2009 Seventh International Conference on Advances in Pattern Recognition

DA-IICT Cross-lingual and Multilingual Corpora for Speaker Recognition

Hemant A. Patil
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT)
Gandhinagar, India
hemant_patil@daiict.ac.in

Sunayana Sitaram
Computer Engineering Department
S. V. National Institute of Technology
Surat, India
sunayana.sitaram@gmail.com

Esha Sharma
Infosys Technologies Limited
Pune, India
esha_sharma@infosys.com

Abstract— **In this paper the design and development of the DA-IICT Cross-lingual and Multilingual Speech Corpora is presented which includes unconventional sounds like cough, whistle, whisper, frication, idiosyncrasies, etc. from bilingual subjects (i.e., who can speak Hindi and Indian English) and trilingual subjects (who can speak Hindi, Indian English and mother tongue) for the development of Automatic Speaker Recognition System. Thirteen Indian languages and the Nepali language are considered as the subjects' mother tongue/native languages. Unconventional sounds are considered to examine how much speaker-specific information they carry. Finally, an ASR system based on spectral or cepstral features (i.e., LPC, LPCC, MFCC) and polynomial classifier of 2^{nd} order approximation is presented to evaluate the developed corpora.**

Keywords- Cross-lingual and multilingual speaker recognition, data collection and corpus design, linear prediction, Mel cepstrum, and polynomial classifier.

I. INTRODUCTION

Automatic Speaker Recognition (ASR) deals with the identification of person's identity with the help of machines. Speaker recognition by humans is extremely robust, often having the ability of recognizing speakers from their voice using multiple levels of speaker information conveyed in the speech signal including a unique laugh, idiosyncrasies or whispered speech. We can improve the reliability and accuracy of speaker recognition systems by exploiting these sources of information in the speech signal [8].

ASR has mainly been explored in American English for mono-lingual speaker recognition. The success rate in ASR degrades for the cross-lingual and multi-lingual tasks as opposed to monolingual task [11]. The potential of unconventional sounds, which also carry speaker-specific information, has not been explored in ASR literature. In this paper, we describe the design and development of the Cross-lingual and Multilingual Corpora which include sounds like cough, whisper, whistle, frication, idiosyncrasies etc. which we term as "unconventional" sounds. Recognizing a speaker with the help of these unconventional sounds could be useful in situations where the speaker is not co-operative. For example, a person whispers when he does not want to be heard or recognized. Also, finding out the relative amount of speaker-specific information present in these unconventional sounds is an interesting exercise in itself.

A. Organization of the paper

The paper is organized as follows: Section 2 describes some of the features of other publicly available corpora for ASR tasks. Section 3 describes the features of the corpora; sections 4 and 5 describe details of the corpora. Section 6 describes the results obtained for ASR. Section 7 summarizes the paper.

II. PREVIOUS WORK

Recently, Campbell *et al.* [2] developed the *MMSR* cross-lingual and cross-channel corpora. In their work, for developing cross-lingual corpora, they have considered American English as the default language and Arabic, Mandarin, Russian or Spanish as the second language. The MMSR corpus is limited to bilingual speakers only.

The YOHO corpus is suitable for speaker verification experiments [5]. However, the text material is limited to prompted phrases and the acoustic environment is limited to an office. The OGI speaker recognition corpus consists of prompted phrases, digits and prompted monologue. However, the acoustic environment is limited to the home and office environments [6]. The switchboard corpus is suitable for text-independent ASR work (including the NIST evaluation subsets [7], [9] of conversational speech). However, it has a limited acoustic environment of an office.

The DA-IICT corpora, consisting of speech from 137 subjects have been specifically designed for cross-lingual and multilingual speaker recognition tasks. They consist of different types of speech including read speech, combination-lock phrases and spontaneous speech. In addition to this, we have also incorporated unconventional sounds of the subjects in the corpus. They contain a wide range of acoustic environments with varying levels and types of noise. To the best of authors' knowledge, there are no publicly available corpora having all these features.

978-0-7695-3520-3/09 $25.00 © 2009 IEEE
DOI 10.1109/ICAPR.2009.72

III. FEATURES OF THE CORPORA

In this section, different features of the DA-IICT corpora such as languages covered, unconventional sounds, age, gender profiles of the subjects and acoustic environments have been described.

