Hierarchical Clustering Experiments for Application to Audio Event Detection

Thomas Pellegrini[1]*, José Portêlo*[1]*, Isabel Trancoso*[12]*, Alberto Abad*[1]*, Miguel Bugalho*[12]

[1]INESC-ID Lisboa, Portugal
[2]IST, Lisboa, Portugal
thomas.pellegrini@l2f.inesc-id.pt

Abstract

In previous work, it has been shown the feasibility of using an isolated sound effect corpus to train Audio Event Detectors (AED) for *real life* data. Thus, one can avoid the time-consuming task of manually annotating large amounts of movies, documentaries, TV shows or any other kind of data of interest. However, obtaining a quality sound effect corpus is still a tough task particularly when a large number of acoustic events is considered. In this case, unsupervised techniques able to classify semantic concepts can be very useful to avoid as much as possible the need for listening to the audio samples. In this paper, preliminary experiments involving hierarchical clustering of sound patterns are described. Both intra- and inter-concept clusterings have been carried out, at pattern level and at concept level by using means and variances of the set of selected features. Parameters of single mixture Gaussian models have also been used to identify audio concept similarities. Clusters of concepts with strong similarities between them have been obtained, both from the perceptual point of view and the semantic point of view. Additionally, these results are planned to be used to design an AED system that would have a hierarchical architecture.

1. Introduction

This paper describes preliminary work done in the framework of the VIDIVIDEO European project, whose goal is to substantially enhance semantic access to video, implemented in a search engine, by detecting instances of audio, visual or mixed-media content. The Audio Event Detection (AED) task consists of detecting all kinds of audio events, for example an emergency car passing by, a gun shot, animal cries, etc. Since a main research direction in the project is to try to model many distinct semantic concepts [1], hierarchical clustering appeared to be a suitable approach to avoid the time-consuming task of listening to all the audio samples, used to train detectors.

Hierarchical clustering (HC) is also called *agglomerative clustering* in the case of iteratively merging patterns into bigger clusters, or *divisive clustering* in the inverse case. It tries to discover structure in a data set without *a priori* knowledge, so that data items within each cluster are more closely related to one another than to items assigned to other clusters, in terms of a predefined distance [2, 3]. For instance, HC has been used in speech recognition to cluster texts collected from the Web, to build topic dependent text corpus, without supervision [4, 5, 6]. Many variants of the clustering algorithm exist, but no method is known to perform universally better than others [3]. By using the agglomerative approach, once the clustering is finished, all patterns will be clustered into a single cluster. The clusters that best describe the structure of a data set can be identified by stopping the clustering process when the model is satisfying. In our case, clustering was performed entirely, and clusters have been identified by cutting dendrograms at heights that gave "natural" results.

HC is commonly used to build a taxonomy of classes. In [7], hierarchical taxonomies for musical instrument classification are automatically derived from clustering. HC can also be used to detect outliers. In our case, outliers are sound patterns that are not representative of a semantic concept. For example, a Woodpecker that would peck instead of sing should be removed from the training set of the *bird* concept, if this concept is supposed to describe only singing birds.

In the literature, the sound detection and identification system named SOLAR [8], and systems for musical instrument classification [9, 10], use a hierarchical architecture. Large categories of sounds are considered first, for example sustained and non-sustained sounds, and then more precise classes are detected. In this study, agglomerative clustering aimed also at defining without knowledge the sound classes that could be used in a hierarchical detection system.

After describing briefly the data set and the features used in this study, intra- and inter-concept clustering results are given, by showing dendrograms. Then clustering results on means and variances of one mixture Gaussian models, trained on all sound patterns of every concept are analyzed. Finally, a brief description of how these results could be used to design a hierarchical AED system is given.

