DELIVERABLE IDENTIFICATION

Identification number	King-ASR-435
Type	Technical Report
Title	Definition of Corpus, scripts, standards and Specifications of environment/speaker coverage for the Hindi language.
Status	
Date	2019-05-24
Version	1.0
Number of pages	9
Author	Ping PAN, 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	Telephony 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 speech database collected over telephone for the Hindi language.
Status of the abstract	Public

Contents

List of Tables

1. Introduction

This is a telephony Hindi conversational speech database, which is collected over fixed telephone. This database is collected and owned by Beijing Haitian Ruisheng Science Technology Ltd (SpeechOcean, www.speechocean.com).

The corpus contains 210 pairs of Hindi spontaneous conversational speech, which were from 419 speakers. For this collection, 2 speakers of each group performed the recording in separate quiet rooms. 24 topics were contained in this database. The pure recording time is about 433.8 hours, including the reasonable leading and trailing silence. The total size of this database is 78.2G.

1.1 Speech File Format

The utterance waves of each channel are stored as 8 KHz, 16 bit, Mono channel; Windows uncompressed PCM format. All the wave files are stored in \DATA\WAVE directory.

1.2 Directory Structure

The 2-level directory structure under DATA folder defined as:

WAVE\Conference<ID>

Where each part is defined in Table 1-1:

Conference<ID>	Defined as Conference< AAAA > where <AAAA> is the conference number from 0002 to 7236; (numbers were used discontinuously)

Table 1-1 Telephony Speech Directory Structure

Usually, each group of speech consists of 3 audio files: two of them are from single speaker separately and the third one is from the mixed channel. These 3 files were recorded simultaneously. While for the data of Conference6022, one of the single channels was discarded for some reason.

For each group of dialog, the speech from each speaker was stored as “<speakerID>.wav”; where <speakerID> is defined as conf_<AAAA>_<AAAA>_<BBBBBBBB>, where <BBBBBBBB> is a number from 00601020 to 03781293. The speech of mixed channel was stored as “conf_< AAAA >_< AAAA >_all.wav”

1.3 Transcription Files

The speech audio files from the separated channel of each group were transcribed. The transcription files are under \DATA\SCRIPT folder, and their names are defined as “Conf< AAAA >_< AAAA >_<BBBBBBBB>.TextGrid”, where BBBBBBBB is the speaker ID. (E.g. conf_7221_7221_00602841.TextGrid)

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:

```
    size = 2
    item []:
      item [1]:
          class = "IntervalTier"
          name = "conf_7221_7221_00602841"
          xmin = 0
          xmax = 8950.615
          intervals: size = 2928
          intervals [1]:
               xmin = 0
               xmax = 26.357
               text = "<S>"
          intervals [2]:
               xmin = 26.357
               xmax = 27.371
               text = "हेलो"
          intervals [3]:
               xmin = 27.371
               xmax = 29.071
               text = "<S>"
          intervals [4]:
               xmin = 29.071
               xmax = 30.721
               text = "हाँ आपका नाम क्या है"
....
                 item [2]:
          class = "IntervalTier"
          name = "conf_7221_7221_00602841"
          xmin = 0
          xmax = 8950.615
          intervals: size = 5
          intervals [1]:
               xmin = 0
               xmax = 1468.38
               text = "Family"
          intervals [2]:
               xmin = 1468.38
               xmax = 2682.562
               text = "Films"
          intervals [3]:
               xmin = 2682.562
               xmax = 3212.493
               text = "Restaurant/Food"
          intervals [4]:
               xmin = 3212.493
               xmax = 5911.354
               text = "Plans-Hopes-Dreams"
          intervals [5]:
               xmin = 5911.354
```

```
                xmax = 8950.615
                text = "Restaurant/Food"
        intervals [4]:
                xmin = 5056.343
                xmax = 5675.237
                text = "Music"
        intervals [5]:
                xmin = 5675.237
                xmax = 8253.774
                text = "Online-Mobile"
```

Table 1-2 Transcription Example

There are two items in every transcription files.

The “text” in “item [1]” 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 long recording was segmented into short phases, “xmin” and “xmax” mean the starting and ending time of each sentence.

The “text” in “item [2]” shows the topics of each speaker. “xmin” and “xmax” mean the starting and ending time of each topic.

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 recording platform is based on an IVR system; each speaker dialed in the system and input a unique conference ID to participate in the recording.

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

Almost each group (2 speakers) was asked to talk with each other around 2.5 hours, based on certain topics.

The collection was mainly 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

24 topics were included in this database, which were commonly mentioned in Indian daily life. Each pair of speakers could select 4~5 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 topic distribution of the database.

Topic Name	Frequency	Content
Restaurant/Food	69	food & drink, cooking, restaurant
Work	65	work
Politics/News	39	politics/news
Shopping	48	shopping, fashion
Plans-Hopes-Dreams	32	plans, hopes, dreams
Education/Study	82	parenting/own experience, theory & practice, kids
Social Networking/Friends	78	daily life, recent update, commodity price, environment
Films	91	movie
Family	88	family members
Cooking	72	cooking
Mood	14	mood
Health/Medical	52	health, medical
Sports	42	sports
Hobbies	21	entertainment, amusement, video game
Weather	28	Weather, climate
Online-Mobile	20	computers, hardware, software
Music	51	music
Finance/Insurance	8	finance, insurance
Hotel	17	hotel
Pets	31	animals, pets
Tourism	41	travel
Lottery	4	lottery
Humor	2	humor
Immigration	1	immigration

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.

5. Time-stamp\Topic 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.

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

6. 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.

6.1 Gender Balance

The database consists of 241 male speakers (57.5%) and 178 female speakers (42.5%).

6.2 Age Distribution

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

Age group	# Speakers	# Speakers (%)
18 – 30 years	224	53.5%
31 – 45 years	158	37.7%
46 – 60 years	37	8.8%
Total	**419**	**100%**

Table 6-1 Speakers' Age Distribution

6.3 Dialectal Regions

According to the characteristics of the Hindi language, Hindi is divided into 2 dialect areas: Eastern dialect and Western dialect.

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

Region	# Speakers	# Speakers (%)
Eastern	94	22.4%
Western	325	77.6%
Total	**419**	**100%**

Table 6-2 Speakers' Region Distribution

7. Pronunciation Lexicon

A pronunciation lexicon with a phonemic transcription in hi-in_xsampa.phset 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 55,632 entries (55,044 words).

8. 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;

9. Reference

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