A. Languages covered

Out of the 137 subjects considered, 122 of them were trilingual and their voice was recorded in Indian English, Hindi and their mother tongue. 15 subjects were bilingual and were recorded in Indian English and their mother tongue. Figure 1 illustrates the languages used as mother tongues and their percentages in the corpus.

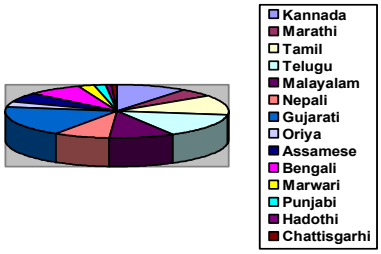

Figure 1: Languages used as mother tongues in the DA-IICT corpora.

Due to the number of different languages considered, the corpora can also be used for Language Identification (LID) experiments.

Table 1: Unconventional sounds in the DA-IICT corpora.

Sound	Number of files
Cough	126
Whisper	111
Whistle	90
Laugh	81
Teasing	79
Fricative	37
Other	137

B. Unconventional sounds

During the recording, subjects were asked to produce unconventional sounds. The unconventional sounds we have considered are cough, laugh, whisper, whistle, fricative, teasing (which is the clearing of the throat or a suggestive cough) and idiosyncrasies of the speaker which we have classified as "other". Table 1 shows the numbers of unconventional sounds in the corpora.

C. Acoustic enviroments

A wide range of acoustic environments having varying levels of noise were used while creating the corpora. Subjects were recorded at engineering college in the classrooms, labs, hostels, cafeteria, garden and staff club. We did not record any subject in a closed sound booth so that we could create a natural environment for recording. Background noise included other people voices, noise due to computers, mobile phones, vehicles and birds (in case of speech recorded outdoors). Noise levels have been classified into three broad categories based on the location and observed experiences during recording and during playing back the files. The noise level varied from high to low, a summary of which is given in Table 2 below.

Table 2: Distribution of noise levels in the DA-IICT corpora.

Noise level	Percentage in corpora
Low	46%
Medium	38%
High	16%

D. Age and gender profiles of subjects

The subjects considered were undergraduate students of DAIICT Gandhinagar and NIT Surat. The age of the subjects ranged from 17-22 years. 86 females and 51 males were considered as subjects in the corpus.

IV. BUILDING THE CORPORA

Recording was done using Sony Sound Forge 9.0 with the help of Creative Headset HS-300 with noise canceling microphone. The subjects were not paid; their participation in the data collection was purely voluntary. The subjects were recorded in a single session mostly during the evening or night hours.

A. Text material used

A list consisting of questions, isolated words, digits, combination-lock phrases and sentences was prepared in all the languages. The speaker was asked to speak spontaneously on any topic of their choice at the end of the recording for each language. The contextual speech consisted of a description of the speaker, his/her family and friends, native place or some memorable event. Due to the varied nature of the topics, the speech was mostly conversational. Speakers were posed with some questions about any of the above topics to motivate them to speak fluently. The interview was started with a few questions to know about the speaker such as his/her name, age, education, profession, etc. After that the list was given to the speaker to read in his or her own way. During the contextual speech, speakers were also asked to produce unconventional sounds. The data was recorded with 10 repetitions except for the contextual speech and unconventional sounds, to track all the possible variations in speech. Table 3 shows the distribution of the type of speech in the corpus. Table 4 gives the distinct features of the corpora.

Table 3: Approximate distribution of type of speech in the DA-IICT corpora.

Type of Speech	Percentage in corpora
Read sentences	20
Read words	15
Read numbers	25
Spontaneous	30
Unconventional	10

Table 4. Description of the DA-IICT corpora.

Item	Details

No. of speakers	137
Specialty of speakers	Most of the speakers studied in English medium
No. of sessions	1 or 2
Data type	Speech
Sampling rate	22,050 Hz
Sampling format	1-channel, 16-bit resolution
Type of speech	Read sentences, isolated words and digits, combination-lock phrases, questions, contextual speech of considerable duration, unconventional sounds
Application	Text-independent Multilingual and Cross-lingual ASR system, ASR system using unconventional sounds
Training language	Bengali, Kannada, Tamil, Marathi, Telegu, Gujarati, Malayalam, Oriya, Assamese, Hadothi, Marwari, Chattisgarhi, Punjabi, Nepali, Hindi and Indian English
Testing language	-do-
No. of repetitions	10 except for contextual speech and unconventional sounds
Microphone	Creative Headset HS-300 noise canceling microphone
Recording Software	Sony Sound Forge 9.0
Acoustic environment	Cafeteria, hostel, classroom, staff club, lab, garden

B. Data acquisition

The speech files recorded using Sound Forge were stored in '*.wav'* format. While recording, the dc level was adjusted in the software in order to eliminate the dc offset produced by the audio hardware. The interviewer's voice was deleted from the speech file so that models of the actual speaker can be made.