2. Data description

All the sound patterns are part of a subset of a sound effect corpus provided by B&G, one of the partners of the VIDIVIDEO project. Sound effects differ from real sounds found in videos, in that they are isolated audio events, and quite homogeneous from the beginning to the end of each file. In real videos, audio events are very often simultaneous, such as singing birds with sea noise in the background for example. Typically, sound effects are also recorded with very good acoustic conditions, which is not always the case in common video recordings. Nevertheless, the use of sound effects allows to avoid the morose, time-consuming and expensive task of manual labeling videos. Furthermore, previous detection experiments showed the feasibility of this approach, by successfully detecting birds, machines, traffic, water and steps, with training one-against-all detectors with a pilot sound effect corpus [1].

All the sound effect files are sampled at 44.1kHz, although the original sampling frequency was not always that high. In this study, twenty-three semantic "concepts" have been used. Table 1 shows the number of patterns and the total duration, for each concept, excluding the silence (very low energy) frames. As can be seen, some concepts are less represented in comparison to others. More than one hour of data is available for *water*, whereas there are only a few minutes for both types of ring-

tones (*analogic* and *digital telephone*).

The concepts *bird, cat, chicken, dog* consist of vocal sounds of these animals, whereas *horse* is comprised of the noise produced by walking or trotting horses. *Applause* stands for crowd applause. *Siren* combines different kinds of emergency and police car sirens, modern and old ones. Many distinct water sounds are assigned to the *water* concept: heavy rain on asphalt, water falls, fountains, rivers, etc.

Concepts	Abb.	Nb of patterns	duration
jet airplane	jet	18	11min 16s
propeller airplane	pro	39	19min 13s
applause	ap	20	11min 3s
bird	bir	62	41min 22s
bus	bus	23	28min 46s
car	car	64	32min 0s
cat	cat	26	5min 2s
chicken	chi	11	2min 40s
dog	dog	30	13min 11s
fire	fir	35	45min 57s
gun	gun	74	10min 14s
helicopter	hel	17	11min 14s
horn	ho	53	6min 25s
horse	hor	57	29min 30s
insect buzz	buz	19	17min 10s
insect chirp	ch	22	32min 0s
siren	sir	32	9min 58s
analogic telephone	bell	11	3min 26s
digital telephone	dig	9	2min 1s
thunder	thu	34	6min 39s
traffic	tra	21	43min 35s
typing	typ	21	46min 50s
water	wat	48	65min 42s

Table 1: Number and durations of patterns given for each of the 23 concepts. Column 2 gives the abbreviations used in tables and figures.

3. Feature set

As a feature set, 19 Perceptual Linear Prediction (PLP) coefficients and their derivatives, Zero Crossing Rate (ZCR), Brightness (BN) and Bandwidth (BW), that are respectively the first and second order of spectrograms, have been used. All the 41 features have been computed within 0.50s frames with a 0.25s shift. For the clustering at pattern-level, with results given in sections 4.1 and 4.2, feature means and variances over all signal duration are used. At concept-level, the clustering, with results given in the second part of section 4.2, is computed over means and variances of single mixture Gaussian models trained over frame-based features.

4. Experiments

This section illustrates the clustering results with the help of dendrograms, which consist of many 'U-shaped' lines connecting objects in a hierarchical tree. The height of a 'U' represents the distance between the two connected patterns or clusters.

At each iteration of the clustering, two patterns or clusters of patterns are merged, and the distance between the new cluster and the other patterns need to be updated. We chose to update distances using the group averaged link method [2]. With this

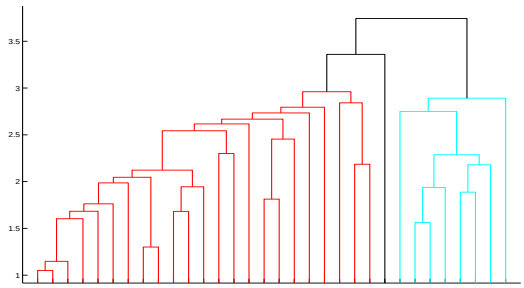

Figure 1: Dendrogram for the *siren* concept. Y-axis represents cluster distances.

