DOI: 10.17469/O2110AISV000005

FRANCESCO SIGONA, SONIA D'APOLITO, COSIMO IAIA,
BARBARA GILI FIVELA, MIRKO GRIMALDI

Forensic Automatic Speaker Recognition with dialectal speakers: a pilot study on the Taranto and Brindisi varieties

In the Forensic Speaker Recognition, the choice of the reference linguistic population plays a key role in evaluating the typicality of the voice recordings to be compared, and therefore the strength of the evidence, within the Likelihood Ratio framework. In the present study we carry out multiple comparison tests among voice recordings of the dialectal speech of Taranto and Brindisi varieties (Puglia, Italy), through an Automatic Speaker Recognition system, using two reference populations (databases): one of "standard" Italian, by professional speakers, and one of dialectal speakers of the same variety as the voices to be compared. The aim is to observe whether the accuracy of the recognition system improves if the dialect reference population is used instead of the Italian spoken one.

Keywords: forensic voice comparison, likelihood ratio, reference population, spoken dialect, accuracy.

1. *Introduction*

1.1 The forensic voice comparison

In the Forensic Voice Comparison (FVC), one or more audio recordings of the voice of a known speaker (hence, known samples) are compared to an audio recording of the voice of a speaker of questioned identity (hence, questioned samples)[1]: the goal is to understand to what extent the two samples of voices can probably be attributed or not to the same person.

To this purpose, a Bayesian approach has been widely established, and recently recommended by European Network of Forensic Science Institutes (Drygajlo et al., 2016). Accordingly, the task of the forensic scientist is to provide the court with a strength-of-evidence statement in answer to the question: "How much more likely are the observed differences between the known and questioned samples to occur under the hypothesis that the questioned sample has the same origin as the known sample than under the hypothesis that it has a different origin?" (Morrison, 2009). The answer to this question is quantitatively expressed as a Likelihood Ratio (LR): the LR represents the relationship between similarity and typicality of the compared voice samples and quantifies how similar the characteristics of the

[1] Also, Forensic Speaker Recognition (FSR) is frequently used.

recorded signal of the two voice samples are to each other, relative to the diffusion of the same characteristics in a reference linguistic population. The reference linguistic population, therefore, must be defined with particular attention and must be as homogeneous as possible (gender, age, language of the speakers, etc.) with respect to the characteristics of the recorded samples to be compared (Rose, 2002, 2005; Jessen, 2008). Failure to implement an adequate reference population is one of the main factors that makes the estimate of the LR, as well as the strength of the evidence, inaccurate (Robertson, Vignaux, 1995; Rose, 2006).

A crucial issue for FVC is represented by dialectal variation (and also micro-variation)[2], not yet fully addressed within this field of research. At the moment, the databases of speech populations used in semi- and full-automatic systems refer to idealized speakers: that is, speakers broadly identified in respect of a national language (English speakers, Italian speakers, German speakers, etc.). Conversely, we know very well that national languages are abstract (administrative) entities and that speakers normally use regional varieties (more or less markedly) in their real life: this is particularly true for the Italian linguistic area. Furthermore, the varieties used by speakers are, in many cases, characterized by different phonological systems and different suprasegmental patterns. Also, phonological systems may drastically vary within a limited linguistic space, showing puzzling micro-variation; and each phonological system may show systematic phonological processes that change phonemes at the phonetic surface. So, what we today compare in forensic practice are, at best, common (general) features eventually shared by linguistic systems.

We do not know what biases this fact introduces within the forensic approaches (for instance, the Bayesian approach). Finally, this issue is inherently linked to socio-phonetics aspects: in fact, speaker's features may vary in respect of sociolinguistics variables (age, sex, literacy, contexts of use, etc.). The semi-automatic approach, indeed, allows an expert to select general properties of the vowels analyzed, avoiding that macro- and micro-variation have a drastic impact on the LR. On the contrary, this is impossible when an automatic approach is used.

For what concerns the Italian domain, previous contributions have highlighted the necessity to appropriately develop speech databases for FVC taking care of dialectal variation (cf. Romito et al., 2009; Romito, Galatà, 2008). An example is given by the *Primula corpus*, which contains over 900 recordings of 4 Calabrian speakers. It is characterized by three types of recording channels: high fidelity, environmental and telephone recordings. The recordings have been captured under different conditions that determine their quality: silent room, tapping in and out of a car, calls effected in the car, in the street, and in the classroom (Romito, Galatà, 2008). Unfortunately, this corpus is not, at the moment, available for research scopes.

[2] With the term 'micro-variation' we refer to the fact that dialects may often manifest subtle and irregular variations in respect of a general phonological or morpho-syntactic phenomena.