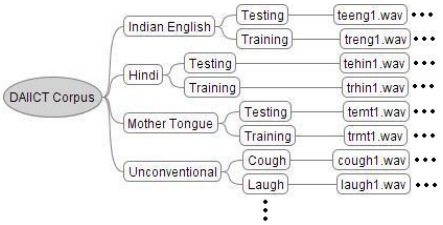

Figure 2: Organization of DA-IICT corpora

C. Organization of the Corpora

After the data was collected, each speaker's file was split into training and testing files corresponding to each language. Separate files were created for each of the unconventional sounds. The files were organized as shown in Figure 2. The corpora consist of a total of 20-25 minutes of speech of each of the 137 subjects. In each language, the testing file consists of the speaker's introduction based on questions asked and read sentences. The training file consists of spontaneous speech, combination-lock phrases and read words.

V. EFFORTS AND EXPERIENCES DURING DATA COLLECTION

In this section, different challenges faced and experiences during field recording have been described in brief.

A. Challenges faced

A few subjects were initially reluctant to spare approximately 30 minutes required for recording their voice in three languages. Some subjects hesitated in giving the unconventional sounds and in speaking their mother tongue. Different dialects, temperamental changes of subjects across different sessions, boredom in repeating same text material, and intersession variability are some of the distinct challenges observed during the data collection.

B. Observations

Almost all subjects had studied in the English medium till their higher secondary. Most subjects had studied Hindi as a subject in school, while some of them, especially the subjects from Kerela and Tamil Nadu had not studied it or even spoken it before coming to Gujarat. Subjects who had resided in states other than their native state had more difficulty in pronouncing the numbers in their mother tongue, possibly because they had not studied it as a subject in school. Similarly, subjects who had not studied Hindi in school had difficulty in pronouncing the numbers in Hindi. While playing back the files, it has been observed that it was relatively easy to identify the speaker with the help of their whispered speech and laugh to a certain extent, compared to the other unconventional sounds.

VI. EXPERIMENTAL RESULTS

In this paper, polynomial classier of 2^{nd} order approximation is used as the basis for all the experiments [3]. Feature analysis was performed using a 23.2 ms duration frame with an overlap of 50%. Hamming windows was applied to each frame and subsequently, each frame was pre-emphasized with the filter *(1-0.97z-1)*. A *12th* order LPC were extracted for frame of 23.22ms (512 samples) duration after pre-processing. LPCC was calculated from roots of LPC polynomial [1]. The standard 12 MFCC computations per frame were performed as per method suggested in [4]. Table 5 shows speaker recognition results in the form of average success rates (computed over testing segment of 1s, 3s, 5s 7s, 10s, 12s, and 15s) for in mono-lingual and cross-lingual task with for population size of 97 speakers (with 1 min. training) whereas Table 6 shows success rates when the unconventional sounds are used to test against normal speech in English language for 34 speakers (with 1 min. training).

Table 5. Average success rates (%) for monolingual, cross-lingual, and multilingual ASR task for 1min. training.

Train/Test	LPC	LPCC	MFCC
MT/MT	**53.90**	**64.94**	**56.84**
MT /H	46.09	53.90	51.54
MT/E	45.80	54.19	46.53
E/E	**53.60**	**66.42**	**56.11**
E/H	53.75	63.32	63.03

E/MT	43.44	50.07	47.12
H/H	**53.75**	**63.32**	**63.03**
H/E	42.41	53.90	50.36
H/MT	44.62	54.19	50.07

MT =Mother Tongue, E = English, H = Hindi, MT/E = Mother tongue is used for training and English is used for testing.

Table 6. Success rates (%) for unconventional sounds (Training duration=1min.).

