Feature Extraction Assessment for an Acoustic-Event Classification Task using the Entropy Triangle

David Mejía-Navarrete, Ascensión Gallardo-Antolín,
*Carmen Peláez-Moreno, Francisco J. Valverde-Albacete**

Department of Signal Theory and Communications, Universidad Carlos III Madrid, Spain

gallardo, carmen, and fva at tsc dot uc3m dot es

Abstract

We assess the behaviour of 5 different feature extraction methods for an acoustic event classification task—built using the same SVM underlying technology—by means of two different techniques: accuracy and the entropy triangle. The entropy triangle is able to find a classifier instance whose relatively high accuracy stems from an attempt to specialize in some classes to the detriment of the overall behaviour. On all other cases, fair classifiers, accuracy and entropy triangle agree.

Index Terms: Classifier evaluation, accuracy entropy triangle, acoustic event classification, SVM classifiers.

1. Introduction

In recent years, the problem of automatically detecting and classifying acoustic non-speech events has attracted the attention of numerous researchers. Although speech is the most informative acoustic event, other kind of sounds (such as laughs, coughs, pen handwriting, etc.) can give relevant cues about the human presence and activity in a certain scenario (for example, in an office room). This information could be used in different applications, mainly in those with perceptually aware interfaces such as smart-rooms [1], automotive applications [2], mobile robots working in diverse environments [3] or surveillance systems [4].

Usually, Acoustic Event Classification (AEC) systems are composed of two main processing stages: the feature extraction and classification modules. With respect the first module, several features have been proposed in the literature, such as MFCC (*Mel-Frequency Cepstral Coefficients*) [5], PLP (*Perceptual Linear Prediction*) [6], log-energy, spectral flux, fundamental entropy and zero-crossing rate [1]. For the classifier itself, various machine learning approaches have also been considered, mainly Gaussian Mixture Models (GMM) [1], [7], Hidden Markov Models (HMM) [5] or Support Vector Machines (SVM) [1], [6], [7].

The most usual measure to assess classifier performance is *accuracy*, but it has often been challenged on the basis of non-desirable properties [8]. With this frame of mind, [9] argued for entropic measures that take into account the information transfer through the classifier, like the *expected mutual information* between the input and output distributions

$$MI_{P_{XY}} = \sum_{x,y} P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} \qquad (1)$$

and provided a contrived example with three confusion matrices with the same accuracy but clearly differing performances, in their opinion due to differences in mutual information. Such examples are alike those put forth by [8] to argue for Cohen's kappa as an evaluation metric for classifiers.

In [10] a tool for the evaluation and comparison of classifiers was first introduced, the (de Finetti) entropy triangle which tries to address several shortcomings of traditional measures like accuracy [8], Mutual Information, Variation of Information [11] or the AUC (VUS) [12, 13] of the ROC curve for the assessment of multiclass classifiers [14].

In this paper we have carried out several comparative tests on an SVM-based AEC system using different feature sets—all of them derived from the short-term MFCCs—and the performance of the resulting classifiers has been compared in terms of the classification accuracy and the entropy triangle.

2. Methods

2.1. SVM classification of acoustic events

The AEC system is based on a one-against-one SVM with RBF kernel and a majority voting scheme for the final decision. This classifier has been tested using several feature sets. All of them have been obtained by applying different temporal feature integration techniques over the conventional short-term MFCCs. Temporal feature integration is the process of representing an audio segment by a single feature through the combination of the frame-by-frame feature vectors belonging to it [15]. This kind of *segment-based* parameters are commonly used in tasks related to audio classification (AEC, speech-music discrimination, etc.). In particular, the 5 sets of segment-based features considered in this work are:

- Features derived from the statistics of the short-term vectors computed over each audio segment:
  - ***MeanStd***. Mean and standard deviation of MFCCs and log-energy.
  - ***MeanStdSk***. Mean, standard deviation and skewness of MFCCs and log-energy.
  - ***MeanStdSkKu***. Mean, standard deviation, skewness and kurtosis of MFCCs and log-energy.
- ***FC***. Filter Bank Coefficients, which summarize the periodogram of each short-term feature dimension (computed over each audio segment) in four frequency bands: a dc-filter, 1-2 Hz modulation energy, 3-15 Hz modulation energy and 20-43 Hz perceptual roughness [15]. In contrast to the statistics-based features, FC takes into account the temporal structure of the MFCC vectors contained in the corresponding audio segment.
- ***MeanStdSk+FC***. The concatenation of such features.

