A NEW DATABASE FOR SPEAKER RECOGNITION

Ling Feng and Lars Kai Hansen

Informatics and Mathematical Modelling, Technical University of Denmark
Richard Petersens Plads, Building 321, DK-2800 Kongens Lyngby, Denmark
phone: (+45) 4525 3888,3889, fax: (+45) 4587 2599, email: lf,lkh@imm.dtu.dk
web:http://isp.imm.dtu.dk

ABSTRACT

In this paper we discuss properties of speech databases used for speaker recognition research and evaluation, and we characterize some popular standard databases. The paper presents a new database called *ELSDSR* dedicated to speaker recognition applications. The main characteristics of this database are: English spoken by non-native speakers, a single session of sentence reading and relatively extensive speech samples suitable for learning person specific speech characteristics.

1. INTRODUCTION

Over the last two decades there has been an increasing interest in speaker recognition. In order to get adequate amounts of speech to train and test the speaker recognition system, speech databases are needed. There are several applications of speaker recognition, leading to a diversity of the structure and content of speaker recognition databases. The most obvious benefit of using standard and readily available (public) databases is that system performances using different techniques on the same database become comparable, hence, enabling quantitative evaluation of methods and speaker recognition protocols. In a search we have found 36 existing databases including both public and proprietary bases that have been used in speaker recognition studies, a comprehensive review of databases has earlier been given in the project report [1]. We here provide an overview and aspects of a taxonomy of speech databases, in order to facilitate future case studies and new database design. We also describe a new database *ELSDSR* which was created to meet the needs of our own recent effort in speaker recognition, and which will be freely available for research.

In this paper, section 2 gives a general taxonomy of speech databases used in speaker recognition research. Moreover a brief description of existing databases will be given. The database, *ELSDSR* for speaker recognition is introduced in section 3. Section 4 concludes and summarizes characteristics of *ELSDSR*.

2. OVERVIEW OF CURRENT SPEECH DATABASES

A taxonomy of speaker recognition databases may be based on features such as the recording protocol, the population of participating subjects, the recording device, language(s), type of verbal statement, and the intended use, etc. The taxonomy of speech databases bases mainly on the database survey of *COST250 Working Group 2* (Lindberg et al., 1996), and rather detailed overview of current publicly available databases for speaker recognition research and evaluation by (Campbell Jr. and D. A. Reynolds, 1999). We here also provide a brief overview of the *TIMIT*, *Polycost*, and *YOHO* databases.

2.1 Taxonomy of Existing Speech Databases

The intra-speaker and inter-speaker variability are important parameters for a speech database. Intra-speaker variability can be very important for speaker recognition performance and can be estimated if the same sentence is read several times by the same subjects. The intra-speaker variation can originate from a variable speaking rate, changing emotions or other mental variables, and in environment noise. The variance brought by different speakers is denoted inter-speaker variance and is caused by the individual variability in vocal systems involving source excitation, vocal tract articulation, lips and/or nostril radiation [2]. If the inter-speaker variability dominates the intra-speaker variability speaker recognition is feasible.

Speech databases are most commonly classified into single-session [3] and multi-session [4,5]. Multi-session databases allow estimation of temporal intra-speaker variability. Combination sets are also possible including single-session recording with a larger set of speakers and multi-session recordings with a smaller set of speakers, for instance, *SpeechDat*, *Switchboard-1*, *SIVA* and *Gandalf,* consult [1] for references. For sampling of low inter-speaker variability subjects, which is relevant, e.g., for admission control systems, some databases even include close relatives among speakers [6, 7], or human mimicry and technical mimicry [8].

With respect to input devices the most common means of recording are microphones or telephone handsets, the latter can be modified by being over local or long distance telephone lines. According to the acoustic environment, databases are recorded either in noise free environment, such as in the sound booth, or with office/home noise. Moreover, according to the purpose of the databases, some corpora are designed for developing and evaluating speech recognition, for instance *TIMIT* [3], and some are specially designed for speaker recognition, such as *SIVA*, *Polycost* and *YOHO* [1]. Many databases were recorded in one native language of recording subjects; however there are also multi-language databases with non-native language of speakers, in which case the language and speech recognition become the additional use of those databases.

