals [2]:

                    xmin = 2.78028066547375

                    xmax = 4.33862602084094

                    text = "hello □□□□□"

            intervals [3]:

                    xmin = 4.33862602084094

                    xmax = 6.48060077147404

                    text = "<Z>"

            intervals [4]:

                    xmin = 6.48060077147404

                    xmax = 9.98278516986669

                    text = "□□ □□□ □□□□ □□□ class □□ □□□ □□□"

            intervals [5]:

                    xmin = 9.98278516986669

                    xmax = 14.0451792160189

                    text = "<Z>"

            …
```

Table 1-2 Transcription Example

The long recording was segmented into short phases, “xmin” and “xmax” mean the starting and ending time of each sentence.

In addition to the previous structure, additional directories are used to store some other files, as defined in Table 1-2.

/	README file with overview of database

	COPYRIGHT file
/DOC	Documentation includes Technical Documentation, Speaker Information, Conversation Information transcription guidelines, phone set and Pronunciation Lexicon.

Table 1-3 Non-Speech Related Directory Structure

2. Database Design and Collection

2.1 Recording Devices

The devices used to collect data are show in table 2-1.

Device	Model
Computer	Acer Aspire
Microphone	Shure WH20

Table 2-1 Recording Devices

2.1.1 Microphone

Shure WH20 is a close-talk headset microphone and the technical specification is:

Microphone Type	Dynamic, close-talk
Frequency range	50 to 15,000 Hz
Polar pattern	Cardioid, uniform with frequency, symmetrical about axis
Impedance	Rated at 150 ohms (200 ohms actual)
Hum Sensitivity	38.4 dB equivalent SPL in a 1 millioersted field
Polarity	Positive sound pressure on diaphragm produces positive voltage on pin 2 with respect to pin 3 of microphone output XLR connector.
Net Weight	98 g (3.5 oz)

Table 2-2 Technical Specification of Shure WH20

2.2 Recording Environment

The speakers recorded in different environments. The detailed information of each speaker could be found in *ConversationInfomation.xlsx* in /DOC directory.

- Inside an elevator / small office with the door closed
- Living room at home with soft music in the background
- On the balcony with the sound of heavy wind and rain in the background

- At a children’s playground with many kids playing
- Open-plan office within a cubicle on a working day
- Restaurant with cutlery noises and people conversing at nearby tables
- On a footpath with many pedestrians beside a bus stop on a road busy with traffic
- In a car near a traffic intersection during peak hour with windows open & engine running

2.3 Speaker Recruitment

Each group (4 Speakers) was asked to talk with each other around one hour, based on certain topics.

The entire collection was mostly performed in India.

Many recruitment methods are adopted: Posters spread in the Universities, newspaper publicity as well as co-operation with a country-wide HR agency, etc. We made a strict balance control on the age, gender and regional distribution when hiring, all the regional accents were evaluated by our linguists before final recording. And all the speakers or their guardians signed an agreement to assign over to Beijing Haitian Ruisheng Science Technology Ltd. all rights in any speech samples collected in connection with his/her participation.

3. Database Contents Definition

There are 17 topics were included in this database, which were commonly mentioned in daily life. Each pair of speaker could select several of them by their interests.

