information

MDPI

Article

Investigation of Spoken-Language Detection and Classification in Broadcasted Audio Content

Rigas Kotsakis, Maria Matsiola, George Kalliris and Charalampos Dimoulas *

School of Journalism and Mass Communications, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; rkotsakis@jour.auth.gr (R.K.); mmat@jour.auth.gr (M.M.); gkal@jour.auth.gr (G.K.)

* Correspondence: babis@eng.auth.gr; Tel.: +30-2310-9942454

Received: 1 March 2020; Accepted: 14 April 2020; Published: 15 April 2020

Abstract: The current paper focuses on the investigation of spoken-language classification in audio broadcasting content. The approach reflects a real-word scenario, encountered in modern media/monitoring organizations, where semi-automated indexing/documentation is deployed, which could be facilitated by the proposed language detection preprocessing. Multilingual audio recordings of specific radio streams are formed into a small dataset, which is used for the adaptive classification experiments, without seeking—at this step—for a generic language recognition model. Specifically, hierarchical discrimination schemes are followed to separate voice signals before classifying the spoken languages. Supervised and unsupervised machine learning is utilized at various windowing configurations to test the validity of our hypothesis. Besides the analysis of the achieved recognition scores (partial and overall), late integration models are proposed for semi-automatically annotation of new audio recordings. Hence, data augmentation mechanisms are offered, aiming at gradually formulating a Generic Audio Language Classification Repository. This database constitutes a program-adaptive collection that, beside the self-indexing metadata mechanisms, could facilitate generic language classification models in the future, through state-of-art techniques like deep learning. This approach matches the investigatory inception of the project, which seeks for indicators that could be applied in a second step with a larger dataset and/or an already pre-trained model, with the purpose to deliver overall results.

Keywords: radio production; machine-driven modelling; language discrimination

1. Introduction

Tremendous evolution in digital technology and mobile computing has allowed multiple users to capture and share audiovisual content (the so-called User Generated Content—UGC), thus increasing the capacities of massive audio production and distribution. Given the progress in storage media and the efficiency of state-of-the-art compression algorithms, the problem has shifted from warehousing needs to the documentation and management strategies [1–3]. In this context, new audio recognition and semantic analysis techniques are deployed for General Audio Detection and Classification (GADC) tasks, which are very useful in many multidisciplinary domains [4–16]. Typical examples include speech recognition and perceptual enhancement [5–8], speaker indexing and diarization [14–19], voice/music detection and discrimination [1–4,9–13,20–22], information retrieval and genre classification of music [23,24], audio-driven alignment of multiple recordings [25,26], sound emotion recognition [27–29] and others [10,30–32]. Concerning the media production and broadcasting domain, audio and audio-driven segmentation allow for the implementation of proper archiving, thus facilitating content reuse scenarios [1–3,11,12,16–20,31]. Besides internal (within the media organization) searching and retrieval, publicity metrics of specific radio stations and programs can be associated with the presence of various audio classes (of both speakers and music species), providing valuable feedback

to all involved in the broadcasting process (i.e., producers, advertisers, communication professionals, journalists, etc.) [33].

Extending the above notes, contemporary technological developments that are employed in the categorization of the plethora of data, found in various forms, may provide media agencies with the fastest and most reliable outcomes regarding information retrieval. In this framework, language identification/discrimination can be deployed in a large amount of multilingual audio recordings that are collected daily in newsrooms and need to be monitored and further analyzed. Nowadays, we have a polyglot worldwide environment within nations, as well as at the international level, so audio-driven Natural Language Processing (NLP) is not easily deployable in physical everyday communication [34]. The same applies to the detection of words belonging to different languages, mixed within a sentence, a very common feature in conversational speech (including radio broadcasting), e.g., to present the right terms, titles or locations. In this context, language identification and discrimination, as an automated process, can become a valuable tool, especially for content management, labeling and retrieval purposes. The incoming data streams from worldwide radio stations would be automatically classified, creating large digital audio repositories in the newsrooms. Afterwards, the assorted files could be forwarded to an additional kind of elaboration, leading to the creation of numerous content archives. For instance, term recognition processes, encountered in audio and audiovisual content, and generically speech-to-text (STT) services can be expedited by stressing, first, the preprocessing language detection problem.

The present paper investigates language discrimination of audio files derived from real-world broadcasted radio programs. The aim is the differentiated patterns that appear in everyday radio productions, such as speech signals, phone conversations and music interferences to be initially detected and afterward classified. A small dataset works as the inception of the project, since it is an investigatory approach, which will be gradually enhanced with more experiments. At this point, we are not seeking for a generic solution rather than indicators that could be applied in a second step, with a larger dataset that would deliver overall results. The above-stated limits make the problem under discussion even more demanding and difficult to deal with. To the best of our knowledge, this is the first time that such a challenging task has been confronted within the restrictions of narrow dataset availability and the associated training difficulties, even within the investigative character of the current work. The rest of the paper is organized as follows. Section 2.1 presents problem definition and motivation. The proposed framework is analyzed in Section 2.2, Section 2.3, Section 2.4, Section 2.5, Section 2.6, considering all the involved data selection, configuration and training aspects. Experimental results are analyzed and discussed in Section 3, followed by discussion and future work remarks.

2. Materials and Methods

2.1. Problem Definition and Motivation

This paper investigates machine learning techniques and methodologies that will support multilingual semantic analysis of broadcasted content, deriving from European (and world-wide) radio organizations. The current work is an extension of previous/already published research that is continually elaborated [1–3,33], aiming at meeting the demands of users and industries for audio broadcasting content description and efficient post-management. Since the interconnectivity and spread of Web 2.0 is increasing (moving to Semantic Web 3.0 and beyond), more potential features are engaged for radio networks to continue to develop cross-border broadcasting. Given the achieved progress in separating the main patterns of the radio broadcasted streams (Voice, Phone, Music and main speakers, telephone conversations and music interferences (VPM) scheme) [1–3], the proposed framework investigates language detection mechanisms, targeting the segmentation of spoken content at a linguistic level. Focused research was conducted in speaker diarization/verification problems or sentiment analysis via Gaussian Mixture Modeling with cepstral properties [8,17]. Moreover,

innovative Deep Learning and Convolutional Neural Networks architectures are deployed in this direction [4,5,35] with 2-D input features [7,15]. In addition, several experiments were conducted either on singing voices [18] or in utterance level [19]. Despite the progression in algorithmic level, many related efforts shifted to the development of mechanisms for archiving increased language material [36–40]. Furthermore, though the first trials in resolving this problem are dated many years ago, nonetheless, the multicultural and differentiated linguistic nature of the radio industry deteriorates the possibility of achieving effective recognition scores. Therefore, new methodologies need to be developed for language identification demands.

The aforementioned task is quite demanding because of the diversity of audio patterns that appear in radio productions. Specifically, common radio broadcasting usually includes audio signals deriving from the main speakers' voices (frequently, with strong overlapping), telephone dialogues (exposing distinctive characteristics that depend on the communication channel properties), music interferences (imposed with fade in/out transitions or background music), various SFX and radio jingles or even noise segments (ambient/environmental noise, pops and clicks, distortions, etc.) [1–3,33]. It has to be noted that broadcasted music content depends mainly on the presenters' or audiences' preferences and, consequently, may not be in accordance with the respective language of origin (i.e., a German radio broadcasting may include either German or English songs). The same applies to the corresponding live dialogues and comments that precede or follow music playback, making it rather inevitable to encounter multilingual terms. Most of all, it is not possible to ascribe universal linguistic labeling to the entire recording based on the main language of the associated program.

