knowledge

MDPI

Article

Extending Radio Broadcasting Semantics through Adaptive Audio Segmentation Automations

Rigas Kotsakis [1] and Charalampos Dimoulas [2,*]

[1] Department of Information and Electronic Engineering, International Hellenic University, 57001 Thessaloniki, Greece; rkotsakis@gmail.com
[2] School of Journalism & Mass Communications, Aristotle University, 54124 Thessaloniki, Greece
* Correspondence: babis@eng.auth.gr

Abstract: The present paper focuses on adaptive audio detection, segmentation and classification techniques in audio broadcasting content, dedicated mainly to voice data. The suggested framework addresses a real case scenario encountered in media services and especially radio streams, aiming to fulfill diverse (semi-) automated indexing/annotation and management necessities. In this context, aggregated radio content is collected, featuring small input datasets, which are utilized for adaptive classification experiments, without searching, at this point, for a generic pattern recognition solution. Hierarchical and hybrid taxonomies are proposed, firstly to discriminate voice data in radio streams and thereafter to detect single speaker voices, and when this is the case, the experiments proceed into a final layer of gender classification. It is worth mentioning that stand-alone and combined supervised and clustering techniques are tested along with multivariate window tuning, towards the extraction of meaningful results based on overall and partial performance rates. Furthermore, the current work via data augmentation mechanisms contributes to the formulation of a dynamic Generic Audio Classification Repository to be subjected, in the future, to adaptive multilabel experimentation with more sophisticated techniques, such as deep architectures.

Keywords: audio semantics; content analysis; radio broadcasting

Citation: Kotsakis, R.; Dimoulas, C. Extending Radio Broadcasting Semantics through Adaptive Audio Segmentation Automations. *Knowledge* **2022**, *2*, 347–364. https://doi.org/10.3390/knowledge2030020

Academic Editor: Gabriele Santoro

Received: 25 June 2022
Accepted: 15 July 2022
Published: 18 July 2022

1. Introduction

The remarkable progress of web technologies and multimedia services has promoted the quick, easy and massive interchange of multimodal digital content. The involvement of plural heterogenic resources and customized user preferences require applicable content description and management mechanisms [1]. Moreover, content recognition, semantic interpretation and conceptualization attempts are currently being deployed, thus generating further difficulties and challenges, especially for time-based media (TBM) [2]. From the media organization point of view, media assets management (MAM) automation and intelligent multimedia processing technologies are needed for proper content archiving with optimum exploitation of both human resources and infrastructures, thus facilitating content reuse scenarios and audiovisual production in general. The same applies to the individual producers, the freelancers that are involved in the media and the contributors to the so-called user-generated content (UGC). In fact, their needs in audiovisual content management and archiving services are even harder to meet [3], considering that they do not usually have at their disposal professional MAM software equipped with radio and audiovisual broadcasting automation utilities. Considering the media consumer site, content classification, summarization and highlighting are pursued for multimedia content indexing, searching and retrieval automation. New trends regarding content recognition refer to topic classification, story understanding and/or enhanced semantic interaction, thus requiring adaptive audiovisual feature extraction and selection engines along with machine learning methods. These services rely on the utilization of extended multivariate

databases, usually demanding applicable content annotation and/or semantic tagging. However, there are issues regarding the inhomogeneity of labeling meta-data, while in some cases, ground-truth training pairs are difficult to obtain (or are even completely unavailable). Hence, combinations of supervised, semi-supervised and unsupervised data mining algorithms are utilized to serve the specific necessities of various real-world multimedia semantics [1,4–9].

Sound recognition plays an important role in most of the encountered audio and audiovisual pattern analysis cases, where related content is massively produced and uploaded (i.e., digital audio broadcasting, podcasts and web radio, but also video on demand (VoD), web-TV and multimodal UGC sharing in general). Specifically, there are various pattern recognition and semantic analysis tasks in the audio domain, including speech–music segmentation [8], genre recognition [10], speaker verification and voice diarization [11], speech enhancement [12,13], sound event detection [14], phoneme and speech recognition [15–17], as well as topic/story classification [18–20], sentiment analysis and opinion extraction [4,5,21], multiclass audio discrimination [22], environmental sound classification [23] and biomedical audio processing [24]. Audio broadcast is generally considered to be one of the most demanding recognition cases, where a large diversity of content types with many detection difficulties are implicated [1]. In addition, audio broadcasted content can be easily accessed, while new productions are massively and continuously created and uploaded/distributed. Hence, smart multi-purpose audio semantic and associated future Web 3.0 services can be built and progressed upon such audio broadcasting scenarios.

The current paper focuses on the investigation of various audio pattern analysis approaches in broadcast audio content. Following the results of previous research [1,25], stress test evaluation procedures and assessment of real-word scenarios are conducted, investigating the impact of the involved feature engines, the windowing configuration and the formulated classification schemes in both supervised and unsupervised strategies. The main target is to highlight the most effective parameterization at each stage of the entire modeling, aiming at assembling hybrid smart systems featuring optimum behavior without excessive computational load and resources demand (like in deep neural architectures).

The rest of the paper is organized as follows. The subsequent section addresses the problem definition and background work, with the corresponding particularities of the current experimental approach in radio productions. The literature state of the art follows, presenting previous research related to the topic under investigation. The implementation section describes the configuration and modeling aspects, including pre-processing actions, definition of classification taxonomies, ground truth data acquisition and feature engine formulation. Thereafter, experimental results of various methods and classification schemes are analyzed with the use of appropriate performance metrics. Finally, validation and optimization aspects regarding the whole implementation are addressed, followed by the discussion and conclusion section.

2. Materials and Methods

2.1. Background Work and Problem Definition

The current work investigates efficient and easy-to-implement adaptive strategies for voice detection, segmentation and classification in audio broadcast programs. Such audio signals usually comprise multiple segments—events, implicating various patterns, such as speakers' voices, phone correspondences, recorded reports, commercial jingles and other sound effects. Thus, efficient treatment and management of the broadcasted content involve demanding semantic analysis tasks. There are many issues that deteriorate the efficiency of audio recognition, requiring special attention. A common difficulty that must be faced in typical radio programs relies on the temporal overlapping of events and patterns, where music usually coexists with voice components. In addition, background noise and/or reverb contamination (mostly in non-studio recordings) deteriorate the recognition accuracy, while fade in/out operations and similar creative (/mixing) effects further complicate the speech detection task. Other unwanted matters include various recording

and preprocessing artifacts, speech rate variability, pronunciation effects and subjective speech degradation issues in general.

Motivated by the results of a previous research on program-adaptive pattern analysis for Voice/Music/Phone [1] and Language Discrimination taxonomies [6], the presented methodology functions as an add-on module towards the formulation of a dynamic Generic Audio Classification Repository. Hence, following already adopted hierarchical classification strategies, new schemes were adapted based on clustering techniques, but also their combination with supervised training methods. In this context, semi-supervised hierarchical and hybrid pattern recognition systems are proposed for light-weighted speech/non-speech segmentation, noise detection and further discrimination of male/female voices, independently of the involved speakers. Several experiments were conducted for the determination and validation of adaptive audio feature engines at every classification level of the involved hierarchies. Another issue that is addressed in the current work is the investigation of the impact of the temporal segmentation accuracy in the overall classification performance. Several stress tests were conducted in this direction, using different window lengths and segmentation resolution, offering windowing efficiency insights.

Semantic analysis procedures in radio content mainly involve voice detection, speech recognition and speaker identification tasks. Machine learning approaches based on clustering techniques that determine speech/non-speech frames were implemented for voice activity detection via Gaussian mixture models, Laplacian similarity matrices, expectation maximization algorithms, hidden Markov chains and artificial neural networks [26–30]. A more specific and interesting audio pattern that can be detected in audio signals, i.e., in broadcast programs, refers to phone line voices, due to the contained particular spectral audio properties [1,25,26].

