# SUPPORT VECTOR MACHINE FUSION OF IDIOLECTAL AND ACOUSTIC SPEAKER INFORMATION IN SPANISH CONVERSATIONAL SPEECH

*D. García-Romero, J. Fiérrez-Aguilar, J. González-Rodríguez and J. Ortega-García*

Speech and Signal Processing Group (ATVS) – Universidad Politécnica de Madrid (UPM)
email: dgromero@atvs.diac.upm.es ,{jfierrez, jgonzalez, jortega}@diac.upm.es

**ABSTRACT**

This paper proposes a Support Vector Machine (SVM) based combining scheme that incorporates ideolectal and acoustic characteristics for speaker recognition. Two statistical model paradigms, namely GMM for acoustic modeling and Bigrams for language modeling, provide multilevel speaker information that affords a better classification performance when SVM-based fusion is accomplished. This combining approach is useful for all speaker recognition tasks where a considerable amount of data is available. Motivated by the absence of Spanish databases that made feasible our research experiments, more than nine hours of Spanish conversational speech was collected and manually transcribed from broadcasted radio talk shows.

## 1. INTRODUCTION

Current speaker recognition systems rely almost exclusively on short-time acoustic information. MAP-adapted Gaussian Mixtures Models [1] represent the state-of-the-art technique in text independent speaker recognition and achieve a very good performance but is susceptible to acoustic corruptions such as channel variability and noise. However, there are other levels of information in speech which convey speaker identity and are not under the influence of those corrupting factors. These higher-level information sources are the subject of intensive research efforts at present, as shown with the inclusion of the "extended data" speaker detection task in the NIST [2] yearly evaluation and the recently held SuperSID workshop [3] whose main focus was to analyze, characterize, extract and apply high-level information to speaker recognition tasks.

Among other long term speech patterns, idiolectal differences between speakers have proven to be one of the most promising speaker information sources [3,4]. To exploit these potential benefits, a considerable amount of speech is required in order to perform accurate Bigram models [4]. Therefore, only a subset of speaker recognition tasks presents a propitious scenario for multilevel speaker information fusion. Good examples are some forensic areas, specifically those where fully-automated remote acquisition can be easily accomplished. Particularly fitting into that situation we find the Spanish scenario, where once a judge has emitted a legal authorization for an investigation, substantial amount of data can be automatically collected by means of wire tapping the telephone line. Also, human transcriptions of the speech information are required by the judge to follow the investigation.

Hence, multilevel speaker information fusion affords a promising departure point in Spanish forensic scenarios. No prior work has been done on Spanish conversational speech so this provides an important motivation for the present work.

## 2. BASELINE SYSTEMS

### 2.1. Acoustic system

Acoustic scores have been calculated with our GMM system used in the 2002 NIST evaluation [5]. A gender-independent 256 mixtures UBM was trained with approximately 2 hours (gender balanced) of microphonic speech extracted from Ahumada corpus [6]. Target speaker models have been trained via MAP adaptation of the UBM. Score normalization, namely TNORM and ZNORM [1] was performed to develop a common likelihood scale.

### 2.2. Idiolect system

A language model for each target speaker has been trained using word bigrams [4] and a conventional likelihood ratio test was used to calculate test segments scores. The scoring formula is expressed as follows:

$$Score = \frac{\sum_k \{N_{tokens}(k) \cdot \log[\Lambda_{TS}(k)/\Lambda_{BG}(k)]\}}{\sum_k \{N_{tokens}(k)\}} \quad (1)$$

where $N_{tokens}(k)$ is the number of occurrences of bigram type $k$ in the test segment and $\Lambda_{TS}(k)$, $\Lambda_{BG}(k)$ are the

estimated likelihood of bigram k for the target speaker model and a background speaker model respectively. Nine hours of speech were used to train the background model.

## 3. FUSION STRATEGIES

Kittler *et al.* considered in [7] the task of combining classifiers in a probabilistic bayesian framework. Several ways to combine individual scores were obtained (sum, product, max, min, …), based on the Bayes theorem and certain hypothesis, from which the *Sum Rule* (adding individual scores normalized to range {0,1}) proved to be the best in the experimental comparison (in a multilevel biometric fusion problem). This perspective will be referred as to *rule-based fusion.*

Multilevel fusion can also be treated as a *pattern classification* problem if the scores given by individual classifiers are considered as input patterns to be labeled as accepted/rejected (for the verification task). Under this point of view, any learning machine approach can be applied as a fusion strategy. In a recent contribution following this approach [8], the paradigm of *Support Vector Machines* (SVMs) has been proved to outperform other *learning-based* (including Neural Networks) and rule-based fusion strategies. This objective result motivates the SVM-based idiolectal-acoustical combination proposed in this paper.

