DELIVERABLE IDENTIFICATION

Identification number	King-ASR-282
Type	Technical Report
Title	Definition of Corpus, Scripts, Standards and Specifications of recording device, environment/speaker coverage for Hindi language.
Status	
Date	2017-4-10
Version	1.0
Number of pages	10
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, ASR database, Database Collection, Contents Design, Transcriptions, Speaker Information.
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

List of Figures

1. Introduction

This is a 3-channel Hindi mobile phone 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 the recordings of 235,725 utterances of Hindi speech data which were from 180 speakers. Each speaker was designed to record 2 sessions, totally 440 utterances in both quiet (office/home) and noisy environment (restaurant/street). The total pure recording time is about 412 hours for all 3 channels and 137 hours per channel, including the leading silence (about 500 milliseconds) and the trailing silence (about 500 milliseconds). The total size of this database is 44.2 GB.

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\ CHANNEL<ID>\WAVE directory.

1.2 Speech File Directory Structure

The 4-level directory structure under DATA folder defined as:

CHANNEL<ID>\WAVE\SPEAKER<ID>\SESSION<ID>

Where each part is defined in Table 1-1:

CHANNEL<ID>	Defined as CHANNEL <A> where <A> is the channel number from 0 to 2; CHANNEL0: recordings from Android phones; CHANNEL1: recordings from iPhones; CHANNEL2: recordings from Windows phones
SPEAKER<ID>	Defined as SPEAKER <BBBB> where <BBBB> is a number from 0051 to 0350 (numbers were used discontinuously)
SESSION<ID>	Defined as SESSION <C> where <C> is the session number 0 or 1;

Table 1-1 Mobile Speech Directory Structure

The wave files are under the SESSION <ID> folder, and their names are defined as <ABBBBCDDD>.WAV, where A is the channel ID, BBBB is the speaker ID, C is the session ID and DDD is the audio file ID, which is from 000 to 439.

1.3 Prompt Files

The script files are under CHANNEL<ID>\SCRIPT folder, and their names are defined as <ABBBBC>.TXT, where A is the channel ID, BBBB is the speaker ID, C is the session ID.

The prompt files contain both the original prompts, which were shown to the speakers, and the transcriptions, which were provided by our transcribers. The transcriptions contain the noise labels and reflect the real reading content.

An example is shown below:

100510000

Where:

100510000 is the wave file's name, which is followed by a <tab> and then the original prompt. The transcriptions and labels of this channel are after it.

1.4 Other Table Files

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

/	README.TXT file with overview of database COPYRIGHT.TXT file
/DOC	Documentation includes Technical Documentation, Transcription Guidelines and Phoneme Definition.
/TABLE	Pronunciation Lexicon, Speaker Information, Speech Acoustic Statistical information Session information and Transcription content information.

Table 1-2 Non-speech Related Directory Structure

1.4.1 Speech Acoustic Features File

The SAMPSTAT.TXT file contains the following acoustic characteristics for each speech file in the database and it contains the following fields:

SPKID	Speaker ID
SESID	Session ID
UTTID	Utterance ID
SAMPRATE	Sampling rate
BITS	Bits rate
DUR	Speech duration (seconds)
SNR	SNR value
CLP	Clipping rate
MAXAMP	Maximum sampling value (db)
MEANAMP	Mean value of sampling

Table 1-3 Acoustic Features File Definition

The file is generated by running a software on the database to calculate those values.

1.4.2 Session Information

SESSION.TXT stores information about the recording session proper.

SES	Session number
SCD	Unique speaker code
CHN	Channel ID
DVC	Recording device model
REP	Recording place
ENV	Environment
NSC	Noisy condition: noisy or quiet
UTN	Number of utterances

Table 1-4 Session Information File Definition

1.4.3 Contents Index File

The contents index files stores a transcription field and relates it to properties of the signal file and speaker.

SCD	Speaker ID
SES	Session ID
UID	Utterance ID
TRS	Speech transcription

Table 1-5 Content Index File Definition

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 NOTE4
iOS	iPhone4,iPhone5, iPhone4S, iPhone5S
Windows Phone	Nokia 620, Nokia820

Table 2-1 Recording Devices

2.2 Recording Environment

Each speaker recorded in 2 different environments: quiet and noisy. SESSION0 was recorded in quiet environment, such as office conference room or bedroom in home. SESSION1 was recorded in noisy environment, including restaurant and street.

The quiet environments are:

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

While the noisy environments included:

1. Restaurant (has background noise, other people speaking in vicinity)
2. Outside on busy street (has background noise from cars and passerby)

2.3 Speaker Recruitment

Each Speaker was to record around 440 sentences which were selected from a pool of phonetically rich sentences in approximate 70 minutes as natural as possible.

