Combination of Rule-based and Data-driven Fusion Methohdologies for Different Speaker Verification Modes of Operation

Saeid Safavi, and Iosif Mporas
School of Engineering and Technology,
University of Hertfordshire, College Lane, AL10 9AB,
Hatfield, UK.
{s.safavi, i.mporas}@herts.ac.uk

Abstract— **In this paper we present three methodologies for the fusion of different speaker verification modes of operation. Specifically, we investigate a knowledge-based (rule-based) method, based on biometrics and security knowledge, a data-driven method, based on machine learning fusion models and a combination of them. The experimental results indicate that the hybrid fusion architecture, which is the combination of knowledge-based and data-driven based fusion, offers both robustness against spoofing and improvement in speaker verification performance.**

Keywords— **Speaker verification; spoofing; fusion; machine learning; knowledge-based; data-driven**

I. Introduction

Automatic speaker recognition is basically the intersection of two areas, first is speech technologies and the second is biometrics. The general area of speaker recognition is divided into three specific tasks. These are authentication, surveillance and forensic speaker recognition. Depending on the applications, speaker recognition also divided into three specific categories, namely the identification, verification, and segmentation [1]. The goal of the speaker identification task is to determine which speaker out of a group of known speakers produces the input voice sample. While in speaker verification, the aim is to determine whether a person is who he/she claims. It is considered as a true-or-false binary decision problem. Speaker segmentation techniques are used in multiple-speaker scenarios.

The focus of this paper is speaker verification. Speaker verification could be either text-dependent or text-independent. Text-dependent speaker verification constraints the spoken text, i.e. same text is spoken on both training and test phases, while in the text-independent case where there is no text restriction, it could be different in training and test phases. Furthermore, text-dependent systems could be divided in two sub-categories, namely the unique pass-phrase and the text-prompted. In the unique pass-phrase scenario speaker authentication is performed by pronouncing a pass-phrase, while in the text prompted scenario the prompts came from a closed set of predefined utterances. In the case of text-dependent, the verification does not depend only on the voice characteristic, but also on the linguistic content of the utterance. In general, the unique pass-phrase systems are vulnerable to replay attacks but at the same time very accurate in absence of replay attacks. In real-life application, text-independent systems are more commercially attractive than text-dependent systems as it is more difficult to mimic an unknown phrase than a known one. Speaker verification systems can be used as a biometric engines in variety of real life applications, for example they can be used to safe guard the users in social media, online banking, safe locations, and etc. Different real-life applications need different levels of vulnerability of the system against spoofing attacks, performance, and the user's convenience.

The most widely used approach in speaker recognition is to decompose speech signals to feature vectors and model the distribution of feature vectors using a Gaussian Mixture Model (GMM) [2]. In this approach a speaker-independent model, called Universal Background Model (UBM) is estimated using training data. Speaker-dependent GMMs are then created by retraining the UBM model using speaker specific training data and Maximum A Posteriori (MAP) adaptation [3]. In the GMM-UBM approach for speaker verification, the probability of utterance given speaker-specific model for claimed speaker, is divided by the probability of same utterance given UBM, and the result is compared with a threshold. The GMM-UBM approach has been superseded by methods that map an utterance into a vector space and then apply a static classifier such as a Support Vector Machine (SVM) [4]. For example, MAP adaptation of the UBM is performed using an utterance, and the mean vectors of the components of the resulting GMM are concatenated to form a GMM `supervector' [5]. Intersession variability, including inter-channel variability or channel factors, is a significant source of error for these technologies. Intersession variability compensation (ISVC) [6] and nuisance attribute projection (NAP) address this issue by isolating this variability in a linear subspace of the supervector space. In more recent approaches based on Joint Factor Analysis (JFA) [7-8], and i-vectors [9], the GMM supervector is mapped into a lower dimensional discriminative total variability space. These methods have been applied to a number of problems, including recognition of speaker [10-12], language, regional accent and dialect [13-15], age and gender [16-18].

The primary motivation for fusing scores from different modes of operation for speaker verification is that different modes of speaker verification not only carry different underlying information and offer different levels of accuracy, but also exhibit different levels of vulnerability to various types of spoofing. For example, a unique passphrase speaker verification system offer the highest level of accuracy compared with other modes of operation and also seem to pose a comparatively low risk for voice conversion attacks, but at the same time it is very vulnerable to replay attacks. A text-independent speaker verification system offers the lowest level of accuracy, but offers a low risk for replay attacks, which is the most accessible approach to spoofing.