1.2 The aim of the present work

Along this line of research, we collected a dialectal database from speakers of two different varieties spoken in Southern Apulia (Italy). In this pilot study, we aim to investigate to what extent the use of a dialectal population, compared with a standard Italian speech population (produced by professional speakers), can influence the performance of state-of-the-art Forensic Automatic Speaker Recognition (FASR) systems. We assume that the accuracy of the comparison can improve when dialectal variation is taken under consideration.

This pilot study focuses only on male speakers for the following reasons. Firstly, it is well known that the production of speech sounds from male and female speakers is strongly dependent on biological differences, such as inner dimension of the mouth, throat, and vocal folds (Simpson, 2009; Hillenbrand & Clark, 2009). These important differences therefore justify a differentiated experimentation between the two genders, using two distinct reference population samples. On the other hand, collecting and analyzing audio samples (and especially dialect audio samples) is time-consuming. So, having a limited amount of time, we decided to focus our efforts on getting a large enough sample of one gender, rather than risk getting two too few samples for both genders. In the near future, however, it will always be possible to repeat the study with a sample of the other gender, integrating the results also in a comparative perspective between the two genders. The choice to begin with the male gender, instead of female, was essentially random.

Furthermore, the study has been carried out with audio samples at the classic telephonic audio quality, as this kind of signal happens very frequently in the forensic field (as in the case of wiretapping).

2. *Methods*

2.1 Datasets

2.1.1 The Italian dataset

A dataset made of 150 Italian male professional speakers has been collected by recordings of freely available audio samples (https://www.audible.it). The characteristics of the collected samples are summarized in Table 1.

Table 1 - *Characteristics of the recorded audio samples for the Italian spoken dataset*

Average duration	3 minutes
Number of samples	150
Number of samples per speaker	1
Speaker gender	Male 100% (Female 0%)
Audio encoding (format, sampling rate, bit depth)	Mono, wav, 44100 Hz, 16 bit
Background sounds	no
Signal-to-Noise Ratio (SNR)	> 25 dB

In short, the voices of this dataset were those of professional speakers reading audiobooks (in the same case the reader impersonates two characters conversing on specific arguments). The average age of the speakers could not be determined.

Each audio recording was band-pass filtered between 300 and 3400 Hz and downsampled to 8000 Hz, using the Praat software (Boersma, Weenink, 2020), when necessary, after the sample collection.

2.1.2 The dialect datasets

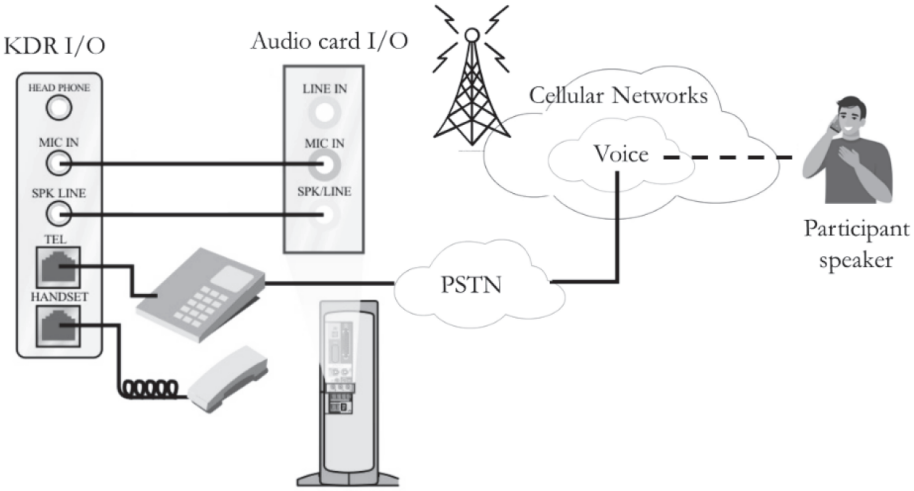
Speakers from Taranto and Brindisi areas were interviewed in order to collect dialectal data. Specifically, during a semi-guided telephonic interview, they were asked to translate, as spontaneously as possible, into dialect some Italian target phrases proposed by the interviewer. The speakers were called to their own mobile phone from a landline connection using a dedicated hardware (KDR, by Sistel s.r.l.) connected to a Microsoft® Windows®-based Personal Computer running the software Audacity®[3] which was used to capture and record the audio signal. The whole setup is depicted in Figure 1.

And so, we collected dialectal data as follows:

- 22 male speakers (mean age 44) for the Taranto area: 14 speakers from Taranto city, 4 from Grottaglie, 2 from Monteiasi, 1 from Fragagnano, and 1 from Carosino;
- 22 speakers for the Brindisi area (mean age 40): all speakers come from Francavilla Fontana (see Figure 2).

All the audio samples were resampled to 8000 Hz, and have SNR > 20 dB.

Figure 1 - *Diagram of the setup for the dialect varieties speech samples collection*



[3] Audacity® software is copyright © 1999-2021 Audacity Team. Web site: https://audacityteam.org/. It is free software distributed under the terms of the GNU General Public License. The name Audacity® is a registered trademark.

