FUSING LANGUAGE IDENTIFICATION SYSTEMS USING PERFORMANCE CONFIDENCE INDEXES

Jorge Gutiérrez, Jean-Luc Rouas and Régine André-Obrecht

IRIT – Université Paul Sabatier
118, Route de Narbonne
F-31062 Toulouse Cedex 4

ABSTRACT

In the field of automatic language identification, several mostly-empirical arithmetic fusion operations are currently done to make a consensus decision from a set of acoustics-based identification systems whose estimated performance is taken into account by means of weighting techniques. This paper presents how to apply the Discriminant Factor Analysis method to formally compute and use weighting performance confidence indexes at the expert and class levels. Moreover, the observation level is also explored. These confidence indexes allow us not only to qualify the identification decision with additional insight by means of a certainty degree but also to provide acoustics-based identification systems with powerful uncertainty-based inference techniques where systems' *a priori* performance knowledge is a key heuristic-like element to improve language identification capabilities.

1. INTRODUCTION

The aim of Automatic Language Identification (ALI) systems consists on identifying as soon as possible the language in which an utterance has been pronounced. Several approaches have been studied to take advantage of language-discriminant features. The most classical ones issue forth:

- Acoustic information: vocalic and consonant phones and their frequency of occurrences differ among languages [5];
- Phonotactic information: specific sequences of phonetic units appear with different occurrences for each language[8];
- Prosody: the sound duration, the fundamental frequency, the intensity variation or the rhythm are language discriminant [6].

To take these various linguistic features into account, "primary ALI systems" are built: the acoustic system where the acoustic information of each language is modelled by Gaussian Mixture Models (GMM) or Hidden Markov Models (HMM) [8]; the phonotactic system where bi-gram or tri-gram model traduce the language phonotactic rules; the prosodic system which is based on statistical moments computed on the rhythm and the fundamental frequency, and so on. To take several sources of information into account, an ALI system is composed of several primary ALI systems which are now called **experts**, but the problem of merging decisions or decision scores appears. Empirical techniques have been implemented in order to fuse identification decisions coming out from several experts. Very often, good performance is obtained by weighting properly the decision scores [3]. The difficulty is to define the weights which represent **the expert performance confidence**, and to determine for each language decision its confidence which is called **the class performance confidence**.

We propose an original way to precise them automatically and we study alternative fusion methods based on them.

One way of formally computing performance confidence indexes consists of extracting language-discriminant information by processing a development speech corpus and using the Discriminant Factor Analysis (DFA) method in the decision score field. The DFA projection is used to obtain the confusion matrix and to provide expert and class performance confidence indexes.

Therefore, with these confidence indexes, three kinds of fusion techniques may be studied: empirical fusion, statistical fusion and uncertainty-based fusing techniques like the one provided by the Theory of Evidence.

In this paper, we present in section 2 how finding the expert and class performance confidence indexes. In section 3, we describe the three fusion techniques. Experiments are explained in section 4.

2. PERFORMANCE CONFIDENCE INDEXES

ALI primary systems or experts accept a speech utterance called the observation, as input, and provide the class (or language) decision as output, after computing language-scores (in many cases, a statistical model is used and the language-score is the language-likelihood); so they handle a vector of language-likelihood values. Given M languages to identify, L_i, $1 \le i \le M$, and N experts, s_j, $1 \le j \le N$, we obtain for each observation, N vectors of M values, each one ranging

from 0 to 1; this global observation is represented as a score matrix $\delta = [\, d_i^j \,]_{1\le i\le M,\ 1\le j\le N}$ (Table I).

To explain our future fusion techniques, it is necessary to define not only the expert performance confidence indexes and the class performance confidence indexes, but also the observation performance confidence indexes which represent for each expert the confidence of the decision taken for the observation. The two first families of indexes are independent of the current observation.

So to define expert and class confidence, we need the score matrices of the development set of acoustic segments for each expert s_j; we apply the DFA and build the confusion matrix (Figure 1); the class confidence indexes ($\beta_{ij}, 1\le i\le M$) can directly be mapped from the diagonal values of the confusion table while the expert confidence index must be computed as an averaged value:

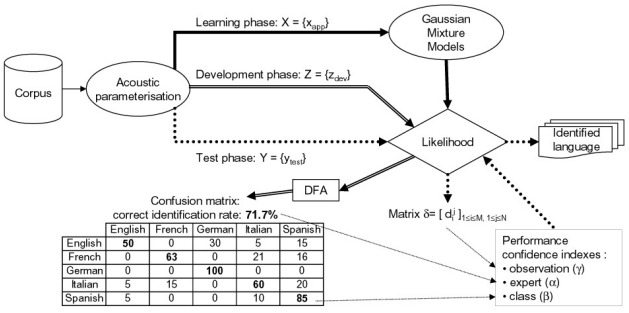
$$\alpha_j = (1/M) \sum_{i\in[1,M]} \beta_{ij}.$$