### 3.1. SVM-based fusion

The principle of SVM relies on a linear separation in a high dimension feature space where the data have been previously mapped, in order to take into account the eventual non-linearities of the problem.

Formally, the training set $X = (\mathbf{x}_i)_{i=1}^{l} \subset \mathbf{R}^R$, where $l$ is the number of training vectors, $\mathbf{R}$ stands for the real line and $R$ is the number of modalities, is labeled with two-class targets $(y_i)_{i=1}^{l}$, where:

$y_i \in \{-1,1\} = \{"Impostors","Clients"\}$. $\Phi : \mathbf{R}^R \to F$ maps the data into a feature space $F$. Vapnik [9] has proved that maximizing the minimum distance in space $F$ between $\Phi(X)$ and the separating hyperplane $H(\mathbf{w},b) = \{\mathbf{f} \in F \mid < \mathbf{w},\mathbf{f} >_F + b = 0\}$, (where $<\cdot,\cdot>_F$ denotes inner product in space $F$), is a good means of reducing the generalization risk. Vapnik also proved [9] that the optimal hyperplane can be obtained solving the convex quadratic programming (QP) problem:

$$\text{Minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l} x_i \qquad (2)$$

$$\text{with} \quad y_i(< \mathbf{w}, \Phi(\mathbf{x}_i) >_F + b) \geq 1 - x_i \quad i = 1,..,l$$

where constant $C$ and slack variables $x_i$ are introduced to take into account the eventual non-separability of $\Phi(X)$ into $F$. Applying the Karush-Kuhn-Tucker conditions to the problem in (2), the following sparse expression is obtained for the optimal hyperplane $H(\mathbf{w}^*,b^*)$:

$$\mathbf{w}^* = \sum_{i \in SV} a_i y_i \Phi(\mathbf{x}_i) \qquad (3)$$

where $SV = \{i \mid a_i > 0\}$ is the set of support vectors. Taking into account that the decision function $D$ that classifies a test pattern $\mathbf{x}_T$ is:

$$D(\mathbf{x}_T) = sign\{< \mathbf{w}^*, \Phi(\mathbf{x}_T) >_F + b^*\} \qquad (4)$$

defining $K(\mathbf{x}_i,\mathbf{x}_j) =< \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) >_F$ as the kernel function and using (3) leads to:

$$D(\mathbf{x}_T) = sign\left\{\sum_{i \in SV} a_i y_i K(\mathbf{x}_i, \mathbf{x}_T) + b^*\right\} \qquad (5)$$

Problem (2) is generally solved for $(a_i)_{i=1}^{l}$ and $b^*$ in its dual form with QP solvers which, together with decision function (5), avoids manipulating directly the elements of $F$ and starting the design of the SVM for classification directly from the kernel function. Typical choices for $K$ are:

Linear Kernel: $K(\mathbf{x}_i,\mathbf{x}_j) =< \mathbf{x}_i, \mathbf{x}_j >_{\mathbf{R}^R}$

Polynomial Kernel: $K(\mathbf{x}_i,\mathbf{x}_j) = (< \mathbf{x}_i, \mathbf{x}_j >_{\mathbf{R}^R} + 1)^p$

Gaussian kernel: $K(\mathbf{x}_i,\mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2s^2)$

In [8], the fusion strategy relies on the computation of the decision function $D$. A modification in order to obtain not a final classifier decision, but a combined multilevel score based on the proximity of the test pattern to the separating surface is proposed here. The combined score $s_T \in \mathbf{R}$ of the multilevel pattern $\mathbf{x}_T \in \mathbf{R}^R$ will be calculated as:

$$s_T = \sum_{i \in SV} a_i y_i K(\mathbf{x}_i, \mathbf{x}_T) + b^* \qquad (6)$$

Following this approach, the verification threshold parameter can be adjusted to reach different working points. This modification also permits to compare competing multilevel fusion strategies in terms of DET plots.

