The Unifi-EV2009-1 Protocol for Evalita 2009

Monica Carfagni and Matteo Nunziati

University of Florence, Florence, Italy
{monica.carfagni, matteo.nunziati}@unifi.it

Abstract. This paper presents the Unifi-EV2009-1 protocol employed by the Department of Mechanics and Industrial Technologies (DMTI) of University of Florence during the second evaluation campaign of Natural Language Processing tools for Italian (Evalita 2009). DMTI has attended the Forensic Speaker Identity Verification (SIV) task of Evalita 2009 by using an automatic speaker recognition (ASR) system. Being this the very first time such kind of technologies have been employed by DMTI on real forensic data, this paper is aimed to present both results obtained by and improvements scheduled for Unifi-EV2009-1.

Key words: Forensic speaker recognition, Automatic speaker recognition.

1 Introduction

DMTI has joined the Evalita 2009 Speaker Identity Verification SIV task in order to start a new research project. The aim is to slowly introduce ASR in Italian forensics as well as it has been introduced in other countries. Being this the first time DMTI can approach the ASR problem on an official evaluation campaign featuring real forensic data, a specific protocol named Unifi-EV2009-1 has been defined. The Unifi-EV2009-1 goal is to adopt classical, well assessed (even if obsolete) technologies. We think that this approach allows for a better understanding of any component in modern ASR systems, leading to a smooth implementation path.

2 Unifi-EV2009-1 approach to SIV-forensic Track

The Evalita SIV task is divided into two Tracks: Application and Forensic. The Forensic Track asks for the recognition of two speakers into two test sets: a closed set with a fixed number of speakers (CST) and an open set with an undefined number of speakers (OST). The CST is composed by 16 recordings, each one containing the voice of a single speaker. The OST is composed by one big recording, featuring many speakers talking together. According to the Speaker Identity Verification - Forensic Track Task Guidelines, both tests are recorded in real forensic conditions with A-law compression and 8 kHz sampling rate. In order to perform a recognition, a train set provides 4 samples of the two speakers. Both speakers are recorded in a laboratory by employing high quality equipment in optimal acoustic conditions. Test data are available in both 44.1 kHz and 8 kHz formats with linear encoding (PCM). Development and background data are not provided.

The Unifi-EV2009-1 protocol applies NIST rules [1] to CST, thus, recordings are processed without any human intervention (e.g. listening to recording contents). As a matter of fact, OST can't be automatically processed unless an automatic speaker diarization tool is available. Such kind of tool has not been scheduled for this task, therefore hand made segmentation has been planned for the OST test. This implies that any classical forensic analysis is planned on OST too (e.g. attended SNR estimation). Segmented speakers are expected to be treated as CST recordings.

3 System Description

The Unifi-EV2009-1 system is based on the ALIZE/SpkDet software developed and distributed under LGPL by University of Avignon [2]. Speech processing is performed by means of Spro [3], a signal processing software specifically developed for speech parametrization. Automatic signal amplitude normalization is applied to each recording in background, development, train and test sets. No automatic SNR estimation or automatic noise removal/masking has been planned for this test. The recognition front-end is based on 13 MFCC plus delta and delta-delta features. Log-energy has been retrieved for each frame, in order to perform a speech/non-speech classification.

All features have been normalized (null mean, unitary variance) by means of cms and variance normalization. SNR is estimated in an attended manner for OST due to limitations described in the above paragraph.

Basic speech/non-speech detection is performed by means of statistical log-energy clustering. A two-component GMM is employed and the higher energy cluster is retained and classified as speech. As no noise estimation/remotion/masking is applied (excluding OST) noise could be mixed with voice.

Recognition is based on a simple GMM-UBM stage [4]. 512 components are used for UBM and MAP is applied to GMM means only. Employed UBM is gender- and language-independent. GMM-UBM scores are retrieved for a development population (other than train and test sets). Scores are T-normed [4] by using UBM speakers: only the ten best results per speaker are retained. Development Target and non-Target scores are approximated with two normal distributions: N_t and N_{nt} respectively. Given a score *s*, LR is computed for each comparison as: $LR(s)=N_t(s)/N_{nt}(s)$.