2.1.3 Dialectal properties of the Taranto and Brindisi varieties

The Taranto varieties shows clear features of the Apulian dialects: i.e, diphthongization of stressed vowels within open syllables ([aˈʧejtə] *vinegar*; [ˈpajlə] *hair*), weakening of the final unstressed vowels ([ˈtandə] *many*; [kapeḍːə] *hair*), reduction of the metaphonic diphthongs [je], [we] to [i], [u] ([ˈpjedə] > [ˈpidə], *foot*, [ˈmwertə] > [ˈmurtə]), etc. (Mancarella, 1998).

On the other hand, the Francavilla Fontana variety, although characterized by the Apulian features, preserves many Salentino features, as the metaphony affecting the mid-high vowels [e], [o] developed by the Latin vowels ĭ, ē and ŭ, ū. The outcome is the raising of [e], [o] to the high vowels [i], [u]: [la ˈpera] *the pear* Sg F / [lu ˈpiru] *the pear* Sg M; [lu kuˈlore] *the color* Sg / [li kuˈluri] *the colors* Pl, etc. Also, the vowels [ɛ], [ɔ] derived by the Latin vowels ĕ, ŏ may be affected by metaphonic diphthongization: i.e., [lu ˈpɛti] *the foot* Sg. / [li pjeti] *the feet* Pl; [la ˈnotti] *the night* Sg. / [li ˈnwetti] *the nights* Pl. (Ribezzo, 1912).

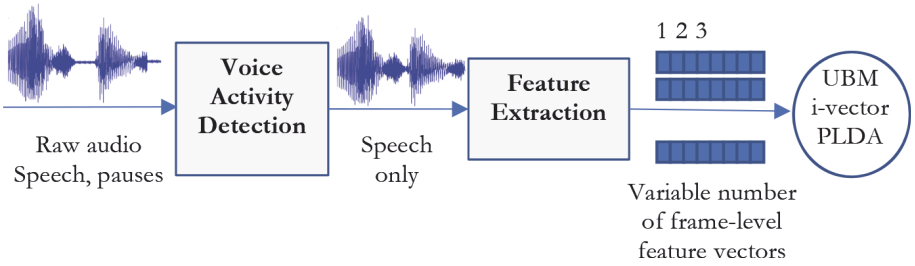
Figure 2 - *Maps of the investigated dialect areas*

It is interesting to note that for what concerns the Taranto area within the Monteiasi, Carosino, and Fragagnano varieties, Apulian features coexist with Salentino features (Ribezzo, 1912; Mancarella, 1998). Thus, this fact introduces further dialectal variation within the speech population, which is at the core of our aim (cf. Section 1.2).

2.2 Front-end audio processing

The comparison algorithm that has been chosen in the current study requires a preliminary conversion of the recorded audio samples into a sequence of vectors of numbers, which give an alternative representation of the information contained in the waveform of the voice signal. These "feature vectors" are the output of a front-end audio processing stage, that for the current study has been simplified into a chain of only two processing blocks: the Voice Activity Detection (VAD) and the real Feature Extractor (see Figure 3).

Figure 3 - *Block diagram of the front-end audio processing, which transforms the input audio file into a variable number of feature vectors required by the modelling block (the UBM/i-vector/PLDA)*



2.2.1 Voice Activity Detection

Voice Activity Detection (VAD), often referred to as Speech Activity Detection (SAD) or simply Speech Detection, is the task of locating speech segments within an audio recording. VAD plays a key role in any speech processing system, including speaker recognition applications where, at least for those approaches based on short-term spectral features, it is required to prevent unnecessary processing of non-speech segments. To this purpose, many methods have been proposed. Classic digital signal processing methods usually classify voiced/unvoiced frames based on scalar features such as short-term energy, zero-crossing rate (Benyassine et al., 1997), periodicity (Tucker, 1992) or spectral divergence (Ramirez et al., 2004); these methods are quite simple and effective on clean condition, but the classification accuracy tends to suffer on low SNR. Statistical model-based approaches have been explored (Sohn et al., 1999; Shin et al., 2010), assuming that the spectral coefficients follow a particular parametric distribution, where the VAD decision is sought by calculating the likelihood ratio based on the hypothesized models. The statistical methods often outperform the classic methods in the presence of stationary noise, but non-stationary noise conditions remain challenging. Supervised models have also been studied, using machine learning techniques and leveraging prior knowledge in large, annotated audio collections (Ng et al., 2016; Plchot et al., 2016; Wu, Zhang, 2011; Zhang, Wu, 2013; Thomas et al., 2015). Such VAD approaches tend to be sensitive to acoustic mismatch between the training and test. Adaptive supervised VADs, such as Huijbregts et al. (2007) and Kinnunen and Rajan (2013) have also been considered, that represent a compromise between the powerful supervised approaches, such as neural networks, and statistical model-based methods which require no prior training but whose parametric modelling assumptions might be over-simplistic.