## 4. EXPERIMENTS

### 4.1. Database

There are not publicly available Spanish conversational speech databases that allow exploring the benefits of fusing acoustic and idiolect speaker information. Therefore, more than nine hours of Spanish conversational speech were collected and manually transcribed from broadcasted radio talk shows, during a period of time of

two months, ensuring enough variability in both acoustic and idiolect characteristics. Nine speakers compound the database (7 male, 2 female) and each speaker data was divided into training and testing subset. For the idiolect experiments the training subset of associated transcription files were divided into three text segments of an equivalent duration of 30, 15, and 10 minutes of speech. The testing subset was dived into 5 segments of 3 equivalent minutes of speech. Regarding to the acoustic data, a NIST [2] oriented scheme was followed. Two minutes segments were created for training each speaker model and 20 seconds of speech were extracted from each of the 5 test segments.

Bigram models were trained for each of the nine speakers with the 30 minutes-equivalent text training segment of the database. GMM were trained using the 2 minutes training segments corresponding to each speaker. To assess both systems performance the same testing strategy was conducted in each of them. Scores for each target were calculated with 5 target trials and 40 non-target trials.

### 4.2. Asymptotic performance

A statistically-motivated experimental procedure has been applied to the resulting DET curves in order to smooth them. Due to the inherent difficulty to collect a larger database in Spanish some scarcity data side effects seem to appear in the DET plots, adding an arduous effort to accomplish a visual comparison. The smoothing procedure works as follows: two GMM with 4 components each are estimated respectively from client and impostor score histograms using the EM algorithm [10]. Then 10,000 points from the resulting distributions are generated and used as input for the performance testing DET plots. Figure 1 shows an example of a DET plot without smoothing (top) and with the smoothing procedure (bottom).

### 4.3. Rule-Based Fusion

For rule-based fusion strategies, all test scores will be used for testing the verification performance.

Figure 1 shows the results from both systems and also the resulting scores of two rule-based combination strategies, namely sum and product (where idiolectal and acoustic scores were previously mapped linearly into {0,1} range). As may be observed, the acoustic system performance is at least twice better than the idiolect system in all the operating points. Therefore, ideolectal speaker information may be considered as a potential boosting factor for acoustic speaker recognition performance but not as an isolated system.

The rule-based combination strategies accomplished were not capable to exploit the potential synergy between speaker language characteristics and acoustic traits.

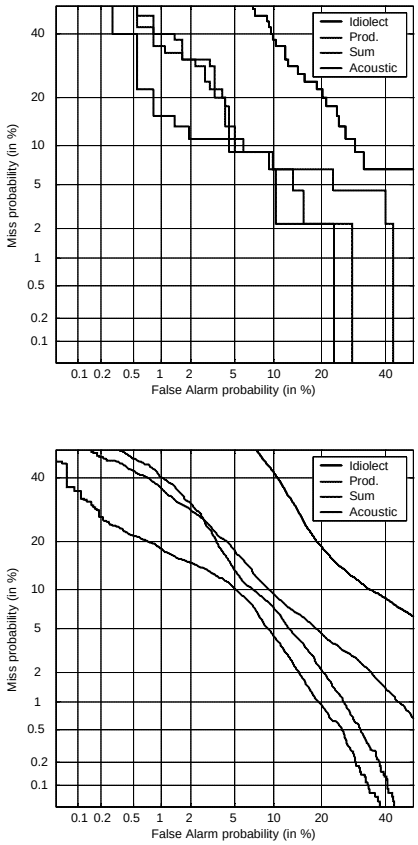


*Figure* 1.- Baseline systems performance without smoothing (top) and with smoothing procedure (bottom).

### 4.4. SVM-Based Fusion

For the SVM-based fusion experiments, the leave-one-out method [7] will be used to maximize the size of the training and testing data of the learning machine, while maintaining their independence. Idiolectal and acoustic scores from one user will be combined with a SVM trained on other users, generating 5 client and 40 impostor combined scores. This strategy is carried out on the remaining 8 subjects, yielding 5×9 = 45 client and 40×9 = 360 impostor combined test scores. A Gaussian kernel (Radial Basis Function) has been used, and $s^2$ parameter has been varied in order to find a good generalization point. Figure 3 shows the detection performance of the SVM-based fusion approach with three different kernel parameters.