method, the distance $d(A, B)$ between two clusters A and B, given in equation 1, is the mean of the distances between every pair of patterns $\mathbf{x}$ and $\mathbf{y}$, with one pattern from each cluster.

$$d(A, B) = \frac{1}{\text{card(A)card(B)}} \sum_{\mathbf{x}\in A, \mathbf{y}\in B} d(\mathbf{x}, \mathbf{y}) \qquad (1)$$

Various distances $d(\mathbf{x}, \mathbf{y})$ have been tested: two Euclidean distances, the weighted Gaussian distance used to cluster Hidden Markov Model states characterized by multivariate single mixture Gaussians in the HTK toolkit ([11], p.268), and the common Mahalanobis distance. The first Euclidean distance, given in equation 2, computes the Euclidean distance between the mean vectors μ_x and μ_y, and will be called the "classical" Euclidean distance in this article. Used in section 4.2 only, a variant takes into account the standard deviation, as if it were a parameter similar to the means. This second Euclidean distance, called "modified" Euclidean distance, is given in equation 3, with σ_x and σ_y being the standard deviation matrices. It should be noted that in this study, only diagonal covariance matrices are used.

$$d(x, y) = (\mu_{\mathbf{x}} - \mu_{\mathbf{y}})^t(\mu_{\mathbf{x}} - \mu_{\mathbf{y}}) \qquad (2)$$

$$d(x, y) = (\mu_{\mathbf{x}} - \mu_{\mathbf{y}})^t(\mu_{\mathbf{x}} - \mu_{\mathbf{y}}) + (\sigma_{\mathbf{x}} - \sigma_{\mathbf{y}})^t(\sigma_{\mathbf{x}} - \sigma_{\mathbf{y}}) \qquad (3)$$

To avoid large feature values (typically ZCR, BW and BN values) swamping the small ones (typically the PLP and PLP derivative values), a linear data scaling to the range $[0, 1]$ is performed before clustering. Input values feeding the HC algorithm are equalized by a scaling coefficient that is different for each feature: for a feature value x_i, its equalized value is $(x_i - m_i)/(M_i - m_i)$, where m_i and M_i are respectively the minimum and maximum values of the i^{th} feature to which x_i belongs to.

The various distances gave different clusters. The two Euclidean distances gave the most satisfying clusters, in a semantic and perceptual point of view, thus only the results achieved with these two distances will be reported here.

4.1. Intra-concept clustering

Intra-concept clustering has been performed to see if characterizing concept examples with only feature means and variances is precise enough to identify different types of sounds within a

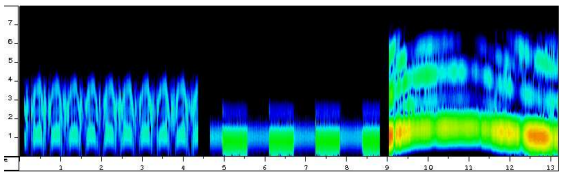

Figure 2: Spectrograms of the siren types I, II, III.

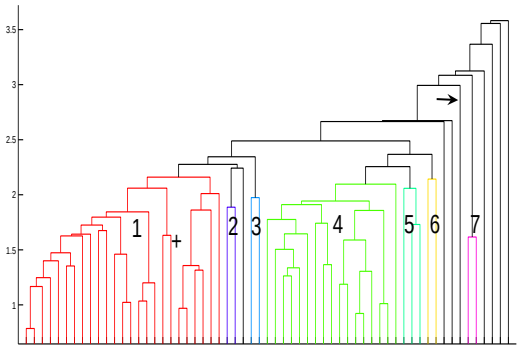

Figure 3: Dendrogram for the *bird* concept.

same concept. As an illustration, clustering results for the *siren* and *bird* concepts will be presented.

