IDIAP COMMUNICATION REPORT

ICB 2013 - COMPETITION ON SPEAKER RECOGNITION IN MOBILE ENVIRONMENT USING THE MOBIO DATABASE: THE EVALUATION PLAN

Elie Khoury Sébastien Marcel Manuel Günther

Idiap-Com-04-2012

DECEMBER 2012

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

ICB 2013 - Competition on speaker recognition in mobile environment using the MOBIO database: The Evaluation Plan

Elie Khoury, Manuel Günther, Sébastien Marcel

I. Introduction

The Biometric group at the Idiap Research Institute is organizing the second competition on *text independent* speaker recognition for the 2013 International Conference on Biometrics[1] (ICB-2013) to be held in Madrid, Spain on June 4-7, 2013. This competition is the second in an ongoing series of speaker and face recognition evaluations conducted on mobile environment using MOBIO database.

As for NIST-SRE[2] evaluations, the basic task is speaker detection, *i. e.*, to determine whether a specified target speaker is speaking during a given segment of speech.

Participation in the evaluation is open for all academic and industrial organizations that find the task interesting for their research goals. Participants must comply with the evaluation rules presented in this plan. In addition, participants will be invited to co-author a paper that describes the different submitted systems. This paper will be presented at ICB-2013 in Madrid, Spain on June 4-7, 2013. To register, please follow the instructions on the official webpage of the competition[3] in order to fill out the registration form before January 14, 2013.

Goal: This evaluation is intended to help researchers measuring their progress in the last couple of years, particularly in mobile environment. This point makes the major difference with other competitions.

II. Task Definition

As briefly described in the competition, the task is speaker detection. For each *target* speaker (also known as *client*), several samples of speech are provided to the speaker recognition system. These samples are used by the system to create a model of the target speaker. Then a test audio sample is provided to the system. Performance is judged according to how accurately the test segment is classified as containing (or not) speech from the target speaker. This is expressed by a score such as the log-likelihood ratio (LLR). The higher the score is, the more the test segment is similar to the target speaker. Furthermore, since the detection threshold may be determined from the scores on the Development set (see section III-B), system output will not include a detection decision.

[1] http://atvs.ii.uam.es/icb2013/
[2] http://www.nist.gov/itl/iad/mig/sre12.cfm
[3] http://www.beat-eu.org/evaluations/icb-2013-speaker-recognition-mobio

III. The MOBIO Database

A. Description

The competition will be carried out on the MOBIO database. MOBIO is a challenging bimodal (face/speaker) database recorded from 152 people. The database has a female-male ratio of nearly 1:2 (100 males and 52 females) and was collected from August 2008 until July 2010 in six different sites from five different countries. This led to a diverse bi-modal database with both native and non-native English speakers.

In total 12 sessions were captured for each client: 6 sessions for Phase I and 6 sessions for Phase II. The Phase I data consists of 21 questions with the question types ranging from: Short Response Questions, Short Response Free Speech, Set Speech, and Free Speech. The Phase II data consists of 11 questions with the question types ranging from: Short Response Questions, Set Speech, and Free Speech.

The database was recorded using two types of mobile devices: mobile phones and laptop computers. The mobile phones used to capture the database were NOKIA N93i mobiles while the laptop computers were standard 2008 MacBook computers. The laptop were only used to capture part of the first session. The MOBIO database is a challenging database since the data is acquired on Mobile devices possibly with real noise, and the duration of the speech segments is relatively very short (around 10s or less).

More technical details about the MOBIO database can be found in [1] and on its official webpage[4]. This webpage details the procedure to follow in order to download the database.

B. Database partitioning

In order to have an unbiased evaluation, the clients of the database are split up into three different sets:

1) Background training set: The data of this set are used to learn the background parameters of the algorithm (UBM, subspaces, *etc.*). This also can be used for score normalization (cohort, *etc.*). At the beginning of the competition (January 14, 2013), participants will be provided with the corresponding list. This list consists of n lines where n is the number of targets in the set. In each line, the first element is the target ID, while the remaining elements are file names belonging to this target. If a gender-specific training should be performed,

[4] https://www.idiap.ch/dataset/mobio

participants can restrict the targets used for training to the desired gender. This can be easily achieved since the target IDs start with "m" for male targets, and for "f" for female targets. Note that the file names are stored relative to the MOBIO database directory and without file extension. It is worth noting that participants can use extra data in their background training, however they should explicitly precise it in their system description.

2) Development set (DEV): The data of this set are used to tune meta-parameters of the algorithm (*e. g.* number of Gaussians, dimension of the subspaces, *etc.*). For each client of the development set, the audio files are divided into files to be used for target model training (also known as model enrollment), and files that serve as test segments (also known as probes). At the beginning of the competition, participants will be provided with two lists per gender (there are no cross-genders trials), one for the enrollment, and the other for the test. The enrollment lists have a structure similar to the training list. The probe lists do not contain the target IDs anymore. Hence, there is only one probe file per line. In order to help computing the performance and, thus, tune meta-parameters in an easy way, the target name is explicitly given as a part of the audio file name.