It must be noted that an extended study was carried out in [1] for content analysis and description purposes of broadcast radio programs, aiming at the formulation of adaptive classification taxonomies (speech/non-speech and voice/music discrimination, speaker verification, noise detection) with the utilization of various direct, hierarchical and combined hybrid schemes implementations. In this direction, efficient annotation and segmentation processes were applied in the radio signals formulating the ground truth database for the subsequent corresponding semantic analysis. During supervised classification, based on the developed taxonomies, several algorithms were employed from the statistical domain (i.e., linear, logistic regressions), decision trees (i.e., J48 tree models), support vector machine techniques (i.e., SMO) and artificial neural network modeling. The comparison between the respective classification performances indicated that neural network implementations provided the highest discrimination efficiency in almost all cases/schemes. Moreover, a thorough feature evaluation was conducted in [25] to investigate the saliency of an initial augmented extracted audio feature set. Ranking algorithms were employed, aiming at the detection of the most efficient feature subsets for classification purposes, based on the above speech/non-speech segmentation and speaker discrimination taxonomies. In this context, the current paper aims to extend the previously conducted work, trying to integrate unsupervised classification potentials via clustering techniques in the content of radio programs. Indicative comparisons of the previously and currently employed classification methods are presented in the following sections.

2.2. Proposed Framework

Figure 1 presents the proposed methodology for the program-adaptive classification problem. In real-world cases, the sound source could be consisted of a short length broadcasted signal (for example 10 min duration), which is thereafter implicated in semantic analysis tasks, based on pattern recognition operations, aiming to formulate efficient/automated content description/labeling mechanisms for archiving purposes. As mentioned above, the conducted work/experiments investigate the feasibility of the proposed architecture via supervised classification and clustering strategies, mainly in voice content. The audio signal that triggers the initiation of the process in Figure 1 could derive

from radio streams, either traditionally broadcasted or hosted on web radio platforms. Taking into consideration multiple speakers' voices coexisting in different radio shows with differentiated characteristics/structure, it is anticipated that the initially formed Ground Truth Repository could not function efficiently towards multilayer classification, due to reduced data acquisition. On this basis, the current work proposes an initial experimentation with a small input dataset of a specific radio stream in order to examine the potential voice discrimination rates. Thereafter, the Ground Truth Repository will be gradually/iteratively augmented with other instances of the same radio program (therefore retaining the same content characteristics/structure), reinforcing the confidence in the classification results. In this context, each group of records of the same radio program functions as a sub-(ground truth) dataset in the Generic Repository, justifying the adaptive character of the proposed framework (dotted lines in Figure 1). The same operation is followed for other radio shows along with their respective diversified instances, in an iterative way, leading to effective data augmentation of the Generic Ground Truth Repository, which subsequently can be utilized for experimentation with more sophisticated deep learning strategies. As anticipated, the most crucial step (and demanding one) towards the feasibility of the proposed architecture has to do with the classification effectiveness in the initial reduced duration radio stream. The aforementioned topic/step constitutes the main research objective of the current work, namely, to investigate if traditional light-weighted machine learning methods could support efficient discrimination rates before proceeding into more complex and with increased computational load methods on augmented data.

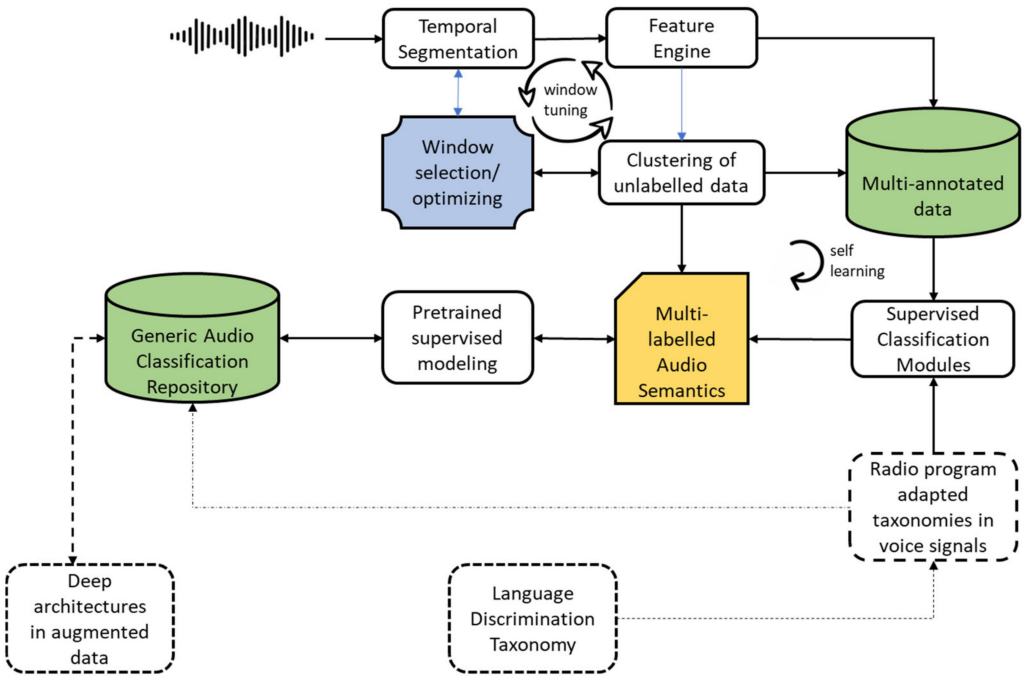

Figure 1. Block diagram of the proposed methodology.

Analyzing Figure 1, the block diagram of the suggested framework is initiated with the presence of the aforementioned reduced duration radio signal. Thereafter, the content is subjected to multiple looping operations for window selection/tuning and feature extraction. At this point, the experimentation could begin with unsupervised/clustering methods without any data annotation, based on hierarchical taxonomies for voice discrimination (the classification schemes will be discussed in next section). The labeling procedure based on specific radio program adaptation leads to confident supervised machine learning modeling in the same taxonomies. Both classification strategies are tested separately and combined in multilayer/hierarchical discrimination schemes, given the investigative character of the presented work. It must be noted that the dataflow withholds an iterative

character for the determination of optimized windowing in terms of classification rates of stand-alone and combined machine learning topologies in every layer of the hierarchical framework. The multivariate experimentation in window lengths, feature engine and classification algorithms, along with meaningful results extraction, could thereafter lead to the formulation of pretrained models in the Generic Repository before testing in future implementations (i.e., deep architectures).

2.3. Data Collection—Content Preprocessing

For the demands of the model under development, audio content from different broadcast shows is combined and transcoded in the common uncompressed PCM WAV format (16 bit, 44,100 Hz). During the conducted experiments, the stereo information is not taken into consideration, since the differentiations in channels properties have been thoroughly studied in previous work [26]. Moreover, the selected audio content involves the typical audio patterns in radio productions, namely, speaker voices, different kinds of noise and music, and phone conversations are also included.

Thereafter, the synthesized audio file is segmented into smaller audio frames. In order to quantitatively investigate the impact of the time windows duration, especially in the unsupervised classification performance, four different temporal lengths are employed: 1000 ms, 500 ms, 250 ms and 125 ms. Table 1 presents the formulated audio database, including the label of the samples, of the 3.5 min (210 s) synthesized audio file.

Table 1. Population of annotated audio samples.

Window Length	1000 ms	500 ms	250 ms	125 ms
P	100	200	400	800
V	300	600	1200	2400
$\mathbf{V_{MV}}$	120	240	480	960
$\mathbf{V_{FV}}$	120	240	480	960
$\mathbf{V_{GV\text{-}GV,\ G=M,\ F}}$	30	60	120	240
$\mathbf{V_{MV\text{-}FV}}$	30	60	120	240
R	200	400	800	1600
M	120	240	480	960
J	40	80	160	320
N	40	80	160	320
Sum	**600**	**1200**	**2400**	**4800**

In Table 1, the phone conversation samples are notated with P, the voice signal with V and the residual non-speech segments with R. The voice signal implicates different male and female voices, V_{MV} and V_{FV}, respectively, while $V_{GV\text{-}GV,G=M,F}$ represent speech overlapping between same gender speakers (G = M, F) and $V_{MV\text{-}FV}$ withholds speech overlapping between different gender speakers, which commonly occur in radio productions. The residual signal includes the music content and the noisy interferences. In the music content (M), different music genres are selected, such as rock, lounge, hip-hop and classical music, with both male and female singers, which are usually heard in radio programs. Moreover, the audio content includes representative jingles of radio programs (J). Finally, noise reductions (N) refer to reverb, hiss effects, silence and other noisy interferences. In this way, the collected data contain all the typical audio patterns in radio productions.