As shown in Table 1, the *siren* concept includes 32 patterns, totaling almost ten minutes of non low-energy audio signal. These patterns come from different types of police car, fire truck, ambulance and other types of emergency vehicle sirens. Figure 1 shows the dendrogram obtained for this concept. Two distinct clusters, colored in red and blue, emerge from this classification. By listening to the audio examples, three types of sirens, named type I, II and III, can be identified. Spectrograms of three representative excerpts are given in Figure 2. Type I corresponds to two tone sirens, like those of old emergency cars. Type II are characterized by a fast continuous variation between two tones, and type III consists of slow continuous variations up and down in the frequency spectrum, like some typical USA police car sirens. Perceptually, types I and II are somehow similar. The red cluster in Figure 1 actually re-groups sound patterns of these two types, whereas the blue cluster corresponds to type III. Between the two clusters, stands an isolated pattern. It corresponds to a two-tone siren of a car passing by very fast, with a strong 'Doppler effect' that makes the frequency spectrum unique among the 32 examples. This pattern may be considered as an outlier and may be left outside the training corpus. More generally, the last items to be merged are potential outliers.

Figure 3 shows the dendrogram of the *bird* concept. Several clusters have been highlighted with colors, and numbered. Clusters 1, 2, 3 (red, dark blue and blue) correspond to low-pitched bird singing like Crows and Ducks. For example cluster 2 re-groups two Squawk samples. The other clusters include higher pitched sounds, like Robins, or Black Birds, for which the three samples of our database were clustered together (the light blue cluster, number 5).

Among the 62 sound samples of this concept, there are at least two true outliers: a pecking Woodpecker with no singing (pointed by an arrow in Figure 3) and the sound of a sea wave

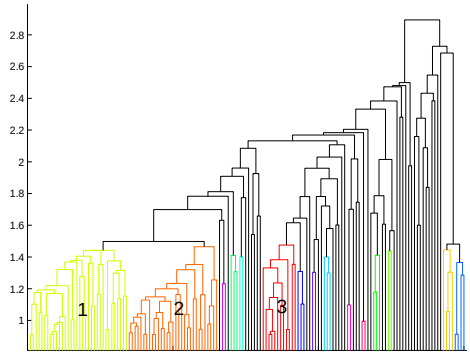

Figure 4: Dendrogram for the 23 concepts, limited to the 150 highest nodes.

(indicated by a '+' sign). These two outliers should be clustered with the biggest distances with the other patterns. The pecking Woodpecker is indeed clustered among the last ones, but it is not the case for the sea wave sound. It has been clustered with a pattern of a singing bird with a continuous noise of wind in the background, that could explain the similarity between both sounds. This example shows the limit of considering only the last patterns to be merged corresponding to the black lines in the dendrograms, as potential outliers. Furthermore, the pecking Woodpecker is not the last one to be merged. Some singing bird patterns were found less similar to the other sounds than the Woodpecker pattern, like cluster 7 for instance, in violet, which clustered two Owl samples.

4.2. Inter-concept clustering

At pattern level

With all the 746 concept patterns listed in Table 1, the entire dendrogram shows 745 distinct nodes, that make the dendrogram very "crowdy" and thus not easy to analyze. Therefore, it is interesting to look at different heights of the dendrogram to try to identify clusters.

Figure 4 shows the highest 150 nodes of the entire dendrogram. We chose this threshold since two big clusters, shown in yellow and orange in the left lower corner, number 1 and 2, are well distinguishable. Cluster 1 and 2 totalize respectively 14.6% and 66.9% of all the patterns. The other colored clusters are smaller, clustering from two to ten patterns of the same concept. For example, cluster number '3' in the figure (in red), is comprised of 10 bird patterns, totalizing 16% of the total number of bird patterns.