Taking the above aspects into consideration, and the experience gained from previous research, hierarchical classification approaches are best suited to resolve this task, allowing the initial, highly accurate VPM scheme to isolate non-voice segments, i.e., to exclude them from the subsequent language classification process. Furthermore, both theory and previous experimentation showed that supervised machine learning solutions outperform the accuracy of unsupervised/clustering methods, with the counterbalance of the need for time-accurate standardized semantic annotation, which is rather an old fashioned and time-consuming human-centric procedure [1]. While large-scale labeled audio repositories are currently available to be involved in deep learning processes (e.g., audio books), again, past testing revealed that spontaneous speech and non-stopping sound of real-world radio streams do pose some distinctive features that would not be easily confronted with generic solutions. In fact, program-adaptive training solutions proved to be even more advantageous, since they provide adaptation and generalization mechanisms to the specific speakers, the jingles and the favorite music of each radio show [1–3,33]. In this context, the grouping of multiple shows with similar characteristics may also be feasible. Concerning the specific problem under study, it is further expected that speaker information can be associated with the different multi-lingual patterns, thus facilitating language recognition through similar adaptations (i.e., a speaker would have specific voicing and pronouncing features while speaking different languages, which could be more easily detected and associated).

The idea and the motive behind the current work is the examination of whether it is possible to train such a language detection system with a small initial dataset. After that, following the strategy that is originally deployed in [33], ensemble learning by means of late integration techniques could be applied [41], combining multiple unsupervised and supervised modalities and with different time windowing configurations. Hence, with the only restriction that the initially trained model should be speaker-independent, matching labels between the different modalities would offer ground-truth data augmentation mechanism through semi-automated annotation (i.e., even by requiring users' feedback to verify highly-confidence instances). Thereafter, sessions of semi-supervised learning [42] could be iteratively deployed, thus offering the wanted gradual growth of both generic and program-adaptive repositories (the principle data-flow of this process is given in [33], while more specific details on the current task are provided in the corresponding implementation sections).

Media organizations, communication researchers and news monitoring service providers could be benefited by such an information retrieval system. Distinct spoken words that are recognizable in

a multilingual environment could accelerate the processes of content documentation, management, summarization, media analytics monitoring, etc. Classified audio files, subjected to transcription and/or translation, could be propelled to other media rather than radio. In addition, they might be used for automatic interpretation of worldwide radio streams. On the next level, the categorized and annotated data, extracted through the implementation of semantic analysis and filtering, could be exploited for content recommendations to users or journalists, enhancing their work while avoiding time-consuming personal quests in vast repositories. An additional outcome of this research, in its fully developed phase, could be the valuable provision of feedback to the radio producers themselves, through the analysis of their own used vocabulary, for instance, by using term-statistics on the frequency of foreign words in their dialogues. This user-centric metric approach could lead them to personal and professional improvements. In another context, it could give linguists significant content-related data on the path of a nation's heritage, based on the alterations of the spoken language through the additions of foreign vocabulary.

2.2. Proposed Use Scenario Framework

The far-reaching use scenario of the recommended methodology is depicted in Figure 1. In terms of problem definition, a new long-term sound record (i.e., of 1-h duration or more) is subjected to pattern analysis, segmentation, and indexing, targeting proper documentation and semantic labeling that could offer efficient content-management automation. As already explained, the current work focuses on the semi-supervised classification and annotation of linguistic information. The input audio signals refer to the associated streams of specific radio programs, which are captured along with their live transmission (via RF-broadcasting or webcasting). Two types of different ground-truth datasets are involved. The first one refers to a "generic multilingual audio repository," with the time-marks of the different language segments, which is not expected to match well the individualities of the different radio shows (as argued in the previous section). For this reason, a small dataset is initially formed for each different radio show, investigating the feasibility of the proposed classification scheme through the subsequent experimentation. After this initiation step, the iterative process led to the augmentation of the starting repository, within the analysis of the same recording (i.e., the same broadcasting stream). Given that the proposed procedure will be deployed on multiple episodes of the same radio-show, a modular repository is created, containing multiple records of the same program (or groups of programs). Proceeding to other shows, the dataset will be enhanced by ingesting additional groups of respective records. Hence, the generic repository of Figure 1 is composed by the smaller ones (as the connecting dotted line implies). In simpler terms, a small set is involved in the beginning of the training sessions, posing the most demanding part of the whole system. This is the principal motive and aim of the current paper, to answer whether such kind of traditional machine learning is feasible. Next, provided that the whole iterative process can rely on this assumption, the further looping operations involve the augmentation of the collection with new audio stream entries. Hence, a well-organized and self-indexed database is promoted. This gradually incremented repository can be utilized for media assets management and retrieval purposes, also facilitating future research experimentation on the broader multidisciplinary domain. The smaller dataset contains instances of a specific radio show or group of shows with similar characteristics, aiming at facilitating the training of program-adaptive pattern recognition modules. The main idea is that both the "local" and the "global" ground-truth databases will be elaborated, as the method is deployed with multiple broadcasting sound sequences as inputs. Likewise, the formed training pairs will be further exploited in the implementation of associated "generic" and "specific" language recognition systems. Hence, the anticipated results are equally related to the materialization of the iterative self-learning procedures and the formation of a well-organized big-data structure (that is assembled by multiple featured sub-groups).

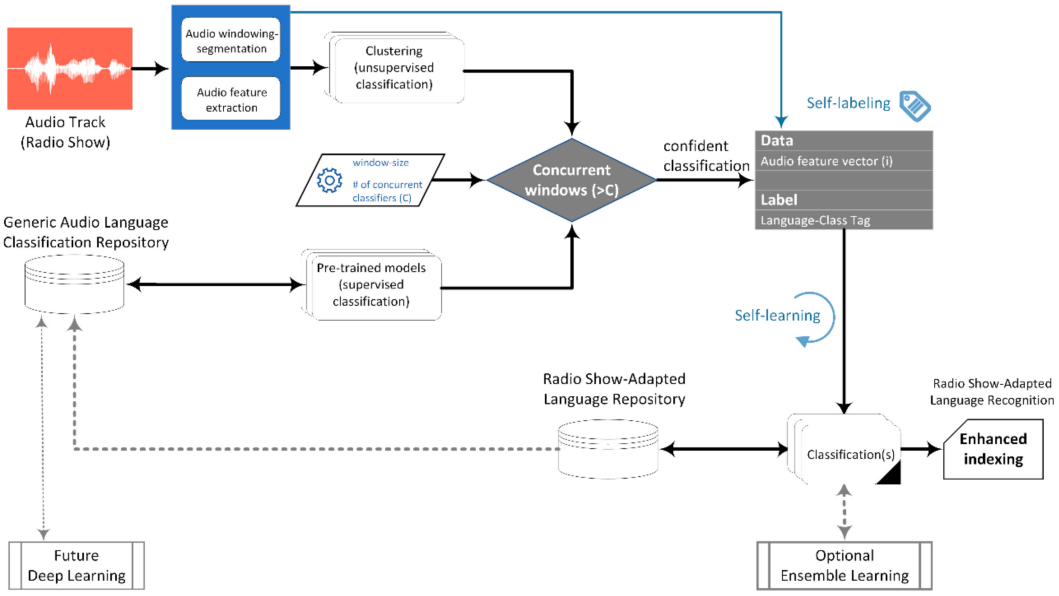

Figure 1. Block diagram of the proposed methodology.

