English Language Speech Database for Speaker Recognition

(*ELSDSR*)

ELSDSR corpus of read speech has been designed to provide speech data for the development and evaluation of automatic speaker recognition system. *ELSDSR* corpus design was a joint effort of the faculty, Ph. D students and master students from department of Informatics and Mathematical Modeling (IMM) at Technical University of Denmark (DTU). The text language is English, and is read by 20 Danes, one Icelander and one Canadian. No formal rehearsal has been done, thus perfect pronunciation is not obtained, however not necessary for getting the specific and uniquely identifiable characteristics from individuals.

1. Recording Environment

The recording work has been carried out in a chamber (room 133) in building 321, 1st floor at DTU. The chamber is an 8.82*11.8*3.05 m^3 (width*length*height) computer room (classroom), with 22 monitors and 34 tables. The recording is manipulated in, approximately, the middle of this chamber, with one microphone, one 70*120*70 cm^3 table in front of speakers. In order to deflect the reflection, two deflection boards with measure of 93*211.5*6 cm^3 were placed at tilted angles facing each other, and were in front of the table and speakers. For details please see the setup drawing, drawing of the room and position of recording, etc., in appendix.

2. Recording Equipment

The equipment for recording work is MARANTZ PMD670 portable solid state recorder. PMD670 can record in a variety of compression algorithm, associated bit rate, file format, and recording type (channels recorded) parameters. It supports two kinds of recording format: compressed recording, which includes MP2 and MP3; uncompressed recording, which includes linear pulse code modulation (PCM). The recording type can be stereo, mono or digital, and the file can be recorded into .wav .bwf .mpg or .mp3 format.

In this database, the voice messages are recorded into the most commonly used file type--.wav. And the algorithm used is PCM. The sampling frequency is chosen 16 kHz with a bit rate of 16. Table 1 shows the initial setup for the recorder, for detail see PMD670 user guide.

Table 1: Recorder Setup

Input	**Setup**								
	Auto Mark	**Pre Rec**	**Analog Out**	**MIN Atten**	**Repeat**	**ANC**	**EDL Play**	**Level Cont.**	**S. Skip**
MIC (MONO)	OFF	ON	OFF	20dB	OFF	FLAT	OFF	MANUAL	ON 20dB

3. Corpus Speaker Set

ELSDSR contains voice messages from 22 speakers: 10 female, 12 male, and the ages are covered from 24 to 63. Most of them are faculty and Ph. D students working at IMM, and 5 of them are master students including 1 international master student.

No a priori control of the speaker distribution by nationality and age has been done, except for the gender. Due to the practical problem of uneven gender distribution at the experiment site, the average age of female subjects is higher than that of male, and around half of the female subjects are secretaries in IMM. 84% male speakers were between 26 and 37 years old; however the ages of female speakers spread in a large scale. Table 2 shows the speakers ID, speakers' ages with average for each gender, and nationalities.

The subjects of this database are from different countries and different places of one country, the dialect of speaking or reading English language in this database does not play a significant role for the purpose of speaker recognition, since the features which are interesting for this particular intention are language independent. Moreover it might be possible to use this database for accent recognition.

Table 2: Information about Speakers

Speaker ID	**Age**	**Nationality**
FAML	48	Danish
FDHH	28	Danish
FEAB	58	Danish
FHRO	26	Icelander
FJAZ	25	Canadian
FMEL	38	Danish
FMEV	46	Danish
FSLJ	24	Danish
FTEJ	50	Danish
FUAN	63	Danish
Average	40.6	
MASM	27	Danish
MCBR	26	Danish
MFKC	47	Danish
MKBP	30	Danish
MLKH	47	Danish
MMLP	27	Danish
MMNA	26	Danish
MNHP	28	Danish
MOEW	37	Danish
MPRA	29	Danish
MREM	29	Danish
MTLS	28	Danish
Average	31.3	

Speaker ID is constructed and started by F or M, indicating the gender, and followed by 3 letters of speaker initials.