4 Decision Threshold

The decision threshold setup is a non-problem for forensic trials [5]: decision has to be based on posterior probabilities, which rely on both LR and prior probabilities. Being the latter province of the court or investigation offices, the expert has not to define any threshold. Nonetheless, according to task rules, a threshold has been adopted at P_{post}=50%. Thus, we consider two recordings as coming from the same speaker if P_{post} is higher than 50%. In order to define a posterior probability, a prior probability has to be fixed on its turn. The number *n* of cohort speakers has been considered for P_{prior}, along with a flat prior distribution: each speaker in the set has an associated P_{prior} of *1/n*. According to Bayes' theorem, the following threshold has been retrieved:

$$\frac{P_{post}}{(1-P_{post})}=LR\frac{P_{prior}}{(1-P_{prior})}\Rightarrow 1<\frac{LR}{(n-1)}\Rightarrow LR>(n-1)\,. \tag{1}$$

Thus, we consider two recordings as coming from the same speaker if LR is higher than (n-1). For CST, the following quote from Speaker Identity Verification - Forensic Track Task Guidelines has been taken into account: *"[...] a recording session in a noisy place including the four speakers present in the speech corpus, together with a large number [...]"*. Therefore, the CST cohort has been fixed to 4 speakers. The OST threshold depends on the number of speakers retrieved from the OST file.

5 Background and Development Populations

Unifi-EV2009-1 corpora are based on CSLU data. The background population is a subset of the "22 Language" corpus from CSLU. According to CSLU description [6], the number of male and female speakers has been taken as balanced as possible, thus we can consider UBM as 50% male and 50% female. There isn't any age specific information in the corpus documentation. Very different languages are represented with a similar amount of data: from English to Chinese, including Italian. Recordings have been acquired on a land telephone line with 8 kHz sampling rate and mu-law encoding. An average recording duration of 20 s has been selected for the UBM. Spontaneous speech has been included, any other kind of speech has been removed (e.g. word or digit repetitions). The number of employed recordings is 1977, with an average of 94 speakers per language.

A subset of the "Speaker Recognition" corpus from CSLU [7] has been adopted for false acceptance (FA) and false rejection (FR) estimation, as well as for LR computation. The same considerations of "22 Languages" can be applied here about sex, age, kind/duration of speech and acquisition methodologies. English language is adopted by speakers. Speakers are both L1 and L2 speakers, but no information is available about the percentage of L2 speakers on the total. The number of employed recordings is 1048, with an average of 8 recordings per speaker.

6 Results

Figure 1 reports the discriminatory capability of the Unifi-EV2009-1 system as obtained against the development dataset: EER is approximatively 15%. 8 kHz linearly encoded train data have been employed for the task, being 44.1kHz sampling frequency not comparable with forensic recordings. A preliminary test has been conducted in order to define the accuracy over the train set: the 8 recordings have been tested against themselves, generating 56 comparisons. During the actual CST, the defined threshold has never been surpassed: the highest LR was *0.91*. Synthetic results for development, train and CST conditions are reported in Table 1 in terms of both FA and FR. Development values are retrieved by applying a working point equal to LR=*3*.

Table 1. Recognition errors for Unifi-EV2009-1 system.

	FA	FR
Devel. (μlaw, phone quality)	4.1%	37.2%
Train (PCM, high quality)	12.5%	0.0%
CST (Alaw, unk. quality)	0.0%	100.0%