2.4. Classification Taxonomies and Ground Truth Data

As Table 1 exhibits, there are many audio patterns that can be investigated/classified in the audio content of radio broadcasting. An initial experimental procedure was conducted in [1], implicating several direct, hierarchical and hybrid classification schemes

with the utilization of supervised classification algorithms and the subsequent comparison of results. The current work attempts to extend the semantic analysis process for speech detection/discrimination. In this context, three classification schemes are employed in an hierarchical mode, in order to disintegrate the initial complex pattern recognition problem into more efficient layers. Figure 2 presents the classification schemes with the respective audio content labeling. The first layer includes the voice discrimination of speaker and phone conversations (VPR scheme). It must be noted that phone voice is considered as a distinct audio speech signal because of its specific audio and spectral properties, as [1,25] present. In this way, the voice and phone signals can be classified from music, jingles and other noise content (residual signal). The second layer includes a single speaker vs multiple speakers scheme (SM scheme) that attempts to discriminate a speakers' voice from speech overlapping between them. Finally, the subsequent third layer presents the speaker genre diarization problem, aiming to classify male/female voices (MF scheme). It must be stated that the whole semantic analysis is conducted with both supervised and unsupervised classification algorithms, and a combination of them, but the procedure can be also served by the solely automatic clustering process between the layers/schemes.

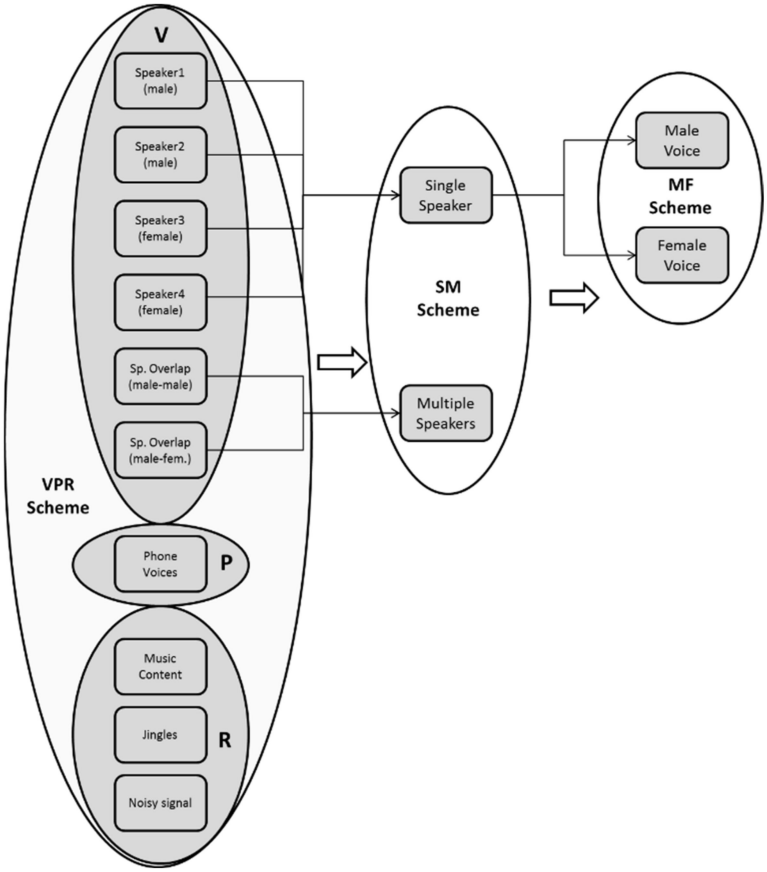

Figure 2. Proposed classification schemes.

According to the previous mentioned discrimination schemes, the annotation procedure assigns the corresponding class to the respective segmented audio frames, with the notation of Table 1. In this context, a ground truth database is formulated only for supervised classification purposes, since unsupervised classification utilizes only the initial non-labeled audio samples for clustering detection. The annotation procedure is also essential in order to evaluate the discrimination rates of the unsupervised automated classification, comparing the clustering structures to the assigned classes of the ground truth formulated database.

2.5. Feature Engine

In the current paper, 90 different features were initially selected and extracted via bibliographic suggestions, empirical observations and trial and error experiments. Hence, we utilize standard spectral audio features that are frequently used in audio content classification, such as spectral centroid (SPc) and its MPEG-7 perceptual version (audio spectral centroid, ASC), audio spectral spread (ASS), spectral flatness measure (SFM), roll-off frequency (RF), bandwidth (BW), spectral irregularity and brightness, spectral flux (SF) and delta spectrum magnitude (DFi) [1,25].

Similarly, popular time–domain audio features are employed (i.e., low short time energy ratio—LSTER; crest factor—CF; logarithmic expressions of normalized recording level, average power and dynamic range—loudness, pav and DR, respectively), in combination with time and frequency–domain signal statistics (audio entropy, RMS energy, temporal and spectral skewness and kurtosis, etc.) [1,25]. In addition, the first thirteen mel frequency cepstral coefficients (MFCCs) are selected due to their increased discriminative power in speech and speaker recognition. Audio envelope thresholding and peak analysis is also performed for the estimation of additional features, such as global and efficient signal to noise ratio (GSNR/ESNR); envelope level that has been exceeded in 85% of the signal length (E85 [dB]); estimation of the total number envelope peaks (nPeaks), including the average and variance values of their magnitudes and their time distances (PKSavr, PKSvar, PKS-Distavr, PKS-Distvar, respectively); peak transition measure (PTM) and its normalized version (nPTM); and estimation of the number of significant samples (nss) exceeding the envelope threshold level of E85, providing also the average and variance of the length of the silent (insignificant) segments (meanL-SP, varL-SP) [1,25]. Finally, spectral bands' comparison features are extracted by means of FFT and 9-level DWT/UWT analyses. Hence, band energy ratios (BER) of low (LF, <250 Hz), medium (MF, 250 Hz–4 kHz) and high frequencies (HF, >4 kHz) to the overall spectral energy formed using the FFT sequences. Similarly, wavelet average power and crest factor sequences of all the k = 10 formed scales of the WT coefficients are estimated (WPav-k, WCF-k), also allowing the extraction of wavelet power centroid (WPc) and time variance (WPcv), the energy ratio of the lowest band to the total energy expressing the significance of the lowest band (WLBsign) and the energy concentration to the highest level wavelet band (WLBconc).

The above initially extracted feature set was engaged and tested in previous experiments in [25], and consequently, useful remarks and conclusions referring to their efficiency emerge from comparisons in the current work. Figure 3 exhibits the main categories of the features.

Another aspect that must be taken into consideration refers to the evaluation of the extracted features, because each audio property contributes at a different level to the discrimination efficiency. Moreover, the exploitation of the whole feature set leads, as it is anticipated, to increased computational load and processing needs and therefore, smaller and efficient subsets are sought in order to resolve these issues. Several experiments have been carried out in [1,25] concerning the saliency and ranking of the feature set while employing supervised classification in different implementations and schemes. The computed cross-correlation matrices and principal component analysis revealed in [25] a feature vector with dimension/rank equal to 36 for supervised classification with artificial neural system implementations. For this reason, the subsequent experiments in the next section employ the salient feature vector that has been determined for supervised classification purposes.

Since unsupervised classification algorithms do not use predefined classes and only investigate clusters of data/values, the whole feature set cannot be evaluated strictly in terms of the implemented scheme (VPR, SM, MF). Consequently, the attributes' saliency is determined by their respective discrimination impact in the classification efficiency. Nevertheless, an initial indicative ranking of the audio properties (the first 30) is presented in Table 2 for each classification scheme, while utilizing the "InfoGainAttributeEval" algorithm on the audio feature values in the WEKA environment. This technique evaluates the importance of each attribute individually by estimating the information gain with respect to the class using entropy measures. Furthermore, in the next section, several subsets of the

initially extracted features are tested on the grounds of their effectiveness in the clustering procedure while employing the unsupervised classification algorithm each time.

