# Feature Selection Based on Information Theory for Speaker Verification

Rafael Fernández[1,2], Jean-François Bonastre[2], Driss Matrouf[2], and José R. Calvo[1]

[1] Advanced Technologies Application Center, Havana, Cuba
[2] Laboratoire d'Informatique d'Avignon, UAPV, France
{rfernandez,jcalvo}@cenatav.co.cu
{jean-francois.bonastre,driss.matrouf}@univ-avignon.fr

**Abstract.** Feature extraction/selection is an important stage in every speaker recognition system. Dimension reduction plays a mayor roll due to not only the *curse* of dimensionality or computation time, but also because of the discriminative relevancy of each feature. The use of automatic methods able to reduce the dimension of the feature space without losing performance is one important problem nowadays. In this sense, a method based on mutual information is studied in order to keep as much discriminative information as possible and the less amount of redundant information. The system performance as a function of the number of retained features is studied.

**Keywords:** mutual information, feature selection, speaker verification.

## 1 Introduction

The task of feature extraction/selection is a crucial step in an automatic speaker recognition system. The performance of the later components –speaker modeling and pattern matching– is strongly determined by the quality of the features extracted in this first stage [1,2]. Most of the efforts today are doveted to the classification stage, and little to find optimal representations of the speakers.

Methods like Principal Component Analysis (PCA), Discrete Cosine Transform (DCT) and Linear Discriminant Analysis (LDA) have been employed extensively in the literature for reducing the dimension of the feature space. These methods rely on maximazing the data distribution variance –globally or per class– or minimizing the reconstruction error by selecting only a subset of the original feature set. However, this does not mean necessarily high speaker discrimination or low redundancy in the feature set. Others like the *knock-out* method [3,4] proposes to evaluate all the subsets of $n-1$ coefficients in order to keep the best subset in each iteration. One problem here is that the selection of the best subset strongly depends on the classification method employed.

Various combinations of LFCC-based features, the energy, and their deltas and delta-deltas were tested in [5] to obtain the best configuration. In this work,

E. Bayro-Corrochano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856, pp. 305–312, 2009.

we study the use of an information theory based method [6] in order to select automatically the best subset of acoustic coefficients. This method was applied in [7] for selecting the best Wavelet Packet Tree (WPT) in a speaker identification system. The main principle of the selection algorithm is to maximize the discriminative information while minimizing the redundancy between the selected features.

A study of the system performance as a function of the dimension of the feature space is presented. A state-of-art speaker verification system [8] is used in order to evaluate the usefulness of the method.

The remainder is organized as follows: Section 2 shows the basis of the proposed information theory oriented method; Section 3 shows the details of the implemented algorithm; database description, experimental work and results are summarized in Section 4; last section is devoted to conclusions and future works.

## 2 Feature Selection Based on Mutual Information

Reducing the dimensionality of feature vectors is usually an essential step in pattern recognition tasks. By removing most irrelevant and redundant features, feature selection helps to improve the performance of learning models by: alleviating the effect of the *curse* of dimensionality, enhancing generalization capability, speeding up learning process and improving model interpretability.

Methods based on Information Theory can act as a general criterion, since they consider high order statistics, and can be used as a base for nonlinear transformations [9]. With these methods, low information redundancy is achieved and the discriminative information is intended to be kept while reducing the dimensionality.

In probability theory and information theory, the mutual information of two random variables is a quantity that measures their mutual dependence [10].

Let $\mathcal{S}$ and $\boldsymbol{X} \in \mathbb{R}^N$ be the variables for the speaker class and the speech feature vector respectively. The mutual information between $\mathcal{S}$ and $\boldsymbol{X}$ is given by:

$$I(\mathcal{S}, \boldsymbol{X}) = H(\mathcal{S}) + H(\boldsymbol{X}) - H(\mathcal{S}, \boldsymbol{X}) = H(\mathcal{S}) - H(\mathcal{S}|\boldsymbol{X}), \tag{1}$$

where $H(\cdot)$ is the entropy function, which is a measure of the uncertainty of the variable. For a discrete-valued random variable $\boldsymbol{X}$, it is defined as:

$$H(\boldsymbol{X}) = -\sum_m p(\boldsymbol{X} = \boldsymbol{x}_m) \log_2 p(\boldsymbol{X} = \boldsymbol{x}_m), \tag{2}$$

where $p(\boldsymbol{X} = \boldsymbol{x}_m)$ is the probability that $\boldsymbol{X}$ takes the value $\boldsymbol{x}_m$.