In this paper we study the effectiveness of a number of fusion architectures, employed on the task of speaker verification, for fusing scores from different modes of operation.

The remaining of this paper is organized as follows: Section II presents the fusion architectures. Section III contains description of corpora, protocols for different modes of operation, and description of speaker verification framework. In Section IV the performance of single mode speaker verification is described. The score-level fusion methodologies are also evaluated. Finally, Section V summarizes the major results and conclusions of this research.

II. Fusion Architectures

In real-world voice based biometrics applications the user is asked by system to provide speech input in order to verify the claim of speaker about his/her identity. The mode of the input can vary, i.e. a fixed pass-phrase (mode A), a prompted text-dependent (mode B) or a prompted text-independent (mode C), based on the application's needs and the user's convenience. The use of different modes results in a trade-off between the speaker verification performance and the vulnerability to spoofing attacks, with mode B and C being the most robust against spoofing attacks and mode A presenting the lowest equal-error rate. The fusion of these three modes of operation of the speaker verification can results in a robust to spoofing attacks system while retaining low equal error rate. The fusion can be performed either using a knowledge-based (also known as rule-based) scheme or using a data-driven fusion model, as illustrated in Figure I.

In the case of knowledge-based fusion, rules based on the decision of each mode-specific speaker verification engine are set. These rules are based on previous knowledge in biometrics and security issues [21]. For example, acceptance of the claimed speaker identity by mode A and rejection from modes B and C is a probable replay attack (i.e. the impostor replays the fixed passphrase of mode A, but is unable to replay the unknown prompted text of mode B and C). Similarly, acceptance of the claimed speaker identity only from mode C could be a voice conversion based spoofing attack (i.e. a voice conversion system can cheat the high equal-error rate of mode C, but not of mode A or B). The knowledge-based rules for decision level fusion of different speaker verification modes of operation are tabulated in Table I.

TABLE I. Knowledge-based fusion of different modes of speaker verification operation. RP and VC stands for replay and voice conversion attacks.

	Mode A	**Mode B**	**Mode C**	**Decision**
1	Accept	Accept	Accept	*Accept*
2	Accept	Accept	Reject	*Accept*
3	Accept	Reject	Accept	*Accept*
4	Accept	Reject	Reject	*Reject RP*
5	Reject	Accept	Accept	*Accept*

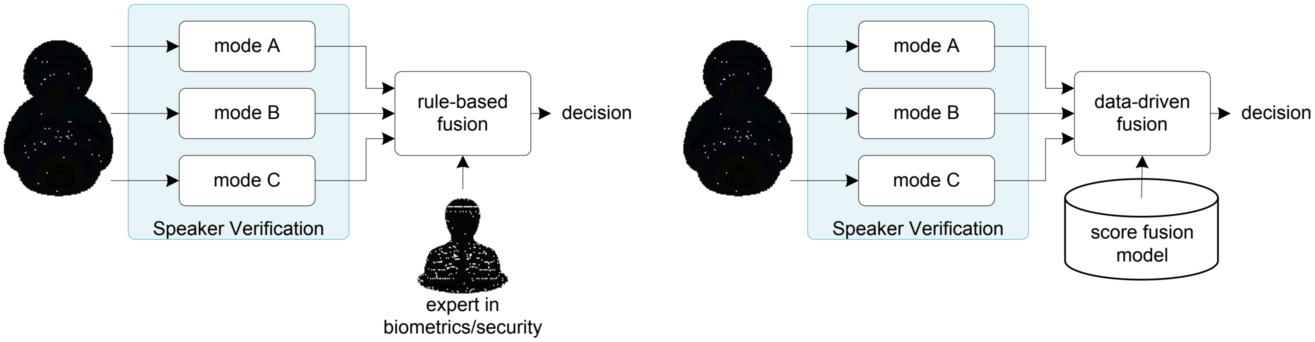

Figure I. Block diagrams of the knowledge-based (left) and data-driven (right) methodologies for fusion of scores from different speaker verification modes of operation.