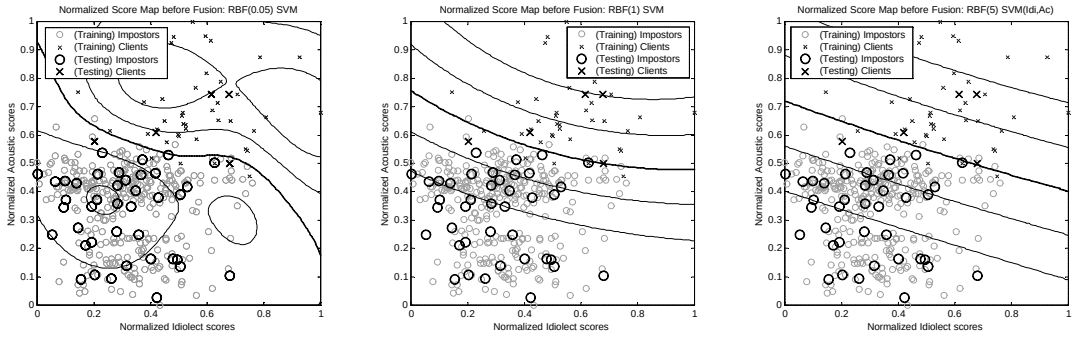


Figure 2.-User and impostor acoustic-idiolect normalized score map used for training and testing one SVM with the trained separating hyperplane (bold) and curves of equal combined score (dashed).

In order to visualize the discrimination capability of the RBF SVM fusion approach, client and impostor maps of idiolectal and acoustic scores before the fusion are plotted in Figure 2. Decision boundaries and curves of equal combined score for one user of the leave-one-out procedure, whose client and impostor scores have been enlarged, have also been included.

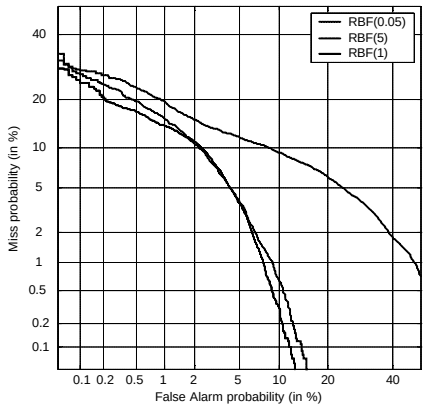


*Figure* 3.- RBF SVM fusion results varying the $s^2$ parameter

As a result, the performance comparison of idiolectal, acoustic and SVM-based combined verification systems is plotted in Figure 4.

## 5. CONCLUSIONS

SVM fusion techniques provide an excellent framework for combining ideolectal and acoustic speaker information in Spanish conversational speech. Compared to the best performing baseline system (acoustic), an improvement of at least 40% in all the operating points was achieved by the SVM fused system. Ruled-based fusion techniques were not able to exploit the potential benefits of combining two different level information sources.

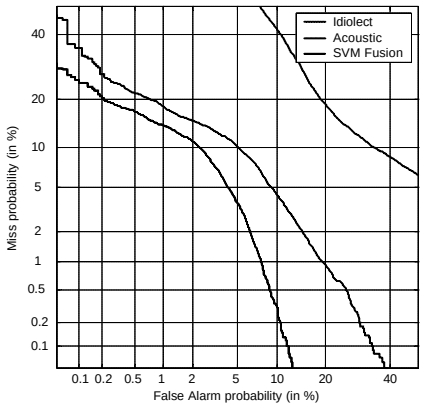


*Figure* 4.- Performance improvement of the SVM fusion

## 6. REFERENCES


[1] Douglas A. Reynolds et al., "Speaker Verification using Adapted Gaussian Mixture Models", Digital Signal Processing, vol. 10, pp. 19-41 (2000).

[2] The extended data task definition is included in NIST's 2001 evaluation at www.nist.gov/speech/tests/spk/2001

[3] SuperSID: Exploiting High-level Information for High-performance Speaker Recognition. Maryland, July 2002.

[4] Doddington," Speaker Recognition based on Ideolectal Differences between speakers."Proc. Eurospeech'01.

[5] D. García-Romero et al., "ATVS-UPM Results and Presentation at NIST'2002 Speaker Recognition Evaluation", Vienna, VA, 2002.

[6] J. Ortega-Garcia, J. Gonzalez-Rodriguez, and V. Marrero-Aguiar, "AHUMADA: a Large Speech Corpus in Spanish for Speaker Characterization and Identification", Speech Communication, vol. 31, pp. 255-264, June 2000.

[7] J. Kittler, et al. "On Combining Classifiers", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, No. 3, pp. 226-239, 1998.

[8] B. Gutschoven, P. Verlinde, "Multi-modal identity verification using Support Vector Machines (SVM)", *Proc. of the 3rd Int. Conf. on Information Fusion*, 2000.

[9] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, 1999.

[10] S. Theodoridis, *Pattern recognition*, Academic Press, 1999.