From (1), mutual information measures the uncertainty reduction of $\mathcal{S}$ knowing the feature values. Those features with low speaker information have low values of mutual information with the speaker class. Following this criterion, the best $K$ coefficients from the original set $\boldsymbol{X} = \{X_1, \ldots, X_N\}$ are those $\boldsymbol{X}' = \{X_{i_1}, \ldots, X_{i_K}\} \subset \boldsymbol{X}$ which maximise the mutual information with the speaker class:

$$\boldsymbol{X}' = \underset{\{X_{j_1}, \ldots, X_{j_K}\}}{\arg\max} I(\mathcal{S}, \{X_{j_1}, \ldots, X_{j_K}\}). \tag{3}$$

If the features were statistically independent, the search in (3) would be reduced to find those features iteratively. If we know the first $k-1$ features, the $k$-th is obtained using this recursive equation:

$$X_{i_k} = \underset{X_j \notin \{X_{i_1},\ldots,X_{i_{k-1}}\}}{\arg\max} I(X_j, \mathcal{S}), \qquad k = 1, \ldots, K. \tag{4}$$

In the case of statistically dependent feature –very frequent in real life problems– the latter is not true. Here, the problem of finding out the best subset (see Eq. (3)) becomes a search for all the $\binom{N}{K}$ combinations.

In order to select the best coefficients, the sub-optimal method [6,11] was applied. If we have the first $k-1$ coefficients $\boldsymbol{X}_{k-1} = \{X_{i_1}, \ldots, X_{i_{k-1}}\}$, the $k$-th is selected according to:

$$X_{i_k} = \underset{X_j \notin \boldsymbol{X}_{k-1}}{\arg\max} \left[ I(X_j, \mathcal{S}) - \frac{1}{k-1} \sum_{s=1}^{k-1} I(X_j, X_{i_s}) \right]. \tag{5}$$

The idea is to look for the coefficients with high mutual information with the speaker class and low average mutual information with the features previously selected. Last term in (5) can be thought of as a way to reduce the redundant information. Here, mutual information between two variables is the only estimation needed, which avoids the problem of estimating the probability densities of high dimension vectors. We used histogram method to calculate the probability densities.

## 3 The Algorithm

Based on the stated above, we developed the following algorithm to withdraw the worst features in an original 60-feature LFCC configuration.

**Algorithm 1**. Proposed method

```
k = N;
SearchList = {1, ..., N};
while k > K do
    foreach n ∈ SearchList do
        C = SearchList \ {n};
        F(n) = I(X_n, S) − (1/(k−1)) Σ_{m∈C} I(X_n, X_m);
    end
    n* = arg min_m(F(m));
    SearchList = SearchList \ {n*};
    k = k + 1;
end
```

At every stage, the coefficient to be eliminated is selected according to (5). This cycle is repeated until the desired number of $K$ features is reached.

# 4 Experiments and Results

## 4.1 Database

All the experiments presented in section 4 are performed based upon the NIST 2005 database, 1conv-4w 1conv-4w, restricted to male speakers only. This condition consists of 274 speakers. Train and test utterances contain 2.5 minutes of speech on average (extracted from telephone conversations). The whole speaker detection experiment consists in 13624 tests, including 1231 target tests and 12393 impostors trials. From 1 to 170 tests are computed by speaker, with an average of 51 tests.

## 4.2 Front End Processing

All the experiments were realized under the LIA_SpkDet system [12] developed at the LIA lab. This system consists in a cepstral GMM-UBM system and has been built from the ALIZE platform [8]. Depending on the starting set of coefficients, two cases –described below– were analyzed. The feature extraction is done using SPRO [13]. Energy-based frame removal –modelled by a 3 component GMM– is applied as well as mean and variance normalization. The UBM and target models contain 512 Gaussian components. LLR scores are computed using the top ten components. For the UBM, a set of 2453 male speakers from the NIST 2004 database was used.

The performance is evaluated through classical DET performance curve [14], Equal Error Rate (EER) and Detection Cost Function (DCF).

## 4.3 Experiments

Two experiments were done in order to select the best coefficients. In the first one a 60-feature set was taken as starting set, and in the second one, a 50-feature subset was considered.

**Experiment 1:** The starting set considered in this experiment consists in a 60-feature set composed of 19 LFCC, the energy and their corresponding first and second derivative.

**Experiment 2:** In this case the selection started from a 50-feature subset –derived from the previously described set– composed of the first 19 LFCC, their first derivative, the first 11 of the second derivative and the delta energy. This configuration is the result of large empirical experience based on human expert knowledge [5] and will be taken as the baseline in this work.

The results of the EER and DCF for each experiment are shown in Figures 1 and 2.

In order to analyze the eliminated features at each stage, the rank order for each coefficient for both starting sets is shown in figure 3. For better understanding they were divided in three classes: static, delta (D), and delta-delta (DD).