The dataflow of the proposed architecture begins with the engagement of an initial broadcast audio content that is subjected to audio windowing and feature extraction processes, as Figure 1 depicts. In this process, a prime clustering (unsupervised classification) is feasible for the investigative linguistic data grouping (according to hierarchical VMP schemes, as it will be explained later). Moreover, feature representations of the windowed sound signals are fed to pre-trained "generic" and "radio show-adaptive" language recognition models (provided that the latter would be available; if not, they will be omitted, relying entirely on the broad modules). Supervised and unsupervised classification outcomes are then produced for different windowing-lengths and algorithms, in which confident pattern recognition can be indicated based on the number (C) of the classifiers that provide identical or consistent results [13]. The term "consistent" is used because clustering methods do not actually deliver specific pattern information but only their grouping, so the class is implied by the comparison with the pre-trained modalities. These confidently classified audio segments can be subsequently annotated (semi-autonomously) for the acquisition of voice and language class tags, to support the show-adaptive supervised classification and the formulation of the associated ground-truth database. It has to be noted that the implementation of this work follows an iterative character of a multivariate experimental setup, especially in terms of window lengths, aiming at identifying common grounds (agreement) in the classification results deriving from the different paths of machine learning. In this way, the pre-trained models would contribute to the formulation of a Generic Audio Language Repository that could be further tested in future implementations such as Deep Architectures.

2.3. Data Collection—Content Preprocessing

For the conducted experiments, radio content was collected from broadcasts in four different languages, namely native Greek, English, French, and German, which are among the most commonly spoken foreign languages in Europe (at least from a Greek radio audience point of view). The audio signals were formatted (transcoded) to PCM (Pulse-Code Modulation) Wav files (16-bit depth, 44,100 Hz sample rate). At the same time, the stereo property was discarded, since it could serve only for the music/genre discrimination and not for voice (and language) recognition, as it was thoroughly studied in [16–20].

Since the radio content derives from broadcasting in different countries, it is anticipated that several differentiations in structure/fashion may appear. In order to overcome this obstacle, only the common audio patterns were retained, therefore aiming at avoiding homogeneity problems in the conducted semantic analysis. Specifically, each radio segment of the data collection had a 10-min

duration, which involved 8 min of main speakers' voices (male and female), 1-min phone conversations and 1-min of music clips, forming a quite representative proportion/ratio of a data sequence with temporal length of 1 h, which is a typical duration of a broadcast show. In this context, the entire multilingual audio signal had an overall duration of 40 min (4×10). As already justified, this small duration was a somewhat conscious choice, since it reflects the real-world nature of the given scenario, depicted in Figure 1.

2.4. Definition of Classification Taxonomies

As described in the problem definition section, the methodology that was implemented in the current work was constituted by the development of two classification taxonomies, in a hierarchical fashion. Specifically, the first level involves an initial scheme in the radio content, aiming to separate the voice signals of the main speakers (V), the telephone conversations (P) and the music interferences (M), therefore forming the VPM discrimination scheme. It has to be noted that the VPM taxonomy was firstly introduced and examined in [14–20] for Greek radio broadcasts, to isolate the main voices for validation and indexing purposes. Hence, the adoption of the correspondent VPM scheme in the current work serves as an extension to [14–20], in the context of generic classification of speech/non-speech radio signals, deriving from an augmented database of multilingual radio broadcasts.

Thereafter, the second discrimination level includes the hierarchical classification of the spoken language sub-categories, constituting the Language Detection (LD) scheme. As mentioned before, the LD taxonomy was based only on the speakers' voice samples (that were successfully discriminated via the previous VPM scheme), because of the degradation/deterioration that could be implicated by the unique properties of the phone signals [14,15] and the disorientation by the multinational character of the music data. Moreover, music language can be easily identified in textual processing terms, using the associated meta-data of the song files (title, artists, etc.). The phone category, on the other hand, is not present in all radio streams and it holds a much shorter duration. Hence, it might not be worth complicating the LD hierarchy in such a way that it will not be applicable in the presented scenarios. Even within the same radio program, telephone voices would have significant differentiation, emanated by both the speech features of the involved persons and the nature/quality of the connecting lines. Given that the detection of P windows has proved quite easy [14–20], these segments could be further processed to match the V patterns (i.e., in terms of spectral shaping and dynamic range accommodation). Thus, in case that language recognition of the few phone events is considered necessary, mapping to the used LD scheme could be an alternative approach. Overall, the implemented hierarchical taxonomy was formed in a trade-off kind of concept (to balance between genericity and complexity) and is presented in Figure 2.

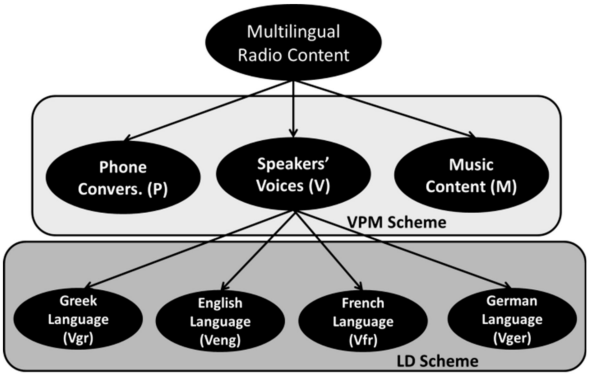

Figure 2. Hierarchical Classification Scheme.

2.5. Windowing Process—Annotation

One of the most crucial steps in audio data mining problems is the signal windowing, since the segmentation length influences the performance of the classification models. In the current

work, the audio content was segmented in short frames via Python script coding [21]. Moreover, several lengths were utilized and more specifically 100 ms, 500 ms, 1000 ms and 2000 ms, aiming at investigating the impact of the temporal windowing, indicating the most effective ones in the current language discrimination process. In addition, the differentiated window length analysis was considered necessary, since two hierarchical schemes were involved in the subsequent classification experiments. Table 1 presents the population of the input samples in each category with the correspondent labeling of Figure 2, according to the following notations:

- M, P, V for Music, Phone and Voice samples, respectively.
- Vgr, Veng, Vfr, Vger for Greek, English, French, German Voice samples, respectively.

Table 1. Distribution of the input samples.

Notation	100 ms	500 ms	1000 ms	2000 ms
M	2400	480	240	120
P	2400	480	240	120
V	19,200	3840	1920	960
Vgr	4800	960	480	240
Veng	4800	960	480	240
Vfr	4800	960	480	240
Vger	4800	960	480	240
Sum	24,000	4800	2400	1200

2.6. Feature Engine—Feature Evaluation

After the completion of the segmentation step, the formulated datasets of the audio samples were subjected to the feature extraction procedure. Specifically, from each audio frame, an initial set of audio properties was extracted. In the current work, 56 features (Table 2) were computed, taking into consideration previous experience, trial and error tests and bibliographic suggestions [1–11], while the extraction process was conducted via the MIRToolbox specialized software in the Matlab environment [43]. The audio properties include time-domain variables (number of peaks, RMS (Root Mean Square) energy, number of onsets, rhythmic parameters, zero-crossing rate, etc.), spectral characteristics (rolloff frequencies, brightness, spectral statistics, etc.) and cepstral features (Mel Frequency Cepstral Coefficients). A thorough description of the extracted feature set can be addressed in [13–15,22–26].