Our aim is to compare these sets of features for AEC.

2.2. The entropy triangle

The (de Finetti) entropy triangle (ET)[1] is a tool for the assessment of multiclass classifiers using the decomposition of the joint entropy of two random variables [10].

Let $V_X = \{x_i\}_{i=1}^n$ and $V_Y = \{y_j\}_{j=1}^p$ be sets of input and output class identifiers, respectively, in a multiple-class classification task. The behavior of the classifier can be sampled over N iterated experiments to obtain a count matrix N_{XY} where $N_{XY}(x_i, y_j) = N_{ij}$ counts the number of times that the joint event $(X = x_i, Y = y_j)$ occurs. We say that N_{XY} is the *(count-based) confusion matrix or contingency table* of the classifier.

A better ground for discussing performance than count confusion matrices may be empirical estimates of the joint distribution between input and outputs, like the maximum likelihood estimate used throughout this paper $P_{XY}(x_i, y_j) \approx \hat{P}_{XY}^{\mathbb{MLE}}(x_i, y_j) = N(x_i, y_j)/N$, so let:

- $P_{XY}(x, y)$ be an estimate of the joint probability mass function (pmf) between input and output with marginals $P_X(x) = \sum_{y_j \in Y} P_{X,Y}(x, y_j)$ and $P_Y(y) = \sum_{x_i \in X} P_{X,Y}(x_i, y)$.
- $Q_{XY} = P_X \cdot P_Y$ be the pmf[2] with the same marginals as P_{XY} considering them to be independent (that is, describing independent variables).
- $U_{XY} = U_X \cdot U_Y$ be the product of the uniform, maximally entropic pmfs over X and Y, $U_X(x) = 1/n$ and $U_Y(y) = 1/p$.

Then the loss in uncertainty from U_{XY} to Q_{XY} is the difference in entropies:

$$\Delta H_{P_X \cdot P_Y} = H_{U_X \cdot U_Y} - H_{P_X \cdot P_Y} \tag{2}$$

Intuitively, $\Delta H_{P_X \cdot P_Y}$ measures how far the classifier is operating from the most general situation possible where all inputs are equally probable, which prevents the classifier from specializing in an overrepresented class to the detriment of classification accuracy in others. Since $H_{U_X} = \log n$ and $H_{U_Y} = \log p$, $\Delta H_{P_X \cdot P_Y}$ may vary from $\Delta H_{P_X \cdot P_Y}^{\min} = 0$, when the marginals themselves are uniform $P_X = U_X$ and $P_Y = U_Y$, to a maximum value $\Delta H_{P_X \cdot P_Y}^{\max} = \log n + \log p$, when they are Kronecker delta distributions.

We would like to relate this entropy decrement to the expected mutual information $MI_{P_{XY}}$ of a joint distribution. For that purpose, we realize that the mutual information formula (1) describes the decrease in entropy when passing from distribution $Q_{XY} = P_X \cdot P_Y$ to P_{XY},

$$MI_{P_{XY}} = H_{P_X \cdot P_Y} - H_{P_{XY}} . \tag{3}$$

And finally we invoke the well-known formula relating the joint entropy $H_{P_{XY}}$ and the expected mutual information $MI_{P_{XY}}$ to the conditional entropies of X given Y, $H_{P_{X|Y}}$ (Y given X, $H_{P_{Y|X}}$, respectively),

$$H_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}} + MI_{P_{XY}} . \tag{4}$$

Therefore $MI_{P_{XY}}$ ranges from $MI_{P_{XY}}^{\min} = 0$ when $P_{XY} = P_X \cdot P_Y$, a bad classifier, to a theoretical maximum $MI_{P_{XY}}^{\max} = (\log n + \log p)/2$ in case the marginals are uniform and input and output are completely dependent, an excellent classifier.

