information

MDPI

Article

Web Radio Automation for Audio Stream Management in the Era of Big Data

Nikolaos Vryzas *, Nikolaos Tsipas and Charalampos Dimoulas

Multidisciplinary Media & Mediated Communication Research Group (M3C), Aristotle University of Thessaloniki, 541 24 Thessaloniki, Greece; nitsipas@auth.gr (N.T.); babis@eng.auth.gr (C.D.)

* Correspondence: nvryzas@auth.gr

Received: 12 March 2020; Accepted: 9 April 2020; Published: 11 April 2020

Abstract: Radio is evolving in a changing digital media ecosystem. Audio-on-demand has shaped the landscape of big unstructured audio data available online. In this paper, a framework for knowledge extraction is introduced, to improve discoverability and enrichment of the provided content. A web application for live radio production and streaming is developed. The application offers typical live mixing and broadcasting functionality, while performing real-time annotation as a background process by logging user operation events. For the needs of a typical radio station, a supervised speaker classification model is trained for the recognition of 24 known speakers. The model is based on a convolutional neural network (CNN) architecture. Since not all speakers are known in radio shows, a CNN-based speaker diarization method is also proposed. The trained model is used for the extraction of fixed-size identity d-vectors. Several clustering algorithms are evaluated, having the d-vectors as input. The supervised speaker recognition model for 24 speakers scores an accuracy of 88.34%, while unsupervised speaker diarization scores a maximum accuracy of 87.22%, as tested on an audio file with speech segments from three unknown speakers. The results are considered encouraging regarding the applicability of the proposed methodology.

Keywords: web radio; big data; web application; speaker recognition; speaker diarization; convolutional neural networks

1. Introduction

In the past decade, there has been a breakthrough concerning the production and distribution of digital content on the web and in social media. This outbreak of available data has highlighted the importance of the development of suitable tools and frameworks for editing and management of the created content, aiming at more efficient distribution and consumption. Meanwhile, there have been important developments in data-driven techniques for automated semantic information retrieval of broadcast content, based on machine and deep learning models [1–11]. For the aims of radio and television production, validation and media asset management are essential for semantic annotation and archiving of content. Radio producers can also benefit from semantic analysis tools for more efficient content production [12,13].

1.1. Radio in an Evolving Ecosystem

The transition from analog radio to the digital world has followed a different path than digital television. There was not a global mandatory transition from analog to digital format broadcasting in the radio spectrum. One big factor that shaped the landscape was the rise of web radio. A big part of radio consumption is via personal computers through web radio station websites. Smartphone devices, which comprise a major choice for mobile radio listening, are no longer equipped with radio receivers, but depend on audio transmitted over 3G/4G networks [14]. While traditional radio is supplemented

rather than substituted by web radio, experts in the media industry favor the rise of on-demand radio and non-linear broadcasting in the next decade [14].

Convergence across different media is expected, leading to an enhancement of audio content with more visual and diverse content that can be accessed in the social media and across multiple platforms [14–17]. Cross-platform distribution aims at intensifying and prolonging wider audience engagement [17]. Radio stations often make their content available for on-demand listening. Meanwhile, podcasts have emerged as a genre. The difference between audio-on-demand and podcast is blurred and mostly defined by aesthetic and empirical aspects, so the terms are often used interchangeably [15,18,19]. Trending directions set by the industry want to make radio content more podcast-friendly [18]. This requires supporting shareability, discoverability, repeatability, reproducibility, and asynchronous access [15]. Personalization and customization can also play an important role in the viability of modern radio [16].

Podcasting is not seen as an alternative to broadcasting, but the two of them are meant to coexist, in order to address different audiences [19]. While broadcast radio is still relevant in many scenarios, younger listeners are more likely to pursuit content on demand. Despite the new possibilities and challenges in the collection, storage and distribution of radio content, the traditional values of radio concerning mobility, ease-of-access, real-time content, and interactivity should not be abandoned [20].

1.2. Big Audio Data

The explosion of the data production rate has not only created new potential and possibilities in many domains, but has also set some new challenges, and the rise of a new field, which is commonly referred to as big data [21–29]. While there is not one single and universal definition, big data mostly involves data volumes that exceed traditional processing capabilities [21]. Descriptive models can be found in the literature, using the 3Vs model [22–24], which set the main characteristics that define big data. The model is often extended with the addition of supplementary dimensions [21,25–29]. Commonly found big data attributes include:

Volume: This refers to the magnitude of data. Large amounts of data are gathered from a range of sources. The size of big data volumes is not clearly and universally defined and measurable. It evolves through time, since thresholds rise, and depends on the type of data [24].

Variety: There are multiple kinds of data for the analysis of a situation or event, with structural heterogeneity [24]. Structured, unstructured and semi-structured data, including text, audio, video, etc., can be part of the same volume.

Velocity: This refers to the rate at which data is gathered. Besides gathering and stocking data, rapid decision-making processes are often required concerning the next data to be gathered and analyzed [21].