Table 2. Extracted Audio Properties (Features) of the Windowed Signals.

Time-Domain	Spectral-Domain
rms	Npeaks_spectral
Npeaks_temporal	flux_avr, flux_std
lowenergy	rolloff_0.2, 0.5, 0.8, 0.9, 0.99
Nonsets	bright_500, 1000, 1500, 2000, 3000, 4000, 8000
event_density	sp_roughness, sp_irregularity
rthythm_clarity	Npitch, pitch_avr, pitch_std
zerocross	fundam_freq, inharmonicity
attacktime_avr,std	mode
attackslope_avr,std	Nhcdf, hcdf_avr, hcdf_std
	sp_centroid, sp_spread
Cepstral-Domain	sp_skewness, sp_kurtosis
mfcc1 . . . 13	sp_flatness
	entropy

The values of the computed audio properties from each audio frame (100 ms, 500 ms, 1000 ms, 2000 ms) were combined to the respective annotations of Table 1, in order to formulate the ground-truth

database, which is necessary for the subsequent data mining experiments, and specifically for training the supervised machine learning models.

The extracted audio properties usually implicate different discriminative performance, with their efficiency and suitability to be strongly related to the specific task under investigation. For this reason, their impact is examined in the current work via an evaluation process, that algorithmically ranks the most powerful features of the 2-layer classification problem. Specifically, the "InfoGain Attribute Evaluation" method was utilized in the WEKA environment [44], which estimates the importance of each property separately, by computing the related achieved information gain with entropy measures (for the correspondent classification scheme).

Table 3 presents the feature ranking that was formulated for the first discrimination scheme (VPM), with the implementation of the differentiated window length during the segmentation step. It has to be noted that Table 3 exhibits only the first 10 properties that prevailed during the evaluation tests, while the hierarchy/ranking continues for the whole 56-dimensional feature set.

Table 3. Feature Ranking for the main speakers, telephone conversations and music interferences (VPM) Classification Scheme.

#	W = 100 ms	W = 500 ms	W = 1000 ms	W = 2000 ms
1	bright_4000	bright_4000	bright_4000	mfcc1
2	mfcc7	rolloff_0.9	bright_8000	bright_8000
3	mfcc2	mfcc2	rolloff_0.9	hcdf_avr
4	rolloff_0.9	mfcc7	sp_flatness	sp_flatness
5	bright_3000	bright_8000	mfcc1	mfcc2
6	bright_8000	bright_3000	mfcc2	bright_4000
7	mfcc1	mfcc1	rolloff_0.99	rolloff_0.99
8	rolloff_0.2	rolloff_0.99	mfcc7	mfcc7
9	mfcc10	sp_flatness	sp_centroid	rolloff_0.9
10	rolloff_0.8	sp_centroid	bright_3000	sp_centroid

As Table 3 presents, the feature ranking for the VPM discrimination scheme involves slight variations in relationship with the respective temporal lengths. The supremacy of the spectral properties is evident, since the associated features are placed in the first ranking positions, i.e., the brightness (for 3000 Hz, 4000 Hz and 8000 Hz threshold frequencies), the rolloff frequencies (for all threshold energies) and the spectral statistical values of centroid, spread, kurtosis and flatness. Furthermore, the high discriminative power of the cepstral coefficients (mfccs) in speech/non speech classification problems is validated, as these properties hold high order ranking in the feature hierarchy.

The extracted feature vector was also subjected to the evaluation process, on the basis of the second discrimination layer LD. Table 4 presents the feature ranking that was formulated via the InfoGain Attribute algorithm. However, the hierarchy of LD scheme involves both spectral properties (brightness measures, rolloff frequencies, spectral centroid and spread) and temporal features (rms energy, zerocross, attacktime, flux, rhythm_clarity) in the first ten prevailing positions. Since the LD scheme is focused on the discrimination of (multilingual) voice signals, the prevalence of cepstral coefficients (mfccs) is somehow diminished, compared to their impact in the VPM classification layer.

Table 4. Feature Ranking for the Language Detection (LD) Scheme.

#	W = 100 ms	W = 500 ms	W = 1000 ms	W = 2000 ms
1	rolloff_0.99	rolloff_0.99	rolloff_0.99	rolloff_0.99
2	bright_8000	bright_8000	rolloff_0.9	rms
3	rolloff_0.9	rolloff_0.9	rms	sp_spread
4	bright_4000	rms	bright_8000	rhythm_clarity
5	rolloff_0.8	sp_spread	sp_spread	attackslope_avr
6	sp_centroid	rolloff_0.8	rhythm_clarity	rolloff_0.9
7	bright_3000	sp_centroid	attackslope_avr	bright_8000
8	Rms	sp_flatness	rolloff_0.8	attackslope_std
9	Entropy	flux_avr	flux_avr	flux_avr
10	sp_flatness	entropy	attackslope_std	sp_roughness

The aforementioned feature rankings of Tables 3 and 4 are based on the computation of entropy metrics, hence, they represent a comparative feature analysis (an initial indication of their impact) rather than a strict evaluation of their actual efficiency. Performance and suitability of the selected feature vector are derived from the discrimination rates of the subsequent machine learning experiments, in which all the attributes are simultaneously implicated/exploited in the classification process. Because of the investigatory nature of the current research along with the restricted sample size, the aforementioned hierarchy of audio properties cannot be generalized at this step, towards the extraction of solid conclusions, since potential differentiations could occur while moving to the augmented multilingual audio repository. Nevertheless, the feature ranking results can be useful toward combined, early and late temporal (feature) integration decision making, combining multiple modalities/machines.

The specific ranking was conducted entirely quantitatively, based on the used information evaluation algorithms. A potential explanation of the higher ranking of the spectral attributes might be found on the basis of the differentiated letters/phonemes distribution [45–48] in the various languages (for example, the increased energy containing "t" and "p" phonemes/utterances). The same effect can be explained on the fact that some languages favor explosive-like speech segments and/or instances. For instance, the rolloff_0.99 parameter (that is constantly first in all LD time-windows) efficiently detects such transients and their associated high spectra (not solely but in combination with other features). Again, the focus of the current paper and its investigative character does not leave room for other related experiments, beyond the provided ranking with the associated trial and error empirical observation and justification comments. In truth, it would be risky to attempt such an interpretation within the small-size audio dataset used, where slight variations in the recording conditions and the particularities of the different broadcasted streams could have a stronger effect than the speaker or the language attributes.

3. Results

3.1. Configuration and Validation of the Training and Evaluation Procedures

The audio samples that were collected, segmented and annotated in the previous steps, along with the respective values of the extracted features, constitute the training pairs for the machine learning and testing experiments, based on both the formulated classification schemes (VPM, LD). Extensive experiments were conducted in [14], aiming at comparing supervised classification methods (decisions trees, artificial neural systems, regressions, etc.) based on their overall and partial discrimination rates, in various implementations and schemes of broadcast audio content. In this context, the utilization of artificial neural networks (multilayer perceptrons) achieved increased and more balanced classification rates. Consequently, artificial neural systems (ANS) were selected as the main supervised machine learning technique in the current work. Several experiments were conducted regarding the network topology, in order to achieve efficient training performance, leading in the structures of one hidden

layer (with sigmoid trigger function) and an output linear layer, while an approximate number of 20–23 neurons was engaged in the intermediate layer (via trial and error tests). Furthermore, the k-fold validation method was implemented for training purposes, which divides the initial input data set into k-subsets and thereafter, the (k-1) subsets are exploited for training the classifier and the remaining subset is utilized for model validation, while the whole process is repeated k times iteratively [13,14]. The k-fold validation technique aims to evaluate the performance of the ANS topologies, and furthermore, contribute to the formulation of generalized classification rules. The value of k = 10 was selected in the supervised machine learning experiments.