Table 3-1 shows the examples of the topics’ name:

NO.	Topic
1.	Restaurant
2.	Hotel
3.	Education
4.	Family
5.	Work
6.	Finance
7.	Cooking
8.	Health
9.	Films
10.	Music
11.	Pets
12.	Politics

13.	Shopping
14.	Social networks
15.	Sports
16.	Tourism
17.	Others

Table 3-1 Topic List

4. Transcription

All audio files were manually transcribed and annotated by our native transcribing team based on the transcribing conventions, i.e. the file TRANSCRIP.PDF in the folder /DOC. A strict evaluation work was made on all the transcribing files by our QA Team.

A professional transcription tool was developed by SpeechOcean to support this transcription work and some new short-cut functions were embedded into the tool such as the button for the non-speech acoustic events.

While calculating the recording hours, the segmentations marked with <S> or (Z) will be excluded.

5. Speaker Demographic Information

For this database, qualified speakers were carefully selected by considering gender, age and dialectal region balance. The detailed information of each speaker could be found in *SpeakerInformation.xlsx* in /DOC directory.

5.1 Gender Balance

The database consists of 789 male speakers (49.9%) and 792 female speakers (50.1%).

5.2 Age Distribution

For this project, speech data were collected in the following age categories:

Age group	# Speakers	# Speakers (%)
< 25 years	627	39.7%
25 – 40 years	517	32.7%
41 – 55 years	309	19.5%
>55 years	128	8.1%
Total	**1,581**	**100.0%**

Table 5-1 Speakers' Age Distribution

5.3 Dialectal Regions

This database target speakers are Native Hindi speakers in India.

Figure 5-1 shows the area where Hindi is spoken natively.

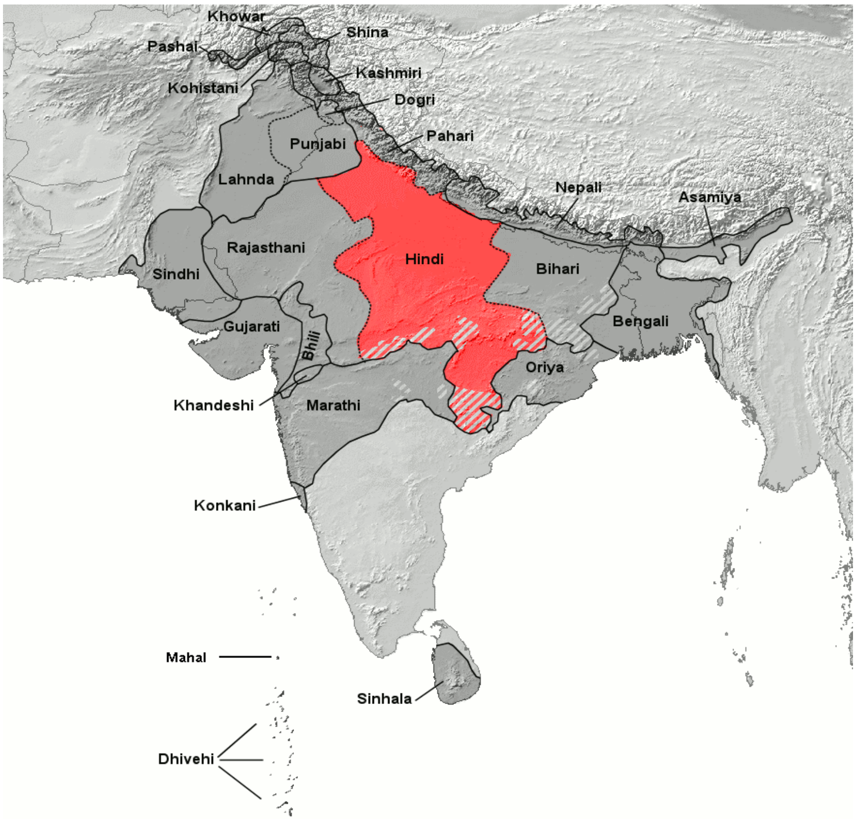

Figure 5-1 Hindi Spoken Area

Table 5-2 shows the average speakers' Region distribution details:

Region	# Speakers	# Speakers (%)
Western	328	20.7%
Eastern	325	20.6%
Central	335	21.2%
North	297	18.8%
North East	296	18.7%
Total	**1,581**	**100.0%**

Table 5-2 Speakers' Region Distribution

6. Pronunciation Lexicon

A pronunciation lexicon with a phonemic transcription in XSAMPA was carefully generated by covering all the words in the transcription files. The Phoneme definition file is PHONEME.PDF in "\DOC" directory.

The lexicon is generated with an L2S rule automatically.

The lexicon file is LEXICON.LEX in “\DOC” directory. It is UTF-8 encoding, including 122,451 entries (118,926 words).

7. QA Process

This corpus has passed the QA process of the following check points:

1. Structure:
   a) All the required specified files exists and with correct format;
   b) The transcriptions and the wave files is one-to-one corresponding;
2. Speech audio files:
   a) The sampling rate and sampling precision meet the requirement;
   b) The header of the WAVE file is the standard WAVE format, with the size of 44 bytes;
   c) There is no clipped audio file, which means for each WAVE file, the value of each sampling point is not bigger than 32700 (with 16bit sampling precision).
3. Transcriptions:
   a) The sentences error rate of the word is lower than 5%.
   b) The sentence error rate of the non-speech labels is lower than 10%, because the labelling of non-speech events is more subjective.
4. Lexicons:
   a) All the phoneme used in the lexicon has definition in the phone definition file;
   b) All the words from the transcriptions has corresponding entry in lexicon, i.e. there is no missing word;

8. Reference

http://en.wikipedia.org/wiki/Hindi_languages