In this work we decided to use a VAD technique which was adequate to the quality of the available speech recordings and for which a software implementation was already available. We found both requirements met in VoiceBox[4], a freely

[4] VOICEBOX: Speech Processing Toolbox for MATLAB. Web site: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

available toolbox for Matlab®, that comes with an implementation of Sohn et al.'s (1999) approach.

2.2.2 The Feature Extractor

The output of the VAD is an audio waveform that is input to the real Feature Extractor algorithm. Even in this case, we decided to choose baseline features, such as the Mel-Frequency Cepstral Coefficients (MFCCs). MFCCs are short-time spectral features, meaning that the speech utterance is first divided into (usually overlapped) fragments (frames), that are small enough to be able to assume that in each one the characteristics of speech are time-invariant, then a MFCCs vector is computed for each frame. MFCCs were introduced in the early 1980s by Davis and Mermelstein for speech recognition, and then adopted in many studies on speaker recognition. MFCCs are computed with the aid of a psychoacoustically motivated filter bank, followed by logarithmic compression and discrete cosine transform (DCT). The step required to compute MFCCs are:

- pre-emphasis: this step refers to a filtering operation that emphasizes the speech signal at higher frequencies and is considered in many speech processing applications.
- framing: this operation has been described above. The frame length is usually fixed, but pitch-synchronous analysis has also been (Nakasone et al., 2004; Zilca et al., 2006; Gong et al., 2008) and is still studied (Chen and Miller, 2020).
- windowing: this operation aims at tapering the signal to zero at the beginning and end of each frame, to deal with the finite-length effect of the Discrete Fourier Transform (DFT). The Hamming window is commonly used but the choice of the window function is not considered critical (Kinnunen, Li, 2010).
- DFT: the very well-known technique. It is computed frame by frame to get the spectral amplitude of the signal inside each frame.
- Mel filter bank processing: the output of the DFT is multiplied by a bank of filters to achieve the so-called mel-spectrum. A mel is a unit of measure based on the human perception of tones. The human auditory system apparently does not perceive pitch linearly, so the mel does not correspond to the value of the physical frequency.
- Logarithmic compression: since voiced sounds can be modelled by a source signal filtered by the resonance cavity of the vocal tract, which is a multiplication in the frequency domain, applying logarithm operation allows to get multiplied factors of the spectrum (in this case, the mel-spectrum) into additive ones. This results in a signal in the cepstral domain with a quefrequency peak corresponding to the pitch of the signal and several formants representing low quefrequency peaks.
- Discrete Cosine Transform (DCT): since the vocal tract is smooth, the energy levels in adjacent bands tend to be correlated. Also, the filters in the filter-banks are overlapped, so the energy from ones next to each other is being spread between two. DCT applied to the transformed mel frequency coefficients produces a set of cepstral coefficients.

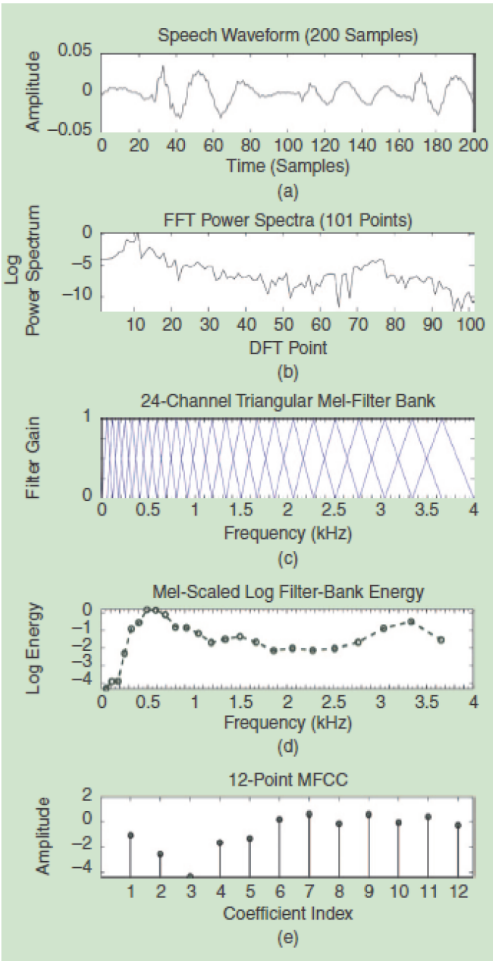
Figure 4 summarizes the MFCCs computation processing, while further details can be found in any speech processing manual.

The specification of the MFCCs extraction used in this study are:

- Frame duration: 16 ms (i.e., 128 samples @ 8000 Hz);
- Frame overlap: 50% (i.e., 64 samples)
- 14 static coefficients, including 0-order and log energy
- 28 dynamics coefficients (delta, delta-delta), corresponding to the first- and second-order derivatives of the static coefficients.
- F = Total number of MFCCs = 42.