In particular, the eight-minute duration of audio speech signal involves eight different speakers, at each language (one minute each). Regarding the k-fold validation, the process was involved both at the Voice/Music/Phone taxonomy (with k = 10) and in the language discrimination task (with k = 8). However, it has to be clarified that in all cases the training process was completely different at each session, while feature treatment (offset, scaling, etc.) was also applied individually, i.e., only the training samples were involved in every case. Furthermore, a randomization of the samples was applied at the audio level, with the precaution of avoiding the same speakers to be engaged both in training and validation tasks. Thus, seven out of the eight speakers were used for training, leaving the remaining one for testing/evaluation purposes. The same operation was conducted eight times, with each training loop leaving aside as unseen data a speaker for each language, also ensuring the avoidance of co-existence of same data in both training/testing subsets. In this concept, we were also able to tackle the partial scores in each language.

Additional experiments were conducted under the hold-out validation approach, this time forming pairs of two different speakers and languages (21 in total) as evaluation sets, while using the remaining data for training purposes (again, initiating entirely different/isolated training sessions for each assessment cycle). The results showed that the difference (with the k=fold validation) in the observed accuracy scores varied between 1% and 3%, thus validating the soundness of the approach. While this analysis was considered adequate for the investigative character of the paper, still, indicative testing was performed on additional (entirely unseen) audio recordings, revealing almost identical performance. Among others, mixed language sentences (i.e., Greek with English words) spoken by the same person were included, also anticipating the ability of the system to monitor language alteration at the word level. While, again, the recognition scores were attained at the same levels, the configuration of a proper window and hop lengths was tricky, depending on the speech rate of the involved speaker and related timing characteristics of the associated recordings and languages.

On the other hand, the popular K-Means algorithm was selected as the respective unsupervised learning method (clustering process), aiming at inspecting and detecting the formulation of groups of data (clusters of feature values), according to a similarity metric. The measure that determines the integration of each sample into a cluster is usually related to a distance metric (Euclidean, Manhattan, Chebyshev, Min-Max, etc.), that estimates the nearness of each sample to the cluster center. The most commonly used metric of Euclidean Distance was also used in the current work (for simplicity reasons), in order to investigate the possible formulation of data clusters, based on the defined discrimination categories of the proposed schemes.

The overall pattern recognition performance (P) of the ANS modules was estimated for each of the implemented schemes by the generated confusion matrices. Specifically, the classification rate is represented by the % ratio of the number of the correctly classified samples to the total number of the input samples [1–3]. In the same way, the partial recognition rate P(X) of class X was based on the % ratio of the correctly classified samples (within the class X) to the total number of samples that X class includes, according to the ground-truth annotations. On the contrary, the clustering models attempt to detect the formulation of data groups directly. Hence, the previous metrics cannot be utilized for the performance estimation, since the unsupervised method (K-Means) does not take into account the ground-truth dataset. It has to be noted that one of the main objectives of the current investigation is to examine the feasibility of the automatic unsupervised classification process in audio signals

through clustering strategies and compare the results with the corresponding ones by supervised machine learning (ANS). For this reason, the formulated data clusters of K-Means were compared to the respective annotated classes of ANS, to evaluate the cluster concentration solely. Consequently, the partial integrity/performance P(U) of a cluster U is calculated as the % ratio of the number of samples of the ground-truth class X (that were assigned to cluster U), to the total number of class-X samples. This metric essentially represents a % estimation measure of the resemblance between cluster U and class X (also known as cluster purity). In this way, the overall performance of the clustering process (P) was computed through the % proportion of the correctly gathered samples (summed for each class X) to the total number of samples. These metrics also favored the performance evaluation independently of the size of the formed clusters, which, in real-world radio content, are expected to have unbalanced size distributions. The number of clusters in the unsupervised classification experiments was configured manually to four (in accordance to the attempted formulation of groups of data), based on trial and error experiments, while also employing the expectation maximization (EM) algorithm [44], implicating the probability distributions on the involved samples. Moreover, this parameter could also be set by the user based on prior knowledge on specific shows, supporting this way the program-adaptive classification process in a real-world scenario that this research investigates.

3.2. Performance Evaluation Results

The supervised and clustering techniques (ANS and K-Means) were deployed independently in the first discrimination scheme VPM. Thereafter, the machine learning techniques could be combined in the second classification layer LD, promoting either a strict supervised/unsupervised character of the classification modules or a hybrid coupling. Figure 3 presents the above combinations for the two layers of the adopted taxonomy. For simplicity reasons the letters S, U are used for Supervised and Unsupervised classification, respectively. It has to be noted that the clustering "path" leads in the integration of more automation in whole semantic analysis process, compared to the prerequisites of ground-truth data formulation that supervised learning demands. However, it cannot stand alone, because it does not offer a class labeling outcome (either comparison with a supervised module is needed or some subjective labeling). Given the hierarchical scheme with the two classification layers (VPM, LD) and the two different types of classifiers (S, U), four combinations of the machine learning methods, namely SS, SU, US, UU, are formed according to each path (Figure 3).

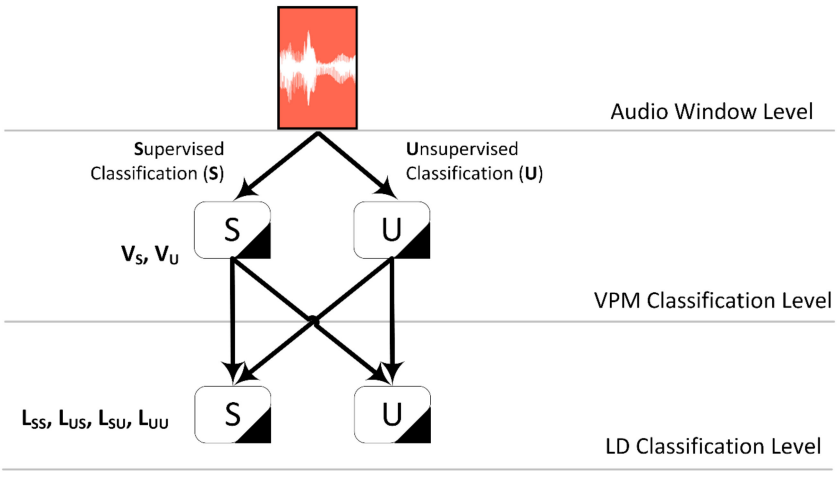

Figure 3. Configuration of the different classification paths, combining the two different classifiers (S, U) and the two taxonomy levels (VPM, LD).