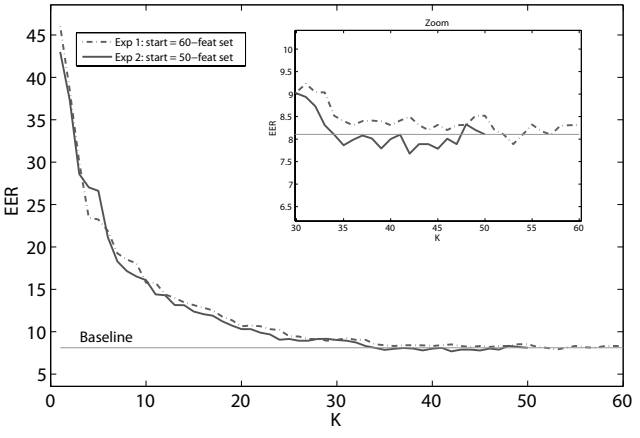


**Fig. 1.** EER as a function of the feature space dimension

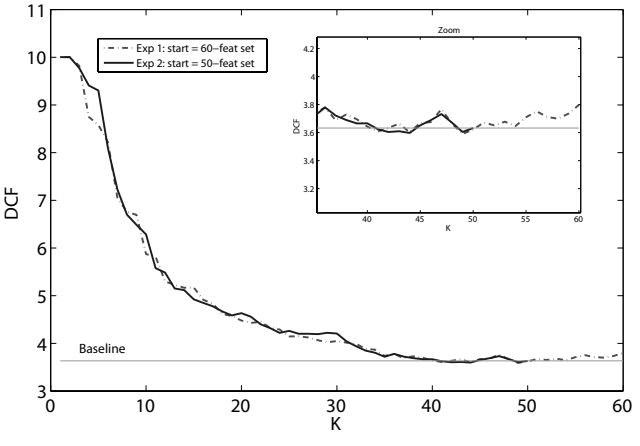


**Fig. 2.** DCF as a function of the feature space dimension

A general observed behavior is that even when all the features were used in the selection (Experiment 1), almost all the delta-delta features were the first eliminated. This is in accordance with the experience accumulated that state that these features have a weak contribution to the speaker verification task, since they do not carry a significant amount of new information. Meanwhile, by and large the static parameters show the highest relevance for both experiments. Static coefficient 'zero' was the least ranked among the statics coefficients as expected, although it was the best ranked among the delta-deltas.

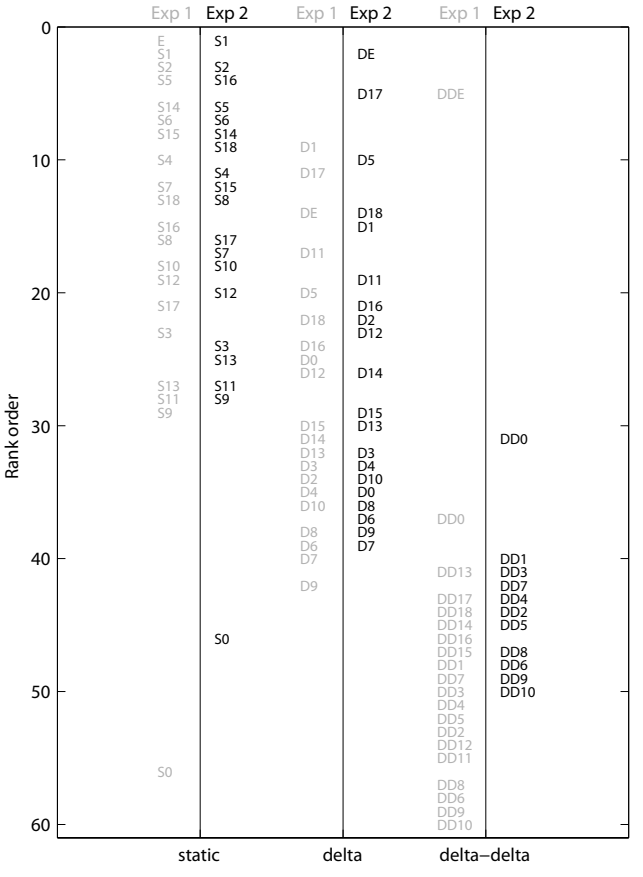


**Fig. 3.** Feature rank order for both experiments

One significant difference is the first place of the energy when all the features are considered, which was not used in the 50-feature starting set. However, when the selection starts from this configuration (Experiment 2), the system achieved the best results. This may be determined by all the *a priori* information that it includes, which is based in strong experimental basis and human expert knowledge. However, for both starting sets, the feature selection method was able to detect better configurations with a lower dimension.