The actual computation has been performed using the VoiceBox toolbox.

Figure 4 - *Steps in MFCC feature extraction from a speech frame: (a) 200-sample frame representing 25 milliseconds of speech sampled at a rate of 8 kHz, (b) DFT power spectrum showing first 101 points, (c) 24-channel triangular Mel-filter bank, (d) log filter-bank energy outputs from Mel-filter, and (e) 12 static MFCCs obtained by performing DCT on filter-bank energy coefficients and retaining the first 12 values (adapted from Hasan, Hansen, 2015)*



2.3 Automatic Speaker Recognition Algorithm

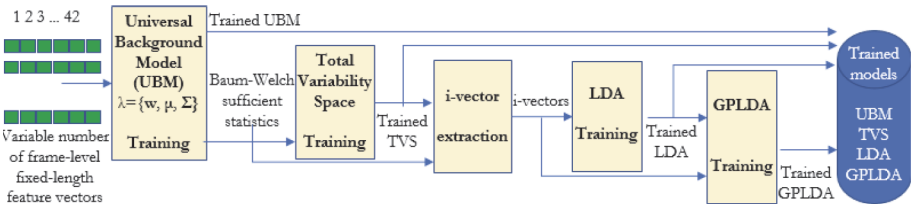
The chosen algorithm for the voice comparison is based on the UBM/i-vector/PLDA paradigm, which is a well-known state-of-art approach. In short:

- the Universal Background Model (UBM; Reynolds, 1997) is a probability density function in the form of a Gaussian Mixture Model (GMM) with K components, which is trained with all the feature vectors coming from all the speakers in the reference population. The output of the training is the set of UBM parameters (one estimated mean vector and covariance matrix for each Gaussian component in the mixture) as well as Baum-Welch sufficient statistics, that will play a role in the next blocks of the chain.
- i-vectors refers to a particular vector representation of the single utterance on a vector space called "Total Variability Space" (TVS, Dehak et al., 2011), which is particularly convenient for the purpose of voice comparison. The basic idea is that every single utterance of a generic speaker contains information that is a combination of a bias due to the target population, plus speaker-dependent and speaker-independent information. Speaker independent information depends on external factors such as the quality of the recording device, the amount and type of noise superimposed on the voice, etc., and is usually referred to as "channel / session dependent". The i-vector approach has proven to be a viable and effective way to represent speaker-dependent and independent information, without any bias due to the target population, using far fewer elements than the amount needed to work directly with Gaussian Mixed Models.
- GPLDA (Gaussian Probabilistic Linear Discriminant Analysis; Kenny 2010; Garcia-Romero and Espy-Wilson, 2011) is a back-end stage, that extracts the speaker-dependent information from the i-vectors (possibly after a further dimensionality reduction by Linear Discriminant Analysis), providing at the same time the framework to compare two i-vectors in terms of a Likelihood Ratio-based score.

Figure 5 depicts the processing steps that occur to train all the mentioned models, while Figure 6 depicts the processing steps that occur in the computation of the models of the reference speaker (the known speaker in a real FVC case) and the test speaker (the questioned sample, in a real FCV case), up to the scoring stage.

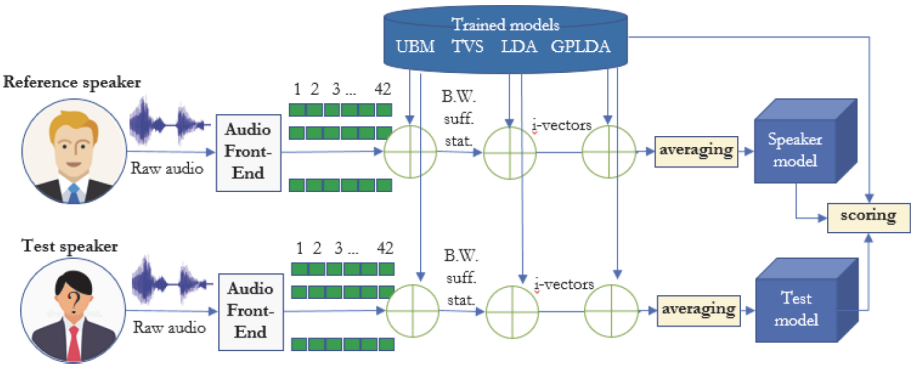
The computation of the models has been done by means of the Microsoft® MSR Identity Toolbox[5] for Matlab®.