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 Design

The prompts were the phonetically rich sentences. Due to the potential cognitive load of saying these sentences by the subjects, we took care to choose natural sentences of length between 7 and 20 words. The raw sentences are selected from different domain: news, conversations, twitter and etc. We did remove a number of sentences

that includes offensive or negative words or phrase. In order to achieve a good phone balance, we choose sentences from the sentences list to fill out our number. Finally, we had around 40,000 unique sentences in our list of sentences, that we generated the prompt sheets from with no more than 4 times for each.

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.DOC 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. Transcribers will leave the prompts blank if some utterances are missed or empty (without valid speech in it). Blank prompts and the corresponding wave files will be discarded.

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 SPEAKER.TXT in /TABLE directory.

5.1 Gender Balance

The database consists of 99 male speakers (55%) and 81 female speakers (45%).

5.2 Age Distribution

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

Age group	# Speakers	# Speakers (%)
18 – 30 years	106	58.89%
31 – 45 years	47	26.11%
>45 years	27	15.00%
Total	**180**	**100%**

Table 5-1 Speakers' Age Distribution

5.3 Environment Distribution

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

Environment	# Sessions	# Sessions (%)
Office	180	50.0%
Restaurant	82	22.78%
Street	98	27.22%
Total	**360**	**100%**

Table 5-2 Recording Environment Distribution

5.4 Dialectal Regions

According to the characteristics of the Hindi language, India is divided into 2 dialect areas: West and East.

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

Region	# Speakers	# Speakers (%)

West	90	50.00%
East	90	50.00%
Total	**180**	**100%**

Table 5-3 Speakers' Region Distribution

Figure 5-1 shows the Hindi Dialectical Region.

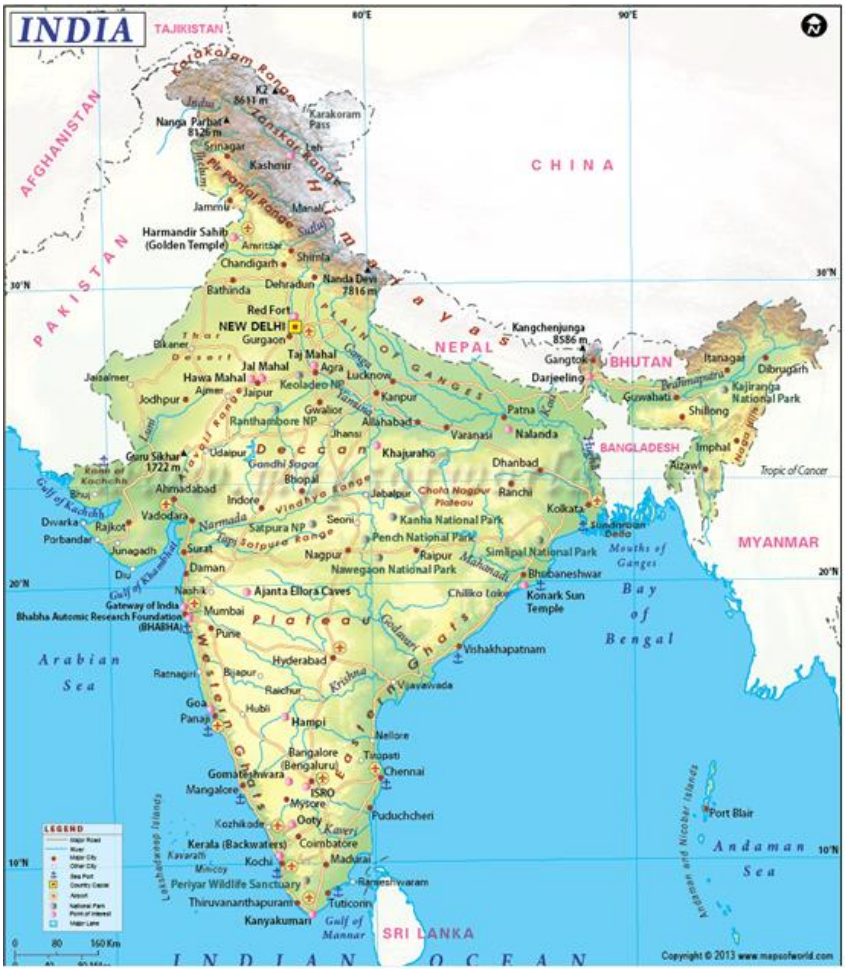

Figure 5-1 Hindi Dialectical Region

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 and then manually checked by our linguists.

The lexicon file is LEXICON.LEX in "\TABLE" directory. It is UTF-8 encoding, including 26,195 entries (25,960 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 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_language
http://www.digigram.com/products/product_infos.php?prod_key=11500