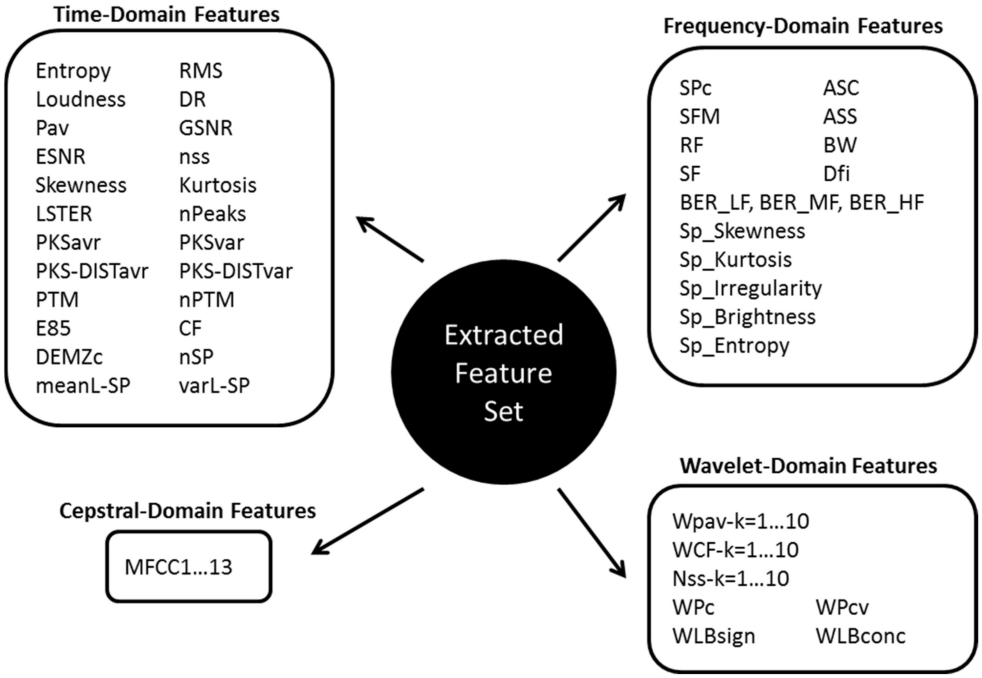

Figure 3. Extracted feature set.

Table 2. Feature evaluation.

	VPR Scheme	SM Scheme	MF Scheme
1	BER_MF	nss-k = 8	Wpav-k = 9
2	Wpav-k = 3	LSTER	MFCC13
3	Wpav-k = 10	nss-k = 7	PKSavr
4	RF	WCF-k = 7	MFCC9
5	SFM	nss-k = 4	RMS
6	BER_LF	WCF-k = 8	Loudness
7	MFCC1	nss-k = 5	Pav
8	Wpav-k = 9	WLBconc	PKS-DISTavr
9	ESNR	PKSvar	CF
10	nss-k = 6	nss	nPTM
11	MFCC2	Sp_Kurtosis	Sp_Kurtosis
12	Wpav-k = 2	WCF-k = 4	PTM
13	Sp_Entropy	E85	GSNR
14	BER_HF	WCF-k = 5	WCF-k = 10
15	Sp_Kurtosis	CF	nss-k = 10
16	ASS	MFCC11	MFCC12
17	WCF-k = 6	Dfi	RF
18	ASC	nss-k = 9	nPeaks
19	Wpav-k = 4	WCF-k = 9	MFCC7
20	E85	nss-k = 6	nss
21	BW	Sp_Skewness	Wpav-k = 2
22	Sp_Skewness	nPeaks	Wpav-k = 4
23	nss-k = 5	MFCC10	BER_HF
24	Wpav-k = 7	WCF-k = 6	MFCC5
25	nss-k = 7	PKS-DISTavr	MFCC2
26	GSNR	DR	nss-k = 2
27	Dfi	DEMZc	Wpav-k = 5
28	Sp_Brightness	ESNR	MFCC8
29	RMS	PKS-DISTvar	WCF-k = 2
30	Pav	Wpav-k = 6	MFCC10

3. Results

3.1. Classification Techniques and Performance Metrics

Several experiments have been conducted in [1] in order to compare supervised classification techniques (decisions trees, artificial neural systems, regressions, etc.) in terms of their overall and partial discrimination efficiencies in various implementations and schemes. One of the most balanced discrimination rates emerged from the employment of artificial neural training for the development of the supervised classifier. Consequently, in the current work, artificial neural systems (ANS) are solely utilized for supervised classification purposes. Several topologies were tested in order to achieve efficient training performance leading to network implementations of two sigmoid hidden layers and an output linear layer, while an approximate number of 20 nodes are engaged in the hidden layers. Furthermore, the k-fold validation method is utilized, dividing the initial audio samples set into k-subsets and thereafter using the $(k - 1)$ subsets for training requirements and the remaining subset for validation purposes; the whole procedure is iteratively repeated k times. The k-fold validation technique is employed for the performance evaluation of the ANS classifiers and, moreover, favors the formulation of generalization classification rules. For the current experiments, we selected k = 10, ensuring that for each of the 10 iterations, 90% of the audio frames' feature values are engaged in the model training process and 10% for validation purposes.

The performance rates for each parameters' combination (window length, k-fold iteration, etc.) are based on the extracted confusion matrix of the respective model. Specifically, the confusion matrix represents an array of values of the correctly classified instances and the misclassified ones. Table 3 presents an example of the output confusion matrix for the temporal length 125 ms for the taxonomy voice (V), phone (P) and residual (R) according to VPR scheme. As shown, the correctly classified samples are on the main diagonal of the array, while above and below are the erroneously classified samples.

Table 3. Example of confusion matrix.

		Predicted		
	Class	**V**	**P**	**R**
Actual	V	2255	99	46
	P	32	755	13
	R	27	28	1545

The overall pattern recognition performance PS of the ANS for each of the implemented schemes is evaluated by the % ratio of the number of the correctly classified samples Ncor to the total number of the input samples N. In the same way, the partial discrimination rate PSci of a class Ci is defined as the % ratio of the correctly classified samples Ncci in the class Ci to the total number of samples Nci that Ci class includes. The above definitions are described in Equations (1) and (2).

$$PS = \frac{N_{COR}}{N} \times 100\% \quad (1)$$

$$PS_{Ci} = \frac{N_{CCi}}{N_{Ci}} \times 100\% \quad (2)$$

Applying Equations (1) and (2) in our example of Table 3, the numbers of correctly classified samples in each class are N_{CCV} = 2255 (for class V), N_{CCP} = 755 (for class P) and N_{CCR} = 1545 (for class R), while the total number of correctly classified instances in the model is Ncor = 2255 + 755 + 1545 = 4555. Furthermore, as Table 1 exhibits, the input dataset had N = 4800 samples in total, and for each class, we have N_V = 2400, N_P = 800 and N_R = 1600. Consequently, the partial recognition rates for each class are:

$$PS_V = (2255/2400) \times 100 = 93.96\%$$

$$PSP = (755/800) \times 100 = 94.38\%$$

$$PSR = (1545/1600) \times 100 = 96.54\%$$

On the other hand, the clustering process in the current work is being implemented through the k-means classification algorithm that aims to detect the formulation of group of data/feature values, according to a similarity metric. The criteria that defines the integration of a sample into a cluster, usually refers to a distance metric such as Euclidean, Manhattan, Chebyshev or min–max distance from the cluster center/average. The experiments that are carried out in the present work utilize both Euclidean and Manhattan distance in the k-means implementation for additional comparison purposes.

Since the clustering process only detects data groups, the classification performances cannot be directly evaluated with Equations (1) and (2). One of the main objectives of the current work is to investigate the feasibility of the automatic unsupervised classification in audio data through clustering methods and compare the results with the respective ones by supervised ANS. In this way, we can compare the data clusters formulations of k-means with the corresponding classes in ANS in order solely to evaluate the clusters. Table 4 presents the example of the output cluster formulation of the k-means algorithm for the same window length of 125 ms.

Table 4. Example of cluster formulation.