[1] http://www.mathworks.com/matlabcentral/fileexchange/30914-entropy-triangle

[2] We drop the explicit variable notation in the distributions from now on.

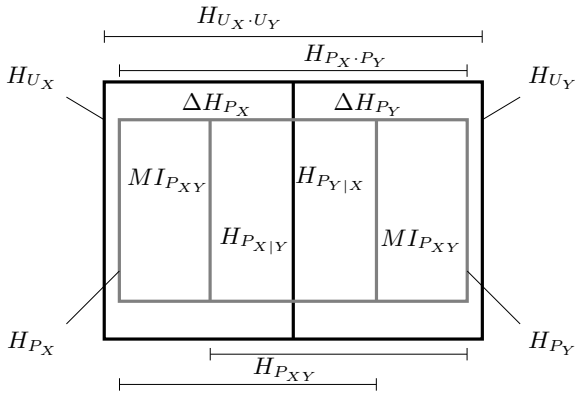


Figure 1: **Extended information diagram of entropies related to a bivariate distribution**. The expected mutual information appears *twice*.

The variation of information is defined in [11] as

$$VI_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}} . \tag{5}$$

For optimal classifiers with deterministic relation from the input to the output, and diagonal confusion matrices $VI_{P_{XY}}^{\min} = 0$, e.g., all the information about X is borne by Y and vice versa. On the contrary, when they are independent $VI_{P_{XY}}^{\max} = H_{P_X} + H_{P_Y}$, the case with inaccurate classifiers which uniformly redistribute inputs among all outputs.

Collecting Eqs. (2)–(5) results in the *balance equation for information related to a joint distribution*,

$$H_{U_{XY}} = \Delta H_{P_X \cdot P_Y} + 2MI_{P_{XY}} + VI_{P_{XY}} . \tag{6}$$

The balance equation suggests an *information diagram* as depicted in Figure 1. In this diagram we distinguish the familiar decomposition of the joint entropy $H_{P_{XY}}$ as the two entropies H_{P_X} and H_{P_Y} whose intersection is $MI_{P_{XY}}$. But notice that the increment between $H_{P_{XY}}$ and $H_{P_X \cdot P_Y}$ is yet again $MI_{P_{XY}}$, hence the expected mutual information appears *twice* in the diagram. Further, the interior of the outer rectangle represents $H_{U_X \cdot U_Y}$, the interior of the inner rectangle $H_{P_X \cdot P_Y}$ and $\Delta H_{P_X \cdot P_Y}$ represents their difference in areas.

To gain further understanding of the entropy decomposition suggested by the balance equation, from Eq. (6) and the paragraphs following Eqs. (2)–(5), we obtain

$$\begin{aligned} H_{U_{XY}} &= \Delta H_{P_X \cdot P_Y} + 2MI_{P_{XY}} + VI_{P_{XY}} \\ 0 &\le \Delta H_{P_X \cdot P_Y}, 2MI_{P_{XY}}, VI_{P_{XY}} \le H_{U_{XY}} \end{aligned}$$

imposing severe constraints on the values the quantities may take, the most conspicuous of which is that given two of the quantities the third one is fixed. Normalizing by $H_{U_{XY}}$ we get

$$\begin{aligned} 1 &= \Delta H'_{P_X \cdot P_Y} + 2MI'_{P_{XY}} + VI'_{P_{XY}} \\ 0 &\le \Delta H'_{P_X \cdot P_Y}, 2MI'_{P_{XY}}, VI'_{P_{XY}} \le 1 . \end{aligned} \tag{7}$$

This is a 2-simplex in normalized $\Delta H'_{P_X \cdot P_Y} \times 2MI'_{P_{XY}} \times VI'_{P_{XY}}$ space. A de Finetti diagram is a projection of this simplex 2-dimensions like that depicted in Figure 2.a), whereby each classifier with joint distribution P_{XY} can be characterized by its *joint entropy fractions*, $F_{XY}(P_{XY}) = [\Delta H'_{P_{XY}}, 2 \times MI'_{P_{XY}}, VI'_{P_{XY}}]$.