The DET curves for some interesting configurations are shown in figure 4. Three configurations derived from the Experiment 2 are shown: the first one is the configuration with the smallest dimensionality that achieves at least the same performance as the baseline (K=34), the second one achieved the best performance among all the analyzed combinations (K=42), and the the third one is the baseline (K=50). The result for the full 60-feature set (K=60) is also presented.

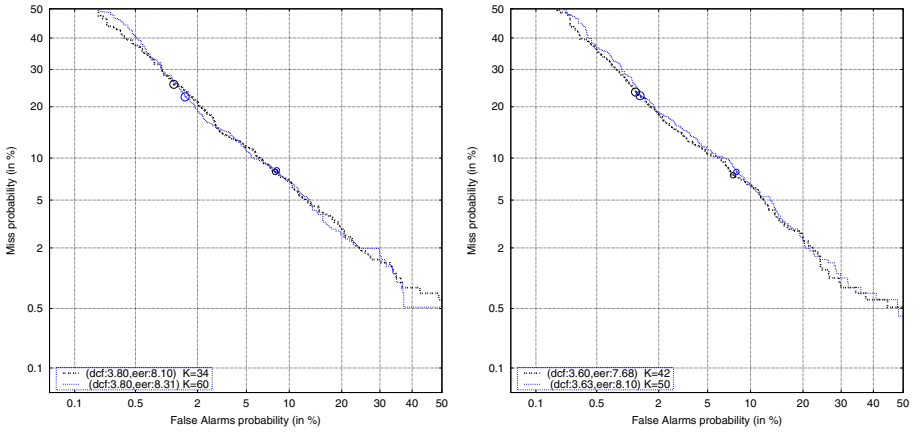


**Fig. 4.** DET curves for the configurations K=50 (baseline), K=34 and K=42, corresponding to the 50-feature set as a starting set for the selection

For almost all the operation points the configuration corresponding to K=42 outperforms the baseline (slightly, though). More significant is the configuration corresponding to K=34, which leads to the same EER as the baseline with a reduced number of features.

## 5 Conclusions

The problem of selecting the best LFCC subspace for speaker verification is discused in this work. A mutual information criterion has been studied to select the best LFCC features. The proposed feature selection method showed good capabilities for reducing the feature space dimension without losing performance. The experiment starting with *a priori* information finds better configurations. Better models must be studied in order to look for the optimal configuration with no *a priori* information. Other ways to find the most informative time-spectral regions will be analysed in the future.

## References

1. Campbell, J.P.: Speaker recognition: A tutorial. Proceedings of the IEEE 85(9), 1437–1462 (1997)
2. Kinnunen, T.: Spectral features for automatic text-independent speaker recognition. Lic. Th., Department of Computer Science, University of Joensuu, Finland (2003)
3. Sambur, M.R.: Selection of acoustic features for speaker identification. IEEE Trans. Acoust. Speech, Signal Processing 23(2), 176–182 (1975)
4. Aha, D.W., Bankert, R.L.: A comparative evaluation of sequential feature selection algorithms. In: Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics, pp. 1–7. Springer, Heidelberg (1995)

5. Fauve, B.: Tackling Variabilities in Speaker Verification with a Focus on Short Durations. PhD thesis, School of Engineering Swansea University (2009)
6. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy. IEEE Trans. Patt. Anal. and Mach. Intel. 27(8), 1226–1238 (2005)
7. Fernández, R., Montalvo, A., Calvo, J.R., Hernández, G.: Selection of the best wavelet packet nodes based on mutual information for speaker identification. In: Ruiz-Shulcloper, J., Kropatsch, W.G. (eds.) CIARP 2008. LNCS, vol. 5197, pp. 78–85. Springer, Heidelberg (2008)
8. Bonastre, J.F., et al.: ALIZE/spkdet: a state-of-the-art open source software for speaker recognition, Odyssey, Stellenbosch, South Africa (January 2008)
9. Torkkola, K., Campbell, W.M.: Mutual information in learning feature transformations. In: Proc. Int. Conf. on Mach. Learning, San Francisco, CA, USA, pp. 1015–1022. Morgan Kaufmann Publishers Inc., San Francisco (2000)
10. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience, Hoboken (1991)
11. Lu, X., Dang, J.: Dimension reduction for speaker identification based on mutual information. In: Interspeech, pp. 2021–2024 (2007)
12. LIA_SpkDet system web site:
`http://www.lia.univ-avignon.fr/heberges/ALIZE/LIA_RAL`
13. Gravier, G.: SPRO: a free speech signal processing toolkit,
`http://www.irisa.fr/metiss/guig/spro/`
14. Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The DET curve in assessment of detection task performance. In: Eurospeech, Rhodes, Greece, September 1997, pp. 1895–1898 (1997)