6	Reject	Accept	Reject	*Reject*
7	Reject	Reject	Accept	*Reject VC*
8	Reject	Reject	Reject	*Reject*

As regards the data-driven fusion, this is performed using a machine learning algorithm, which will assign to each set of A, B and C mode-dependent scores a binary decision, i.e. acceptance or rejection of the claimed identity. The fusion algorithm uses as input the speaker verification scores as inputs and estimate a fusion score. The training of the fusion parameters and the acceptance threshold are estimated using a training dataset. During the test phase for each of the three modes of operation, a speaker verification score is produced and the three values are used as 3-dimentional vector input to the fusion model which decides whether the speaker is an authorized user or not.

Except the knowledge-based and the data-driven fusion methodologies, we propose a combination of these two (hybrid fusion). Specifically, in the knowledge-based methodology described above we adopt hard decisions, i.e. either acceptance or rejection, without taking into account the confidence level of each decision. Moreover, there are knowledge-based cases where it is not clear whether they correspond to acceptance or rejection. On the other hand, in the data-driven methodology there are cases where apriori knowledge of the hypothesis is not exploited and due to misclassifications of the fusion algorithm, error is introduced to the final decision. The combination of the two methodologies (hybrid fusion), which is illustrated in Figure II, uses a case dependent selector (indicated as case identifier) to forward each set of A, B, C mode scores to the knowledge-based or the data-driven fusion scheme.

As can be seen in Figure II, based on the scores of the three modes of operation a case identifier switches the operation to the knowledge-based or the data-driven approaches. Based on existing knowledge on biometrics and security applications [21] we selected the case identification setup that is tabulated in Table II.

TABLE II. SELECTION OF KNOWLEDGE-BASED OR DATA-DRIVEN FUSION METHODOLOGY. *FOR THIS SPECIAL CASE THE USAGE OF BOTH DATA-DRIVEN AND KNOWLEDGE-BASED ARE INVESTIGATED IN SECTION IV.

	Mode A	**Mode B**	**Mode C**	**Selection**
1	Accept	Accept	Accept	*knowledge-based*
2	Accept	Accept	Reject	*data-driven*
3	Accept	Reject	Accept	*data-driven*
4	Accept	Reject	Reject	*data-driven**

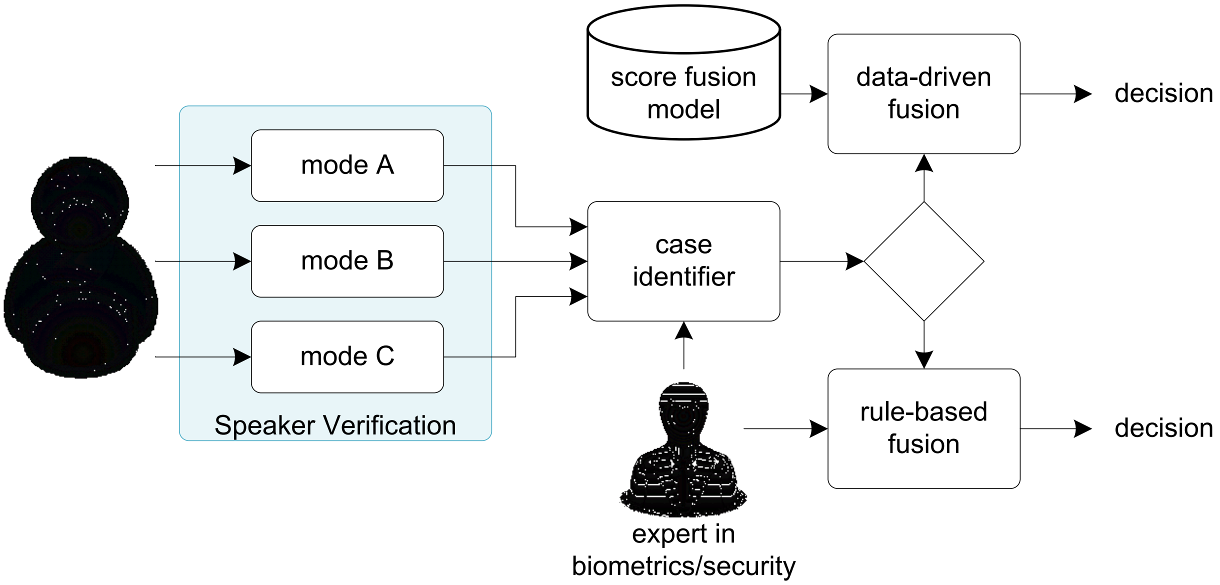

FIGURE II. COMBINATION OF KNOWLEDGE-BASED AND DATA-DRIVEN SPEAKER VERIFICATION MODES FUSION METHODOLOGIES.