Notice that, since both U_X and U_Y and P_X and P_Y are independent as marginals of U_{XY} and Q_{XY}, respectively, we may write:

$$\Delta H_{P_X} = H_{U_X} - H_{P_X} \qquad \Delta H_{P_Y} = H_{U_Y} - H_{P_Y} \tag{8}$$

what suggests writing separate balance equations,

$$\begin{aligned} H_{U_X} &= \Delta H_{P_X} + MI_{P_{XY}} + H_{P_{X|Y}} \\ H_{U_Y} &= \Delta H_{P_Y} + MI_{P_{XY}} + H_{P_{Y|X}} \,. \end{aligned} \tag{9}$$

Their normalization produces a *split entropy triangle* where the marginal contributions to the entropy can be plotted side by side as in Figure 2.b).

3. Results

3.1. Experimental data

The database collected for the experiments consists of a total of 1 279 sounds belonging to 20 different acoustic classes as is shown in Table 1. The composition of the whole database was intended to be similar to the one used in [1]. Audio files were obtained from different sources: websites, the ShATR database (for speech sounds) [16] and the FBK-Irst database [17]. In this latter case, sounds were previously segmented and extracted from the original audio records of the database. All sounds were converted to the same format and sampling frequency (8 KHz).

Table 1: *Database used in the experiments.*

Class number	Event	Files (train+test)	Frames (test)
1	Chair moving	39	860
2	Clapping	86	860
3	Cough	134	2620
4	Door slam	94	1320
5	Pen/pencil handwriting	31	3480
6	Keyboard	92	8360
7	Laugh	160	7420
8	Liquid pouring	63	2360
9	MIMIO pen buzz	36	1360
10	Music	44	9260
11	Paper crumple	46	1480
12	Paper tear	98	1380
13	Paper wrapping	36	1080
14	Phone ring	66	2180
15	Sneeze	57	880
16	Sniffing	22	380
17	Speech	70	20440
18	Puncher/stapler	35	900
19	Steps	38	1060
20	Yawn	32	760

3.2. Assessing feature extraction using accuracy

Several experiments were carried out to assess the classification performance of the AEC system considering the different feature sets mentioned in subsection 2.1. In all cases, the first five MFCCs were computed every 10 ms using a Hamming analysis window of 25 ms long and 20 mel-spaced spectral bands. Also, the log-energy of each frame was computed and added to the MFCC parameters. Temporal feature integration was applied over segments of 2 s length with overlap of 1 s.

The SVM-based classifier was trained following the experimental protocol used in [1]. The acoustic samples were randomly numbered within each class, assigning odd samples to training and even ones to testing. Each experiment was repeated 20 times following this procedure and results were averaged afterwards for obtaining the overall classification performance. For each one of these experiments, a 5-fold cross validation was used for computing the optimal values of RBF kernel parameters.

Table 2 shows the results achieved in terms of classification accuracy (percentage of frames correctly classified). The inclusion of the skewness parameter outperforms significantly the baseline system (***MeanStd***). The use of a higher order statistic—the kurtosis—does not produce further improvement. Results obtained with the ***FC*** approach are similar to those achieved with mean, standard deviation and skewness. However, as gleaned from their confusion matrices either system produces very different kinds of errors. In fact, the concatenation of ***MeanStdSk*** and ***FC*** features outperforms the corresponding individual systems, yielding the best results.