Many solutions may be proposed to define the observation confidence indexes. We retain two formulas to be applied on test-set matrices: given an identification system s_j and $\hat{\imath}$ the decision class, $d_{\hat{\imath}j} \ge \max_k (d_{kj})$, $k\in[1,M]$,

- $g_j = d_{\hat{\imath}j} - \max_{k\neq\hat{\imath}} d_{kj}$;
- $g_j = d_{\hat{\imath}j} - \frac{1}{M-1}\sum_{k\neq\hat{\imath}} d_{kj}$

L\S	s1	s2	…	sj	…	sN
L1	d11	d12	…	d1j	…	d1N
L2	d21	d22	…	d2j	…	d2N
…	…	…	…	…	…	…
Li	di1	di2	…	dij	…	diN
…	…	…	…	…	…	…
LM	dM1	dM2	…	dMj	…	dMN

Table I. Matrix $\delta = [\, d_i^j \,]_{1\le i\le M,\ 1\le j\le N}$, of scores obtained for each observation.



	English	French	German	Italian	Spanish
English	**50**	0	30	5	15
French	0	**63**	0	21	16
German	0	0	**100**	0	0
Italian	5	15	0	**60**	20
Spanish	5	0	0	10	**85**

Figure 1. Obtaining confidence indexes

3. FUSION TECHNIQUES

3.1. Empirical fusion

Summing and multiplying score values are the most current operations to empirically fuse decision scores. Sometimes estimated performance indexes are taken into account to weight each system's decision score. The concept of weighting performance indexes matches ours regarding the expert confidence index α described above. Thus, a language is considered as the identified one if it corresponds to the greatest value computed with the following weighted rules:

- Sum $\quad L^* = \arg\max_{i\in[1,M]} [\, \sum_{j\in[1,N]} \alpha_j d_{ij} \,]$
- Product $\quad L^* = \arg\max_{i\in[1,M]} [\, \prod_{j\in[1,N]} (d_{ij})^{\alpha_j} \,]$

3.2. Statistical Fusion

The GMM fusion

The occurrence of scores matrices can statistically be modelled by Gaussian Mixture Models (GMM). One model is learned for each language L_i from the matrices issued from the development-set acoustic segments. We initialize by Vector Quantification and we apply the iterative Expectation-Maximization algorithm to optimise Gaussian components.

Let δ be the score matrix corresponding to the acoustic segment y. The probability that the segment y belongs to language L_i is given by:

$$P(\delta \mid L_i) = \sum_{n\in[1,Q_i]} \omega_n N(\delta, \mu_n, \sigma_n)$$

where n is the Gaussian component number and Q_i the total number of components for the language L_i. The language the most likely for matrix δ is the one corresponding to the maximum likelihood:

$$L^* = \arg\max_i [\, P(\delta \mid L_i) \,].$$

The DFA fusion

As the dimension $N \times M$ of the score matrices space is relatively large, we try to reduce their dimension and search a better representation space. We apply the DFA on the set of score matrices obtained from the development set of acoustic segments. We use the $M-1$ factorial axis corresponding to the $M-1$ eigen-values (different to zero) and we project the score matrices on this subspace. Besides, we take advantage of this projection step to implement the DFA-based classifier by applying on each score matrix the following identification decision rule:

if $\mathrm{Dist}(\delta \mid L_i)$ represents the Euclidean distance between the projected matrix δ and the projected centre of gravity of language L_i,

$$L^* = \arg\min_i [\mathrm{Dist}(\delta \mid L_i)].$$

3.3. Theory of Evidence

Let $L = \{L_1, L_2, \ldots, L_i, \ldots, L_M\}$ denote the finite set of possible languages to be identified; this set L is composed of M exhaustive and exclusive hypotheses of the decision process and we assume every union of hypotheses may be a response of the decision process. The set 2^L of all possible events A based on L is the set of all subsets of L: $2^L = \{A \mid A \subseteq L\}$; $|2^L| = 2^M$, that is to say:

$2^L = \{\emptyset, \{L_1\}, \{L_2\} \ldots, \{L_M\} \ldots, \{L_1, L_2\} \ldots, \{L_{M-1}, L_M\} \ldots, L\}$

For each unknown utterance, and for each expert s^j, we define a basic belief mass function m_L^{Sj}, which explains how the decision L* belongs to the subset A of L: $m_L^{Sj} : 2^L \rightarrow [0, 1]$

with the constraints: $\sum_{A \subseteq L} m_L^{Sj}(A) = 1$ and $m_L^{Sj}(\emptyset) = 0$.