Cluster 1 (yellow) merges patterns from the *bird*, *cat*, *dog* and *siren* concepts, with respective percentages 40.3%, 19.2%, 90.0% and 25.0% of patterns for each concept. Cluster 2 merges patterns from all the other concepts, except telephone patterns, with very high percentages (>95%) for *airplane*, *applause*, *bus*, *gun*, *helicopter*, *horse*, *traffic* and *water* concepts. The low percentages observed for cluster 1 may indicate that the likeliness between the involved concepts is limited. Cluster 2 is much more representative of its involved concepts. The presence of horse patterns in cluster 2 is natural since they represent walking or trotting horse sounds, and may be similar to some applause, helicopter or even gun sounds. Some bird and siren patterns

are also present in this cluster, with respective percentages of 11.3% and 28.1%. This shows the limit of our representation involving only means and variances. Also, the cluster analysis depends on the height where the dendrogram is cut. Looking at smaller heights, one of the node includes 159 patterns (32.6%) in total, with patterns from the *airplane*, *bus*, *helicopter*, *horse* and *traffic* concepts, which is more satisfying from a perceptual or semantic point of view.

Even if the two clusters are very general since many concepts are involved, the clear distinction between both clusters seems rather natural, since sounds from the two clusters are very different. Cluster 1 merges sounds that have a pitch, such as animal cries and sirens, whereas cluster 2 merges sounds mainly produced by engines. For a hierarchical AED system architecture, it would be useful to have a first step that differentiates these two categories.

At concept level

In the experiments reported above, feature means and variances were computed for each sound example. We report now an experiment where we trained Gaussian mixture models (GMM) with only one mixture, with one model for each concept, using the Torch library [12]. There are 41 features so 41 means and variances are estimated. For each concept, a vector containing the 82 parameters is used as a representation of the concept. The 23 vectors fed the same hierarchical clustering algorithm as used previously.

We have used a GMM training tool because initially we wanted to use likelihoods to do the clustering. So far no concluding results have been found with likelihoods, and this is still ongoing work. Since the training of single mixture GMM consists in a likelihood maximization, it is equivalent to simply compute means and variances over all sound patterns of each concept.

Figure 5 shows the dendrogram for the 23 concepts, with the use of the classical Euclidean distance. The first important cluster, has been highlighted in red. It is comprised of the *bus*, *traffic*, *car*, *propeller airplane*, *helicopter*, *jet airplane*, *horse*, *fire*, *thunder*, *insect buzz*, *applause*, *water*, *typing* and *gun* concepts. There is also a second interesting cluster, which involves bigger distances, and which merges the *dog*, *chicken* and *horn* concepts. The first cluster, which is very general since it involves many concepts, corresponds mainly to sounds produced by engines or by mechanical movements, such as flying insects or walking horses. These sounds can be viewed as non-pitched sounds. Figure 5 shows a clear distinction between non-pitched sounds, produced by an engine (buses, airplanes, etc.) or by a mechanical behavior (horse walking, water falling), and pitched sounds like animals cries, siren or ring-tones. The most similar concepts that have been found are *bus*, *traffic* and *car*. These concepts are similar from a semantic point of view, and a generic concept may be useful. The first big cluster also contains another meaningful cluster, which is *applause*, *typing* and *water*.

Finally, *digital telephone* ring-tones, and *sirens* form an isolated cluster, nevertheless with a high clustering distance of all the clusters.

Figure 6 gives the dendrogram found with the modified Euclidean distance. Clusters are globally similar to those of figure 5, nevertheless there are interesting differences. The *helicopter* concept is now clustered to *propeller airplane*, which is more satisfying than *jet airplane*. The *dog*, *chicken* and *horn* concepts form a cluster, as in the previous experiment, but in a distinct clustering order. The *dog* concept has been found closer to *horn* than to *chicken*.

FDR rank	*Features*
1-6	PLP: 1, 3, 2, 0, 6, 4
7-8	BR, BW
9	PLP: 5
10	ZCR
11-19	PLP: 7, 15, 17, 9, 8, 16, 13, 12, 11

Table 2: Fisher Discriminative Ratio (FDR) ranks and the corresponding features.