5	Reject	Accept	Accept	*data-driven*
6	Reject	Accept	Reject	*data-driven*
7	Reject	Reject	Accept	*knowledge-based*
8	Reject	Reject	Reject	*knowledge-based*

In this hybrid setup there are three cases where the knowledge-based methodology is activated and five cases where data-driven fusion is activated, except in case 4 which both data-driven and knowledge-based are compared. The combination of the two fusion methodologies is performed on the basis of the decisions per speaker verification mode. In the combination methodology the case selection is knowledge-based. The use of the case selector will result to speaker verification decisions which are data-driven for the cases where the knowledge-based decision is not obvious.

III. Experimental Setup

A. Data Description

The RSR2015 [19] speech corpus was used for our evaluation purposes. This corpus involves 300 speakers (157 males, 143 female). For each speaker, there were 3 enrolment sessions of 73 utterances each and 6 verification sessions of 73 utterances each. In total there were 657 utterances in 9 sessions for each speaker. The sample rate for the speech files was 16 kHz, and the sample coding was 16 bits linear. Speech files were stored in raw format.

TIMIT [20] was used for training a background model. TIMIT contained broadband recordings of 630 speakers, each reading ten phonetically rich sentences. The sample rate for the speech files was 16 kHz, and the sample coding was 16 bits linear. Speech files were stored in waveform format.

B. Setup for Fusion

This research provided a tentative set of results using the recent RSR2015 corpus intended for benchmarking different modes of automatic speaker verification and their fusion. In particular, training and trial lists (definition of speaker pairs) were designed to simulate system evaluation of three different configurations associated with speech content; (A) unique pass-phrase (B) text-prompted phrases and (C) text-independent engines. The first protocol included a unique phrase shared by all the users; the second protocol referred to a scenario whereby a system prompted a randomly selected phrase out of a close subset of pass-phrases. The last scenario was essentially a text-independent scenario with arbitrary enrolment and test phrases. The single mode biometric engine was evaluated for three different circumstances, (A), (B), and (C).

To assess the performance in all three protocols, different enrolment and trial lists were designed. The experiments were conducted on a subset of the male section of a recently released RSR2015 dataset [19]. It consisted of recordings of the same 30 unique pass-phrase sentences across all the speakers captured over 9 sessions.

For all three protocols 43 identical male speakers were used. In the protocol (A), recording of the same sentence from session 01, 04 and 07 were used for creating target models whereas utterances from the other five sessions were used in testing. In this protocol, all the speakers spoke the same utterance. We had a total of 30 unique phrases. Therefore, 30 sub-conditions for this protocol were created.

In the protocol (B), speakers were enrolled with 15 different pass-phrases. For each speaker, sentences 01 to 05 were taken from session 04, sentences 06 to 10 were taken from session 01 and rest of the sentences, 11 to 15 were taken from session 07. All of the 15 sentences used in the enrolment were prompted during testing.

Finally, for protocol (C), the enrolment was done in a similar way as the previous protocol. But the test data was exclusively different from the enrolment data. Here, the rest of the 15 sentences (from 16 to 30) were used in testing.

C. Single Mode Speaker Verification Framework

Feature extraction was performed as follows. Periods of silence were discarded using an energy-based Speech Activity Detector (SAD). The speech was then segmented into 20-ms frames (10-ms overlap) and a Hamming window was applied. The short-time magnitude spectrum, obtained by applying an FFT, was passed to a bank of 24 Mel-spaced triangular bandpass filters, spanning the frequency region from 0 Hz to 8000 Hz. The outputs of all 24filters were transformed into 19 MFCCs plus 19 deltas and 19delta-deltas. In order to reduce the effect of handset mismatch and to make the feature more robust, RASTA [15] and CMVN processing were applied on the MFCC features.

Our speaker verification system was based on the state-of-the-art GMM-UBM [13] method. In the GMM-UBM approach, a UBM (the same was used for all three protocols) with 128 mixture components were built using all the utterances from the 630 speakers from TIMIT. The speaker models were obtained by MAP adaptation (adapting the means only) of the UBM, using the speaker-specific enrollment data, for each of the three protocols described in Section III.B.

