DELIVERABLE IDENTIFICATION

Identification number	King-ASR-60
Type	Technical Report
Title	Definition of Corpus, scripts, standards and Specifications of environment/speaker coverage for Hindi languages
Status	
Date	2013-04-02
Version	1.0
Number of pages	9
Author	Xingtong Ge, Ke Li, 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, ASR database, contents, design, description, speaker coverage.
Abstract	This document provides a specification of the contents, speaker and environment coverage of the speech database to be collected over the desktop for the Hindi language.
Status of the abstract	Public

Contents

1. Introduction

This is a 4-channel Hindi desktop speech database, which is collected in India over a close talk and a desktop microphone simultaneously. This database is owned by of Beijing Haitian Ruisheng Science Technology Ltd (SpeechOcean, www.speechocean.com) and performed in an environment of quiet office.

The corpus contains the recordings of 239,736 utterances of Hindi speech data which were from 200 speakers. The pure recording time is about 481 hours (4-channel), including the leading silence (about 500 ms) and the trailing silence (about 500 ms). The total size of this database is 142.0 G.

1.1 Speech File Format

The utterance waves of each channel are stored as 44.1 KHz, 16 bit, Mono channel; Windows uncompressed PCM format.The wave files from each channel are stored separately.

1.2 Directory Structure

The three-level directory structure is defined as:

<ChannelID>\wav\<SpeakerID>

where each part is defined in Table 2-1.

<ChannelID>	Defined as channel<n> where <n> is the channel number from 0 to 3;
<SpeakerID>	Defined as Speaker<nnn> where <nnnn> is a number from 002 to 304 (numbers were randomly used)

Table 1-1 Desktop Speech Directory Structure

The wave files are under <SpeakerID> folder, and their names are defined as <n>.wav where n is from 1 to 300. 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 and Pronunciation Lexicon
/<ChannelID>/script	Prompts Transcriptions

Table 1-2 Non-speech related directory structure

1.3 Prompt Files

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 in Table 1-3:

Prompt	1 □□ □□ □□□□□□ □□□□□□□□ □□ 1 □□□□□□ □□ □□ □□□□ □□□□ □□□□□□
Channel 0 Transcription	□□ □□ □□□□□□ □□□□□□□□ □□ □□ □□□□□□ □□ □□ □□□□ □□□□ □□□□□ <NON/>
Channel 1 Transcription	<SPK/> □□ □□ □□□□□□ □□□□□□□□ □□ □□ □□□□□□ □□ □□ □□□□ □□□□ □□□□□ <NON/>
Channel 2 Transcription	<SPK/> □□ □□ □□□□□□ □□□□□□□□ □□ □□ □□□□□□ □□ □□ □□□□ □□□□ □□□□□ <NON/>
Channel 3 Transcription	<SPK/> □□ □□ □□□□□□ □□□□□□□□ □□ □□ □□□□□□ □□ □□ □□□□ □□□□ □□□□□ <NON/>

Table 1-3 Prompt/Transcription example

Where:

1 is the wave file's name, which is followed by a <tab> and then the original prompt. The transcriptions and labels of these channels are after them.

2. Database Design and Collection

2.1 Recording Platforms

The recording tracks are made in Windows XP SP2 by a multi-channel recording software, AudioRec, which is developed by SpeechOcean. The recording equipments are listed as below. (Table 3-1)

Computer	HP 2510p	1
Microphone	Shure SM10A /AKG D660S/Sennheiser MKE2/AKG C400bl	1/1/1/1
Audio Card	ECHO Audio fire4	1

Table 2-1 Recording Platform

2.2 Microphone

2.2.1 Shure SM10A (C0)

Shure SM10A is a close-talk microphone and the technical specification is:

Microphone Type	Dynamic, close-talk
Polar pattern	Cardioid, uniform with frequency, symmetrical about axis
Frequency range	50 to 15,000 Hz
Sensitivity	–65.0 dBV/Pa* (0.45 mV)
Polarity	Positive pressure on diaphragm produces positive voltage on pin 2 of microphone connector.

Connector	2 x 3.5mm

Table 2-2 Technical Specification of Shure SM10A

2.2.2 AKG D660S (C1)

AKG D660S is a desktop microphone and the technical specification is:

Polar pattern	hypercardioid
Frequency range	70 to 20,000 Hz
Sensitivity	2.0 mV/Pa (-54 dBV) at 1,000 Hz
Equivalent noise level	20 dB (A) to DIN 45412
Impedance	<=500 ohms
Recommended load impedance	>=1,200 ohms
Connector	3-pin male XLR
Finish	matte black enamel
Net weight	240 g / 8.5 oz.
Shipping weight	423 g / 0.9 lb.

