Weighted Loss Functions to Make Risk-based Language Identification Fused Decisions

Abstract

Making a pattern recognition decision with the maximum-likelihood rule is a particular case of the risk-based Bayesian decision rule which is simplified when the loss function is zero–one symmetrical and classes are equally a priori probable. In the case the recognition system is composed of several experts, we can take into account their estimated performance at the class level as a key heuristic-like factor to weight the loss function and drive the recognition process while fusing their decisions. Such indices are formally computed by applying the Discriminant Factor Analysis method. The experiments are carried out in the automatic language identification domain with a system composed of several identification experts. Fusion of expert decisions is achieved by building statistical classifiers.

1. Introduction

Automatic Language Identification (ALI) systems can be composed of several experts or primary systems, also known as sources of decision information, whose aim is to identify as soon as possible the language in which an utterance has been pronounced. The architecture of an ALI expert can be based on the extraction of Acoustic [7], Phonotactic [10] or Prosodic [9] features of languages.

An ALI system faces the problem of fusing, in a suitable way, the identification decisions issued from experts. Most current fusion techniques are rather empirical (average, addition, multiplication, etc.) whose weighted versions take into account heuristic-like information about the performance of experts by applying estimated confidence indicators [8] to expert decisions. Good performance is often obtained though [4]. So, great efforts have started to be deployed to try to formally justify such techniques [2] [5].

We propose an original way of: a) making risk-based fusion decisions by weighting the loss function with performance confidence indices; and b) formally computing performance confidence indices, at the expert and class levels, by extracting language discriminant information in processing a development speech corpus, applying the Discriminant Factor Analysis (DFA) method in the decision score field, and using the DFA projection to obtain the confusion matrix.

At least two kinds of fusion approaches may be studied along with these confidence indices: empirical and statistical fusion. Thus, in section 2 we explain how to compute the performance confidence indices. In section 3, we describe the fusion approaches. The decision making process is covered in section 4. Experiments are treated in section 5.

2. Performance Confidence Indices

ALI experts accept a speech utterance called the observation, as input, and provide the class (language) decision as output, after computing language score values; mostly a statistical model is used and the language score is the language likelihood; so that the experts 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; the higher the value, the more confident the expert is that the corresponding language is the right one.

This global observation is represented as a score matrix (Table 1): $\delta = [\, d_{ij} \,]_{1 \le i \le M,\ 1 \le j \le N}$, where $L = \{L_1, L_2, \ldots, L_i, \ldots, L_M\}$ is the set of languages and $S = \{ s_1, s_2, \ldots, s_j, \ldots, s_N \}$ the set of experts.

Table 1. Score matrix $\delta = [\, d_{ij} \,]_{1 \le i \le M,\ 1 \le j \le N}$.

$L \backslash S$	s_1	s_2	…	s_j	…	s_N
L_1	d_{11}	d_{12}	…	d_{1j}	…	d_{1N}
L_2	d_{21}	d_{22}	…	d_{2j}	…	d_{2N}
…	…	…	…	…	…	…
L_i	d_{i1}	d_{i2}	…	d_{ij}	…	d_{iN}
…	…	…	…	…	…	…
L_M	d_{M1}	d_{M2}	…	d_{Mj}	…	d_{MN}

Estimation of expert performance, with a view to provide the language identification process with heuristic-like information, can be achieved beforehand by means of an evaluation phase where the expert is tested on a set of segments whose language is known.

We split a global speech corpus into three partitions: a learning corpus $X = \{x_{learn}\}$, a test corpus $Y = \{y_{test}\}$ and a development corpus $Z = \{z_{dev}\}$. We use the last one to compute two families of performance indices: the expert and class indices.

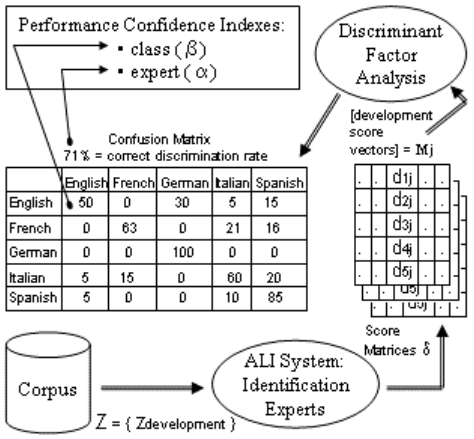

Figure 1. Computing confidence indices.