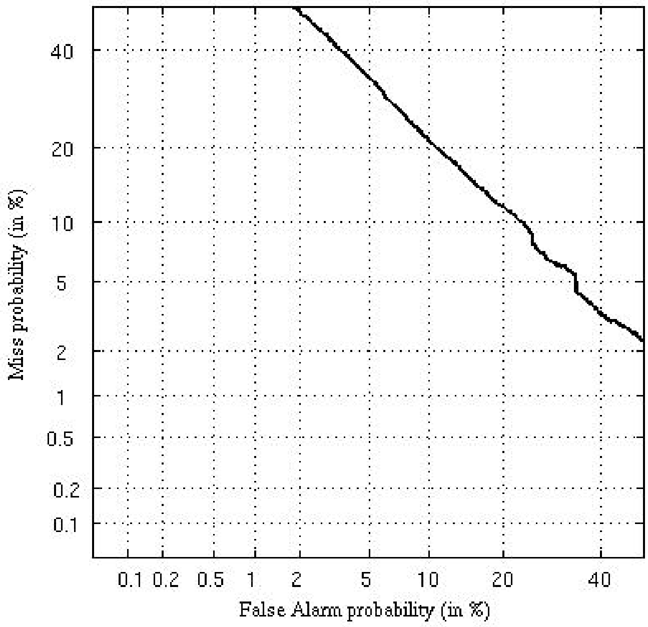

Fig. 1. DET plot for Unifi-EV2009-1 system under development conditions.

As OST has been manually checked, attended noise analysis has been carried out, simulating our common forensic practices. The listening has pointed out a relevant and stationary noise all over the recording. A noise sample of 1.2 s has been kept and SNR has been estimated for the whole recording. Analysis has pointed out a really low SNR for almost the whole duration. Few segments (less than 10 s) go over 6 dB and commonly maintain a stable SNR around 3 dB. Being SNR so low, we have refused any clustering and comparison due to the lack of quality in data, as in our common forensic practice. Therefore, no result is reported for such a trial.

7 Conclusions

The Unifi-EV2009-1 protocol has been applied during the Evalita 2009 Forensic SIV Task. Attended analysis of the OST material has been carried out for clustering, while NIST SRE rules have been applied to CST. OST data have shown a poor quality (average SNR of 3 dB), therefore recognition has not been carried out for them. CST material has been processed obtaining a really poor discriminatory performance. Even though CST set has not been accessed by any human operator, we guess that the same sound quality detected in the OST is detectable in CST too. According to development and test results and having listened to the OST data, we guess that the Unifi-EV2009-1 system performance is mainly induced by the relevant amount of noise present in recordings, being channel mismatch only a secondary element. In our experience, average forensic conditions are better then those proposed in the OST; additionally, the number of both train and CST recordings was too limited for it to be representative. Nonetheless, obtained results show the relevance of an adequate automatic SNR estimation. Moreover, to include quality measures in recognition it has been pointed out in literature as a relevant element for system accuracy [8]. Therefore, we plan to introduce unattended SNR estimation in our system in order to overcome current limitations.

References

1. National Institute of Standards and Technology: http://www.itl.nist.gov/iad/mig/tests/sre/
2. Bonastre, J.F., Scheffer, N., Matrouf, D., Fredouille, C., Larcher, A., Preti1, A., Pouchoulin, G., Evans, N., Fauve, B., Mason, J.: ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition. Available on line at http://mistral.univ-avignon.fr/en/applications.html
3. Gravier G.: Speech signal processing toolkit – release 4.0. Available on line at http://www.irisa.fr/metiss/guig/spro/documentation.html
4. Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau , I., Meignier, S., Merlin, T., Hortega-Garcia, J., Petrovska-Delacretaz, D., Reynolds, D.A.: A tutorial on text-independent speaker verification. EURASIP journal on applied signal processing, vol. 4, pp. 430--451(2004)
5. Champod, C., Meuwly, D.: The inference of identity in forensic speaker recognition. Speech Communication, vol. 31, pp.193--203 (2000)
6. Center of Spoken Language Understanding: http://cslu.cse.ogi.edu/corpora/22lang/
7. Center of Spoken Language Understanding: http://cslu.cse.ogi.edu/corpora/spkrec/
8. Garcia-Romero, D., Fierrez-Aguilar, J., Gonzalez-Rodriguez, J., Ortega-Garcia J.: Using Quality Measures for Multilevel Speaker Recognition. Computer Speech and Language, Special Issue on Odyssey-04: The Speaker and Language Recognition Workshop, vol. 20, pp. 192--209 (2006)