The RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases

Anthony Larcher, Kong Aik Lee, Bin Ma, Haizhou Li

▶ **To cite this version:**

Anthony Larcher, Kong Aik Lee, Bin Ma, Haizhou Li. The RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases. Annual Conference of the International Speech Communication association (Interspeech), Sep 2012, Portland, United States. hal-01927726

HAL Id: hal-01927726
https://hal.science/hal-01927726v1

Submitted on 20 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases

Anthony Larcher, Kong Aik Lee, Bin Ma, Haizhou Li

Institute for Infocomm Research (I2R)
A*STAR, Singapore
alarcher@i2r.a-star.edu.sg, kalee@i2r.a-star.edu.sg

Abstract

This paper describes a new speech corpus, the *RSR2015* database designed for text-dependent speaker recognition with scenario based on fixed pass-phrases. This database consists of over 71 hours of speech recorded from English speakers covering the diversity of accents spoken in Singapore. Acquisition has been done using a set of six portable devices including smart phones and tablets. The pool of speakers consists of 300 participants (143 female and 157 male speakers) from 17 to 42 years old. We propose a protocol for the case of user-dependent pass-phrases in text-dependent speaker recognition and we also report speaker recognition experiments on *RSR2015* database.
Index Terms: database, speaker recognition, text-dependent

1. Introduction

Speaker verification is a binary classification task set to answer the question: *"Is this speaker the person he/she claims to be?"*. To do so, an automatic system compares the speech material provided by the current speaker to a reference of the target user learned during a previous enrolment phase [1]. One of the key challenges in this binary classification task is the variability of speech samples captured by the system. Such variability results from various sources due to the way the signal is recorded (background noise, transmission channel) or to the intrinsic variability of the speech produced by the speaker (language, emotion, duration, lexical content). Many studies have shown that the more mismatched enrolment and test samples are, the more difficult the classification is [1].

In the case of cooperative users, some of the variability sources can be mitigated by placing some reasonable constraints upon the users. Text-dependent speaker verification [2] is the specific case of verification where variability is limited by fixing the text to be pronounced by the speaker during both enrolment and test phases. This constraint reduces both the effects of lexical and duration mismatch. Contrary to text-independent speaker verification that requires at least one minute of speech to reach high accuracy [3], text-dependent verification focuses on short duration utterances. Indeed, reducing the duration and lexical variability improves significantly the performance of automatic systems [2]. The use of short utterances is reinforced by the ergonomic convenience to the user to speak less than a few seconds.

There are several scenarios that constrain the duration and lexical content of speech utterances. The choice of considering one scenario rather than the another mostly depends on the application. In this regard, we distinguishes three main use-case scenarios:

UNIQUE PASS-PHRASE: each client pronounces the same pass-phrase. This is the most constrained scenario as both duration and text are fixed [4].

USER-DEPENDENT PASS-PHRASE: each client pronounces his own pass-phrase (chosen or generated by the system). In this scenario, duration and lexical content vary between speakers [5, 6].

PROMPTED TEXT: each client pronounces a sentence prompted by the system. This scenario does not require the user to remember a specific pass-phrase and reduces the risk of replay attacks. Duration variability can be easily reduced by adding a constraint on the prompts while lexical variability can be decreased by limiting the phonetic content of the prompts. A very common approach consists of using series of randomly ordered digits [7, 8].

In this paper, our focus is on the second use-case where each user has his/her own fixed pass-phrases.

Many databases have been recorded to develop and evaluate text-dependent speaker verification engines [9, 10, 11, 12, 13, 14, 15, 16, 17, 18] under various scenarios. However these databases are designed to fulfil needs in specific research directions, most of them are limited in terms of speaker or lexical variability which are critical in the case of user-dependent pass-phrase application.

