EVALITA 2009

Speaker Identity Verification - Application Track Task Guidelines

Guido Aversano, Parrot SA

1. Introduction

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

Several data sets will be provided by the organizers:

- *Universal Background Model* (UBM) data. Containing speech data from speakers not included in the other sets, it is generally used to train a background speaker-independent model (also known as "world model") for the verification system. Participants are free to use any other data in their possession for UBM preparation.
- *Training* (or *Enrollment*) data. This data, reproducing realistic enrollment of 100 clients, is used to build up genuine user models.
- *Development* data, that is a collection of client access trials, used to tune the verification system by fixing decision thresholds and other system parameters.
- *Test* (or *Evaluation)* data. This data set contains client and impostor access trials. Final evaluation scores will be calculated on this set.

Participants are required to provide a brief description of their system(s) and a full notebook paper describing their experiments, in particular the techniques, the resources used and presenting an analysis of the results.
Participants should also provide full details on additional data (if any) used for UBM training and describe their development/tuning protocol.

2. Corpus description

All distributed data is recorded from landline (PSTN) or mobile (GSM) telephone channels. Recordings are in Italian language, including speakers uniformly selected in all regions.

Data format

Participants will receive raw sound files (samples only, without any header). Encoding will be A-Law PCM, 8 kHz mono.

Files will have the "`.alw`" extension, and their name will be a string of 5 randomly chosen characters (e.g. `nbltu.alw`) .

Universal Background Model (UBM) data

Speech data recorded by 60 speakers (30 female + 30 male), during 20 sessions (10 PSTN calls + 10 GSM calls).
The total duration of UBM speech data is 1200 minutes (~1 minute per call).

Participants are free to use any other data in their possession for UBM preparation.
In this case, details on the data should be given in submission notes.

Sound files for UBM training are stored in the "`ubm`" sub-directory of the distribution support. Organizers will distribute 4 lists of files for UBM training (plain ASCII, containing one filename per line). There will be one list per gender, and per transmission channel (the name of these lists being of the form: `female_pstn.ubm`, `male_gsm.ubm`, etc.).

Training data

Participants should train 100 speaker models (50 female + 50 male), representing all genuine clients of the system.

6 training conditions ("TC") will be considered:

- ***TC1***. "PSTN SHORT" *(1 PSTN call; ~1 minute per client)*.
- ***TC2***. "GSM SHORT" *(1 GSM call; ~1 minute per client)*.
- ***TC3***. "PSTN LONG" *(3 PSTN calls; ~3 minutes per client)*.
- ***TC4***. "GSM LONG" *(3 GSM calls; ~3 minutes per client)*.
- ***TC5***. "MIXED SHORT" *(1 PSTN call + 1 GSM call; ~2 minutes per client)*.
- ***TC6***. "MIXED LONG" *(3 PSTN calls + 3 GSM calls; ~6 minutes per client)*.

Sound files for training are stored in the "`train`" sub-directory of the distribution support.

Organizers will provide 6 training files per gender, corresponding to the above 6 training conditions.
The format of these files follows the NIST-SRE [1] specifications, each line of the file containing: 1) a model identifier (two capital letters); 2) a comma separated list of the speech files (from the "`train`" sub-directory) that are used to train the model.

For example, the `tc3_male.trn` file (training file for condition *TC3*, male gender) will contain 50 lines of the following type:

```
QD bftul.alw,anbia.alw,lffat.alw
```

Test data

Two test conditions will be considered:

- ***TS1***. "SHORT"
  (1 sequence of digits; ~10 seconds).
- ***TS2***. "LONG"
  (1 sequence of digits, 4 short sentences, 2 isolated words; ~ 30 seconds).

Evaluation will be done on more than 2000 test trials, including both genders and both transmission channels.

Participants will receive two lists of test trials (`ts1.ndx` and `ts2.ndx`), with corresponding speech data (found in the "`test`" sub-directory of the distribution support). The format of trial lists is again similar to the NIST-SRE standard. Each line represents an access trial, and contains four fields. The first field is the model identifier (*claimed identity*). The second field is the gender of the model. The third field is the name of the speech file under test, without the "`.alw`" extension. The forth field is the transmission channel ("`P`" for PSTN, "`G`" for GSM, "`X`" for non-specified). Thus, a line of the trial files looks like the following:

```
PA f nttai X
```

Participants should process all the trials, giving a decision (*"client"* or *"impostor"*) for each trial, with the corresponding likelihood score. Details on the submission file, with other required fields, are given in section 4.

