EVALITA 2009
Speaker Identity Verification - Forensic Track
Task Guidelines

Luciano Romito, Laboratorio di Fonetica – Università della Calabria

1. Introduction and description

The following are the guidelines for the Speaker Identity Verification task (Forensic track) of the EVALITA 2009 evaluation campaign.

The Forensic Speaker Identity Verification is characterized by two main points: the first one is related to the individuals involved in the task consisting of suspected individuals that usually have the aim of not being recognized (and therefore not willing to collaborate); the second one is related to a specific balance of the "decision costs" i.e. between wrong identification scores and failed identification scores.

The speech corpus[1] contains recordings of Italian male speakers. The recording channels are of three types: high fidelity, environmental and telephonic. The recordings have been captured under five different conditions that determine their quality: 1) silent room condition (this material will be used as “Training” data set “TR”); 2) wiretapping in and out of car (made possible with the help of police officers by means of a tapping service); 3) phone-calls in a car; 4) in a street, and 5) in a crowded place (part of these last four types of material will be used for the “Closed-set Test” data set “CST”). The whole corpus contains the same material recorded in the above listed conditions. For each recording condition, the recorded material contains: a) reading of 10 phonetically balanced sentences; b) reading of 10 repetitions of 3 phonetically balanced sentences. For the environmental recording condition, spontaneous speech material, both inside and outside the car, is also available.

In the same speech corpus other recordings are present: 1) a recording session in a noisy place including the four speakers present in the speech corpus, together with a large number of other anonymous voices (this file will be used for the “Open-set Test” data set “OST”); 2) four high-fidelity recordings of the previous recording for the four speakers present in the speech corpus.

All sound files are in Italian language and have been conformed in terms of quality to the worst recording condition identified in the environmental (car) wire-tapping condition: 8 kHz – 16 bit – mono in *.wav PCM format. Only for the “Training” data set “TR” the sound files are distributed in a double format: 44.1kHz – 16 bit – mono *.wav PCM and 8 kHz – 16 bit – mono *.wav PCM.

[1] Cfr. http://www.linguistica.unical.it/labfon/home_corpus_primula.html

2. File name description

All files are labelled following the form [data type]_[xx/xxx]_[a]_[b]_[c(d)].wav incorporating their qualitative characteristics, where:

[data type] = the letters identifying the data set:

TR for the Training Set
CST for the Closed-Set Test

[xx/xxx] = two letters [xx] or three digits [xxx] identifying a known or unknown speaker

S1 or S2 in case of TR data set identifying the two known *suspected* speakers
034, 098, n…, a three randomly generated digits sequence in the case of CST data set identifying unknown speakers to identify

[a] = type of recording identifying the recording channel and its acoustic quality:

C (silent room recording)
A (telephone call in crowded place)
S (telephone call in street)
I (wiretapping in car)
X (recording of a telephone call in a car)

[b] = identifies the phonation manner of the speakers:

B (low voice)
N (normal voice)
A (loud voice)

[c] = type of speech material produced by the recorded speaker:

LR (reading with repetition)
LS (one reading)
PS (spontaneous speech)

(d) = identifies the repeated sentence only if [c] = LR:

1 (sentence 1)
2 (sentence 2)
3 (sentence 3)

3. Training data

The “Training” data set (“TR” preceding the file name) reproduces the sample voice of two known *suspected* subjects “S1” and “S2” contained in the speech corpus (e.g.: TR_S1_C _N_LS.wav; TR_S2_C _N_LR3.wav). The voice samples are clean and recorded in a silent room (C) in high fidelity. For each of the two known *suspected* subjects the recorded material contains: a) 1 file containing 10 read phonetically balanced sentences; b) 3 files containing each 10 read repetitions of 1 phonetically balanced sentence. For each of the two known suspected subjects, 4 sample files are provided to the participants for the training task “TR”:

TR_S1_C_N_LR1.wav	(> 40 sec.)
TR_S1_C_N_LR2.wav	(> 50 sec.)
TR_S1_C_N_LR3.wav	(> 40 sec.)
TR_S1_C_N_LS.wav	(> 40 sec.)
TR_S2_C_N_LR1.wav	(> 40 sec.)
TR_S2_C_N_LR2.wav	(> 60 sec.)
TR_S2_C_N_LR3.wav	(> 50 sec.)
TR_S2_C_N_LS.wav	(> 50 sec.)