4. Corpus Text Material and Suggested Training/Test Set Division

Part of the text, which is suggested as training subdivision, was made with the attempt to capture all the possible pronunciation of English language, which includes the vowels, consonants and diphthongs. As for the suggested training and test subdivision, seven paragraphs of text are constructed and collected for training, which includes 11 sentences; with respect to the suggested test subdivision forty-four sentences (each speaker reads two of these sentences) from NOVA Home [1] were collected for test text. In a word, for the training set, 154 (7*22) utterances were recorded; and for test set, 44 (2*22) utterances were provided.

On average, the duration for reading the training data is: 78.6s for male; 88.3s for female; 83s for all. And the duration for reading test data, on average, is: 16.1s (male); 19.6s (female); 17.6s (for all). Table 3 shows the time spend on reading both training text and test text individually.

Table 3: Duration of reading training text and test text

No.	Male		Train(s)	Test(s)	Female		Train(s)	Test(s)
1		MASM	81.2	20.9		FAML	99.1	18.7
2		MCBR	68.4	13.1		FDHH	77.3	12.7
3		MFKC	91.6	15.8		FEAB	92.8	24.0
4		MKBP	69.9	15.8		FHRO	86.6	21.2
5		MLKH	76.8	14.7		FJAZ	79.2	18.0
6		MMLP	79.6	13.3		FMEL	76.3	18.2
7		MMNA	73.1	10.9		FMEV	99.1	24.1
8		MNHP	82.9	20.3		FSLJ	80.2	18.4
9		MOEW	88.0	23.4		FTEJ	102.9	15.8
10		MPRA	86.8	9.3		FUAN	89.5	25.1
11		MREM	79.1	21.8				
12		MTLS	66.2	14.05				

5. *ELSDSR* Directory and File Structure

The voice messages are organized according to the following hierarchy:

CORPUS := = ELSDSR
USAGE := = train | test
SPEAKER ID := = FXXX | MXXX |
where,
F or M indicates the speaker's gender;
XXX indicate the speakers' initials

Sentence ID := = XXXX_SM or XXXX_SrN

where,

XXXX indicate speaker ID;
S indicates training sentence, M indicates the alphabetic number of paragraphs in training text, which is from a to g;
Sr indicates test sentence (randomly chosen sentences), N indicates sentences number in test text, from 1 to 46.

The associated documentation is located in the 'ELSDSR /DOC' directory:

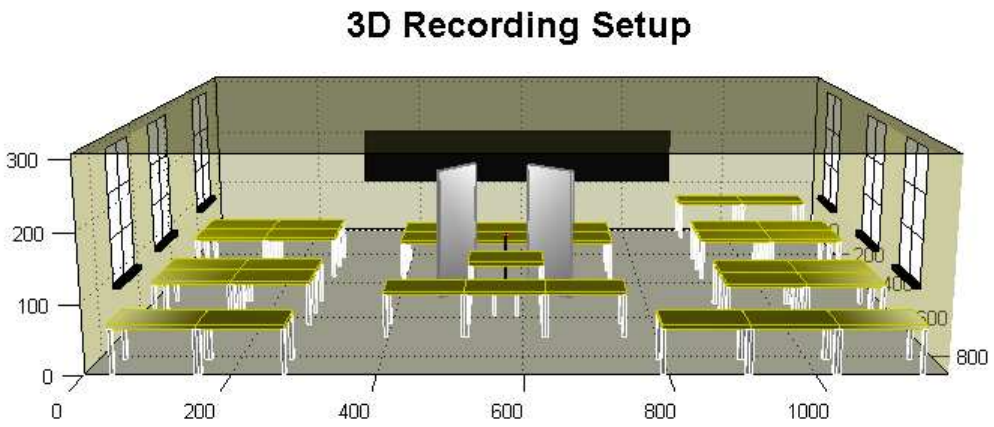
where,

training text.pdf
test text.pdf
phonetic alphabet.pdf(1)
readme.pdf (this file)