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The score-level fusion architectures described in Section II, was evaluated according to the experimental protocols presented in Section III. In this Section we present the experimental results.

A. *Single Mode of Opertaion for Speaker Verification*

The performance of each single mode of operation in terms of sensitivity and specificity are tabulated in Table III. In total there were 165,506 instances, of which 3836 were target and the rest were imposters.

As can be seen in Table III, the best performing single mode operation was the unique-passphrase, while the worst performing was the text-independent. This is in agreement with the literature [19] and is due to the fact that the acoustic model's parameters were more precisely fitted to the acoustic characteristics of the target speaker for known or fixed utterances, following the protocol of the unique pass-phrase mode of operation.

TABLE III. SINGLE MODE SPEAKER VERIFICATION PERFORMANCE IN TERMS OF SENSITIVITY AND SPECIFICITY.

	Sensitivity	***Specificity***
Mode A	**96.35 %**	**96.91 %**
Mode B	84.65 %	97.46 %
Mode C	84.70 %	91.83 %

B. *Fusion of Different Modes of Operation*

In order to exploit the high accuracy of operation mode (A), while retaining the robustness of operation modes (B) and (C) in spoofing attacks, we further examined different fusion methodologies presented in Section II.

TABLE IV. SINGLE MODE AND FUSION OF SPEAKER VERIFICATION PERFORMANCE IN TERMS OF SENSITIVITY AND SPECIFICITY. *FOR THIS HYBRID FUSION ARCHITECTURE THE RULES ARE BASED ON TABLE II, EXCEPT CASE 4 WHICH INSTEAD OF DATA-DRIVEN THE KNOWLEDGE-BASED RULE FROM TABLE I WAS USED.

	Sensitivity	***Specificity***	***Vulnerability to spoofing***
Mode A	96.35 %	96.91 %	High
Mode B	84.65 %	97.46 %	Medium
Mode C	84.70 %	91.83 %	Medium
Knowledge-Based	92.57%	99.23%	Very-Low
Data-driven	98.22%	98.12%	Low
Fusion 1*	92.57%	**99.93%**	Very-Low
Fusion 2	**98.51%**	99.74%	Very-Low

The performance of each operation mode and fusion of them, in terms of specificity and sensitivity are tabulated in Table VI. The last column of Table VI contains information about the level of vulnerability to spoofing attacks for each mode of operation and fusion methodologies. For the last column of this table we have assumed a usage of utterance verification engine before passing the signal to the speaker verification engines. In all proposed fusion methodologies, the specificity improved comparing with the best performing single mode of speaker verification. The best specificity achieved when hybrid Fusion 1 was used. This fusion is based on the hybrid usage of data-driven and rule based fusion. The rules for hybrid Fusion 1 and 2 were the same, as tabulated in Table II, except for case 4. For hybrid Fusion 1 the decision is knowledge based (from Table I) but for hybrid Fusion 2 it was based on data-driven decision.

Different applications need different level of accuracy, security and user convenience. Depends on the application needs, one of the fusion architecture could be used. The best performing architecture for both specificity and sensitivity is the hybrid Fusion 2, with 99.74 % and 98.51 % for specificity and sensitivity, respectively.

V. Conclusion

In speaker verification there is a trade-off between robustness against spoofing attacks and verification performance. The fusion of text-dependent and text-independent modes of operation results in improvement of the verification performance, while the use of text-independent mode reduces vulnerability to spoofing attacks. The presented methodologies, i.e. the knowledge-based, the data-driven and their combination (hybrid fusion) show that the use of apriori knowledge from biometrics security together with machine learning modeling for the cases where the decision about the speaker claimed identity is not obvious can improve the overall performance. We deem the proposed hybrid architecture can result in new real-world voice-based authentication interface applications.

Acknowledgment

This work was supported in part by the OCTAVE H2020 Project entitled “Objective Control for TAlker VErification” funded by the EC with Grand Agreement number 647850. The authors would like to thank Prof Aladdin Ariyaeeinia, Dr Nicholas Evans, and Dr Tomi Kinnunen for their support in this work.