As above reported sound files are in Italian language and are distributed in a double format: 44.1kHz – 16 bit – mono *.wav PCM and 8kHz – 16 bit – mono *.wav PCM.

4. Test data

Two data sets are provided for the Test: “CST” data set for the “Closed-set Test” and “OST” data set for the “Open-set Test”. The two Test data sets are “blind”, i.e. no answer key will be distributed to participants before the submission of results.

The “CST” data set for the “Closed-set Test” is a collection of wiretapping recordings in different environments and in different channels of anonymous speakers. The voices are isolated in 16 different files of different length. The material is composed by read and spontaneous speech and it is distributed to participants in 8 kHz – 16 bit – mono in *.wav PCM format. Following the rules above reported the files are labelled e.g.: CST_008_I_N_PS.wav, CST_035_I_N_LS.wav.

The “OST” data set for the “Open-set Test” consists of a single file of a recording session in a noisy place including the two *suspected* speakers together with a large number of other anonymous voices. Intensity in the file is changing and superimposed voices are possible. If needed, according to the method used, the file should be segmented in single files using the syntax OST_[xxxx] for the file name generation, where:

[xxxx] = a four digits string indicating the voice segment selected from the whole file,

and reporting on a separate *.txt file its position (expressed in hh.mm.ss,000 for starting and ending time) in the whole file together with the name assigned to the selection as follows, with values separated by tabs:

```
OST_0001 tab 00.00.05,345 tab 00.00.56,002
OST_0002 tab 00.01.03,566 tab 00.01.09,241
n…
```

The *.txt file has to be named as [name]_OST_seg.txt, where [name] indicates the name of the participant or the organization. Each *.wav file exported has to report the name given in the segmentation (e.g. OST_0001.wav, OST_0002.wav etc.)

Participants will therefore receive three folders containing the sound files of the three data sets (“TR”, “CST”, “OST”).

Details on the results submission follow in the next section.

5. Submission of results

Participants are allowed to use any of the methods/models nowadays available (automatic, semiautomatic or manual ones), but also new methods or new models not yet tested or verified.

Together with the submission of the results, participants are asked to provide a description of the system, method and statistics used. In addition, if manual or semi-automatic methods are used, participants are asked to produce a point to point protocol (PPP) concerning identification and number of parameters/features, as well as algorithms used and so on. This information has to be provided in an *.rtf file named as [name]_method.rtf.

If a reference population is used for the comparison, it is mandatory for all participants to give all available information about it (e.g. number of speakers, sex, age, features present etc.) in the point to point protocol [name]_method.rtf file.

In the case of semiautomatic or manual methods, participants are also asked to present a full report containing a complete list of the parameters’ values used in a *.txt file named [name]_par_val.txt with strings separated by tab spaces.

Participants should process all files giving a score for identification (same speaker) or missed identification (different speaker) for each trial, with the corresponding thresholds according to the statistic method adopted.

Each trial (reported in a separate *.txt file) should contain a comparison between anonymous and *suspected* voice samples (one of the TR data set), result of the identification (yes/no), score (likelihood ratio, percentage or other according to the statistical method used), and False Rejection (FR) and False Acceptance (FA) if participants use a reference population. The *.txt file, named as [name]_trials.txt, should look as follows (with strings separated by tabs):

```
CST_015_I_N_LS tab TR_S1_C _N_LS tab yes tab 3,98E+05 tab FR Rate (%) tab FA Rate (%)
OST_0002 tab TR_S2_C _N_LR3 tab no tab 1,95E+03 tab FR Rate (%) tab FA Rate (%)
n...
```

or,

```
CST_010_I_N_LS tab TR_S1_C _N_LS tab yes tab 95%
OST_0004 tab TR_S2_C _N_LR1 tab no tab 87%
n...
```

In conclusion, participants should submit the following files in a single folder identified by the name [name] of the participant or the organization:

1. an *.rtf file containing description of the system, method and statistics used, PPP and any other notes. The file is named as [name]_method.rtf;
2. a *.txt file containing the segmentation of the OST.wav file named as [name]_OST_seg.txt according to the rules above described;
3. a *.txt file containing a list of all the trials with the result of the test for each trial named [name]_trials.txt according to the rules above described;
4. a *.txt file containing a complete list of the parameters' values used in the case of semiautomatic or manual methods, named as [name]_par_val.txt with strings in the file separated by tabs.