2.2 An Overview of Standard Speech Corpora

There are numerous corpora for speech recognition. The most popular bases are: *TIMIT* and its derivatives, *Polycost*, and *YOHO*.

2.2.1 TIMIT and Derivatives

The *TIMIT* corpus of read speech has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems [3]. Although it was primarily designed for speech recognition, it is also widely used in speaker recognition studies, since it is one of the few databases with a relatively large number of speakers. It contains 630 speakers' voice messages (438 M/192 F), and each speaker reads 10 different sentences. It is a single-session database recorded in a sound booth with fixed wideband headset. The derivatives of *TIMIT* are: *CTIMIT*, *FFMTIMIT*, *HTIMIT*, *NTIMIT*, *VidTIMIT*. They were recorded by playing different recording input devices, such as telephone handset lines and cellular telephone handset, etc. *TIMIT* and most of the derivatives are single-session, and are thus not optimal for evaluating speaker recognition systems because of lack of intra-speaker variability. *VidTIMIT* is an exception, being comprised of video and corresponding audio recordings of 43 subjects. It was recorded into 3 sessions with around one week delay between each session. It can be useful for research involving automatic visual or audio-visual speech recognition or speaker verification [9].

2.2.2 Polycost

Establishing the *Polycost* corpus was an activity of the so-called COST 250 European project. It includes both native and non-native English from 134 speakers (74 M/60 F) from 13 European countries. Therefore it can not only be used in speaker recognition, but language and accent recognition as well. It has more than 5 sessions recorded over weeks in home/office environment by variable telephone handsets through digital ISDN.

2.2.3 YOHO

The *YOHO* corpus was designed for evaluating speaker verification in text-dependent situation for secure access applications. It consists 138 speakers' speech messages (106 M/32 F). It was recorded in multi sessions over a three months period by fixed high-quality handset in the office environment. The text read was prompted digit phrases.

3. ELSDSR

The intention of creating an English language speech database for speaker recognition is to obtain rich voice messages with respect to measure inter and intra speaker variability. Subjects are recruited in a Danish technical university environment. Most of them are non-native speakers of English. This database has been evaluated in a Master project for speaker recognition [10], and did provide a good speaker recognition rate.

3.1 Design and Recording

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.- and master students from department of Informatics and Mathematical Modeling, Technical University of Denmark. The text language is English, and is read by 20 Danes, one Icelander and one Canadian. There was no formal rehearsal, and perfect pronunciation is not obtained, nor necessary, for getting the specific and uniquely identifiable characteristics from individuals.

3.2 Description

In this section, a detailed description of *ELSDSR* based on the taxonomy of speech database is given.

3.2.1 Recording Environment

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

3.2.2 Recording Equipment

The equipment for recording 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 were recorded into the most commonly used file type-.wav (PCM). The sampling frequency is chosen 16 kHz with a bit rate of 16. Table 1 shows the initial setup for the recorder, for further details see the PMD670 user guide. Subjects were recorded in a single session, and new sessions will be recorded as an extension of this database, and will be announced soon. Unlike *TIMIT* which only includes short reading sentences, the speech messages in *ELSDSR* were extensive and comprehensive, therefore the intra-speaker variability was also collected, such as the changing of speaking rate and emotion, etc.

3.2.3 Corpus Speaker Set

ELSDSR contains voice messages from 22 speakers (12M/ 10F), and the age covered from 24 to 63. 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