We collect the score matrices δ corresponding to the acoustic segments of the development corpus; each expert s_j, $1 \le j \le N$, contributes with a score vector corresponding to an acoustic segment and is represented by column j, in each score matrix δ. Then, for each expert s_j, a matrix M_j (Figure 1) will be composed of development score vectors and will correspond to the whole set of acoustic segments.

For each expert s_j, we apply the DFA statistical method to its matrix M_j in order to search for an appropriate representation space for them and a way of obtaining performance confidence indices on a correct discrimination rate basis: we use the M–1 factorial axis corresponding to the M–1 eigenvalues and project the set M_j of score vectors into this subspace. In building the corresponding confusion matrix (Figure 1), the class confidence indices ($\beta_{ij}, 1 \le i \le M$) are directly mapped from the diagonal values while the expert confidence index must be computed as an averaged value:

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

3. Fusion Approaches

3.1. Empirical fusion

Linear (addition) and logarithmic (multiplication) operations are currently employed to empirically fuse expert decision scores. The estimated performance of each expert can be taken into account to weight its decision score in a heuristic-like way.

The concept of weighting by expert estimated performance matches the one of weighting by 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 $L^* = \arg\max_{i \in [1,M]} [\sum_{j \in [1,N]} \alpha_j d_{ij}]$
- Product $L^* = \arg\max_{i \in [1,M]} [\prod_{j \in [1,N]} d_{ij}^{\alpha_j}]$

3.2. Statistical fusion

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

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

where n is the Gaussian component number and Q_i the total number of components for the language L_i. The most likely language 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 NxM 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 eigenvalues 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 test-corpus score matrix the following identification decision rule.

If $Dist(\delta \mid L_i)$ represents the Euclidean distance between the projected matrix δ and the projected gravity center of language L_i, then:

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

4. Risk-based Decisions

Given an identification decision action a_i that classifies the score matrix δ, corresponding to the acoustic segment y, as being in one of the languages L_k, the overall risk is given by [3]:

$$R = \int R(a(\delta) \mid \delta)\, p(\delta)\, d\delta$$

where $d\delta$ denotes a d-space volume element, and the integral extends over the entire reference space of δ.

Performance experience of the N experts can be taken into account to build a particular loss function, that can be represented as a cube-like matrix, and decomposed into single matrices (Figure 2) according to its source j ($1 \le j \le N$) :

$$\Lambda(a_i \mid L_k) = \{\lambda^1(a_i \mid L_k), \lambda^2(a_i \mid L_k), \ldots \lambda^j(a_i \mid L_k), \ldots \lambda^N(a_i \mid L_k)\}$$

Thus, to minimize the overall risk, we compute the conditional risk:

$R(a_i|\delta) = \sum_{k\in[1,M]}\Lambda(a_i|L_k)P(L_k|\delta) =$

$= \sum_{k\in[1,M]}\lambda^1(a_i|L_k)P(L_k|\delta) + \lambda^2(a_i|L_k)P(L_k|\delta) +$

$...\lambda^j(a_i|L_k)P(L_k|\delta) + ...\lambda^N(a_i|L_k)P(L_k|\delta)$

$= \sum_{k\in[1,M]} (\sum_{j\in[1,N]} \lambda^j(a_i|L_k)) P(L_k|\delta)$

where, $\forall i,k = 1,2, \ldots, M$ and $\forall j = 1,2, \ldots, N$, each single loss function $\lambda^j(a_i|L_k)$ can hypothetically be defined as follows:

- Case 1. Wrong actions are to be maximally penalized:

$\lambda^j(a_i|L_k) = 0$ if : $i = k$

$\lambda^j(a_i|L_k) = 1$ if : $i \neq k$

then $R(a_i|\delta) = \sum_{k\neq i} (N) P(L_k|\delta) = N (1 - P(L_i|\delta))$, and as the probability of each language is considered equally *a priori* probable, then the language likelihood $P(\delta|L_i)$ can take the place of the *a posteriori* probability $P(L_i|\delta)$ in building the maximum-likelihood decision rule [3]:

$L^* = \arg\max_{i\in[1,M]} [P(\delta|L_i)]$ ..... (I)

- Case 2 (depicted in Figure 2). Wrong actions are to be penalized with the experience value of the expert at the class level (β_{kj}):

$\lambda^j(a_i|L_k) = 0$ if : $i = k$

$\lambda^j(a_i|L_k) = \beta_{kj}$ if : $i \neq k$

then $R(a_i|\delta) = \sum_{k\neq i} (\sum_j \beta_{kj}) P(L_k|\delta)$, and

$L^* = \arg\min_i [\sum_{k\neq i} P(\delta|L_k) \sum_j \beta_{kj}]$ ..... (II)

- Case 3. Any action is risky. Wrong actions are to be maximally penalized while correct actions are to be penalized with the uncertainty value of the expert at the class level ($1 - \beta_{kj}$):

$\lambda^j(a_i|L_k) = 1 - \beta_{kj}$ if : $i = k$

$\lambda^j(a_i|L_k) = 1$ if : $i \neq k$

then $R(a_i|\delta) = (\sum_j (1 - \beta_{ij})) P(L_i|\delta) + \sum_{k\neq i} (N) P(L_k|\delta)$,

$R(a_i|\delta) = (N) P(L_i|\delta) - \sum_j \beta_{ij} P(L_i|\delta) + N (1 - P(L_i|\delta))$,

$R(a_i|\delta) = N - \sum_j \beta_{ij} P(L_i|\delta)$, and

$L^* = \arg\max_i [P(\delta|L_i) \sum_j \beta_{ij}]$ ..... (III)

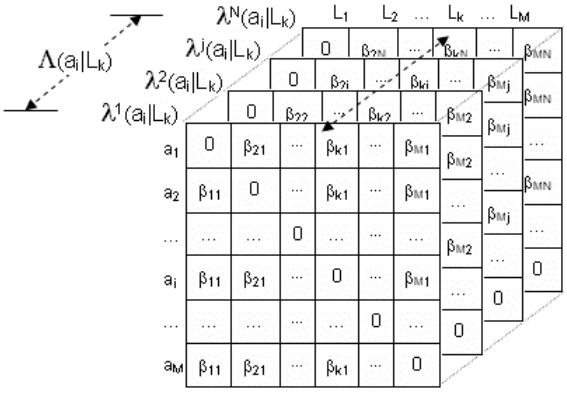

Figure 2. β-weighted loss function $\Lambda(a_i|L_k)$ for case 2.

5. Experimentation

5.1. Fusion system 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 consists 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 global corpus is split into three partitions for each language: the learning corpus, the development corpus and the test corpus (2 speakers: 1 male and 1 female who do not belong to the other corpora).

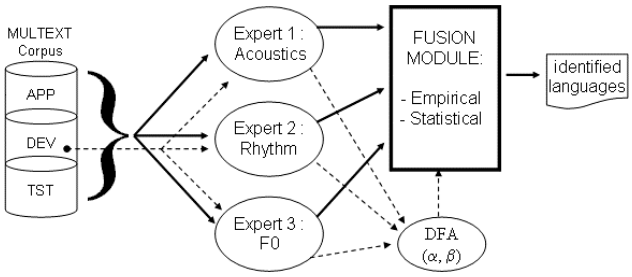

Figure 3. Architecture of the fusion system.

The ALI system is based on three ALI experts and a fusion module (Figure 3):

- Acoustics Expert : 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 underlying segment duration providing a 19-coefficient vector. A cepstral subtraction performs both blind removal of the channel effect and speaker normalisation. For each recording sentence, the average MFCC vector is computed and subtracted from each coefficient.
- Rhythm Expert : Syllable may be a first-rate candidate for rhythm modelling. Nevertheless, segmenting speech in syllables is typically a language specific mechanism; then 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 characterised by its: consonant global duration, vocalic duration, complexity (the number of consonant segments), and energy.
- Fundamental Frequency Expert : The fundamental frequency outlines are used to compute statistics within the same pseudo-syllable frontiers (previously defined) 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 in regard to vocalic onset) and the normalized fundamental frequency bandwidth on each syllable.