		Clusters/Groups		
	Class	**G1**	**G2**	**G3**
Actual Class	V	2240	90	70
	P	0	800	0
	R	106	165	1329

The partial discrimination performance PU_{Gi} of cluster *Gi* is defined as the % ratio of the number of samples of class *Ci* that have been classified to cluster *Gi* to the total number of samples of class *Ci*. The above metric is essentially a % measurement of resemblance of cluster class. In addition, the overall discrimination performance of clustering *PU* is evaluated as the % ratio of the sum of the numbers of samples of each class *Ci* that have been correctly grouped in cluster *Gi* to the total number of input samples *N*. The above metrics are described in Equations (3) and (4). The classification results of the employed supervised/unsupervised techniques in the next section are evaluated based on Equations (1)–(4).

$$PU_{Gi} = \frac{N_{Ci \rightarrow Gi}}{N_{Ci}} \times 100\% \quad (3)$$

$$PU = \frac{\sum_1^{nclusters} N_{Ci \rightarrow Gi}}{N} \times 100\% \quad (4)$$

Applying Equations (3) and (4) in our example of Table 4, the number of clusters is nclusters = 3 and the distribution of class samples in the respective clusters is $N_{V \rightarrow G1} = 2240$ (for class V in Group1), $N_{P \rightarrow G2} = 800$ (for class P in Group 2) and $N_{R \rightarrow G3} = 2240$ (for class R in Group 3). Again, as Table 1 exhibits, the input dataset had $N = 4800$ samples in total and for each class, we have $N_V = 2400$, $N_P = 800$ and $N_R = 1600$. Consequently, the partial recognition rates for each class are:

$$PU_{G1} = (2240/2400) \times 100 = 93.33\%$$

$$PU_{G2} = (800/800) \times 100 = 100\%$$

$$PU_{G3} = (1329/1600) \times 100 = 83.08\%$$

3.2. Performance Results on Combined Taxonomies

The supervised ANS models and the clustering k-means algorithm are implemented independently in the first classification layer of the VPR scheme. Furthermore, the data mining techniques are optionally combined in the subsequent layers, in order to evaluate either a strict supervised or unsupervised character of classification or a hybrid one, while moving down in the classification schemes/layers. Figure 4 exhibits the combinations of classification methods. It must be noted that the clustering "path" leads to a more automated whole semantic analysis process compared to the prerequisites of ground truth databases that ANS classifiers demand.

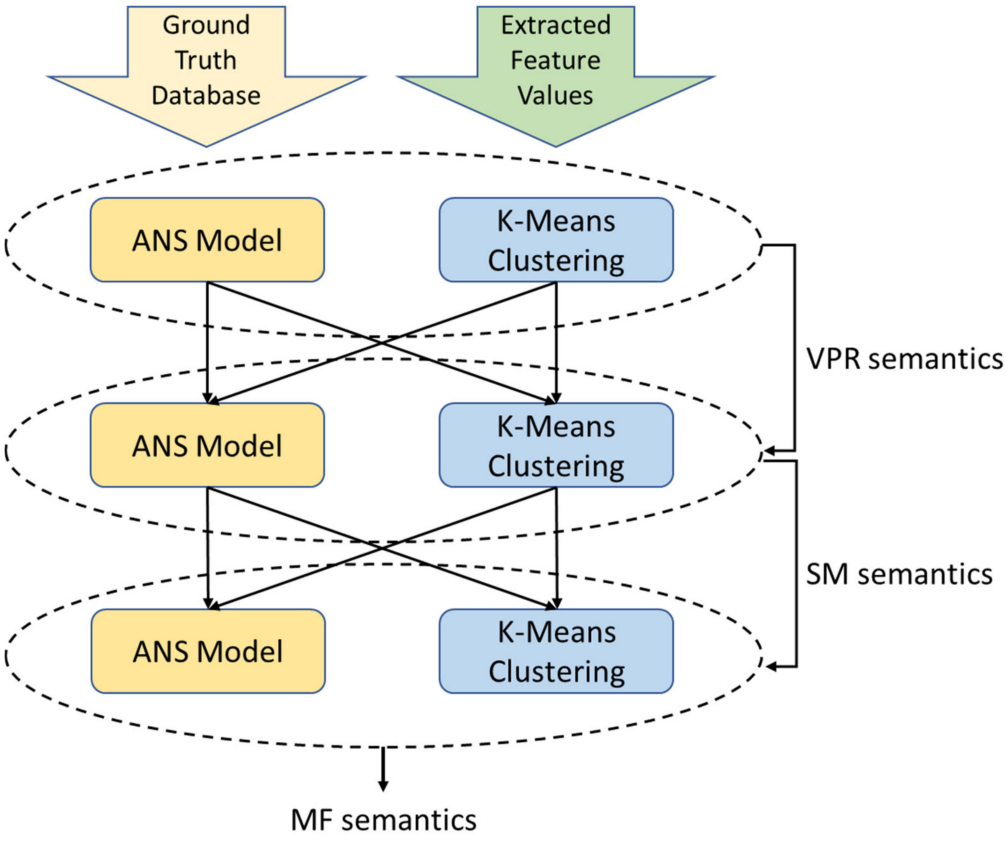

Figure 4. Description of classifier connectivity.

In order to follow the successive implementations of ANS and k-means algorithms, each path in Figure 4 is represented with the initials S (for supervised classification) and U (for clustering) for the three layers, namely, the X-X-X notation where X = S or X = U. For example, the notation S-S-U stands for the combined classification with ANS modeling in VPR and SM schemes and k-means clustering in the MF scheme.

Table 5 presents the performance rates of ANS and k-means for the first VPR classification scheme for several temporal segmentation windows. The overall and partial discrimination rates for supervised classification are quite high for all of the V, P and R classes, reaching even values of 100%. A useful remark derives from the slight decline in performances dependent on the window duration. The unsupervised k-means algorithm also presents high performance rates, i.e., for 1000 ms windowing, 94.76% for overall discrimination, and 96.67%, 100.00% and 93.08% for the cluster formulation according to the corresponding class V, P and R respectively. The phone signal class reaches the 100% discrimination performance, for both algorithms, confirming the initial assumptions of its more specific temporal and spectral audio properties. Moreover, the implementations with Manhattan distance usually lead to slightly increased performance values, but on the other hand, decreases in temporal windowing considerably deteriorate the clustering process

with discrimination values of 71.55% and 63.51% for 250 ms and 125 ms framing windows, respectively. The 1000 ms segmentation window leads to the highest discrimination rates for both supervised/unsupervised techniques, but the impact of temporal differentiations is quite obvious, especially in the clustering process.

Table 5. Classification performances for VPR scheme.

	Window	P	Pv	Pp	P_R
S	1000 ms	99.05	100.00	100.00	98.46
	500 ms	96.43	96.67	100.00	95.77
	250 ms	95.83	94.17	95.00	96.73
	125 ms	95.60	93.96	94.38	96.54
U	1000 ms (Manh)	94.76	96.67	100.00	93.08
	1000 ms (Eucl)	94.29	95.00	100.00	93.08
	500 ms (Manh)	87.62	93.33	100.00	83.08
	500 ms (Eucl)	83.10	88.33	100.00	78.08
	250 ms (Manh)	71.55	61.25	5.00	86.54
	250 ms (Eucl)	73.33	65.00	5.00	87.69
	125 ms (Manh)	63.51	22.71	85.63	78.94
	125 ms (Eucl)	65.77	38.96	10.00	86.73

In order to proceed to the next classification level/layer, the most efficient results of 1000 ms segmentation windowing are employed from the VPR scheme, which provide 100% voice signal discrimination in ANS and 96.67% clustering in k-means. Thereafter, the classification techniques are employed again for the SM scheme, in order to discriminate a single speaker from multiple speakers in the detected voice signal of the VPR scheme. Table 6 exhibits the discrimination rates for the SM scheme.

Table 6. Classification performances for SM scheme.