Table 1. Recording Equipment Setup

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

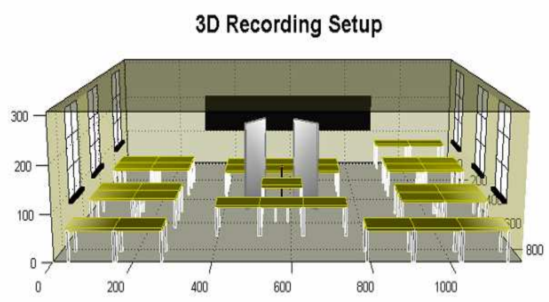

Figure 1. 3D Recording Chamber Setup

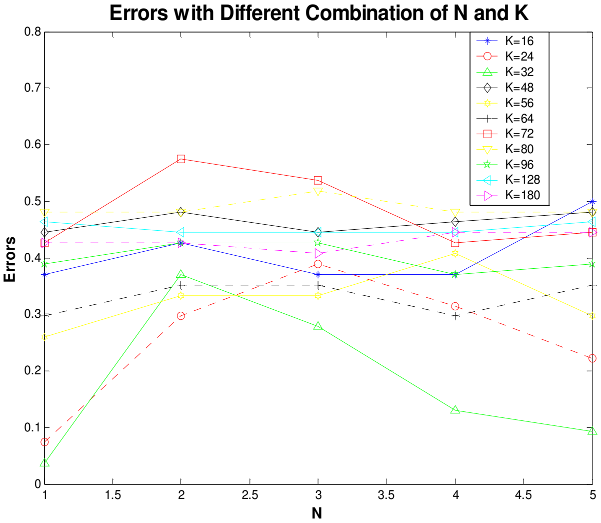

Figure 2. HMM speaker recognition error rates of using a range of HMM hidden state space dimensions (N) and a variable codebook dimensions (K). We generally find good performance for a 1D HMM indicating that there is little useful sequence information in text-independent speaker recognition.

site, the average age of female subjects is higher than that of male. 84% male speakers were between 26 and 37 years old; however the ages of female speakers spread in a large scale. Since the speakers were selected only in IMM, the speaker group exhibits relatively small variation in profession and educational background.

The subjects of this database were from different countries and different places of one country, the dialect of reading English language in this database can probably be used as accent recognition.

3.2.4 Corpus Text 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 including the vowels, consonants and diphthongs, etc. As for the suggested training subdivision, seven paragraphs of text were constructed and collected, which contains 11 sentences; with respect to the suggested test subdivision forty-four sentences (two sentences for each speaker) from NOVA Home [11] were collected. 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; and 83s for all. The duration for reading test data, on average, is: 16.1s (male); 19.6s (female); and 17.6s (for all). Table 2 shows the time consumption on reading both training and test text individually.

3.3 Results from Speaker Recognition Experiments

ELSDSR was used in a speaker recognition modeling master project, which is available in its entirety [10]. In [10] Mel-frequency Cepstral coefficients (*MFCC*) were extracted from text-independent speech. Both *TIMIT* and *ELSDSR* databases proved that 48 dimensional *MFCC* were the desired features for this case. Speaker pruning technique was introduced into the recognition system for the purpose of increasing the recognition accuracy with a little cost of speed. K-nearest neighbor (*KNN*) was implemented to carry out the speaker pruning of most dissimilar known speakers to the unknown speaker in *ELSDSR* within the 22 speakers. The selected (survived) speaker models will then be further recognized by Discrete-Density Hidden Markov Model (*DDHMM*). Within the speech messages from *ELSDSR*, we found out that performance is optimal or near optimal for single state HMM's, which indicates that for this type of text independent application there is little if any useful information in the sequence. This conclusion is consistent with D. A. Reynolds' review [12] and T. Matsui, S. Furui's comparative study [13]. Figure 2 demonstrates the conclusion by investigating the recognition error rates of using different selections of the number of states (N) and the number of codewords (K) in one codebook.

The highest recognition accuracy of the designed speaker recognition system achieved 92.07% with 8 candidates after speaker pruning and with 6s test speech. According to Reynolds's work in 1996 with HMM approach [14] in text-independent speaker verification, the recognition rate with 3s test speech recoded in telephone was 89%; and with 10s test speech accuracy became 94%.

3.4 Availability

A demonstration of the *ELSDSR* database is available from the site www.imm.dtu.dk/~lf/ELSDSR.htm. Academic researchers can contact the authors to obtain a free personal pass code for the complete database.