The Robust Speaker Recognition 2015 (*RSR2015*) database allows for simulation and comparison of speaker verification systems in user-dependent English pass-phrase use-case. Indeed, the security of a large number of commercial applications relies on user-specific passwords. Those existing passwords could be directly used as user-dependent pass-phrases when reinforcing the security level by adding a speaker verification module. The aim of developing *RSR2015* database is to provide the community with a speech corpus that can be used to study the influence of lexical content in short utterances for speaker verification as none of the existing databases offer together as many speaker and lexical variability. The database as described in this paper is the first part of a larger corpus that includes two additional parts not described in this paper[1]. Parts 2 and 3 are designed to develop loaded command-control applications [19] and to evaluate speaker verification engines on randomly prompted digits.

In the next section we first give an overview of databases available for text-dependent speaker verification. Section 3 describes the *RSR2015* database and gives some statistics on the released data. A protocol designed for user-dependent pass-phrases use-case is proposed in Section 4. Section 5 presents initial results obtained on *RSR2015* database. Finally, we discuss future developments and perspectives in Section 6

[1] Part 2 and 3 of *RSR2015* database have not been released yet.

2. Overview of existing databases

During the last 20 years, many databases including constrained speech have been recorded by the scientific community. In this section we give an overview of the most commonly used and publicly available databases for text-dependent speaker verification on fixed pass-phrases. Note that other databases can be found in the literature without being publicly available (e.g. [2, 4]). The rest of this section is organized as follow: Table 1 gives a summary of nine publicly available databases; we then propose some comments on these databases regarding the very specific use-case of text-dependent speaker verification based on user-dependent pass-phrases which requires as many different speakers and pass-phrases as possible.

Table 1: Overview of existing publicly available databases that include text-dependent speech material.

Corpus	#Speakers (male/female)	#Sessions	Lexical content
BANCA [13]	208 (104/104)	12	personal information 1 fixed digit sequence
BIOMET [14]	91 (45/46)	3	personal information
BIOSEC [18]	200 (?/?)	2	fixed digit sequence
BT-DAVID [10]	31+92 imp (15/16)	5	fixed digit sequences and VCVCV sentences
M2VTS [11]	37 (30/7)	5	fixed digit sequence
MIT [17] **MDSVC**	88 (49/39)	6	fixed 2-word phrases
MyIdea [16]	30 (30/0)	3	fixed digit sequence fixed 7-word phrase
Valid [15]	106 (76/30)	5	1 fixed sentence 1 fixed digit sequence
XM2VTS [12]	295 (158/137)	4	fixed digit sequence fixed 7 word-phrase

Speaker representativeness
Evaluation of automatic classifiers requires a large number of trials in order to produce statistically significant results. However, we first notice that six of the considered databases involve less than 150 speakers. The number of sessions and thus number of target trials is limited as only two databases (BANCA and MIT MDSVC) contain more than 5 sessions per speaker. Moreover, in three of these databases (M2VTS, MyIdea and Valid) gender representation is strongly unbalanced. This may be explained by the fact that many of the corpora are collected within institutions were both gender are not equally represented.
It is worth noting that all the databases referred in Table 1 except the MIT MDSVC, include data from multiple modalities. The nature of the corpora probably explains the limited number of recorded subjects as multi-modal acquisition increases the cost and the complexity of the task.

Pass-Phrase variability
When targeting user-dependent pass-phrase application, lexical variability of the speech utterances is another important criteria. Amongst the nine pre-cited databases, six of them designed to supply different scenario, only contain one fixed sequence of digits and one or two pass-phrases. This limited number of possibly used pass-phrases makes impossible any study on the effect of pass-phrase variability.

After this short survey, it appears that most of the existing databases targeted different scenario and do not match our current expectations. Other databases that may be usable for user-dependent pass-phrase scenario contain a limited number of recorded speakers. Section 3 presents *RSR2015* database, a new database designed for user-dependent pass-phrase text-dependent speaker verification.

3. Database description

The *RSR2015* database contains audio recordings from 300 persons, 143 female and 157 male speakers. Participants were selected to be representative of the ethnic distribution of Singaporean population. Selected speakers were between 17 to 42 years old. Each of the participants recorded nine sessions using three portable devices. Each session consists of thirty short sentences.