	Window	P	P_{SS}	P_{MS}
S-S	1000 ms	98.33	97.50	100.00
	500 ms	90.00	91.25	87.50
	250 ms	88.75	92.50	81.25
	125 ms	88.54	92.50	80.63
S-U	1000 ms (Manh)	67.24	52.50	100.00
	1000 ms (Eucl)	68.33	57.50	90.00
	500 ms (Manh)	59.17	41.25	95.00
	500 ms (Eucl)	55.00	35.00	95.00
	250 ms (Manh)	56.25	41.25	86.25
	250 ms (Eucl)	55.42	40.00	86.25
	125 ms (Manh)	58.96	48.75	79.38
	125 ms (Eucl)	59.17	47.81	81.88
U-S	1000 ms	98.28	97.50	100.00
	500 ms	90.52	91.25	88.89
	250 ms	85.34	90.63	73.61
	125 ms	88.36	92.50	79.17
U-U	1000 ms (Manh)	70.69	57.50	100.00
	1000 ms (Eucl)	67.24	52.50	100.00
	500 ms (Manh)	63.79	80.00	61.11
	500 ms (Eucl)	60.34	60.00	61.11
	250 ms (Manh)	65.95	68.13	61.11
	250 ms (Eucl)	56.90	50.63	70.83
	125 ms (Manh)	54.96	43.44	80.56
	125 ms (Eucl)	54.09	40.63	84.03

The selection of temporal window also remains crucial for the ANS and k-means implementations for the SM classification scheme. The 1000 ms framing leads to more efficient overall and partial discrimination rates for all of the S-S, S-U, U-S and U-U combinations. Moreover, the Manhattan distance metric for the k-means results in better clustering performances compared to Euclidean distance. Finally, the most useful remark refers to the 100% discrimination and clustering performance in the combined algorithms' implementation for the multiple speakers class of the voice signal.

Moreover, the U-S and S-S sequences lead to more efficient single speaker discrimination (97.5%), and consequently, the ANS classification is vital in the second layer of the SM scheme. This allows the semantic analysis to proceed in the third hierarchical MF scheme for genre voice classification of the single speaker voice. Table 7 exhibits the classification rates for the third layer of MF scheme.

Table 7. Classification performances for MF scheme.

	Window	P	P_{MV}	P_{FV}
U-S-S **S-S-S**	1000 ms	92.50	95.00	90.00
	500 ms	96.25	95.00	97.50
	250 ms	95.63	95.00	96.25
	125 ms	93.44	93.13	93.75
U-S-S **S-S-U**	1000 ms (Manh)	90.00	100.00	80.00
	1000 ms (Eucl)	90.00	100.00	80.00
	500 ms (Manh)	81.25	95.00	67.50
	500 ms (Eucl)	70.00	60.00	80.00
	250 ms (Manh)	65.62	65.00	66.25
	250 ms (Eucl)	52.50	73.75	31.25
	125 ms (Manh)	51.25	41.88	60.63
	125 ms (Eucl)	50.31	67.50	33.13

As Table 7 exhibits, the male/female voice is discriminated in high performances for both supervised and unsupervised implementations, with overall discrimination values of about 90%. More thoroughly, the ANS modeling offers slightly better and more balanced classification results (92.50%, 95%, 90%) compared to the k-means clustering rates (90%, 100%, 80%). Furthermore, it is quite useful to note that, in the MF scheme, the ANS implementations yield better overall and partial performances for smaller segmentation windows, while the opposite stands for k-means clustering. Finally, the same observation for the better selection of the Manhattan distance metric also remains for the MF classification scheme.

Summarizing the remarks of the overall semantic analysis for hierarchical classification in the three layers of Figure 4, it must be noted that several combinations can be sought for pure or hybrid classification techniques in order to reach efficient discrimination results. The integration of clustering methods in supervised implementations promotes automation and functionality in the whole semantic analysis process.

3.3. Validation and Optimization Issues

As mentioned in Section 2.5, the feature evaluation process is crucial for the overall processing load and time, especially for the supervised classification techniques. Even though the clustering k-means algorithm noted reduced computational load in the previous experiments while exploiting the whole extracted feature set, in this section, a feature evaluation process is conducted especially for the clustering method, while also utilizing the ranking results of Table 2. The k-means algorithm is employed for the three classification schemes VPR, SM and MF, but different numbers of audio properties are exploited in each implementation, based on the ranking of Table 2. Figures 5–7 present the overall and partial discrimination rates while utilizing differentiated numbers of audio features.

From the above diagrams, the optimum number of features is determined based on the best performance rates. Table 8 presents the number of features and the corresponding discrimination rates for the k-means clustering in the hierarchical classification schemes of Figure 2. Comparing the values of Table 8 with the corresponding performances of Tables 5–7, we can observe the positive impact of the diversification of the number of features in clustering, in the context of the whole semantic analysis process performance rates.

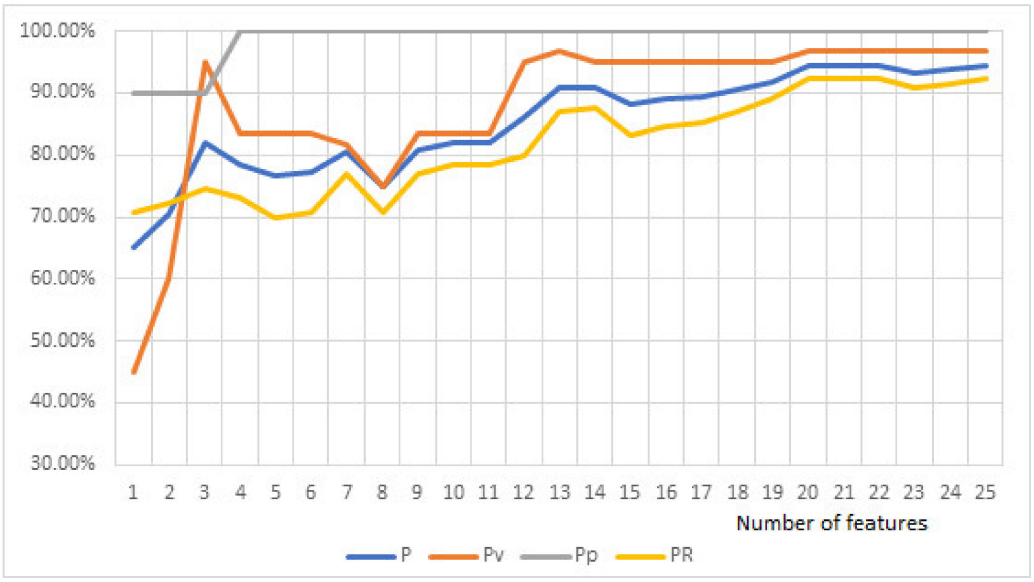

Figure 5. Feature evaluation for VPR scheme.

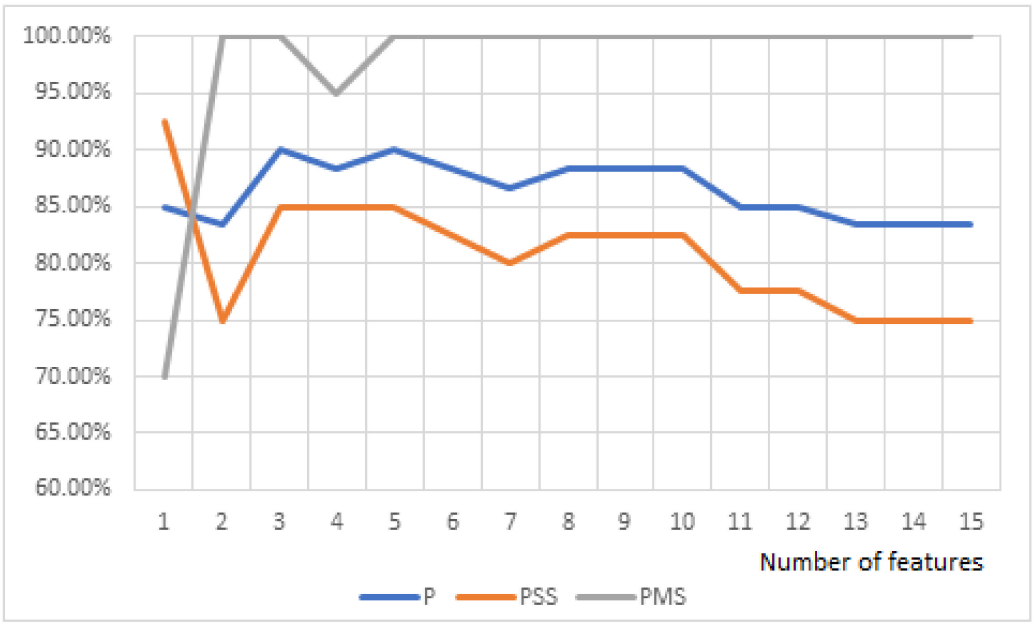

Figure 6. Feature evaluation for SM scheme.