The salience of the acoustic features is evaluated by computing the Fisher Discriminant Ratio (FDR) given in Equation 4. This ratio is evaluated for each feature i of the 41 features in total, and for each pair of concepts C_m, C_n. Terms μ_{i,C_m} and σ^2_{i,C_m} are respectively the concept mean and variance values of feature i, for concept C_m. The larger the FDR is, the more the feature discriminates the two concepts being compared.

$$FDR(C_m, C_n, i) = \frac{(\mu_{i,C_m} - \mu_{i,C_n})^2}{\sigma^2_{i,C_m} + \sigma^2_{i,C_n}} \quad (4)$$

The ratios were computed over the 253 possible pairs of concepts (there are 23 distinct concepts in total) and the feature that yields the biggest ratio is considered to be the more salient feature for the pair of interest. Table 2 shows the most salient features, by showing their rank determined by counting the number of times they had the biggest FDR. For example, the PLP coefficient number 1 (Energy has number 0) appeared to have the biggest FDR for 60 comparisons of distinct concept pairs, so that this feature has rank one. The six most salient features correspond to the first PLP coefficients, followed by brightness (BR), bandwidth (BW) and zero-crossing rate (ZCR). It is interesting to notice that no derivative PLP coefficient appeared in the ranking, which may be due to the large frame interval used to compute the features (250ms).

5. Application to Audio Event Detection

Having identified consistencies and confusions among the concepts, a hierarchical AED system could be designed. This approach seems very interesting, especially if only little training data is available for some concepts.

In a previous study [1], we have used one-against-all classifiers, that give a binary decision for each concept at a frame level. By merging training data of concepts that were found to be similar with the HC, such as *bus*, *traffic*, and *car* for example, a detector trained with this "multi-concept" data, may be more robust. Once a sound to be identified has been detected by the *bus*, *traffic*, *car* detector, three classifiers could be used to assign one of the three concepts to the sound.

In the previous reported experiments, a rough separation between "non-pitched" sounds (concepts: *bus*, *traffic*, *car*, etc.) and "pitched" sounds (concepts: *bird*, *dog*, *chicken*, etc.) has been identified (see Figures 5 and 6). In a hierarchical AED system, a pitched/non-pitched classifier could be trained with these two groups of concepts. After this first distinction, if a sample to be identified was found to be "non-pitched", the next decision would be to classify the sample between the *gun* concept, that stands alone in the red cluster of Figure 5, and all the other "non-pitched" concepts. Then, if for example, this example was not classified as part of the *gun* concept, the next de-

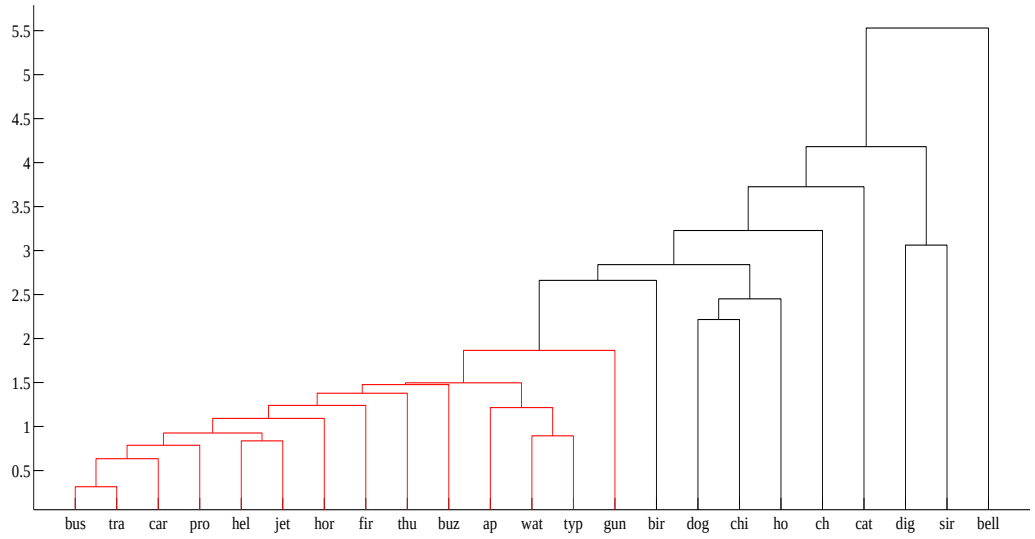

Figure 5: Dendrogram for the 23 concepts, achieved with the classical Euclidean distance. Y-axis represents cluster distances. Abbreviations: tra *traffic*, pro *propeller airplane*, hel *helicopter*, hor *horse*, fir *fire*, thu *thunder*, ap *applause*, wat *water*, typ *typing*, bir *bird*, chi *chicken*, ho *horn*, ch *insect chirp*, dig *digital telephone*, sir *siren*.