Table 5 and Figure 4 show the classification and clustering results for the first discrimination layer (VPM), according to the respective window lengths 100 ms, 500 ms, 1000 ms, 2000 ms. The increased overall and partial discrimination rates for the VPM scheme while the supervised method of ANS

is utilized can be observed. The maximum classification percentage for voice signal is achieved via the segmentation window of 1000 ms. The discrimination percentage of 99.58% refers to the correct classification of 1912 out of 1920 voice data, leaving eight misclassified samples. Moreover, the clustering approach implicated the efficient formulation of data groups, while addressing the respective categories of the ground-truth set. The maximum overall and partial discrimination rates also derive from the utilization of 1000 ms window length, namely 89.75% and 89.95%. In this case, the voice cluster involved 1727 samples out of 1920 previously annotated. It has to be noted that the misclassified data (eight for supervised ANS and 193 for the unsupervised K-Means) had to be removed before proceeding to the next discrimination scheme LD, in order to avoid the transmission of the classification errors in the hierarchical implementation. In this way, only the 1912 samples (voice signal with a duration of 31 min and 52 s) and 1727 samples (voice signal with a duration of 28 min and 47 s) were engaged in the language discrimination process, while also retaining the strategy of parametric window lengths for comparison purposes. While this accommodation was made for purely experimental purposes, such kind of screening is not feasible in real world scenarios, in which error propagation is inevitable. However, this unwanted effect can be diminished by adjusting the confidence classification parameter (C) that was explained with regard to the analysis of Figure 1.

Table 5. Classification Results for VPM Scheme.

	Window-Length	P	P(V)	P(P)	P(M)
S	100 ms	95.81	97.80	85.04	90.67
	500 ms	98.15	99.01	94.58	94.79
	1000 ms	99.25	99.58	96.67	99.17
	2000 ms	98.63	99.06	97.50	96.25
U	100 ms	74.73	74.28	80.71	72.29
	500 ms	77.31	76.46	83.13	78.33
	1000 ms	89.75	89.95	93.33	84.58
	2000 ms	86.58	86.35	90.83	84.17

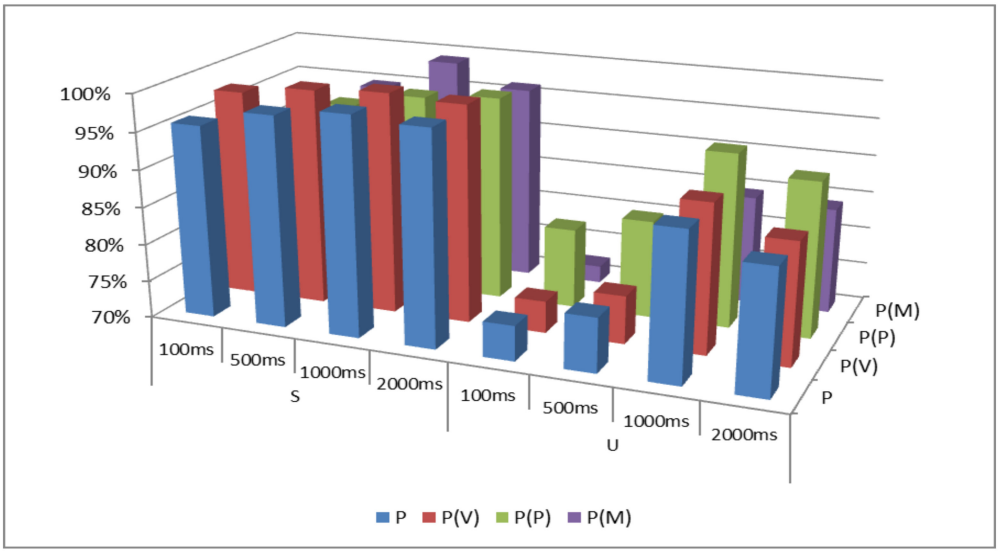

Figure 4. Classification Results for VPM Scheme.

Table 6 and Figure 5 present the overall pattern recognition scores for the combination of the different classifiers in the LD scheme, including the partial recognition rates for each language. Again, the paths starting with supervised language detection (SU, SS) seem to have better performance when compared to the associated unsupervised paths (UU, US). However, it seems that the S and U models do not have significant recognition differences in the LD layer, which can be exploited during the

implementation of late integration and/or ensemble learning approaches. Overall, even with such small datasets, the achieved accuracy is considered quite satisfactory, thus making an initial proof of concept for the proposed methodology presented in Figure 1.

Table 6. Classification Results for LD Scheme.

	Win.	P	P(gr)	P(eng)	P(fr)	P(ger)
SS	100 ms	88.53	91.99	87.97	84.29	89.85
	500 ms	93.59	95.92	92.68	92.47	93.31
	1000 ms	96.08	98.54	94.56	96.03	95.19
	2000 ms	98.01	99.58	96.23	99.58	96.65
SU	100 ms	88.11	91.10	87.72	82.93	90.69
	500 ms	93.72	96.99	92.47	92.71	92.71
	1000 ms	95.60	97.91	93.98	96.76	93.75
	2000 ms	97.80	99.54	95.37	99.07	97.22
US	100 ms	65.98	68.77	65.38	66.99	62.78
	500 ms	76.78	75.84	79.60	77.20	74.48
	1000 ms	81.28	81.80	82.01	81.59	79.71
	2000 ms	84.83	85.77	84.94	84.52	84.10
UU	100 ms	56.35	59.93	56.13	57.57	51.78
	500 ms	60.63	62.80	59.33	60.30	60.07
	1000 ms	73.42	74.01	75.23	72.45	71.99
	2000 ms	80.67	81.02	80.09	82.41	79.17

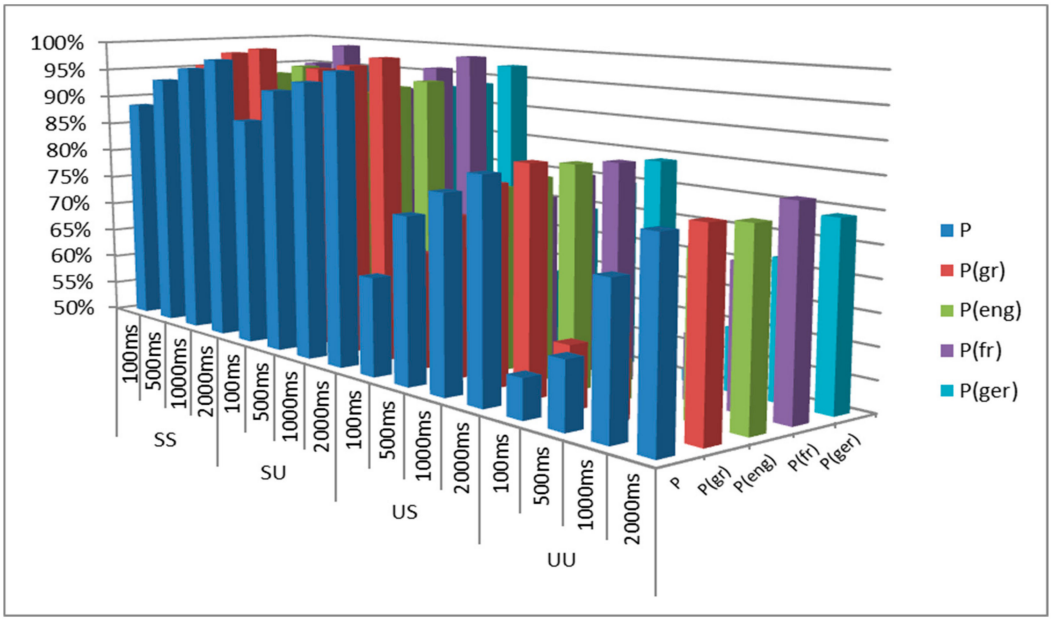

Figure 5. Classification Results for LD Scheme.