Table 2: Duration of Reading Training and Test Material

No.	**ID**	**Train** (s)	**Test** (s)
Male			
1	MASM	81.2	20.9
2	MCBR	68.4	13.1
3	MFKC	91.6	15.8
4	MKBP	69.9	15.8
5	MLKH	76.8	14.7
6	MMLP	79.6	13.3
7	MMNA	73.1	10.9
8	MNHP	82.9	20.3
9	MOEW	88.0	23.4
10	MPRA	86.8	9.3
11	MREM	79.1	21.8
12	MTLS	66.2	14.05
Average		78.6	16.1
Female			
13	FAML	99.1	18.7
14	FDHH	77.3	12.7
15	FEAB	92.8	24.0
16	FHRO	86.6	21.2
17	FJAZ	79.2	18.0
18	FMEL	76.3	18.2
19	FMEV	99.1	24.1
20	FSLJ	80.2	18.4
21	FTEJ	102.9	15.8
22	FUAN	89.5	25.1
Average		88.3	19.6
Total		1826.6	389.55

4. CONCLUSION

Based on several database surveys, we sorted out the current available speech corpus into fundamental taxonomy for speaker recognition studies, and reviewed some of most popular current publicly available speech databases. The new database for speaker recognition, *ELSDSR*, was introduced in detail with the purpose of distributing it to more researchers. It is a single-session database including 22 speakers reading speech (12M/ 10F), and was recorded in a noise free environment with a fixed microphone.

5. ACKNOWLEDGEMENTS

The database recording work lasted for one month, during which time a lot of help from colleagues, students, faculties in DTU was offered voluntarily. The work would not be carried out so smoothly without the help, assistance and support from them. The authors would like to thank all the people who supported the speech database building for their voice messages, time dedication and idea sharing. Warm thanks go to Ms. Ulla Nørhave for recruiting recording subjects and general assistance.

REFERENCES

[1] Melin, H. (2000), "Databases for Speaker Recognition: Activities in COST250 Working Group 2", In: COST250 - Speaker Recognition in Telephony, Final Report 1999 (CD-ROM), European Commission DG-XIII, Brussels, August 2000.

[2] Deller, J.R., Hansen, J.H.L., Proakis, J. G., "Discrete-Time Processing of Speech Signals", IEEE Press, New York, NY, 2000.

[3] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L., "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," *NIST*, 1993.

[4] Petrovska, D., Hennebert, J., Melin, H. and Genoud, D., Polycost: "A Telephone-Speech Database for Speaker Recognition," Speech Communication, 31(2-3), 2000, 265-270.

[5]Campbell, J., "Testing with The YOHO CD-ROM Voice Verification Corpus," *ICASSP*. Detroit, May 1995, p. 341-344.
http://www.biometrics.org/REPORTS/ICASSP95.html

[6] Melin, H. (1996), "Gandalf-a Swedish telephone speaker verification database", *ICSLP'96*, Philadelphia, USA, October, pp. 1954-1957.

[7] Hoge, H., Tropf, H.S., Winski, R., van den Heuvel H., Haeb-Umbach, R. & Choukri, K. (1997). "European Speech databases for telephone applications", *ICASSP'97*, Munich, Germany, pp. 1771-1774.

[8] Cole, R., Noel, M., Noel, V. (1998). "The CSLU speaker recognition corpus", *ICSLP'98*, Sydney, Australia, November 30-Decenber 4, pp.3167-3170.

[9] "Home Page of the VidTIMIT Database", 2001.
http://rsise.anu.edu.au/~conrad/vidtimit/

[10] Feng, L., "Speaker Recognition", Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2004

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

[12] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology", *Proc. ICASSP 2002*, Orlando, Florida, pp. 300-304.

[13] T. Matsui, S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs", *Proc. ICASSP*, vol. II, pp. 157-160, 1992.

[14] J. P. Campbell, JR., "Speaker Recognition: A Tutorial", *Proceedings of the IEEE*, vol. 85, no.9, pp. 1437-1462, Sep 1997.