For each expert, we apply the same learning-testing procedure: for each language, a Gaussian Mixture Model (GMM) is trained using EM algorithm with LBG initialisation [6]. The optimal number of components of the mixture is obtained from experiments on the learning part of the corpus. During the test, the decision relies on a Maximum Likelihood procedure. The performance of these three experts is given in Table 2, and is considered as a reference to be compared with.

5.2. Tests and results

Three sets of the test corpus (2 speakers out of 10: 1 male and 1 female) are selected on a round-robin basis with a view to test over representative expert performance data of good (set 1) and rather-bad examples (sets 2 and 3). Empirical and statistical fusion approaches are experimented to merge the expert decision scores; for the GMM fusion, we use 20 Gaussian components.

All of the fusion techniques are tested in their non-weighted and weighted versions. The development corpus is used to compute the class and expert performance confidence indices whose information is used to drive in a heuristic-like way the decision process for the weighted fusion versions. Statistical fusion decisions, in their weighted version, are made by applying risk-based rules II and III; testing shows both rules deliver the same results.

Most important results in fusing the three experts are shown in Table 2:

- The empirical fusion delivers higher identification rates than those of any expert: up to 83%. Weighted versions are better than non-weighted ones for set 1 (good-example data) only.
- Excepting the statistical fusion, all the other approaches fail in set 2 (bad-example data).
- The weighted statistical fusion delivers the best identification rates: up to 86%.

6. Conclusion

In the field of Automatic Language Identification, both computation of performance confidence indices and application of such indices to make recognition decisions can be done formally by means of the Discriminant Factor Analysis method and the Risk-based Bayes rule. This methodology allows us to weight likelihood values at the class level and appears as a strong alternative to empirical techniques that do weighting at the expert level. That can partially explain why statistical methods delivers better identification rates than the empirical ones. A future direct work, as result of weighted loss functions, could be to explore uncertainty-based techniques such as the ones coming from Possibility and Evidence Theories where we could implement inference processes based on degrees of possibility, compute possibility values of languages and go back to the probability domain to make risk-based decisions.

Table 2. Comparison of fusion strategies.

Language Identification System – Total Success Rate (%) –				
		1st set	2nd set	3rd set
Reference Experts	Expert 1 : acoustics	78	41	62
	Expert 2 : rhythm	70	63	60
	Expert 3 : fundamental frequency	35	37	48
Empirical Fusion Techniques	Addition	80	60	69
	Product	63	49	68
	Weighted Addition	83	58	67
	Weighted Product	65	51	67
Statistical Fusion Methods	Factor Analysis (DFA)	83	68	82
	Gaussian Mixture Model (GMM)	84	66	82
	Weighted DFA	85	70	82
	Weighted GMM	86	67	84

7. References

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

[2] Cooke R.M. *Experts in uncertainty*. Oxford University Press, Oxford, United Kingdom, 1991.

[3] Duda R.O., Hart P.E. and Stork D.G. *Pattern Classification*, John Wiley & Sons Inc, 2nd Edition, USA, 2001.

[4] Hazen T.J. and Zue V.W. "Segmented-based automatic language identification", *Journal of the Acoustical Society of America*, 4(101), 1997.

[5] Kittler J., Hojjatoleslami A.J. and Windeatt T. "Weighting factors in multiple expert fusion". In *Proceedings of BMVC'97*, pages 41-50, Essex University, United Kingdom, 1997.

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

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

[8] Rahman A. and Fairhurst M. "A novel confidence-based framework for multiple expert decision fusion", in *Proceedings of BMVC'98*, University of Southampton, United Kingdom, 1998.

[9] 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*, pages 567-570, Barcelona, Spain, 2003.

[10] Zissman M. and Berkling K.M. "Automatic language identification". In *Speech Communication*, volume 35, pages 115-124, 2001.