More specifically, the employed investigations indicated the feasibility of potential detection and classification of spoken language, even when these audio data derive from short-duration broadcast content. Therefore, the conducted experiments validated the possibility to quickly/efficiently identify spoken-language, based on a small amount of input data, i.e., short-duration annotated files with the voices of the main speakers. Furthermore, previous work has shown that radio program adaptive classification is beneficial and outperforms generic solutions [1,33]. The problem under discussion reflects a real-word scenario, encountered in modern media/monitoring organizations, where semi-automated indexing and documentation are needed, which could be facilitated by the proposed language detection preprocessing. In these grounds, the experimentation with a small dataset

is really essential in the direction of the potential formulation of a quick-decision model (and actually more demanding). Hence, the target of this work was not to implement a generic (i.e., for every purpose) language recognition model, in which elongated ground-truth (e.g., using audio books) could be formed to serve the pattern recognition needs (i.e., exhaustive learning, deep learning architectures, etc.). This attempt would be possibly very useful when the Generic Audio Language Classification Repository will be gradually augmented via the iterative radio-broadcast adaptive operation of Figure 1. However, even for this subsequent step, as described above, the validation of efficient classification performance for reduced size input data is a prerequisite, before moving into further analysis, justifying the experimentation on small duration audio signals as an initial exploring procedure. Consequently, this strategy matches the investigatory inception of the project, which seeks for indicators that could be applied in a second step with a larger dataset and/or an already pre-trained model, with the purpose of delivering overall results.

4. Discussion and Future Work

The current work addressed the problem of audio content classification, derived from European radio productions, based on the spoken language. In order to support this process, a hierarchical structure was proposed and implemented, consisting of two successive discrimination schemes (VPM and LD). Moreover, several segmentation window lengths were tested for comparison purposes, while supervised machine learning and clustering techniques were employed, both independently and combined. In the first step, the conducted experiments achieved high percentages of voice signal detection/isolation (above 99%). The language classification process achieved overall and partial discrimination rates above 90% in most cases, indicating the distinctive characteristics of each spoken language. This implementation was supported by an effective multi-domain feature engine, based on the increased classification performances. Finally, the successful formulations of data clusters favored the integration of automation in the whole semantic analysis process, because of their independence of the ground-truth data.

In the context of future potentials, the presented work could be extended in order to involve radio broadcasts, in more European countries/languages (Spanish, Italian, etc.), or even dialects, which constitute slight variations of the same language, usually dependent on origin. For this purpose, the adaptive classification process could be fed with specialized broadcast data that involve these kinds of language alterations, deriving possibly from radio content in the regions of the dialects. It is also expected that specific radio producers would be attached to specific dialects, which favor the adaptive nature of the approach. In the same direction, another pattern recognition scheme that the model could involve is the identification of a male/female voice in the broadcasted signals (taking advantage of its binary nature and the used hierarchical type). The aforementioned goals can be supported by the computation, evaluation and experimentation within an expanded feature set (even in trial and error tests), because of the complex and specialized character of voice data in the radio productions (articulation, speed, transmission channels, etc.). Furthermore, more thorough research may be conducted on the phonetic level, with very small window lengths, in order to captivate special sound properties, dependent on-air movement (from lungs to mouth, nose, etc.).

Since radio broadcasts appeal to differentiated interests of the audience, the semantic analysis process may be extended in the thematic classification of radio content, dependent on the "character" of the corresponding broadcast (news, athletic, music programs, etc.). These discrimination potentials can be empowered by intelligent pattern recognition systems, aiming to extract and formulate effective metadata mechanisms, aiming for efficient content description, storing, accessing and management, meeting the users' demands and expectations.

As already explained, the motivation behind this work emanates from specific practical needs in documenting and indexing audio broadcasted content, using short-in-duration annotated podcast samples from indicative/past radio program streams. In this context, the whole approach with the

small dataset and the machine training difficulties constitute a very demanding problem, which also signifies the research contribution of the conducted work.

Clearly, the outmost target remains the gradual implementation of the full potentials of the proposed methodology, as depicted in Figure 1. In this context, iterative self-learning procedures can be elaborated for hierarchical language (and generally audio) classification in broadcasting content, taking advantage of the well-organized big-data structure (assembled by multiple program-adaptive sub-groups). Moreover, depending on the size of the "global" and "local" repositories, more sophisticated deep learning architecture can further propel the potentials of this initiative.

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

Funding: This research received no external funding.

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. 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 AES Convention, Budapest, Hungary, 26–29 April 2012; pp. 513–520.
3. 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 2012 Seventh International Workshop on Semantic and Social Media Adaptation and Personalization, Luxembourg, 3–4 December 2012; pp. 68–72. [CrossRef]
4. Vrysis, L.; Tsipas, N.; Thoidis, I.; Dimoulas, C. 1D/2D Deep CNNs vs. Temporal Feature Integration for General Audio Classification. *J. Audio Eng. Soc.* **2020**, *68*, 66–77. [CrossRef]
5. Thoidis, I.; Vrysis, L.; Pastiadis, K.; Markou, K.; Papanikolaou, G. Investigation of an Encoder-Decoder LSTM model on the enhancement of speech intelligibility in noise for hearing-impaired listeners. In *Audio Engineering Society Convention 146*; Audio Engineering Society: New York, NY, USA, 2019.
6. Kostek, B. *Perception-Based Data Processing in Acoustics: Applications to Music Information Retrieval and Psychophysiology of Hearing*; Springer: Berlin/Heidelberg, Germany, 2005; Volume 3.
7. Korvel, G.; Treigys, P.; Tamulevičius, G.; Bernatavičienė, J.; Kostek, B. Analysis of 2D Feature Spaces for Deep Learning-based Speech Recognition. *J. Audio Eng. Soc.* **2018**, *66*, 1072–1081. [CrossRef]
8. Ntalampiras, S. Toward Language-Agnostic Speech Emotion Recognition. *J. Audio Eng. Soc.* **2020**, *68*, 7–13. [CrossRef]
9. Vrysis, L.; Tsipas, N.; Dimoulas, C.; Papanikolaou, G. Extending Temporal Feature Integration for Semantic Audio Analysis. In *Audio Engineering Society Convention 142*; Audio Engineering Society: New York, NY, USA, 2017.
10. Bountourakis, V.; Vrysis, L.; Konstantoudakis, K.; Vryzas, N.N. An Enhanced Temporal Feature Integration Method for Environmental Sound Recognition. *Acoustics* **2019**, *1*, 410–422. [CrossRef]
11. Dimoulas, C.; Vegiris, C.; Avdelidis, K.; Kalliris, G.; Papanikolaou, G. Automated Audio Detection, Segmentation, and Indexing with Application to Postproduction Editing. In Proceedings of the 122nd AES Convention, Vienna, Austria, 5–8 May 2007.
12. Vegiris, C.; Dimoulas, C.; Papanikolaou, G. Audio Content Annotation, Description and Management Using Joint Audio Detection, Segmentation and Classification Techniques. In Proceedings of the 126th AES Convention, Munich, Germany, 7–10 May 2009.
13. Vrysis, L.; Tsipas, N.; Dimoulas, C.; Papanikolaou, G. Crowdsourcing audio semantics by means of hybrid bimodal segmentation with hierarchical classification. *J. Audio Eng. Soc.* **2016**, *64*, 1042–1054. [CrossRef]
14. Künzel, H.J.; Alexander, P. Forensic Automatic Speaker Recognition with Degraded and Enhanced Speech. *J. Audio Eng. Soc.* **2014**, *62*, 244–253. [CrossRef]