This function is built from the score matrix of the utterance by assigning the score values to each singleton: $m_L^{Sj}(\{L_1\}) = d_{1j}$, ..., $m_L^{Sj}(\{L_i\}) = d_{ij}$,..., $m_L^{Sj}(\{L_M\}) = d_{Mj}$, an uncertainty value to A=L corresponding to the one's complement to the observation confidence index: $m_L^{Sj}(L) = 1 - \gamma_j$, the null value to the rest of the events in 2^L, and we normalise all the belief mass values to verify the constraints above.

Let $S = \{ s_1, s_2, \ldots, s_j, \ldots, s_N \}$ the finite set of identification systems; we may combine these experts on a cascade-like pair basis by applying Dempster's orthogonal combination rule:

$$m_L^{Sl,k}(A) = K_L \cdot \sum_{B \cap C = A} m_L^{Sk}(B) \cdot m_L^{Sl}(C)$$

where $K_L = 1/[1 - \sum_{B \cap C = \emptyset} m_L^{Sk}(B) \cdot m_L^{Sl}(C)]$ is a normalisation factor. We obtain so a global belief mass function, noted $m_L^S(A)$, for each event A.

We weight belief mass functions of the events (B, C, etc.) by applying the class confidence indexes β before doing the orthogonal operation:

$m_L^{\beta Sj}(C) = \beta^{Sj} \cdot m_L^{Sj}(C)$, $\forall C \neq L$

$m_L^{\beta Sj}(L) = (1 - \beta^{Sj}) + \beta^{Sj} \cdot m_L^{Sj}(L)$

We use the pignistic transformation to derive a probability on L, from the belief mass values [7]:

$$BetP(Li) = \sum_{Li \subseteq A} m_L^S(A) / |A|.$$

Thus, the decision process can be carried out by maximum pignistic probability [2]:

$$L^* = \arg\max_i [BetP_{Li}(A)].$$

5. EXPERIMENTATION

5.1. Fusion System's Architecture

Acoustic data is provided by the MULTEXT corpus [1] which comprises a set of 20 kHz 16-bit sampled records in 5 languages: English, French, German, Italian and Spanish. Data consist of read passages from the EUROM1 corpus pronounced by 50 different speakers (5 males and 5 females per language). The mean duration of each passage is 20.8 seconds.

The corpus is split into three partitions for each language: the learning set, the development set and the test set (2 speakers: 1 male and 1 female who do not belong to the other sets).

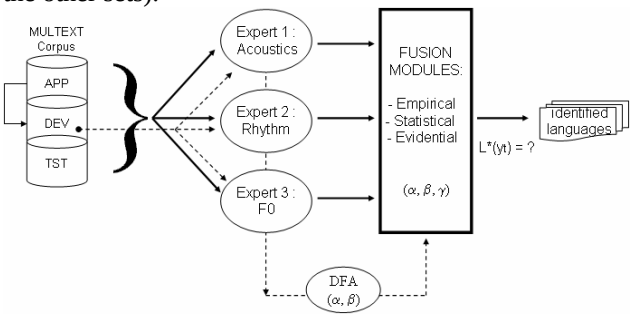


Figure 2. Fusion system architecture.

The ALI system is based on three ALI subsystems and a fusion module (see Figure 2):

- Acoustics Expert [5]: After an automatic vowel detection, each vocalic segment is represented with a set of 8 Mel-Frequency Cepstral Coefficients and 8 delta-MFCC, augmented with the Energy and delta Energy of the segment. This parameter vector is extended with the duration of the underlying segment providing a 19-coefficient vector. A cepstral subtraction performs both blind removal of the channel effect and speaker normalization. For each recording sentence, the average MFCC vector is computed and subtracted from each coefficient.
- Rhythm Expert [6]: Syllable may be a first-rate candidate for rhythm modelling. Unfortunately, segmenting speech in syllables is typically a language-specific mechanism and thus no language independent algorithm can be derived. For this reason, we have introduced the notion of pseudo-syllables derived from the most frequent syllable structure in the world, namely the CV structure. Using the vowel-no vowel segmentation, speech signal is parsed in patterns matching the structure: $.C^nV$. Each pseudo-syllable is then characterized by its consonants global duration, its vocalic duration, its complexity (the number of consonant segments), and its energy.
- Fundamental Frequency Expert [6]: The fundamental frequency outlines are used to compute statistics within the same pseudo-syllable frontiers (previously defined) in order to model intonation on each pseudo-syllable. The parameters used to characterize each pseudo syllable intonation are a measurement of the accent location (maximum f0 location regarding to vocalic onset) and the normalized fundamental frequency bandwidth on each syllable.