Sounds	Test (sec)	LPC	LPCC	MFCC
E/E	Average	64.28	73.52	78.15
Cough	1.85	32.35	47.05	50.00
Whisper	1.85	35.29	55.88	44.11
Whistle	0.78	29.41	35.29	41.17
Frication	1	17.64	29.41	35.29
Idiosyncrasies	1	32.35	47.05	44.11

E/E= English is used for training and English is used for testing, Average= computed over testing segment of 1s, 3s, 5s 7s, 10s, 12s and 15s)

Some of the observations from the results are as follows:

1) The results are better for *monolingual* experiments than the cross-lingual (Table 3). Thus language consistency matters for speaker recognition and for cross-lingual ASR experiments; the system is doing language recognition as much as it is doing speaker recognition [11]. This may be due to the fact that during training, we train and build polynomial models for each speaker with one language (say Bengali) having particular set of sound units. Now during testing with another language (say English), the feature vector falls into different zone of feature space because training feature vectors distribution has different language–specific features,
2) For cross-lingual experiments with Hindi used for training, results are better when mother tongue is used for *testing* than when English is used for testing. This shows that for cross-lingual ASR experiments, there is a significant dependence of testing language on ASR performance,
3) For cross-lingual experiments, the results are better with English as training language as compared to Hindi. This may be due to the fact that the subjects used in this study didn't study Hindi in their school curriculum and hence generally prefer English as their secondary language compared to Hindi (Table 5),
4) It is evident from Table 6 that unconventional sounds also carry significant speaker-specific information. This is one of the significant results in the paper which demonstrates that speaker recognition is possible from unconventional sounds as well and this observation may find its use in forensic science where the subject is highly uncooperative. So a small signature from suspect's voice in the form of cough, whisper, whistle, frication, idiosyncrasies, etc. may be a good clue to find the best match for him,
5) For speaker recognition using unconventional sounds, MFCC performed better than LPC and LPCC in majority of the cases.
6) With respect to testing speech duration, whistle produced by subject carries more dominant speaker-specific information followed by frication as compared to other unconventional sounds.

VII. SUMMARY AND CONCLUSIONS

In this paper, an experimental setup to build up corpora for monolingual, cross-lingual and multilingual or trilingual ASR systems in Indian languages is presented. Such trilingual corpora are rarely available for investigation of cross-lingual and multilingual ASR experiments. An ASR system is presented to evaluate the developed corpora. The degradation in results for cross-lingual speaker recognition is justified using feature clustering in different regions of feature space. In addition to this, major innovative contribution of the paper is ASR experiments to show that speaker recognition is possible from unconventional sounds.

REFERENCES

[1] Atal, B. S., "Effectiveness of linear prediction of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Amer., 55(6):1304-1312, 1974.

[2] Campbell, J. P., *et al.*, "The MMSR Bilingual and Cross Channel Corpora for Speaker Recognition Research and Evaluation", in the Proceedings of the Speaker and Language Recognition Workshop, Odyssey'04, Toledo, Spain, 29-32: 2004.

[3] Campbell, W. M., Assaleh, K. T. and Broun, C. C., "Speaker recognition with polynomial classifiers", IEEE Trans. On Speech and Audio Processing, 10(4):205-212, 2002.

[4] Davis, S. B. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. Acoustic, Speech and Signal Processing, 28(4):357-366, 1980.

[5] Campbell, Jr., J. P., "Testing with The YOHO CD-ROM voice verification corpus" in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP'95:341-344, 1995.

[6] Campbell, Jr., J. P. and Reynolds, D. A., "Corpora for the evaluation of speaker recognition systems" in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP'99:829-832, 1999.

[7] Godfrey, J. J., Holliman, E. C. and McDaniel, J., "Switchboard: Telephone speech corpus for research and development" in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP'92:517-520, 1992.

[8] Jin, Q., "Robust Speaker Recognition", PhD Thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2007.

[9] Martin, F. and Przybocki, M.A., "The NIST speaker recognition evaluations: 1996-2001" A Speaker Odyssey, A Speaker Recognition Workshop, 2001.

[10] Przybocki, M. A., Martin, A. F. and Le A. N., "NIST Speaker Recognition Evaluations Utilizing the Mixer Corpora—2004, 2005, 2006," IEEE Trans. Audio, Speech, and Language Processing, 15(7), 1951-1959, Sept. 2007.