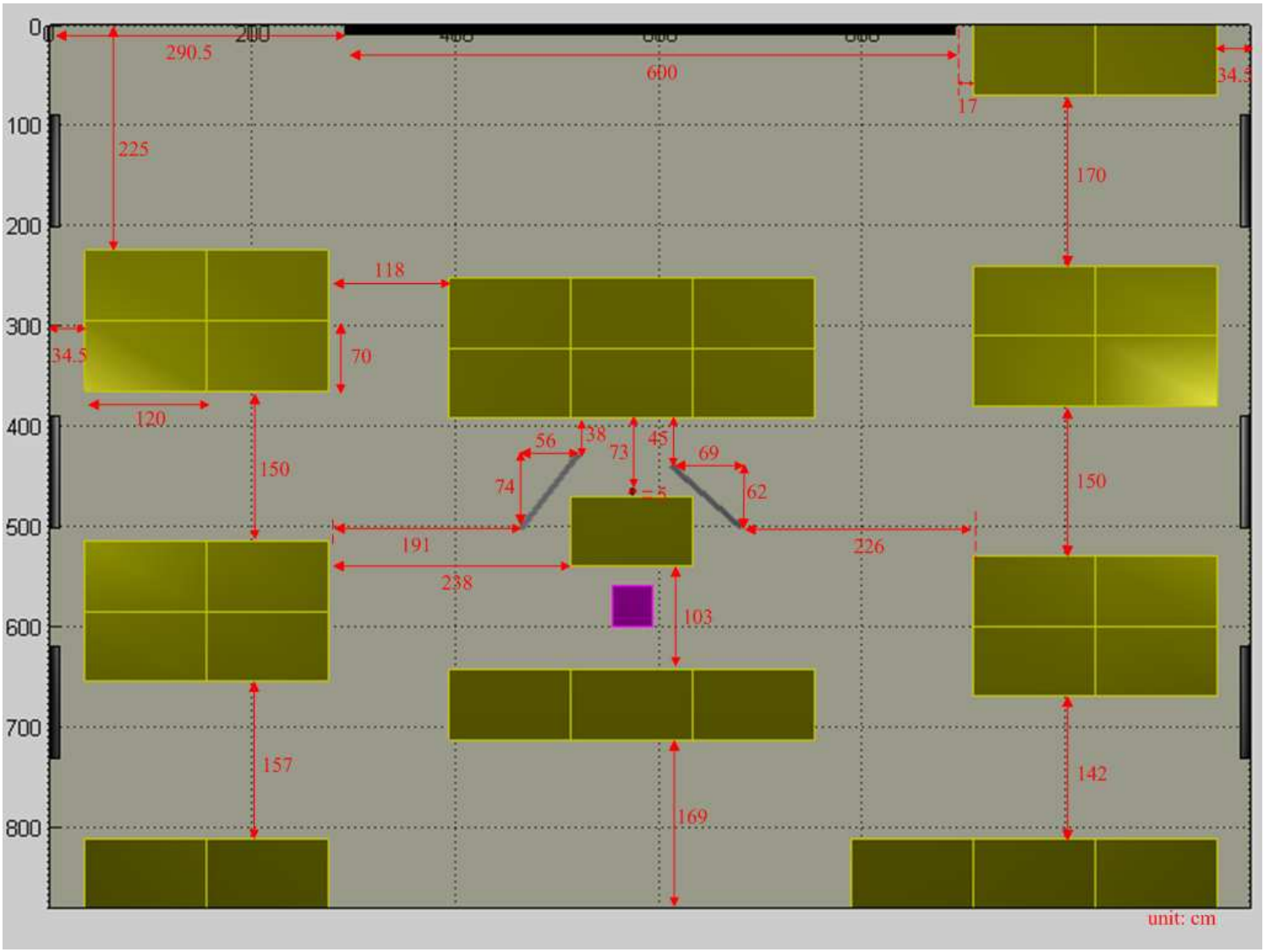
(1) phonetic alphabet.pdf shows the captured vowels, consonants and diphthongs in each paragraph of training data.

Appendix

- 3D Setup of the recording chamber:



- 2D Recording Experiment Setup with measurement:



Reference

[1] NOVA online, WGBH Science Unit, 1997 http://www.pbs.org/wgbh/nova/pyramid/