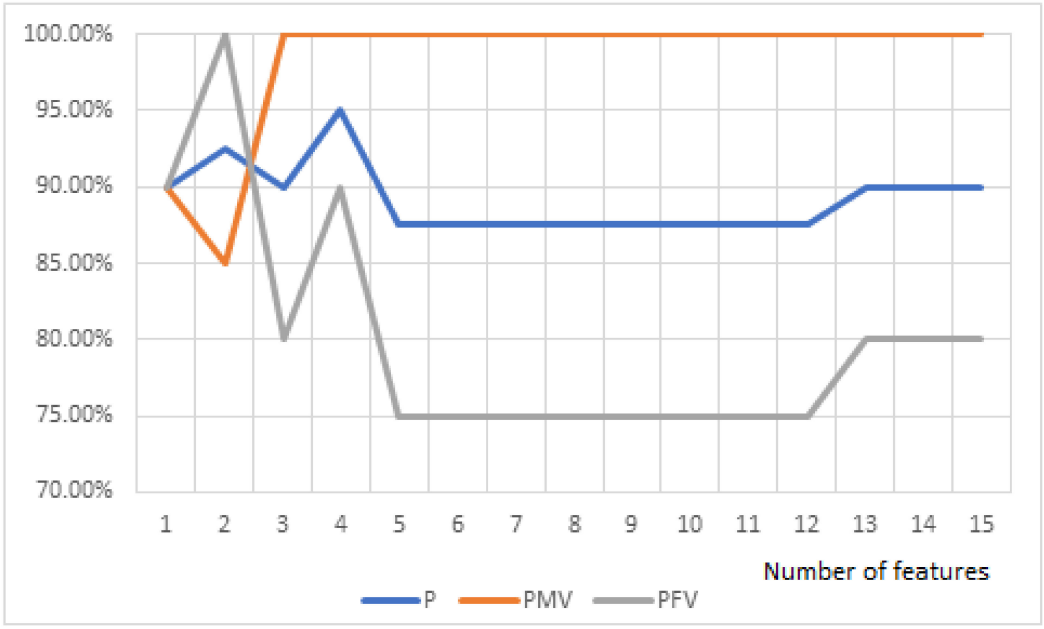

Figure 7. Feature evaluation for MF scheme.

Table 8. Clustering performances with feature subsets.

	$\mathbf{P}$	$\mathbf{P_V}$	$\mathbf{P_P}$	$\mathbf{P_R}$
U(32)	95.24	100.00	100.00	92.31
	$\mathbf{P}$	$\mathbf{P_{SS}}$	$\mathbf{P_{MS}}$	
U(32)—U(10)	90.00	85.00	100.00	
	$\mathbf{P}$	$\mathbf{P_{MV}}$	$\mathbf{P_{FV}}$	
U(32)—U(10)—U(24)	95.00	100.00	90.00	

Another aspect that must be taken into consideration while employing the pattern recognition analysis refers to the selection of the segmentation window, which has a crucial impact on the discrimination performances. More specifically, in almost all of the current experiments and also previous ones in [1,25], the 1000 ms framing length leads to better classification results. One justification may derive from the fact that 1000 ms contains more information data, which is a determinant factor for classification purposes, especially of heterogeneous audio content (i.e., VPR scheme). In order to further investigate and validate the selection of 1000 ms framing length, besides the comparisons in the previous section with various temporal windows (1000 ms, 500 ms, 250 ms, 125 ms) in terms of their corresponding classification results, several experiments are conducted in this section with a sliding 1000 ms segmentation window. In the following analysis, the 1000 ms segmentation begins with 100 ms, 200 ms, 300 ms and 400 ms delays, resulting in successive information loss compared to the initial accurately annotated frames. Moreover, in real conditions, the whole semantic analysis process may suffer from inaccurate annotation tasks, and consequently, the sliding 1000 ms windows may reveal the impact of the selected window, on the grounds of a sensitivity analysis in the classification problem. Table 9 presents the performance values for supervised and unsupervised classification on the VPR scheme with the sliding effect.

Table 9. Temporal sensitivity analysis.

	P	Pv	Pp	P_R
S-no sliding	99.05	100.00	100.00	98.46
U-no sliding (all feat.)	94.76	96.67	100.00	93.08
U-no sliding (32 feat.)	95.24	100.00	100.00	92.31
S-100 ms sliding	99.05	100.00	100.00	98.46
U-100 ms sliding (all feat.)	91.90	96.67	100.00	88.46
U-100 ms sliding (32 feat.)	92.38	100.00	100.00	87.69
S-200 ms sliding	99.05	100.00	100.00	98.46
U-200 ms sliding (all feat.)	92.38	96.67	100.00	89.23
U-200 ms sliding (32 feat.)	94.29	98.33	100.00	91.54
S-300 ms sliding	98.57	100.00	100.00	97.69
U-300 ms sliding (all feat.)	88.57	93.33	100.00	84.62
U-300 ms sliding (32 feat.)	93.33	98.33	100.00	90.00
S-400 ms sliding	97.62	98.33	95.00	97.69
U-400 ms sliding (all feat.)	88.57	95.00	100.00	83.85
U-400 ms sliding (32 feat.)	90.95	96.67	95.00	87.69
S-all together	99.90	100.00	100.00	99.85
U-all together (all feat.)	88.29	91.00	100.00	85.23
U-all together (32 feat.)	92.95	98.33	98.00	89.69

As Table 9 exhibits, the segmentation delay results are different from the respective ones with no sliding. Nevertheless, the impact of the sliding effect is not so obvious for 100 ms, 200 ms or 300 ms temporal delays, but when referring to 400 ms sliding, the classification performances decrease for both ANS and k-means methods, which indicates a significant information loss. Consequently, the 1000 ms framing selection appears to be an efficient segmentation window, which also allows real condition annotation fault margins. Furthermore, in the same table, the discrimination results for k-means with 32 feature implementations are exhibited, which are more increased with the corresponding implementations with the whole feature set. This remark reinforces the previous conclusions of the feature adaptivity potential in the whole semantic analysis process.

4. Conclusions and Future Aspects

This paper presents a framework for quick and light-weighted adaptive segmentation and classification processes in audio broadcasted content. Several combinations of training techniques were implicated in hybrid and hierarchical taxonomies, along with multivariate experiments in feature sets and temporal window tuning. The classification rates (especially in supervised strategies) revealed that such a methodology is feasible for effective content discrimination while choosing a profile of parameters in terms of audio properties, window lengths and machine learning techniques. While moving into more careful conclusions drawn by the whole experimentation setup, it must be mentioned that traditional machine learning strategies can be exploited when limited data exist in order to support initial broadcasted data classification for effective radio content description and archiving purposes. The temporal length of windowing process contributes radically to the taxonomies' performance, favoring mainly medium lengths (around 1 s duration), while the sensitivity experiments with overlapping potentials revealed that sliding operations improve the general classification rates compared to more strict segmentation choices. Moreover, it must be highlighted that clustering methods can facilitate a quick and effective semi-automated "blind" data discrimination without the data annotation step, especially for the initial voice classification layer. In all cases, the audio signals deriving from broadcasted programs can be efficiently processed in hierarchical/hybrid classification implementations since error propagations are treated better while breaking down the content in multilayer discrimination schemes compared to direct ones.

The overarching purpose of this work is the formulation of a dynamic Generic Audio Classification Repository, fed by iterative radio program-adaptive classification pro-

cesses/experimentation. In this context, the audio database is constantly evolving and augmenting, and more taxonomies could be incorporated, either in voice signals (language discrimination, emotion estimation, etc.) or in residual data (music genre classification, noise removal, etc.). In this direction, the semantic analysis process could facilitate more complex and resource-demanding machine learning strategies in rich data content that could involve deep architectures (RNNs, 1d/2d CNNs, etc.). The main target of the presented work is to integrate, step by step, all possible classification schemes based on radio content structures, in order to support effective pretrained models and automated solutions independent of adaptive methodologies.