Table 2: *Classification accuracy of the different feature sets.*

Feature Set	Classification Accuracy (%)
MeanStd	68.16 ± 1.84
MeanStdSk	74.26 ± 1.52
MeanStdSkKu	74.71 ± 1.43
FC	73.57 ± 1.14
MeanStdSk + FC	94.96 ± 0.25

3.3. Assessing feature extraction with the entropy triangle

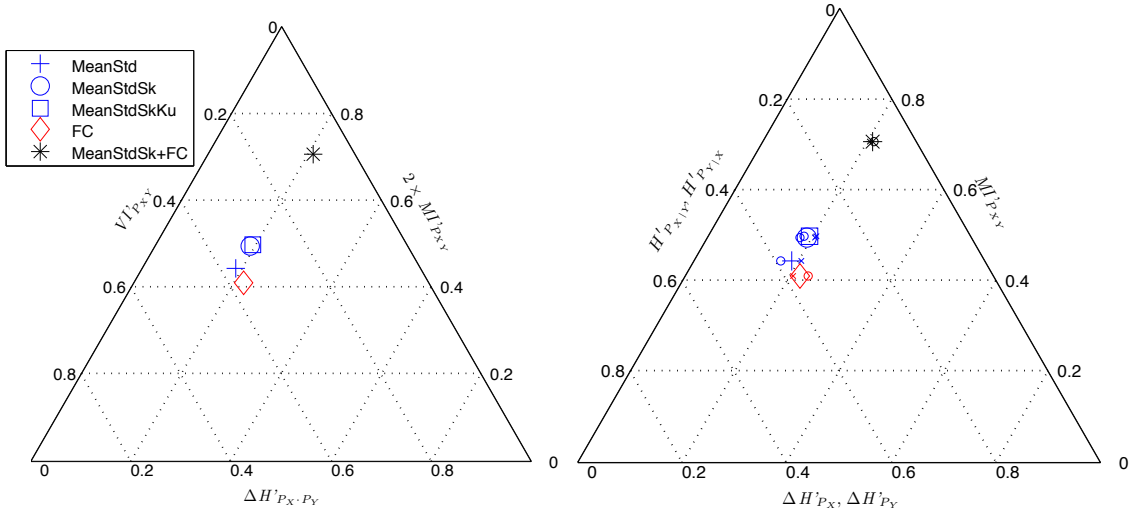
We represented the performance of the classifiers in the entropy triangle in the aggregate and split varieties in Figure 2.

Since a perfect classifier would be at the apex of the triangle—representing diagonal confusion matrices—, the entropy diagram shows clearly that ***MeanStdSk+FC*** (asterisk) outperforms the other sets of features, ***MeanStdSk*** and ***MeanStdSkKu*** (circle and square) are indistinguishable and ***FC*** (diamond) is underperforming, even below ***MeanStd*** (crosshairs), contrary to the accuracy criterion.

A corroboration of the former findings and the last discrepancy can be found in the split triangle of Figure 2.b), where the point related to the P_X marginal is marked with a small cross $\times$ and that related to the output marginal P_Y is marked with a small $\circ$ aligned with the aggregate marker.

In all classifiers, H'_{P_X} is given by the input distribution returning a normalized decrement of entropy $\Delta H_{P_X} = 0.2$. In all logic we would expect a classifier system to increase the entropy of Y with respect to the entropy of X. Otherwise, a classifier may specialize in some classes–for instance majority classes or particularly easy ones—to the detriment of others.

The classifiers with feature sets ***MeanStdSk+FC***, ***MeanStdSkKu*** and ***MeanStdSk*** show very small divergence for $\Delta H'_{P_Y}$, $\circ$ being mostly exactly over $\times$. ***MeanStd*** shows $\Delta H'_{P_Y}$ safely to the left of $\Delta H'_{P_X}$—a smaller $\Delta H'_{P_Y}$ means P_Y is more entropic—but ***FC*** shows a higher $\Delta H'_{P_Y}$ evidencing a less smooth distributions. We take the latter to mean that this classifier is specializing in some decisions to improve accuracy. In this case this is not parallelled by greater transfer of information from input to output—$MI'_{P_{XY}}$ remains constant—but, rather, by a decrement of $H'_{P_{Y|X}}$. Following