Table 2-3 Technical Specification of AKG D660S

2.2.3 Sennheiser MKE2 (C2)

AKG D660S is a collar microphone and the technical specification is:

Polar pattern	Omni-directional
Frequency range	20 to 20000 Hz +- 3 dB
Sensitivity	5 mV/Pa +- 3 dB
Equivalent noise level	26 dB
Impedance	1000 Ohm
Recommended load impedance	>=1,200 ohms
Connection cable	LEMO f. SK50/250
Dimensions	d 4,8 mm

Table 2-4 Technical Specification of Sennheiser MKE2

2.2.4 AKG C400bl (C3)

AKG C400bl is a desktop microphone and the technical specification is:

Polar pattern	hypercardioid
Frequency range	150-15000Hz
Sensitivity	13.5 mV/Pa (-37 dBV)
Equivalent noise level	62dB
Impedance	200 Ohm
Connector	3-pin male XLR
Dimensions	43×24×15mm

Table 2-5 Technical Specification of AKG C400bl

2.2.5 ECHO Audio Fire4

AudioFire4 offers a compact audio interface with all the connections you need for your studio. With the flexibility of FireWire and bus power, you can also take the AudioFire4 on the road. The AudioFire4 can record 24 bit 96 kHz audio with low latency monitoring on any Windows XP/Vista/Win 7 or Mac OS X (10.4 or later) computer (desktop or notebook) with a FireWire port.

AudioFire4 is the perfect centre for any home studio, whether at your desk or on the road. It has 2 universal inputs with mic preamps, phantom power, and trim knobs so you can just plug in your microphone or instrument and record whenever or wherever you want. AudioFire4 also comes with 2 balanced analog inputs (TRS), 4 balanced analog outputs (TRS), a stereo headphone output, 2 FireWire ports, S/PDIF I/O, MIDI I/O, and 6 channels of full duplex 24 bit 96kHz recording and playback.

Hardware Features

FireWire (IEEE 1394a) interface with 8' cable

Bus powered with 6-pin FireWire interface

External 12VDC power supply provided

2 auto-sensing universal inputs (mic/guitar/line):

- Meters, trim knobs & 48V phantom power

2 balanced ¼" analog inputs:

- 112dB (A-weighted) dynamic range
- +4dBu / -10dBV nominal levels

4 balanced ¼" analog outputs:

- 114dB (A-weighted) dynamic range
- +4dBu / -10dBV nominal levels
- S/PDIF I/O at 24-bit/96kHz
- MIDI I/O

Stereo headphone output with volume knob

Sync via S/PDIF

Supports full duplex 6-channel in, 6-channel out operation at 24 bit, 96 kHz sample rate

Heavy-duty aluminum case

Near zero latency hardware monitoring

Software Features

Software console for monitoring, metering, and setting levels

Low-latency drivers support all major pro audio software:

- Low Latency ASIO 2.0
- WDM kernel streaming mode
- GSIF 2.0 (Gigastudio w/low latency MIDI) (32-bit Windows XP, Vista & Windows 7)
- CoreAudio & CoreMIDI

2.3 Speaker Recruitment

Each Speaker was to record around 300 sentences which were selected from a pool of phonetically rich sentences in approximate 35 minutes as natural as possible. The recording was performed in a quiet office environment.

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.

3. Database Contents Definition

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 Hindi domain: Conversations, News, etc. We did remove a number of sentences that includes offensive or negative words or phrase.

Table 3-1 show the phone coverage of the corpus respectively.

Phone	Frequency	Phone	Frequency	Phone	Frequency
@	84	d`_h	443	oU~	43
A:	255967	e	174478	o~	18301
A:~	8180	eI	27734	p	70681
G\	3	eI~	928	p_h	8765
I	123421	e~	39296	q	349
I~	1230	f	7042	r	204007
J	28	g	45682	r`	9250
N	10002	g_h	2933	r`_h	2833
Q	861	h	94586	r`i	2537
S	38404	h\	15471	r`i~	151
U	50392	i:	129548	s	119004
U~	255	i:~	5798	s`	6291
X\	425	j	64618	t	95303
a	348824	j0	1	tS	29467
a~	5237	k	196483	tS_h	8622
b	49272	k_h	13517	t_h	16018
b_h	18398	l	85289	t`	21248
d	75120	m	105718	t`_h	6418
dZ	41220	n	122169	u:	17780
dZ_h	1620	n`	9166	u:~	1452
d_h	7495	o	58082	v\	58000

Table 3-1 Phone Coverage of the Prompt

4. Transcription

All audio files were manually transcribed and annotated by our native transcribing team based on the Transcribing conventions; 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.