Author Contributions: Conceptualization, R.K. and C.D.; Data curation, R.K. and C.D.; Investigation, R.K. and C.D.; Methodology, R.K. and C.D.; Project administration, R.K. and C.D.; Software, R.K. and C.D.; Supervision, R.K. and C.D.; Validation, R.K. and C.D.; Visualization, R.K. and C.D.; Writing—original draft, R.K. and C.D.; Writing—review and editing, R.K. and C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kotsakis, R.; Kalliris, G.; Dimoulas, C. Investigation of broadcast-audio semantic analysis scenarios employing radio-programme-adaptive pattern classification. *Speech Commun.* **2012**, *54*, 743–762. [CrossRef]
2. Veglis, A. Moderation Techniques for Social Media Content. In *Lecture Notes in Computer Science, Proceedings of the 6th International Conference, SCSM 2014 Held as Part of HC International 2014 Heraklion, Crete, Greece, 22–27 June 2014*; Meiselwitz, G., Ed.; Springer: London, UK, 2014; Volume 8531.
3. Kannao, R.; Guha, P.; Chaudhuri, B. Only overlay text: Novel features for TV news broadcast video segmentation. *Multimed. Tools Appl.* **2022**, *81*, 1–25. [CrossRef]
4. Kotsakis, R.; Dimoulas, C.; Kalliris, G.; Veglis, A. Emotional Prediction and Content Profile Estimation in Evaluating Audiovisual Mediated Communication. *Int. J. Monit. Surveill. Technol. Res. (IJMSTR)* **2014**, *2*, 62–80. [CrossRef]
5. Vryzas, N.; Vrysis, L.; Matsiola, M.; Kotsakis, R.; Dimoulas, C.; Kalliris, G. Continuous Speech Emotion Recognition with Convolutional Neural Networks. *J. Audio Eng. Soc. (JAES)* **2020**, *68*, 14–24. [CrossRef]
6. Kotsakis, R.; Matsiola, M.; Kalliris, G.; Dimoulas, C. Investigation of Spoken-Language Detection and Classification in Broadcasted Audio Content. *Information* **2020**, *11*, 211. [CrossRef]
7. Vryzas, N.; Vrysis, L.; Kotsakis, R.; Dimoulas, C. A web crowdsourcing framework for transfer learning and personalized Speech Emotion Recognition. *J. Mach. Learn. Appl.* **2021**, *6*, 100–132. [CrossRef]
8. Gimeno, P.; Viñals, I.; Giménez, A.O.; Miguel, A.; Lleida, E. Multiclass audio segmentation based on recurrent neural networks for broadcast domain data. *EURASIP J. Audio Speech Music. Process.* **2020**, *2020*, 5. [CrossRef]
9. Weerathunga, C.O.B.; Jayaratne, K.L.; Gunawardana, P.V.K.G. Classification of Public Radio Broadcast Context for Onset Detection. *KL Jayaratne-GSTF J. Comput. (JoC)* **2018**, *7*, 1–22.
10. Liu, C.; Feng, L.; Liu, G. Bottom-up broadcast neural network for music genre classification. *Multimed. Tools Appl.* **2021**, *80*, 7313–7331. [CrossRef]
11. Kabir, M.M.; Mridha, M.F.; Shin, J.; Jahan, I.; Ohi, A.Q. A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities. *IEEE Access* **2021**, *9*, 79236–79263. [CrossRef]
12. Kang, Z.; Huang, Z.; Lu, C. Speech Enhancement Using U-Net with Compressed Sensing. *Appl. Sci.* **2022**, *12*, 4161. [CrossRef]
13. Gutiérrez-Muñoz, M.; Coto-Jiménez, M. An Experimental Study on Speech Enhancement Based on a Combination of Wavelets and Deep Learning. *Computation* **2022**, *10*, 102. [CrossRef]
14. Venkatesh, S.; Moffat, D.; Miranda, E.R. You Only Hear Once: A YOLO-like Algorithm for Audio Segmentation and Sound Event Detection. *Appl. Sci.* **2022**, *12*, 3293. [CrossRef]
15. Vandhana, T.S.; Srivibhushanaa, S.; Sidharth, K.; Sanoj, C.S. Automatic Speech Recognition using Recurrent Neural Network. *Int. J. Eng. Res. Technol. (IJERT)* **2020**, *9*. [CrossRef]
16. Nakadai, K.; Sumiya, R.; Nakano, M.; Ichige, K.; Hirose, Y.; Tsujino, H. The Design of Phoneme Grouping for Coarse Phoneme Recognition. In *Lecture Notes in Computer Science, Proceedings of the New Trends in Applied Artificial Intelligence. 20th international Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2007, Kyoto, Japan, 26–29 June 2007*; Okuno, H.G., Ali, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; Volume 4570.
17. Yang, X.K.; Qu, D.; Zhang, W.L.; Zhang, W.Q. An adapted data selection for deep learning-based audio segmentation in multi-genre broadcast channel. *Digit. Signal Process.* **2018**, *81*, 8–15. [CrossRef]

18. Patil, N.M.; Nemade, M.U. Content-Based Audio Classification and Retrieval Using Segmentation, Feature Extraction and Neural Network Approach. In *Advances in Computer Communication and Computational Sciences*; Bhatia, S., Tiwari, S., Mishra, K., Trivedi, M., Eds.; Springer: Singapore, 2017; Volume 924.
19. Haloi, P.; Bhuyan, M.K.; Chatterjee, D.; Borah, P.R. Unsupervised story segmentation and indexing of broadcast news video. *Multimed. Tools Appl.* **2021**, 1–20. [CrossRef]
20. Lopez-Otero, P.; Docio-Fernandez, L.; Garcia-Mateo, C. Ensemble audio segmentation for radio and television programmes. *Multimed. Tools Appl.* **2017**, *76*, 7421–7444. [CrossRef]
21. Seo, J.; Lee, B. Multi-Task Conformer with Multi-Feature Combination for Speech Emotion Recognition. *Symmetry* **2022**, *14*, 1428. [CrossRef]
22. Gimeno, P.; Mingote, V.; Ortega, A.; Miguel, A.; Lleida, E. Generalizing AUC Optimization to Multiclass Classification for Audio Segmentation With Limited Training Data. *IEEE Signal Processing Lett.* **2021**, *28*, 1135–1139. [CrossRef]
23. Guo, J.; Li, C.; Sun, Z.; Li, J.; Wang, P. A Deep Attention Model for Environmental Sound Classification from Multi-Feature Data. *Appl. Sci.* **2022**, *12*, 5988. [CrossRef]
24. Brunese, L.; Mercaldo, F.; Reginelli, A.; Santone, A. A Neural Network-Based Method for Respiratory Sound Analysis and Lung Disease Detection. *Appl. Sci.* **2022**, *12*, 3877. [CrossRef]
25. Kotsakis, R.; Kalliris, G.; Dimoulas, C. Investigation of salient audio-features for pattern-based semantic content analysis of radio productions. In Proceedings of the 132nd Audio Engineering Society (AES) Convention, Paper No. 8663, Budapest, Hungary, 26–29 April 2012.
26. Kotsakis, R.G.; Dimoulas, C.A.; Kalliris, G.M. Contribution of Stereo Information to Feature-Based Pattern Classification for Audio Semantic Analysis. In Proceedings of the Seventh International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Luxembourg, 3–4 December 2012; pp. 68–72.
27. Wu, X.; Zhu, M.; Wu, R.; Zhu, X. A Self-adapting GMM based Voice Activity Detection. In Proceedings of the IEEE 23rd International Conference on Digital Signal Processing (DSP), Shanghai, China, 19–21 November 2018; pp. 1–5.
28. Song, J.H.; LEE, S. Voice Activity Detection Based on Generalized Normal-Laplace Distribution Incorporating Conditional MAP. *IEICE Trans. Inf. Syst.* **2013**, *96*, 2888–2891. [CrossRef]
29. Makowski, R.; Hossa, R. Voice activity detection with quasi-quadrature filters and GMM decomposition for speech and noise. *Appl. Acoust.* **2020**, *166*, 107344. [CrossRef]
30. Mihalache, S.; Burileanu, D. Using Voice Activity Detection and Deep Neural Networks with Hybrid Speech Feature Extraction for Deceptive Speech Detection. *Sensors* **2022**, *22*, 1228. [CrossRef]