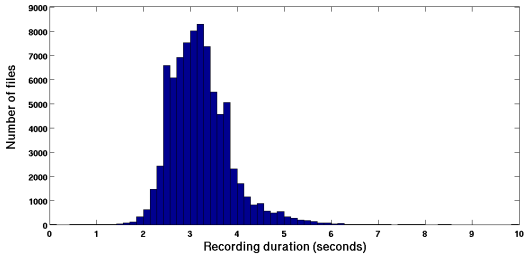
3.1. Recording conditions and protocols

The database was collected in office environments 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. We labelled the three devices as A,B and C. The following recording sequence was applied for each speaker: {A,B,C,A,B,C,A,B,C}. The subject was free to hold the smart phone or tablet in a way he/she was comfortable. Acoustic quality can thus vary significantly. For the recording, a dialogue manager was implemented on the portable devices as an Android© application. This application uses the touchscreen capability of devices to display pass-phrases while controlling the recording. A so-called *push-to-talk* feature was used to allow the user to start the recording and stop it after reading the prompted pass-phrase.

3.2. Lexical and duration variability

For each session, a speaker read thirty short sentences. We selected the sentences from TIMIT database [20] by making sure that the sentences cover all English phones. The number of words per sentence varies from four to eight. During recording, speakers had to pronounce each pass-phrase using the *push-to-talk* mechanism described previously. Due to this protocol, the length of recording was controlled by the user. Thus, recordings contain varying durations of silence before and after the sentences were read. The duration variations over all speakers and all sentences is illustrated by the histogram as shown in Figure 1. The average duration is 3.2 seconds.

Figure 1: Distribution of utterance duration for all 30 sentences from 300 speakers of *RSR2015* database (in seconds).



4. Evaluation protocol

The database is randomly split into three non-overlapping groups of speakers, one for background training, one for devel-

opment and one for evaluation. Each group contains a balanced number of male and female speakers: 50/47 for the background set, 50/47 for the development set and 57/49 for the evaluation set.
The recommended use of the background training set is to derive background models or to be used for normalization techniques [2]. Note that to avoid the use of the 30 pass-phrases in background model training, data from the same 97 speakers recorded for the Parts 2 and 3 of the *RSR2015* database can be used. The development set has to be used to estimate a decision threshold, calibration and fusion parameters that could then be used on the evaluation set. Thus, data from both development and evaluation sets are used for enrolment and test. Development and evaluation data are again split into two parts, three sessions for enrolment and six sessions for test. The enrolment sessions are chosen so that each speaker was recorded with the only one device which is different from the ones used to record the test sessions.

4.1. Enrolment and test

Enrolment of a client is the same for development and evaluation sets. In order to maintain the enrolment duration below 10 seconds which is deemed to be reasonable for commercial applications [2], each user enrols using only three occurrences of a specific pass-phrase (one from each training session). A total of 30 models are trained for each speaker, one for each pass-phrase.

Table 2: Different types of trials considered for experiments. Note the case where impostor speakers pronounce a wrong pass-phrase is not considered as it does not reflect a genuine imposture.

	Correct Pass-phrase	Wrong Pass-Phrase
Target User	CLIENT-true	CLIENT-wrong
Impostor	IMP-true	IMP-wrong

Tests are gender-dependent, meaning that specific threshold and calibration parameters could be derived for each gender. In text-dependent speaker verification context, four types of trial (summarized in Table 2) can be considered given that the test utterance is spoken by the target user or not and that the spoken utterance is the correct pass-phrase or not. Amongst those four types of trials, *IMP-wrong* is not considered in this protocol. Indeed, we consider that an impostor would either pronounce the correct pass-phrase (IMP-true) or play a recording of the target user pronouncing a different lexical content (CLIENT-wrong) as those two types of trials are more representative of a genuine imposture. All six test sessions are used during the test. Each client model is compared to all the other speakers from the same set (dev or eval). The numbers of trial generated are given in Table 3 for each gender and speaker set.

Table 3: Number of tests per speaker-set for each type of trial.