References

[1] Sinith, M.S., Salim, A., Gowri Sankar, K., Sandeep Narayanan, K.V. and Soman, V., 2010, November. A novel method for Text-Independent speaker identification using MFCC and GMM. In Audio Language and Image Processing (ICALIP), 2010 International Conference on (pp. 292-296). IEEE.

[2] Reynolds, D., Rose, R., 1995. Robust text-independent speaker identificationusing gaussian mixture speaker models. IEEE 3, 72-83.

[3] Reynolds, D. A., Quatieri, T. F., Dunn, R. B., 2000. Speaker verificationusing adapted gaussian mixture models. Digital Signal Processing 10 (1-3), 19-41.

[4] Campbell, W., Sturim, D., Reynolds, D., Solomono_, A., 2006a. Svm basedspeaker veri_cation using a gmm supervector kernel and nap variabilitycompensation. In: Proc. IEEE-ICASSP, Toulouse, France. Vol. 1. pp. I-I.

[5] Campbell, W. M., Sturim, D. E., Reynolds, D. A., Solomonoff, A., 2006b. SVM based speaker verification using a GMM supervector kernel and NAPvariability compensation. Proc. IEEE-ICASSP, Toulouse, France.

[6] Vair, C., Colibro, D., Castaldo, F., Dalmasso, E., Laface, P., 2006. Channelfactors compensation in model and feature domain for speaker recognition.In: Proc. Odyssey'06, The Speaker and Language Recognition WorkshopSan Juan, Puerto Rico. pp. 1-6.

[7] Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007. Joint factoranalysis versus eigenchannels in speaker recognition. Audio, Speech, andLanguage Processing, IEEE Transactions on 15 (4), 1435-1447.

[8] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011a. Frontendfactor analysis for speaker veri_cation. Audio, Speech, and LanguageProcessing, IEEE Transactions on 19 (4), 788-798.

[9] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011b. Frontendfactor analysis for speaker verification. IEEE Transactions on Audio,Speech, and Language Processing 19 (4), 788-798.

[10] Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition:From features to supervectors. Speech Communication 52, 12-40.

[11] Safavi, S., Najafian, M., Hanani, A., Russell, M.J., Jancovic, P. and Carey, M.J., 2012. Speaker Recognition for Children'sSpeech. In INTERSPEECH (pp. 1836-1839).

[12] Safavi, S., Speaker Characterization using Adult and Children’s Speech, Ph.D. Dissertation, University of Birmingham, 2015.

[13] Biadsy, F., Hirschberg, J., Ellis, D. P. W., 2011. Dialect and accent recognitionusing phonetic-segmentation supervectors. In: INTERSPEECH.ISCA, pp. 745-748.

[14] Najafian, M., Safavi, S., Weber, P. and Russell, M., Identification of British English Regional Accents Using Fusion of i-vector and Multi-accent Phonotactic Systems. In ODYSSEY 2016.

[15] Safavi, S., Hanani, A., Russell, M., Jancovic, P., Carey, M., 2012. Contrasting the Effects of Different Frequesncy Bands on Speaker and Accent Identification, in IEEE Signal Processing Letters, vol .19, no. 12, pp. 829-832, Dec. 2012.

[16] Bahari, M. H., McLaren, M., van Hamme, H., van Leeuwen, D. A., 2012. Age estimation from telephone speech using i-vectors. In: Proc. Interspeech,Portland, Oregon, USA. ISCA.

[17] Safavi, S., Jancovic, P., Russell, M. J., Carey, M. J., 2013. Identification of gender from children's speech by computers and humans. In: Proc. Interspeech,Lyon, France. ISCA, pp.2440-2444.

[18] Safavi, S., Russell, M.J. and Jancovic, P., 2014. Identification of age-group from children's speech by computers and humans. In INTERSPEECH (pp. 243-247).

[19] Anthony Larcher, Kong Aik Lee, Bin Ma, Haizhou Li, “Text-dependent speaker verification: Classifiers, databases and RSR2015, Speech Communication, Volume 60, May 2014, Pages 56-77,ISSN0167-6393.

[20] Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G. and Pallett, D.S., 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon Technical Report N, 93.

[21] Evans, N., Kinnunen, T., Yamagishi, J., Wu, Z., Alegre, F. and De Leon, P., 2014. Speaker recognition anti-spoofing. In Handbook of Biometric Anti-Spoofing (pp. 125-146). Springer London.