(a) Aggregate entropy triangle

(b) Split entropy triangle

Figure 2: ***(colour online) Entropy triangle representation of the results of the experiments in Section. 3.1:*** *the de Finetti entropy diagram or entropy triangle is a projection of the* 2*-simplex onto a two-dimensional space.*

this line of reason **FC** *a worse classifier than* **MeanStd** *despite its higher accuracy.*

4. Conclusions

We have introduced the entropy triangle as a tool to assess the performance of feature extraction in multiclass classification. We have applied it to the analysis of solution for an acoustic event classification task using SVM classifiers and an inventory of 5 feature sets extracted using different criteria. The richer feature sets show better results in accuracy and in the entropy triangle diagram following a law of diminishing returrns (***MeanStd-MeanStdSk-MeanStdSkKu*** series). An apparently medium-accuracy feature set (***FC***) shows itself as performing below a baseline (***MeanStd***) in the entropy diagram, but its combination outperforms any other (***MeanStdSk+FC***). The entropy triangle explains underperformance suggesting that the classifier using that feature set is specializing in some classes, to the detriment of overall classification.

5. Acknowledgements

This work was supported by the Spanish Government projects 2008-06382/TEC and 2008-02473/TEC, and grant TSI-020110-2009-103, and Comunidad Autonoma de Madrid regional project CCG10-UC3M/TIC-5570.

6. References

[1] A. Temko and C. Nadeu, "Classification of acoustic events using SVM-based clustering schemes," *Pattern Recognition*, vol. 39, pp. 682–694, 2006.

[2] C. Muller, J. Biel, E. Kim, and D. Rosario, "Speech-overlapped acoustic event detection for automotive applications," in *Proceedings of Interspeech'08*, 2008.

[3] S. Chu, S. Narayanan, C. J. Kuo, and M. Mataric, "Where am i? scene recognition for mobile robots using audio features," in *ICME'06*, 2006.

[4] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *ICME'05*, 2005.

[5] X. Zhuang, X. Zhou, T. Huang, and M. Hasegawa-Johnson, "Feature analysis and selection for acoustic event detection," in *Proceedings of ICASSP'08*, 2008.

[6] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro, "Non speech audio event detection," in *Proceedings of ICASSP'09*, 2009.

[7] W.-T. Chu, W.-H. Cheng, J.-L. Wu, and J. Y. jen Hsu, "A study of semantic context detection by using svm and gmm approaches," in *ICME'04*, 2004.

[8] A. Ben-David, "A lot of randomness is hiding in accuracy," *Engineering Applications of Artificial Intelligence*, vol. 20, no. 7, pp. 875—885, 2007.

[9] V. Sindhwani, S. Rakshit, D. Deodhare, D. Erdogmus, J. Principe, and P. Niyogi, "Feature selection in MLPs and SVMs based on maximum output information," *IEEE Transactions on Neural Networks*, vol. 15, no. 4, pp. 937—948, 2004.

[10] F. J. Valverde-Albacete and C. Peláez-Moreno, "Two information-theoretic tools to assess the performance of multi-class classifiers," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1665–1671, 2010.

[11] M. Meila, "Comparing clusterings—an information based distance," *Journal of Multivariate Analysis*, vol. 28, pp. 875–893, 2007.

[12] D. J. Hand and R. J. Till, "A simple generalisation of the Area Under the ROC Curve for multiple class classification problems," *Machine Learning*, vol. 45, pp. 171–186, 2001.

[13] D. J. Hand, "Measuring classifier performance: a coherent alternative to the area under the ROC curve," *Machine Learning*, vol. 77, no. 1, pp. 103–123, 2009.

[14] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul 2009.

[15] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, "Temporal feature integration for music genre classification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1654–1664, 2007.

[16] (2011) The ShATR multiple simultaneous speaker corpus. [Online]. Available: http://www.dcs.shef.ac.uk/spandh/projects/shatrweb/index.html

[17] FBK-Irst database of isolated meeting-room acoustic events. ELRA Catalog number S0296.