3) Evaluation set (EVAL): The data of this set are used for computing the final evaluation performance. It has a structure similar to the development set. The only difference is that the file names are anonymized. This set will be delivered to participants at the Evaluation time (March 1, 2013).

Table I statistically details each of the sets described above, in terms of the number of files, the number of targets, and the number of trials.

IV. Evaluation Performance

To evaluate the speaker recognition performance,the metric that will be used is based on the *false acceptance rate* (FAR) and the *false rejection rate* (FRR). The definition of these rates is dependent on a certain *threshold* θ:

$$\begin{aligned} \mathrm{FAR}(\theta) &= \frac{|\{h_{non} \mid h_{non} \geq \theta\}|}{|\{h_{non}\}|} \\ \mathrm{FRR}(\theta) &= \frac{|\{h_{tar} \mid h_{tar} < \theta\}|}{|\{h_{tar}\}|} \end{aligned} \tag{1}$$

where h_{tar} is a target (client) score, while h_{non} is a non-target (imposter) score.

Particularly, we use the DEV set to define a score threshold θ_{dev}, based on the *Equal Error Rate* (ERR) of the development set. The final evaluation performance is then computed as the *Half Total Error Rate* (HTER):

$$\begin{aligned} \theta_{dev} &= \arg\min_{\theta} \frac{\mathrm{FAR}_{dev}(\theta) + \mathrm{FRR}_{dev}(\theta)}{2} \\ \mathrm{HTER}_{eval}(\theta_{dev}) &= \frac{\mathrm{FAR}_{eval}(\theta_{dev}) + \mathrm{FRR}_{eval}(\theta_{dev})}{2} \end{aligned} \tag{2}$$

The particularity of the evaluation is the restriction to a unique protocol for all participants. This gives a more appropriate and objective comparison between different systems.

V. Submission

Each participant can submit up to 5 systems. They should explicitly precise which of the system is the primary, and which are secondary systems. Their primary system will be reported in the final common paper. Note that both scores for DEV and EVAL should be submitted.

A. Scores submission format

The file lists described in section III-B define the verification protocol, where each probe file is tested against each target of the same gender. The result of the speaker recognition experiment is a text file (standard ASCII format). Each file record must document its trial output with 3 space separated fields:

1) The target ID.
2) The test segment file name.
3) The score. The score is required to represent how the test segment is similar to the target ID.

The DEV scores are used by organizers to retrieve the optimal threshold. This threshold will be used to compute the final HTER on the EVAL scores. Contrarily to NIST-SRE 2012, the computation of the compound Log-Likelihood Ratio is not allowed, *i. e.,* the *a priori* probabilities about known and and unknown targets and non-targets are prohibited. However, any kind of scores normalization and calibration techniques are allowed, but should be mentioned in the system description.

B. System Description

Participants should provide technical details about the different stages of their system. To facilitate this task, organizers are willing to provide a form at the evaluation time. This form will include details about the techniques (if applicable) used for pre-processing, feature extraction, voice activity detection, feature normalization, classifier and modeling, channel normalization, score normalization and calibration, score fusion, *etc*.

Further, since the evaluation is done on data recorded on mobile devices, the processing time and the memory footprint are given special attention. For this reason, participants should provide more details about these at all different stages of their system.

VI. Schedule

As briefly described above, at the beginning of the competition, participants will be provided with the training list of background models, and the enrollment and test lists of the Development set. Additionally, the source code of a baseline speaker recognition system will be published to give an example on how to use these lists. At the evaluation time (March 1, 2013), the equivalent lists of the Evaluation set will be made available to the participants.

At the end of the competition (March 15, 2013), the participants should provide the resulting score files of the Development and the Evaluation set, and give a detailed description of their system(s). These results will be published in a conference paper at the ICB-2013.

TABLE I
PARTITIONING OF THE MOBIO DATABASE

	Background		Development					Evaluation				
			Enrollment		Test			Enrollment		Test		
	Spks	*Files*	*Targets*	*Files*	*Spks*	*Files*	*Trials*	*Targets*	*Files*	*Spks*	*Files*	*Trials*
MALE	37	7104	**24**	120	24	2520	**60480**	38	190	**38**	3990	**151620**
FEMALE	13	2496	**18**	90	18	1890	**34020**	20	100	**20**	2100	**42000**
TOTAL	50	9600	42	210	42	4410	94500	58	290	58	6090	193620

Registration Due: January 14, 2013
Availability of Training and Development sets: January 14, 2013
Availability of Evaluation set: March 1, 2013
Submission of the Results System description: March 15, 2013
Publication of the Results at ICB-2013: April 8, 2013

REFERENCES

[1] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J.-F. Bonastre, P. Tresadern, and T. Cootes. Bi-modal person recognition on a mobile phone: using mobile phone data. In *IEEE ICME Workshop on Hot Topics in Mobile Multimedia*, 2012.