Figure 5 - *Train of the models of reference population required by the UBM/i-vector/GPLDA paradigm*



[5] https://www.microsoft.com/en-us/download/details.aspx?id=52279

Figure 6 - *Processing steps to compute the reference and the test speaker models, up to the scoring*



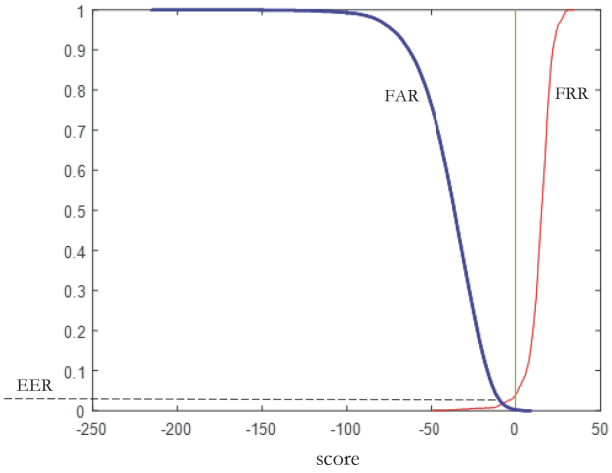
2.4 Performance metric

The main global metrics to assess the accuracy performance of a speaker recognition system are the estimated probabilities that the system response supports contrary-to-fact hypothesis. When the system's response score supports the identification hypothesis of two utterances spoken by two different speakers, a False Acceptance (or False Identification) error occurs; when the system's response score supports the rejection (i.e. no identification) hypothesis of two utterances spoken by the same person, a False Reject error occurs.

By collecting the scores coming from a multitude of comparison tests between couples of utterances spoken by different speakers, it is possible to build a distribution of the scores and the distribution of the False Acceptance error Rate (FAR) as a function of the score, in this different-origin hypothesis. Similarly, by collecting the scores coming from a multitude of comparison tests between couples of utterances spoken by the same speaker, it is possible to build a distribution of the scores, and the distribution of the False Rejection error Rate (FRR), in this same-origin hypothesis. FAR and FRR computed in this way represent an estimate of the respective a priori error probabilities, while the a posteriori estimates depends on the outcome of the real forensic casework.

A very common global metric used to assess the accuracy of the system is then the Equal Error Rate (EER), which is the point where the FAR equals the FRR. The lower the EER, the higher the accuracy of the FASR system. It can be visualized using the Tippet plot as depicted in Figure 7.

Figure 7 - *Example of Tippet plot, showing the FRR (red line) and the FAR (blus line) as functions of the score). The intersection point gives the value of the Equal Error Rate (EER)*



2.5 Design of matched comparison tests

To assess the FASR performances in terms of EER, multiple test comparisons must be performed. In the case of reference population matched with the dialectal variety of the speaker, we proceeded as follows. Firstly, we divided the MFCCs matrix of each speaker in two equally sized matrices, in order to simulate two different utterances (sessions) for each speaker. Then, assuming that N is the number of dialect speakers in the dataset, to assess the distribution of the FAR we adopted a one-leave-out cross-validation, in which we compared the two sessions of each speaker, using the remaining N-1 speakers of the dialectal dataset to build the models of the reference populations. By this way, there is no chance that such models were biased due to any information about the test speaker. To assess the distribution of the FRR, we adopted a similar strategy, in which we compared the first session of each speaker with the second session of any other speaker, using the remaining N-2 speakers to build the models of the reference populations. Even in this way, there is no chance that such models were biased due to any information about the test speakers. Finally, the EER is computed as the score where FAR=FRR. We repeated such a cross validation procedure 100 times, in order to average with respect to some random initializations of the reference population models, occurring during their computation.

2.6 Design of mismatched comparison tests

The mismatched comparison tests are about the comparison of dialectal speaker using the Italian reference models. Since the Italian dataset is much larger than each dialectal dataset, we assumed that was not methodologically correct to build Italian reference models using the whole Italian dataset. Therefore, before to perform the comparison test, we have randomly split the whole Italian dataset into D smaller ones, having the same size (N) of the dialectal datasets. For each of the D sub-dataset,

we computed the reference population models and performed the comparison test between the two sessions of a same dialectal speaker (to assess FAR) a between the first session of each dialectal speaker with the second session of any other dialectal speaker. Moreover, we repeated the whole procedure (including the random composition of the D Italian sub-datasets) 100 times.

3. *Results*

The results achieved in terms of EER in the mismatched and matched conditions are summarized in Table 2 for Taranto speakers, and Table 3 for Brindisi/Francavilla speakers.