Transcriptions are case sensitive. Transcribe abbreviations (e.g. I B M) and word in spelling mode (e.g. N E W Y O R K) spoken as letters using capital letters with spaces. Proper Noun (e.g. names, addresses, countries, organizations, months etc.) begins with capital letter, such as China, Microsoft, August etc. Brand names, trademarks are transcribed as their original format including their case form (e.g. MySpace, NIKE). All other words in the transcription are in lower case.

Truncated words, words that were truncated due to recording errors, mispronunciations, words that are nevertheless intelligible and unintelligible speech, words or stretches of speech that are completely unintelligible, were transcribed by fixed labels.

According to the Transcription Guidelines, we annotated non-speech events as follows.

: Mispronounced words, unintelligible speech, words or stretches of speech that are completely unintelligible, were transcribed by two asterisks: “”. The “**” marker is separated from neighboring intelligible words with spaces.

Filled pause: Marked with <FIL/>. Examples of filled pauses include “uh”, "um", "er” and “mm”.

Speaker noise: Marked with <SPK/>. The various sounds and noises made by the speaker that are not part of the prompted text. These include lip smack, cough, throat clear, tongue click, load breath, and laughing.

Non-human noise: Marked with <NON/>. This is used to mark the occasionally noise, such as mouse clicking, keyboard typing, etc.

Stationary noise: Marked with <STA/>. This is background noise that is not intermittent and has a more or less stable amplitude spectrum over time. We included in stationary noise the ticks from the indicator, as well as loud wiper noise and loud whistling causes by wind at high speeds. Music was also marked as stationary if it was audible in the driving scenario.

Non-Primary Speakers' noise: Marked with <NPS/>. Noise of other human being like lip smack, cough, throat clear, tongue click, load breath, laughing, and speech not from the primary speakers are all included.

If <FIL/>, <SPK/> or <NON/> begins in a word then the symbol was put before the first word affected. The symbols were always separated from the surrounding words by spaces.

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.

5.1 Gender Balance

The database consists of 99 male speakers (49.5%) and 101 female speakers (50.5%).

5.2 Age Distribution

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

Age group	# Speakers	# Speakers (%)
16 – 30 years	125	62.5%
31 – 45 years	55	27.5%
46 – 60 years	20	10.0%

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.

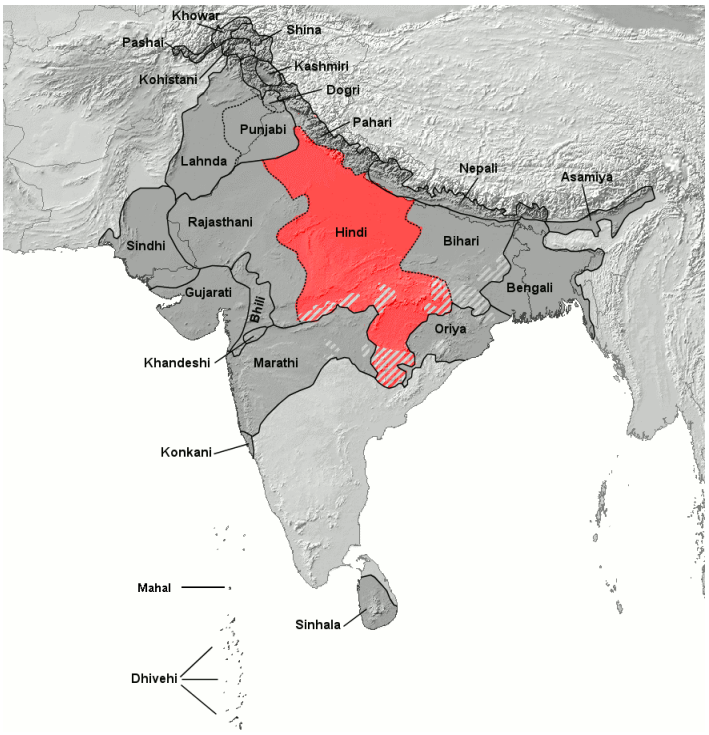

Figure 5-1 Hindi Spoken Area

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

Region	# Speakers	# Speakers (%)
Western	88	44.0%
Eastern	112	56.0%

Table 5-2 Speakers' Region Distribution

6. Pronunciation Lexicon

A pronunciation lexicon with a phonemic transcription in SAMPA was carefully generated by covering all the words in the transcription files.
The lexicon is generated with an L2S rule automatically and then manually checked by our linguists.
The lexicon file is fr-fr.lex in "/doc" directory. It is Unicode encoding, including 9145 words.

7. Reference

http://en.wikipedia.org/wiki/Hindi_languages
http://www.digigram.com/products/product_infos.php?prod_key=11500