Text-dependent Speaker Verification and RSR2015 Speech Corpus

Anthony Larcher and Haizhou Li

RSR2015 (Robust Speaker Recognition 2015) is the largest publicly available speech corpus for text-dependent robust speaker recognition. The current release includes 151 hours of short duration utterances spoken by 300 speakers. RSR2015 is developed by the Human Language Technology (HLT) department at Institute for Infocomm Research (I2R) in Singapore. This newsletter describes RSR2015 corpus that addresses the reviving interest of text-dependent speaker recognition.

WHY ANOTHER CORPUS?

Speaker verification has reached a maturity that allows state-of-the-art engines to be used for commercial applications. During the last decade, the effort of the community has lead to great improvements in terms of performance and usability of speaker verification engines. Nevertheless, it is well known in the community that performance of text-independent speaker verification engines suffers from the lack of enrolment and training data.

In the context of short duration, text-dependency is known to improve accuracy of speaker verification when dealing with short duration speech segments. Despite its strong commercial potential, text-dependent speaker recognition lies on fringes of the main stream speaker recognition. As a result, the speech resources available for such research are either too small or inadequate to take advantage of the technologies develop for the mainstream text-independent speaker verification.

In view of the fact that text-dependent speaker verification and user-customized command and control, that recognizes a user-defined voice command at the same time identifies the speaker, are useful application scenario. RSR2015 is developed for the following objectives.

- To provide a database of reasonable size that supports significance test of text-independent speaker recognition. RSR2015 contains recordings from 300 speakers during 9 sessions in order to create enough target and impostor trials.
- To have a gender-balanced database that allows fair analysis of gender influence in text-dependent speaker recognition. RSR2015 involves 143 female and 157 male speakers.
- To allow for analysis of the phonetic variability in the context of text-constrained speaker recognition [2]. RSR2015 protocol includes more than 60 different utterances spoken by all speakers.

WHAT CAN WE DO WITH RSR2015?

RSR2015 allows for simulation and comparison of different use-cases in terms of phonetic content. For example, the most extreme constraint is to fix a unique utterance for all users of the system all the time. In the case where a larger set of fixed pass-phrases is shared across users, the scenario becomes very similar to user-customized command and control application. On the other hand, it is possible to limit the phonetic content of the speech utterance by randomly prompting sequences of phones or digits. In this case, the context of the phone varies across sessions and especially between enrolment and test.

The choice of a specific scenario depends on what constraints we would like to impose on the users. Unfortunately, no existing database allows for a comparison of speaker recognition engines across scenarios in similar conditions. RSR2015 is designed to bridge the gap. It consists of fixed pass-phrases, short commands and random digit series recorded in the same conditions.

DATABASE DESCRIPTION

RSR2015 contains audio recordings from 300 speakers, 143 female and 157 male in 9 sessions each, with a total of 151 hours of speech. The speakers were selected to be representative of the ethnic distribution of Singaporean population, with age ranging from 17 to 42.

The database was collected in office environment using six portable devices (four smart phones and two tablets) from different manufacturers. Each speaker was recorded using three different devices out of the six. The speaker was free to hold the smart phone or tablet in a comfortable way.

To facilitate the recording, a dialogue manager was implemented on the portable devices as an Android© application. The speakers interact with the dialogue manager through a touch screen to complete the recording. Each of the 9 sessions for a speaker is organized into 3 parts:

PART 1 – Short-sentences for pass-phrase style speaker verification (71 hours)

All speakers read the same 30 sentences from the TIMIT database [3] covering all English phones. The average duration of sentences is 3.2 seconds.
Example: “Only lawyers love millionaires.”

PART 2 – Short commands for user-customized command and control (45 hours)

All speakers read the same 30 commands designed for the StarHome applications. The average duration of short commands is 2 seconds.
Example “Light on”

PART 3 – Random digit strings for speaker verification (35 hours)

All speakers read the same 3 10-digit strings, and 10 5-digit strings. The digit strings are session dependent.

INTERSPEECH 2014 SPECIAL SESSION AND BENCHMARKING

During INTERSPEECH 2014, The Institute for Infocomm Research together with

IBM Research, and the Centre de Recherche en Informatique de Montreal (CRIM) propose a special session on *Text-Dependent Speaker Verification with Short Utterance*. The purpose of this special session is to gather the research efforts from both the academia and industries toward a common goal of establishing a new baseline and explore new directions for text-dependent speaker verification. For ease of comparison across systems, the RSR2015 database, that comes with several evaluation protocols targeting at different scenarios, has been proposed to support research submitted for the INTERSPEECH 2014 special session [4, 5].

WHERE TO GET THIS DATABASE?

The license is available at

http://www.etpl.sg/innovation-offerings/ready-to-sign-licenses/rsr2015-overview-n-specifications

Please contact Dr Anthony Larcher at Email alarcher [at] i2r.a-star.edu.sg or Dr Kong-Aik Lee at kalee [at] i2r.a-star.edu.sg.

REFERENCES:

[1] K. A. Lee, A. Larcher, H. Thai, B. Ma and H. Li, “Joint Application of Speech and Speaker Recognition for Automation and Security in Smart Home”, in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2011, pp 3317-3318.

[2] A. Larcher, P.-M. Bousquet, K. A. Lee, D. Matrouf, H. Li and J.-F. Bonastre, “I-vectors in the context of phonetically-constrained short utterances for speaker verification”, In ICASSP 2012

[3] W. M. Fisher, G. R. Doddington and K. M. Goudie-Marshall, “The DARPA speech recognition research database: specifications and status”, *DARPA Workshop on Speech Recognition,* 1986, pp. 93-99

[4] A. Larcher, K. A. Lee, B. Ma and H. Li, “RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases”, in *Annual Conference of the International Speech Communication Association (Interspeech),* **2012**, 1580-1583

[5] A. Larcher, K. A. Lee, B. Ma and H. Li, “Text-dependent Speaker Verification: Classifiers, Databases and RSR2015,” *Speech Communication,* **2014,** in press