Table 2 - *Descriptive statistics of the set of EER values coming from the multiple comparison tests, for speakers belonging to the Taranto area, using an Italian reference population, then the matched dialectal one. Coefficient of variation % is defined as 100 x standard deviation / mean*

	Mismatched condition (Italian refer. popul.)	*Matched condition (dialectal refer. popul.)*
Number of trials	100	100
EER mean	7.9 %	3.0 %
EER standard deviation	1.3 %	0.43 %
EER coefficient of variation %	16.4 %	14.1 %
EER minimum	5.2 %	2.2 %
EER maximum	12.7 %	4.2 %
ERR range	7.5 %	2.0 %

Table 3 - *Descriptive statistics of the set of EER values coming from the multiple comparison tests, for speakers belonging to the Brindisi area, using an Italian reference population, then the matched dialectal one. Coefficient of variation % is defined as 100 x standard deviation / mean*

	Mismatched condition (Italian refer. popul.)	*Matched condition (dialectal refer. popul.)*
Number of trials	100	100
EER mean	10.0 %	6.1 %
EER standard deviation	1.35 %	0.74 %
EER coefficient of variation %	13.5 %	12.0 %
EER minimum	7.8 %	4.5 %
EER maximum	14.0 %	8.3 %
ERR range	6.2 %	3.8 %

4. *Discussion*

For both the dialectal varieties, we achieved a reduction in every EER-based metric in the matched (dialectal speakers and dialectal reference population) over the mismatched (dialectal speakers and "standard" Italian reference population) condition.

As for the Taranto dialectal area, the mean EER shows that it decreases from 7.9% for the mismatched condition to 3.0% for the matched one. Further, even the standard deviation improves in the matched condition, as it lowers from 1.3% to 0.43%, as well as the coefficient of variation even if quite slightly.

These results are also confirmed considering the Francavilla dialectal data. Indeed, the mean EER improves from the mismatched condition (10%) to the matched condition (6.1%). The standard deviation also improves as it lowers from 1.35% to 0.74% in the matched condition and the coefficient of variation slightly improves too.

The achieved results clearly support our initial research hypothesis that in caseworks where dialectal speakers are involved, FVC systems can provide more accurate results when a reference population of the same dialectal variety is used, instead of a generic one.

However, it should be noted that further studies would be needed to consolidate these results. In fact, the present study, which is only in the preliminary phase, has only two areas in consideration, while it would be appropriate for the study to be repeated for other areas also outside Puglia. Furthermore, more robust results would be obtained by sampling each area more numerously, doubling the number of participants per area, and including an equal number of speakers of the other gender.

5. *Conclusions and future directions*

In the presented study, we have investigated to what extent the use of a dialectal population, compared with a standard Italian speech population, can influence the performance of state-of-the-art Likelihood Ratio-based FASR systems, also assuming that the accuracy of the comparison can improve when dialectal variation is taken under consideration in the reference population.

Although the results obtained in terms of improvement of the EER values are not yet generalizable, due to the small number of dialectal varieties in question, and the limited number of speakers sampled in each area, they clearly support at least the validity of the research hypotheses and establish an excellent starting point for future experiments.

Indeed, we plan to improve the above-mentioned limitations, by increasing the number of investigated areas, the number of speakers for each area, and including female speakers. Also, we plan to test different paradigms for FASR, like recent developments in the field of Deep Learning.

From a different point of view, the achieved results invite to carefully reflect on the importance and need to collect databases of dialectal language, for research purposes, not only at the level of academic institutions, but also and above all in collaboration with the relevant institutions and law enforcement agencies that daily carry out

investigations on the basis of intercepted conversations. By sharing under appropriate non-disclosure agreements, the certainly huge amount of phonic materials collected over years of investigations, it could give a significant boost to research in the field of the study of dialectal variations and their impact in the field of forensic phonetic.

Bibliography

Benyassine, A., Schlomot, E. & Su, H. (1997). ITU-T recommendation G729 Annex B: a silence compression scheme for use with G729 optimized for v.70 digital simultaneous voice and data applications. In *IEEE Commun. Mag*., 35, 64–73.

Boersma, P. & Weenink D. (2020). *Praat: doing phonetics by computer*" [Computer program]. Version 6.2.05. http://www.praat.org/

Chen, J.C., Miller, D.A. (2020). Pitch-Synchronous Analysis of Human Voice. In *Journal of Voice*, 34(4), 494-502.

Davis, S.B. & Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition. In *IEEE Trans. on Acoustics, Speech and Signal Processing*, 28(4), 357–366.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. & Ouellet, P. (2011). Front-end factor analysis for speaker verification. In *IEEE TASLP*, 19, 788-798.

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J., & Niemi, T. (2016). *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition including Guidance on the Conduct of Proficiency Testing and Collaborative Exercises*. https://enfsi.eu/

Garcia-Romero, G. & Espy-Wilson, C.Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. In *Proc. Interspeech*, 249–252.

Gong, W.G., Yang, L.P. & Chen, D. (2008). Pitch synchronous based feature extraction for noise-robust speaker verification. In *Proc. Image and Signal Processing*, 5, 295–298.

Hansen, J.H., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. In *IEEE Signal processing magazine*, 32(6), 74-99.

Hillenbrand, J.M., & Clark, M.J. (2009). The role of f0 and formant frequencies in distinguishing the voices of men and women. In *Attention, Perception, & Psychophysics*, 71(5), 1150–1166.