Type of trial	Male		Female	
	dev	eval	dev	eval
CLIENT-true	8,930	10,244	8,460	8,819
CLIENT-wrong	259,001	267,076	245,340	257,751
IMP-true	437,631	573,664	389,160	423,312

4.2. Performance measures

Performance is evaluated in terms of Equal Error Rate (EER) and Detection Cost Function (DCF). The Detection Cost Function C_{Det} follows the definition proposed in NIST-SRE'08 [2] and given below,

$$C_{Det} = C_{Miss} \times \mathcal{P}(Miss|Target) \times \mathcal{P}_{Target} + C_{FA} \times \mathcal{P}(FA|NonTarget) \times (1 - \mathcal{P}_{Target}) \quad (1)$$

where $C_{Miss} = 10$ and $C_{FA} = 1$ are the relative costs of detection errors (respectively accepting an impostor and rejecting a target user), $\mathcal{P}_{Target} = 0.01$ is the *a priori* probability of the specified target speaker.

5. Initial results

We present below experiments performed on *RSR2015* database by using a real time text-dependent speaker verification system which has been deployed in *STARHome*[3], a fully functional smart home prototype located at the Fusionopolis, Singapore [19].

5.1. System description

Front end processing extracts acoustic features of 50 coefficients (19 Linear-Frequency Cepstral Coefficients, their derivatives, first 11 second derivatives and the delta energy). Acoustic features are computed on a sliding window of 20ms with a shifting of 10 ms. The frequency window is restricted to 300-3400 Hz. Features of lower energy are discarded and simple feature normalization is applied, so that the distribution of each cepstral coefficient is 0-mean and 1-variance for a given utterance. Two gender-dependent 256-Gaussian Universal Background Models (UBM) are trained using respectively 21 and 23 hours of speech from 47 female and 50 male speakers belonging to the Background speaker set of the *RSR2015* database Part 2.

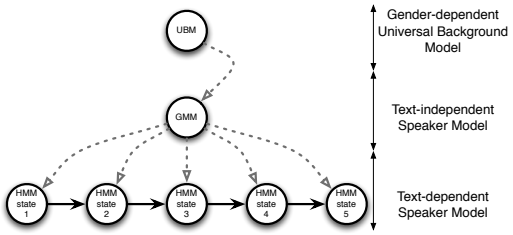


Figure 2: *Acoustic architecture of the HiLam speaker verification system.*

The Hierarchical multi-Layer Acoustic Model (HiLam) speaker verification system is based on a three layer acoustic architecture depicted in Figure 2 . Speaker text-dependent models are 5-state Hidden Markov Models (HMM) obtained through a two-step adaptation process. During the first step, a text-independent Gaussian Mixture Model (GMM) is adapted from the corresponding gender-dependent UBM by using the Maximum A Posteriori (MAP) criteria. A speaker-dependent, text-dependent HMM is then adapted from the GMM text-independent model. Each state of the HMM is a GMM adapted from the text-independent GMM of the speaker by using the MAP criteria. A description of HiLam system is given in [19]. In this work, in order to present the more generic baseline, decision threshold is user-independent and computed on each dataset.

[2] http://www.itl.nist.gov/iad/mig/tests/spk/2008/index.html

[3] http://www.starhome.sg

5.2. Example of performance

Performance of the HiLam system on the development and evaluation sets of *RSR2015* database are given to allow future comparison or references (in Table 4).

Table 4: Performance of the HiLam speaker verification system on the different speaker sets of the in terms of Equal Error Rate (%), minimum Decision Cost Function (min DCF).

Speaker set	Male		Female	
	dev	eval	dev	eval
Equal Error Rate %	1.74	0.93	0.49	0.90
min DCF ($\times$100)	0.75	0.41	0.29	0.53

As expected when dealing with text-dependent speaker verification and such recording environments, Equal Error Rates obtained by our system are low-less than 1% for three of the speaker groups and less than 2% for the fourth group. This performance highlights the importance of providing enough trials to evaluate text-dependent speaker verification systems.

Comparing both genders it is evident that female error rates are lower than male ones. This result may be due to the particular composition of speaker sets. However, it is important to notice that contrary to text-independent verification, for which gender comparison are often presented, such comparison is almost absent from the literature for the case of text-dependent speaker verification. This is probably due to the unbalanced gender representation in text constrained databases (see Section 2) and we hope the *RSR2015* database will provide the opportunity to researchers to analyse speech variability due to lexical content across genders.