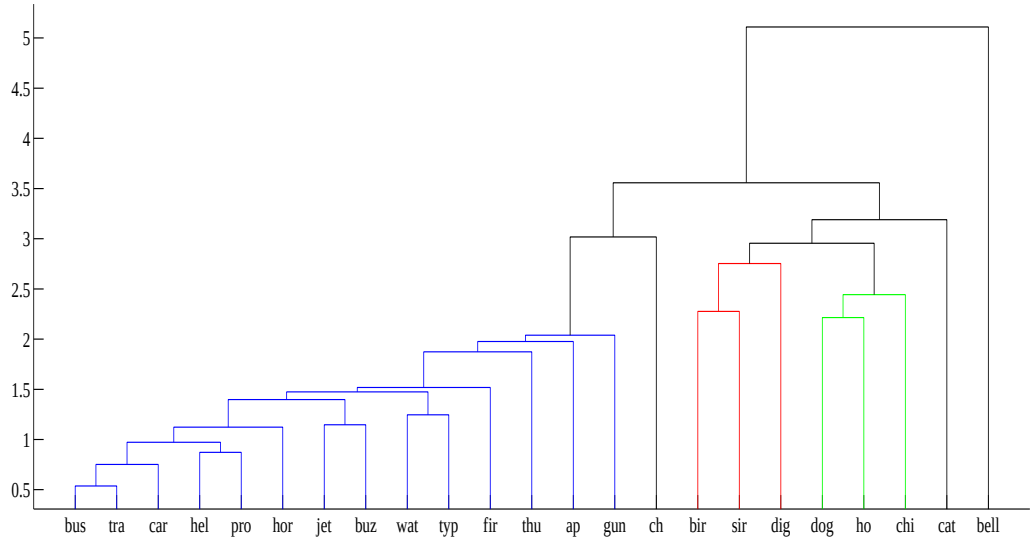

Figure 6: Dendrogram for the 23 concepts, achieved with the modified Euclidean distance. Y-axis represents cluster distances. Abbreviations: tra *traffic*, pro *propeller airplane*, hel *helicopter*, hor *horse*, fir *fire*, thu *thunder*, ap *applause*, wat *water*, typ *typing*, bir *bird*, chi *chicken*, ho *horn*, ch *insect chirp*, dig *digital telephone*, sir *siren*.

cision would be to classify the sample between the cluster *dog*, *chirp* and *horn*, and the remaining ten "non-pitched" concepts, from *bus* to *insect buzz*, etc.

6. Summary and future work

This paper has presented an experimental study assessing the use of hierarchical clustering (HC) to try to identify similarities and dissimilarities between sound samples, represented by common features (PLP, ZCR, etc.), for a task of Audio Event Detection in videos. The experiments were carried out on the sound effect corpus used to train audio event detectors, used in the VIDIVIDEO project.

Even if the sound effect corpus provides sound files with precise titles, there are files that do not correspond to their associated semantic concept. These outliers are not suitable to train detectors or classifiers. HC has been investigated to help finding outliers, to avoid the morose task to listen to each of the audio samples.

