DELIVERABLE IDENTIFICATION

Identification number	King-ASR-323
Type	Technical Report
Title	Definition of Corpus, scripts, standards and Specifications of environment/speaker coverage for the Hindi language.
Status	
Date	2017-4-19
Version	1.0
Number of pages	9
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	iOS, Android, Windows Phone, conversational speech ASR database, contents, design, description, speaker coverage.
Abstract	This document provides a specification of the contents, speaker and environment coverage of the speech database collected over iOS, Android and Windows phone Platform for the Hindi language.
Status of the abstract	Public

Contents

List of Tables

1. Introduction

This is a 3-channel Hindi mobile phone conversational speech database, which is collected over 3 different mobile operating systems simultaneously: iOS, Android and Windows Phone. This database is owned by Beijing Haitian Ruisheng Science Technology Ltd (SpeechOcean, www.speechocean.com)

The corpus contains 100 groups of daily spontaneous conversational speech, which were from 200 speakers. For this collection, 20 people recorded in different rooms and 180 people recorded in a same room. Each group of speakers performed the recording in the same quiet room. 21 topics were contained in this database. The pure recording time is about 308 hours for all 3 channels and 102.7 hours per channel, including the reasonable leading and trailing silence. The total size of this database is 41.2G.

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 3-level directory structure under DATA folder defined as:

Platform\Wav<ID>\Conversation<ID>

Where each part is defined in Table 1-1:

Platform	Android: recordings from Android phones;
	iOS: recordings from iPhones;
	WindowsPhone: recordings from Windows phones

Wav<ID>	wav1: include 20 recordings from 20 speakers who recorded in differen rooms wav2: include 90 recordings from 180 speakers who recorded in a sam room
Conversation<ID>	DATA1: Defined as Conversation <AAAA_C> where <AAAA> is the conversation number from 0050 to 0059; (numbers were used discontinuously) where <C> is the speaker A or B; DATA2: Defined as Conversation < AAAA > where <AAAA> is the conversation number from 1001 to 2019; (numbers were used discontinuously)

Table 1-1 Mobile Speech Directory Structure

1.3 Transcription Files

The speech audio files were transcribed manually after recording. The transcription files were also in the folder of /SCRIPT/script<ID>, named as “<ConversationID>.TextGrid”. The transcription files were transcribed by native speakers, which also contain the noise labels and reflect the conversation content.

All of the audio files were transcribed by native speakers, which also contain the noise labels and reflect the conversation content.

An example is shown as below:

The speech audio files were transcribed manually after recording. The transcription files were also in the folder of /SCRIPT/, named as “<ConversationID>.TextGrid”. 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:

```
…
item []:
        item [1]:
                class = "IntervalTier"
                name = "Speaker"
……
                intervals [1]:
                        xmin = 0
                        xmax = 2.6916215704824977
                        text = "A"
                 ……
                intervals [382]:
                        xmin = 1440.9167235403995
                        xmax = 1444.9324590569624
                         text = "B"
                 ……
        item [2]:
                class = "IntervalTier"
                name = "text"
                ……
                intervals [382]:
                        xmin = 1440.9167235403995
                        xmax = 1444.9324590569624
                        text = "                                                        "
```

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.

script1:

The “text” in “item [2]” shows the transcription of each speaker and several kinds of annotations as below:

- “<S>” means short pause during the conversation.
- “<Z>” means invalid part, including noise or invalid speech.

script2:

And the “text” in “item [1]” shows the speaker ID, which could be “A”/“B”, indicating three different speakers.

The “text” in “item [2]” shows the transcription of each speaker and several kinds of annotations as below:

- “<S>” means short pause during the conversation.
- “<Z>” means invalid part, including noise or invalid speech.

The “item [1]” will be keep blank if the corresponding “text” in “item [2]” contains only tags, without any valid transcriptions.

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, transcription guidelines, phone set and Pronunciation Lexicon.

Table 1-3 Non-Speech Related Directory Structure

2. Database Design and Collection

2.1 Recording Platforms

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

Platform	Phone Type
Android	Samsung Galaxy Note3, Samsung Galaxy S4
iOS	iPhone5, iPhone4S, iPhone 4
Windows Phone	Microsoft Lumia 535, Microsoft Lumia640

Table 2-1 Recording Devices

2.2 Recording Environment

All the speakers recorded in quiet environment, such as office conference room or bedroom in home.

The quiet environments are:

1. Home (quiet, e.g. very low background noise)
2. Office (quiet, e.g. low background noise)

2.3 Speaker Recruitment

Each group of speakers were asked to talk with each other around one hour, based on certain topics.

The entire collection was 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

21 topics were included in this database, which were commonly mentioned in India daily life. Each pair of speaker could select several of them by their interests. The topic information of each conversation could be found in the file “SPEAKERINFO.XLSX” under /DOC directory.

Table 3-1 shows the name, content and coverage distribution of the database.

Topic Name	Content
Family	family members

Health	health
Music	music
shopping	shopping, fashion
Sports	sports
Travel	travel
Work	work
Food	food & drink, cooking, restaurant
education	parenting/own experience, theory & practice, kids
movie	movie
social networks	social networks
friend	friendship, memories of the past, current situation
entertainment	entertainment, amusement, video game
News	news
Pets	animals, pets
computer	computers, hardware, software
TV	TV programs, shows
celebrity	famous people, such as actors
Life	daily life, recent update, commodity price, environment
marriage	marriage, love, date
weather	Weather, climate

Table 3-1 Topic List

4. Transcription and Labeling

4.1 Transcription Guidelines

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

4.2 Timestamp Labeling

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. Speech from each speaker was transcribed separately, and was broken down into short utterances.

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 “SPEAKERINFO.XLSX” in /DOC directory.

5.1 Gender Balance

The database consists of 108 male speakers (54%) and 92 female speakers (46%).

5.2 Age Distribution

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

Age group	# Speakers	# Speakers (%)
18 – 30 years	112	56.0%
31 – 45 years	60	30.0%
46 – 60 years	28	14.0%
Total	**200**	**100%**

Table 5-1 Speakers' Age Distribution

5.3 Dialectal Regions

According to the characteristics of the Hindi language, India is divided into 2 dialect areas: East and West.
Table 5-2 shows the average speakers' Region distribution details

Region	# Speakers	# Speakers (%)
East	92	46%
West	108	54%
Total	**200**	**100%**

Table 5-2 Speakers' Region Distribution

6. Pronunciation Lexicon

A pronunciation lexicon with a phonemic transcription in hi-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 and then manually checked by our linguists.
The lexicon file is LEXICON.LEX in "\DOC" directory. It is UTF-8 encoding, including 27,246 entries (26,932 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 32000 (with 16bit sampling precision).

3. Transcriptions:
    a) The word error rate of the text is lower than 5%.

b) The word 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_language