15. Korvel, G.; Treigys, P.; Tamulevičius, G.; Bernatavičienė, J.; Bożena, K. Borrowing 2D Feature Maps from Audio Signal Analysis to Deep Learning-based Speech Recognition. *J. Audio Eng. Soc.* **2019**. (accepted, in editorial process).
16. Barras, C.; Zhu, X.; Meignier, S.; Gauvain, J.-L. Multistage speaker diarization of broadcast news. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1505–1512. [CrossRef]
17. Zewoudie, A.W.; Luque, I.; Hernando, J. The use of long-term features for GMM-and i-vector-based speaker diarization systems. *EURASIP J. Audio Speech Music Process.* **2018**. [CrossRef]
18. Shi, Y.; Zhou, J.; Long, Y.; Li, Y.; Mao, H. Addressing Text-Dependent Speaker Verification Using Singing Speech. *Appl. Sci.* **2019**, *9*, 2636. [CrossRef]
19. Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-vectors: Robust DNN embeddings for speaker recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5329–5333. [CrossRef]
20. Dimoulas, C.; Kalliris, G. Investigation of wavelet approaches for joint temporal, spectral and cepstral features in audio semantics. In Proceedings of the 134th AES Convention, Rome, Italy, 4–7 May 2013; pp. 509–518.
21. Barbedo, J.; Lopes, A. A robust and computationally efficient speech/music discriminator. *J. Audio Eng. Soc.* **2006**, *54*, 571–588.
22. Tsipas, N.; Vrysis, L.; Dimoulas, C.; Papanikolaou, G. Efficient audio-driven multimedia indexing through similarity-based speech/music discrimination. *Multimed. Tools Appl.* **2017**, 1–19. [CrossRef]
23. Tsipas, N.; Vrysis, L.; Dimoulas, C.; Papanikolaou, G. Content-Based Music Structure Analysis using Vector Quantization. In Proceedings of the 138th AES Convention, Warsaw, Poland, 7–10 May 2015; pp. 419–424.
24. Hellmuth, O.; Allamanche, E.; Kastner, J.H.T.; Lefebvre, N.; Wistorf, R. Music Genre Estimation from Low Level Audio Features. In Proceedings of the 25th AES International Conference, London, UK, 17–19 June 2004.
25. Dimoulas, C.A.; Symeonidis, A.L. Syncing Shared Multimedia through Audiovisual Bimodal Segmentation. *IEEE Multimed.* **2015**, *22*, 26–42. [CrossRef]
26. Lerch, A. *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*; John Wiley & Sons: Hoboken, NJ, USA, 2012; ISBN 9781118393550.
27. Vryzas, N.; Kotsakis, R.; Liatsou, A.; Dimoulas, C.; Kalliris, G. Speech Emotion Recognition for Performance Interaction. *J. Audio Eng. Soc.* **2018**, *66*, 457–467. [CrossRef]
28. Vryzas, N.; Vrysis, L.; Matsiola, M.; Kotsakis, R.; Dimoulas, C.; Kalliris, G. Continuous Speech Emotion Recognition with Convolutional Neural Networks. *J. Audio Eng. Soc.* **2020**, *68*, 14–24. [CrossRef]
29. 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.* **2014**, *2*, 62–80. [CrossRef]
30. Sharan, R.V.; Moir, T.J. An overview of applications and advancements in automatic sound recognition. *Neurocomputing* **2016**, *200*, 22–34. [CrossRef]
31. 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]
32. Strisciuglioa, N.; Ventob, M.; Petkova, N. Learning sound representations using trainable COPE feature extractors. *Pattern Recognition.* **2019**, *92*, 25–36. [CrossRef]
33. Kotsakis, R.; Kalliris, G.; Dimoulas, C. Extending radio broadcasting semantics through adaptive audio segmentation automations. *Multimed. Tools Appl.* **2020**. (prepared for submission).
34. Dweik, B.; Qawar, H. Language choice and language attitudes in a multilingual Arab Canadian community: Quebec-Canada: A sociolinguistic study. *Br. J. Engl. Linguist.* **2015**, *3*, 1–12.
35. Ramaiah, V.S.; Rao, R.R. Speaker diarization system using HXLPS and deep neural network. *Alex. Eng. J.* **2018**, *57*, 255–266. [CrossRef]
36. Kotsakis, R.; Mislow, A.; Kalliris, G.; Matsiola, M. Feature-Based Language Discrimination in Radio Productions via Artificial Neural Training. In Proceedings of the 10th Audio Mostly, ACM, New York, NY, USA, 7–9 October 2015. [CrossRef]
37. Segbroeck, M.; Travadi, R.; Narayanan, S. Rapid Language Identification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 1118–1129. [CrossRef]

38. Leonard, R.G.; Doddington, G.R. *Doddington, Automatic Language Discrimination*; No. TI-08-77-46; Texas Instruments Inc.: Dallas, TX, USA, 1978.
39. Muthusamy, Y.K.; Barnard, E.; Cole, R.A. Reviewing automatic language identification. *IEEE Signal Process. Mag.* **1994**, *11*, 33–41. [CrossRef]
40. Campbell, W.M.; Campbell, J.P.; Reynolds, D.A.; Singer, E.; Torres-Carrasquillo, P.A. Support vector machines for speaker and language recognition. *Comput. Speech Lang.* **2006**, *20*, 210–229. [CrossRef]
41. Kuncheva, L.I. *Combining Pattern Classifiers: Methods and Algorithms*; John Wiley & Sons: Hoboken, NJ, USA, 2004; ISBN 978-1-118-31523-1.
42. Zhu, X.; Goldberg, A.B. Introduction to Semi-supervised Learning. In *Synthesis Lectures on Artificial Intelligence ad Machine Learning*; Brachman, R.J., Dietterich, T.G., Eds.; Morgan & Claypool Publishers: San Rafael, CA, USA, 2009; ISBN 978-1598295474.
43. Lartillot, O.; Toiviainen, P. Mir in matlab (ii): A toolbox for musical feature extraction from audio. In Proceedings of the 8th International Conference on Music Information Retrieval, Vienna, Austria, 23–27 September 2007.
44. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [CrossRef]
45. Zim, H.S. *Codes & Secret Writing: Authorized Abridgement*; Scholastic Book Service: New York, NY, USA, 1967.
46. English Letter Frequency Counts: Mayzner Revisited or Etaoin Srhldcu. 2018. Available online: https://norvig.com/mayzner.html (accessed on 14 April 2020).
47. Corpus de Thomas Tempé. Archived from the original on 30 September 2007. Available online: https://web.archive.org/web/20070930194046/http://gpl.insa-lyon.fr/Dvorak-Fr/CorpusDeThomasTemp%C3%A9 (accessed on 14 April 2020).
48. Beutelspacher, A. *Kryptologie*, 7th ed.; Vieweg: Wiesbaden, Germany, 2014; p. 10. ISBN 3-8348-0014-7.