For each expert, we applied the same learning-testing procedure: for each language, a Gaussian Mixture Model (GMM) is trained using EM algorithm with LBG initialisation [4]. The optimal number of components of

the mixture is obtained from experiments on the learning part of the corpus. During the test, the decision lays on a Maximum Likelihood procedure.

The performance of these three experts is given in Table II, and they are considered as reference performances (three runs were launched where the rate difference showed up at the experts' level because of internal parameter values, e.g. the number of Gaussians or the aleatory initialisation of GMMs). We may observe the relatively bad performance of the fundamental frequency-based expert.

The three techniques of fusion (empirical, statistical and evidential) are experimented to merge the decision scores, outputs of the three experts, as explained in the previous section. For the GMM fusion, we use 20 Gaussian components. The development set is used to fix the class and expert performance confidence indexes.

5.2. Tests and Results

Most important results in fusing the three experts are shown in Table II:

- The empirical fusion delivered higher identification rates than those of any subsystem: up to 85%. Addition or weighted addition give similar performance.
- The statistical fusion delivered acceptable identification rates (keeping in mind no weighting is done). Note that with the GMM fusion, a better identification rate, 84%, was obtained than with the DFA, 83%, but the decision space was quite different (from 15 to 4 axes).
- The best identification rate, 90%, was reached for the fusion system using the Theory of Evidence.

In addition, we tested the influence of 2-expert fusion; as a result, we observed that some combinations delivered better identification rates than the 3-expert combination when fusing empirically. This was not the case for the modelled fusion strategies.

6. CONCLUSION

Formal methods are applied, to do both compute performance confidence indexes and fuse decision information coming out from different language identification systems; it appears as a strong alternative to empirical techniques. Uncertainty-based fusion methods allows us to model properly the language identification problem so that heuristic-like inference techniques can take advantage of weighting values in a more refined way: not only at the expert level but also at the class and observation levels. That can partially explain why the technique based on the Theory of Evidence has delivered better identification rates compared to empirical techniques. Nevertheless, the statistical approach does not allow any direct likelihood-value weighting, then the confidence indexes could only be used so far to qualify the identification decision with a certainty degree; a future work would be to find out, if possible, how to weight such values though. Applying the uncertainty-based techniques based on the Possibility Theory/Fuzzy Logic is also a future work.

Language Identification Systems –Total Success Rate (%) –				
		1st run	2nd run	3rd run
Reference Systems	Expert 1 : acoustics	79	76	78
	Expert 2 : rhythm	71	73	70
	Expert 3 : fundamental frequency	35	35	35
Empirical Fusion Techniques	Addition	83	82	80
	Product	67	68	63
	Weighted Addition (α)	84	85	83
	Weighted Product (α)	68	69	65
Modelled Fusion Methods	Discriminant Factor Analysis	77	83	83
	Gaussian Mixture Model	81	81	84
	Theory of Evidence (α,β,γ)	90	89	87

Table II. Comparison of Fusion Strategies.

7. REFERENCES

[1] Campione E. and Véronis J., "A multilingual prosodic database", In Proceedings of ICSLP'98, Sidney, 1998.

[2] Denoeux T., "Pattern Recognition using belief function", SFC'2002, Toulouse, France, 2002.

[3] Hazen T.J. and Zue V.W., "Segmented-based Automatic Language Identification", Journal of the Acoustical Society of America, 4(101), 1997.

[4] Linde Y., Buzo A. and Gray R.M., "An algorithm for vector quantizer design", IEEE Transaction on Communications, vol. 28, no. 1, pp. 84-95, 1980.

[5] Pellegrino F. and André-Obrecht R., "Automatic language identification: an alternative approach to phonetic modelling" In: *Signal Processing*, Elsevier Science North Holland, V. 80, p. 1231-1244, 2000.

[6] Rouas J.L., Farinas J. and Pellegrino F., "Automatic Modelling of Rhythm and Intonation for Language Identification", In: 15th International Congress of Phonetic Sciences (15th ICPhS), 2003. pp. 567-570, Barcelona, Spain, 2003.

[7] Smets P., "Constructing the pignistic probability function in a context of uncertainty", In *Uncertainty in Artificial Intelligenc*e 5, 29-39, Elsevier Science North-Holland, pp. 29-39, 1990.

[8] Zissman M. and Berkling K.M., "Automatic Language Identification", In: Speech Communication, V. 35, pp. 115-124, 2001.