Unsupervised adaptation of client models is possible, but in this case the tests *must be executed in the order given in the trial lists, and independently for the two lists*. Also, for comparison purposes, participants that submit results running unsupervised adaptation, *must submit an additional set of results obtained without adaptation*.

The test data is "blind", i.e. no answer keys will be distributed to participants before the submission of results.
Note that for these test trials the transmission channel is always "`X`" (non-specified). The actual channel will be indicated in answer keys, after result submission. Systems will be evaluated on the full set of trials, and on the subset of trials matching the training condition.

Development data

For a small subset of the clients, additional access trials (about 300 sound files, in a specific "`dev`" subdirectory) will be provided.

This data is intended for system development and calibration purposes (e.g. fixing decision thresholds and other system parameters).

Corresponding trial lists (`dev_ts1.ndx` and `dev_ts2.ndx`) will be distributed. Development trials are "non-blind", in the sense that they are all guaranteed to be genuine client accesses, with transmission channel ("`P`" or "`G`") indication.
Participants should build-up a set of impostor accesses by their own. For example, they can simulate impostor attacks with the client data provided, using known client files as impostor attempts against other subjects [2].
Details on the development techniques should be given in submission notes.

3. Evaluation Metrics

Performance of the systems will be visualized and compared using standard DET curves [3], calculated on confidence scores. Systems should also provide a decision (whether speaker in the test segment is a client or an impostor). This can raise two kinds of errors:

- *False Rejection* (FR), or *Missed Detection*, when a genuine client is rejected.
- *False Acceptance* (FA), or *False Alarm,* when the system accepts an impostor.

The cost associated to these two kinds of error depends on the considered application. The explicit decisions of the system will be used to measure overall speaker detection performance, with a *Detection Cost* (C_{Det}) computed according to the following function:

$$C_{Det} = C_{FR} \cdot P_{FR/Client} \cdot P_{Client} + C_{FA} \cdot P_{FA/NonClient} \cdot (1 - P_{Client}),$$

where C_{FR} and C_{FA} are respectively the Cost of False Rejection and the Cost of False Acceptance, while $P_{FR/Client}$ and $P_{FA/NonClient}$ are the False Rejection and the False Acceptance Rates. P_{Client} is the *a priory* probability that the speaker is a client.

Participants are requested to use the following decision strategy parameters:

$C_{FR} = 10$
$C_{FA} = 1$
$P_{Client} = 0.5$

This choice corresponds to a "High-Convenience" kind of application, as opposed to a "High-Security" one.

Optionally, participants can prepare additional submissions with different decision strategies. In this case, the value of the parameters should be specified in their submission notes.

4. Submission of results

Each participant site must submit only one "primary" system. Participant are welcome to present additional systems for comparison at the evaluation workshop.
Along with the description of their system(s), participants should return a single *evaluation result file* per system, in ASCII format, following NIST-SRE [1] specifications:

- The file name should be constructed as "`EV09_SIVAP_{`*`SSS`*`}_{`*`N`*`}`", where {*SSS*} identifies the site, and {*N*} identifies the system.
- The file must contain records for all the test trials. Each record (line) has nine fields:
    1. The training condition – **TC1**, **TC2**, **TC3**, **TC4**, **TC5** or **TC6**.
    2. Adaptation mode – "**n**" for no adaptation and "**u**" for unsupervised adaptation.
    3. The test condition – **TS1** or **TS2**.
    4. The sex of the target speaker – **m** or **f**.
    5. The target model identifier.
    6. The test segment identifier (without the "`.alw`" extension).
    7. The transmission channel of the test segment (if channel detection techniques are used) – "**P**" for PSTN, "**G**" for GSM, "**X**" for not detected.
    8. The decision – **t** or **f** (whether or not the target speaker is judged to match the speaker in the test segment).
    9. The confidence score (where larger scores indicate greater likelihood that the test segment contains speech from the target speaker).

The typical content of a submission file looks like the following extract:

```
TC2 n TS1 f PA nttai X f -8
TC2 n TS1 f HY lubft X f -8
TC2 n TS1 f HY naffi X f 0
TC2 n TS1 f HY bltui X t 4
TC2 n TS1 f AK aiabf X t 2
```

References

[1] “The NIST Year 2008 Speaker Recognition Evaluation Plan,” http://www.nist.gov/speech/tests/sre/2008/sre08_evalplan_release4.pdf

[2] E. Bailly-Bailliére et al. “The BANCA Database and Evaluation Protocol.” In Audio- and Video-Based Biometric Person Authentication, Lecture Notes in Computer Science, volume 2688, 2003.

[3] A. Martin et al. “The DET curve in assessment of detection task performance.” In Eurospeech’97, volume 4, 1997.