Huijbregts, M., Wooters, C., Ordelman, R. (2007). Filtering the unknown: speech activity detection in heterogeneous video collections. In *Proc. Interspeech*, 2925–2928.

Jessen, M. (2008). Forensic Phonetics. In *Language and Linguistics Compass*, 2/4, 671-711.

Kinnunen, T. & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. In *Speech communication*, 52(1), 12-40.

Kinnunen, T. & Rajan, P. (2013). A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data. In *Proc. ICASSP*, 7229–7233.

Kenny, P. (2010). Bayesian speaker verification with heavy tailed priors. In *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 69(7), 349-356.

Mancarella, G.B. (1998). *Salento. Monografia regionale della "Carta dei Dialetti Italiani.* Lecce: Edizioni del Grifo.

Morrison, G.S. (2009). Forensic voice comparison and the paradigm shift. In *Science and Justice*, 49, 298-308.

Nakasone, H., Mimikopoulos, M., Beck, S. & Mathur, S. (2004). Pitch synchronized speech processing (PSSP) for speaker recognition. In *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2004)*, 251–256.

Ng, R., Nicolao, M., Saz, O., Hasan, M., Chettri, B., Doulaty, M., Lee, T. & Hain, T. (2016). The Sheffield language recognition system in NIST LRE 2015. In *Proc. Odyssey 2016: The Speaker and Language Recognition Workshop*, 181-187.

Plchot, O., Matejka, P., Fer, R., Glembek, O., Novotn, O., Pesan, J., Vesel, K., Ondel, L., Karafiat, M., Grezl, F., Kesiraju, S., Burget, L., Brummer, N., Swart, A., Cumani, S., Mallidi, S.H. & Li, R. (2016). BAT system description for NIST LRE 2015. In: *Proc. Odyssey 2016: The Speaker and Language Recognition Workshop*, 166-173.

Ramirez, J., Segura, J., Benitez, C., Torre, A.D.L. & Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. In *Speech Communication*, 42, 3-4.

Reynolds, D.A. (1997). Comparison of background normalization methods for text-independent speaker verification. In *Proc. Eurospeech*, 963–966.

Ribezzo, F. (1912). *Il dialetto apulo-salentino di Francavilla Fontana*. Bologna: Forni.

Robertson B., & Vignaux, G.A., (1995). *Interpreting evidence*. London: Wiley.

Romito, L. & Galatà, V. (2008). Speaker Recognition in Italy: evaluation of methods used in forensic cases, In *Actas del IV Congreso de Fonética Experimental (A. Pamies & E. Melguizo, editors*. Granada: Método Ediciones, 229-240.

Romito L, Bove T, Delfino S, Jona Lasinio G, Rossi C (2009). *Specifiche linguistiche del database utilizzato per lo Speaker Recognition in S.M.A.R.T.*. In Romito, L., Rosita, L. & Galatà, V. (Eds.). *La Fonetica Sperimentale: metodo e applicazioni*, Milano: Officinaventuno, pp. 632-640.

Rose, P. (2002). *Forensic Speaker Identification*. New York: Taylor and Francis.

Rose. P. (2005). Forensic Speaker Recognition at the beginning of the twenty-first century – an overview and a demonstration. In *Australian Journal of Forensic Science*, 37, 49-72.

Rose, P. (2006). Technical Forensic Speaker Recognition: Evaluation, types and testing of evidence, In *Computer Speech and Language*, 20, 159–191.

Shin, J., Chang, J.H. & Kim, N. (2010). Voice activity detection based on statistical models and machine learning approaches. In *Computational Speech Language,* 24(3), 515–530.

Simpson, A.P. (2009). Phonetic differences between male and female speech. In *Language and Linguistics Compass*, 3(2), 621–640.

Sohn, J., Kim, N. & Sung, W. (1999). A statistical model-based voice activity detection. In *IEEE Signal Process. Lett*., 6, 1–3.

Thomas, S., Saon, G., Egbroeck, M. & Narayanan, S., (2015). Improvements to the IBM speech activity detection system for the DARPA RATS program. In *Proc. ICASSP*, 4500-4504.

Tucker, R., (1992). Voice activity detection using a periodicity measure. In *IEEE Proc. (Commun. Speech Vis.),* 139(4), 377–380.

Wu, J. & Zhang, X. (2011). Maximum margin clustering based statistical VAD with multiple observation compound feature. In *IEEE Signal Process. Lett*. 18(5), 283–286.

Zhang, X. & Wu, J. (2013). Deep belief networks based voice activity detection. In *IEEE Trans. Audio Speech Lang. Processing,* 21(4), 697-710.

Zilca, R., Kingsbury, B., Navr'atil, J. & Ramaswamy, G. (2006). Pseudo pitch synchronous analysis of speech with applications to speaker recognition. In *IEEE Trans. Audio, Speech and Language Processing*, 14(2), 467–478.