6. Conclusions

We have presented the *RSR2015* database, the associate protocol, analysis and some initial results. This database, which consists of 71 hours of speech data is designed for text-dependent speaker verification using user-dependent pass-phrases. The number of speakers, sessions and pass-phrases recorded provides the opportunity to study the impact of lexical variability on text-dependent speaker verification but also to analyse the effect of lexical content on text-independent speaker verification engines [21]. In the future, two additional parts of the *RSR2015* database will be released; Part 2 is dedicated to user-loaded command control [19] as Part 3 focuses on randomly prompted digit sequences for speaker verification.

7. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[2] M. Hébert, *Text-dependent speaker recognition*. Springer-Verlag, Heidelberg, 2008.

[3] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector Based Speaker Recognition on Short Utterances," in *International Conference on Speech Communication and Technology*, 2011.

[4] O. Toledo-Ronen, H. Aronowitz, R. Hoory, J. Pelecanos, and D. Nahamoo, "Towards Goat Detection in Text-Dependent Speaker Verification," in *International Conference on Speech Communication and Technology*, 2011.

[5] A. Larcher, J.-F. Bonastre, and J. S. D. Mason, "Reinforced temporal structure information for embedded utterance-based speaker recognition," in *International Conference on Speech Communication and Technology*, 2008, pp. 371–374.

[6] M. F. BenZeghiba and H. Bourlard, "User-customized password speaker verification using multiple reference and background models," *Speech Communication*, vol. 48, no. 9, pp. 1200–1213, 2006.

[7] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, vol. 2, 1993, pp. 391–394.

[8] D. Delacretaz and J. Hennebert, "Text-prompted speaker verification experiments with phonemespecific MLPs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, vol. 2, 1998.

[9] J. Campbell and A. L. Higgins, "A. yoho speaker verification corpus ldc94s16 (available on lcd website: http://www.ldc.upenn.edu)," 1994.

[10] J. S. Mason, F. Deravi, C. C. Chibelushi, and S. Gandon, "Project: DAVID (Digital Audio Visual Integrated Database)," Department of Electrical and Electronic Engineering, University of Wales Swansea, Tech. Rep., 1996.

[11] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database (release 1.00)," *Lecture Notes in Computer Science*, vol. 1206/1997, pp. 403–409, 1997.

[12] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The Extended M2VTS Database," in *International Conference of Audio and Video-Based Person Authentication, AVBPA*, vol. 626, 1999.

[13] E. Bailly-Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree *et al.*, "The BANCA database and evaluation protocol," *Lecture Notes in Computer Science (LNCS)*, vol. 2688, pp. 625–638, 2003.

[14] S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. Jardins, J. Lunter, Y. Ni, and D. Petrovska-Delacretaz, "BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities," *Lecture Notes in Computer Science*, vol. 2688/2003, pp. 845–853, 2003.

[15] N. A. Fox, B. A. O'Mullane, and R. B. Reilly, "The Realistic Multi-modal VALID database and Visual Speaker Identification Comparison Experiments," in *International Conference of Audio and Video-Based Person Authentication, AVBPA*, New York (US), July 2005.

[16] B. Dumas, C. Pugin, J. Hennebert, D. Petrovska-Delacrétaz, A. Humm, F. Evéquoz, R. Ingold, and D. V. Rotz, "MyIdea–Multimodal biometrics database, description of acquisition protocols," *Biometrics on the Internet*, vol. 275, pp. 59–62, 2005.

[17] R. H. Woo, A. Park, and T. J. Hazen, "The MIT Mobile Device Speaker Verification Corpus: Data Collection and Preliminary Experiments," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2006.

[18] J. Fierrez, J. Ortega-Garcia, D. Torre Toledano, and J. Gonzalez-Rodriguez, "Biosec baseline corpus: A multimodal biometric database," *Pattern Recognition*, vol. 40, no. 4, pp. 1389–1392, 2007.

[19] 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 *International Conference on Speech Communication and Technology*, 2011.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus linguistic data consortium," *Philadelphia, PA*, vol. 1, 1993.

[21] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2012.