Twenty-three different concepts have been chosen, covering very distinct sounds, from cat meowing to jet airplane engines. First, at pattern-level, means and variances of the features computed over each audio signal have been used to try to point out outliers for each concept. These outliers should be among the patterns that are the most distant from all the other patterns, and therefore clustered in the last iterations of the clustering process. Some perceptually similar patterns were clustered together, and some outliers were found to be clustered with the biggest distances between patterns. But the very simple mean and variance representations seems to be insufficient to detect all the outliers, in particular when the sound patterns have a background noise. Second, means and variances over all the signals of each concept have been used to represent the 23 concepts. Various distance measure have been tested, and two Euclidean distances have been found to provide more consistent clusters in a semantic and conceptual sense than other distances like the weighted Gaussian distance or the Mahalanobis distance. Two large groups emerged from the clustering, one "non-pitched" sound cluster (concepts: *bus*, *traffic*, *car*, etc.), and the "pitched" sounds (concepts: *bird*, *dog*, *chicken*, etc.). Perceptually similar concepts have clustered too, like *applause*, *water* and *typing*, or *dog*, *chicken* and *horn*. This information can be used to design a hierarchical detection system.

HC may be useful also to build a 'top-down' hierarchical detection system, which would first differentiate large categories of sounds, such as pitched and non-pitched sounds, and then use more and more specialized detectors to classify sounds with a chosen granularity. As shown in the litterature, this approach should be more robust than directly using specified detectors, since training data drastically lacks for some semantic concepts. Future work will consider audio event classification experiments.

By using only feature means and variances, the time structure of the audio events has not been used. The approach would clearly beneficiate in precision by using the time structure of the sound patterns, for example to distinguish a car passing by from a stopped car. Silence also provides useful information. For instance, ring-tones are not continuous, otherwise they would be very annoying. On the contrary, emergency or security sirens are continuous. In that case, silences between ring-tones could differentiate them from sirens. In that sense, percentages of low-energy frames could be an additional interesting parameter.

The set of features could be also extended. In particular, temporal shape features like attack-time, temporal increase/decrease and effective duration intuitively seem useful for our task. Finally, adding new concepts and more training patterns is also part of our future work.

7. References

[1] I. Trancoso, J. Portêlo, M. Bugalho, J. Neto, and A. Serralheiro, "Training audio events detectors with a sound effect corpus," in *Proceedings of Interspeech'08*, Brisbane, 2008.

[2] K. Cios, W. Pedrycz, and L. Swiniarski, Kurgan, *Data mining, a knowledge discovery approach*. Springer, 2007.

[3] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. John Wiley and Sons, 2000.

[4] R. Iyer, M. Ostendorf, and H. Gish, "Using out-of-domain data to improve in-domain language models." in *Technical Report ECE-97-001*, Boston University, 1997.

[5] S. Sekine, A. Borthwick, and R. Grishman, "NYU language modeling experiment for 1996 CSR evaluation," in *Proceedings of the DARPA Speech Recognition Workshop*, 1997.

[6] F. Weng, A. Stolcke, and A. Sankar, "Hub4 Language Modeling Using Domain Interpolation and Data Clustering." in *Proceedings of the DARPA Speech Recognition Workshop*, 1997.

[7] S. Essid, G. Richard, and B. David, "Instrument Recognition in Polyphonic Music Based on Automatic Taxonomies," *IEEE Transactions on Speech, Audio and Language Processing*, vol. 14:1, pp. 68–80, Jan. 2006.

[8] D. Hoiem, Y. Ke, and R. Sukthankar, "SOLAR: sound object localization and retrieval in complex audio environments," in *Proceedings of ICASSP*, vol. 5, Philadelphia, 2005, pp. 429–432.

[9] G. Peeters and X. Rodet, "Hierarchical Gaussian Tree with Inertia Ratio Maximization for the Classification of Large Musical Instrument Databases," in *Proceedings of DAFX03*, London, 2003.

[10] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music," in *Proceedings of ICASSP*, Philadelphia, 2005, pp. 245–248.

[11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, and G. Morre, *The HTK Book for HTK Version 3.4*. Cambridge University, 2006.

[12] R. Collobert., S. Bengio., and J. Mariéthoz, "Torch: a modular machine learning software library," *Technical Report IDIAP-RR 02-46*, 2002.