Variation: Velocity is not always consistent. Variation expresses the fluctuation of the rate in data flow [24].

Veracity: With rapid data generation and gathering, the truthfulness of data is a very important concept. Data cannot always be trusted, and analysis requires a data cleansing processes [21].

Volatility: Data storage has a retention period. It expresses the amount of time that it is meaningful and useful to store data [21]. While this concept is not new, and it concerns all databases, it becomes much more relevant in big data scenarios, due to volume, variety, and velocity.

Value: This is the desired outcome. It depends heavily on the applied policy on structuring and managing data [21].

The main objective of big data analytics is to provide the tools which utilize the rest of the Vs to extract value from data [21]. This requires structuring data so that the resulting information that comes from unstructured data can be used to understand a process or interface with another application [22]. The application domain is very important to define the purposes of knowledge discovery [23]. Knowledge extraction, abstraction, and aggregation of information are crucial for discoverability of content and are key features for Semantic Web technologies, in the direction of

creating an Internet of Things [25]. Mobile data and the upcoming 5G network technology promise All-IP Network (AIPN) services for cloud computing [26].

Audio and speech analytics specify the extraction of information from unstructured audio data. Main applications include the domains of customer service, call centers, social media, the media industry, health care, and content-based analytics. These industries produce big audio data streams daily [24].

Deep learning has been associated with big data handling [27–29]. Such techniques utilize a massive amount of data for hierarchical feature extraction to provide complex abstractions and data representations [27–29]. There are many technical challenges to be addressed in managing large-scale, high-dimensional, rapidly changing data. While the field is still considered low in maturity [29], deep learning has been proved to have greater potential for efficient management of big data volumes, compared to traditional ML approaches [27–29].

1.3. Speaker Diarization

Speaker diarization is defined as the problem of deciding "who spoke when?" [30,31], which serves many applications in broadcasting [32], conferencing, and intelligent information retrieval. It includes the sub-tasks of segmenting the input audio and assigning each segment to a certain speaker. Two main approaches are dominant in the literature: top-bottom and bottom-up clustering [30,31]. The number of clusters corresponds to the number of different speakers. In the first approach, the model is initialized with one (or a few) clusters, while in the latter with an excessive number of clusters. In both strategies, the goal is to converge to the optimum number of clusters/speakers. diarization error rate (DER) is commonly used to measure the performance of speaker diarization systems. It is defined as the sum of missed speech error, false alarm speech error and speaker error [31]. The first two errors refer to voice activity detection error, while the third refers to the assignment of speech segments to a wrong speaker. In some cases, DER only refers to speaker error to simplify the evaluation [33].

Identity vector (i-vector) has been the standard feature extraction procedure for speaker recognition and, by extension, speaker diarization. The audio input is segmented in an unsupervised way in 1–2 s segments, from which i-vectors are extracted [34]. A factor-analysis front-end along with principal components analysis of i-vectors is investigated in [35], using data from telephone conversations. Spectral clustering is also proposed as an alternative to K-means for the stage following i-vector extraction and principal components analysis [33]. Another clustering approach is evaluated against agglomerative clustering, incorporating integer linear programming in order to find an optimal clustering solution, leading to a DER decrease [36]. ILP is also integrated into the open-source toolbox for broadcast news diarization LIUM [32]. In the case of the diarization of a big collection of recordings, the clusters which are defined after applying diarization separately in each recording can be used to perform a two-stage clustering approach, and compress the information [37].

In the most common scenario, Gaussian mixture models and factor analysis are used to reduce dimensionality resulting in the compressed representation of i-vectors, which are then compared using probabilistic linear discriminant analysis (PLDA). In [38], a deep neural network (DNN) is trained to learn a fixed embedding and scoring metric to replace the stages of i-vector extraction and PLDA scoring used in baseline techniques and is proved to outperform them. The DNN is used to map speech utterances to fixed x-vector embeddings, having as its input 30-dimensional mel-frequency cepstral coefficients (MFCCs) with a frame-length of 25 ms, mean-normalized over a sliding window of up to 3 s. The experimentation is extended for multi-speaker recordings, achieving state-of-the-art accuracy in common databases [39]. In a supervised approach, an ANN architecture is investigated in [40]. MFCC features are extracted from two audio frames and are used as input in the first layer of the ANN. The ANN is trained as a classifier with two classes, deciding if the two frames belong to the same speaker or not.

As opposed to i-vectors, audio embeddings that come from deep learning approaches are generally named d-vectors [40–47]. Long short-term memory networks (LSTMs), a special type of recurrent

neural network (RNN), have gained much popularity for the diarization task, following the rise of deep learning in the past few years. RNNs and LSTM are considered appropriate for the task, because of their feasibility in capturing the sequential information of audio signals [40–45]. Convolutional neural networks (CNNs) have also been proved efficient in detecting speaker changes and extracting speaker embeddings [46,47].

1.4. Research Aims

The motivation of the current research is to make use of domain knowledge in the field of radio production and state-of-the-art machine learning practice to enhance the processes of radio production, distribution and consumption. The directions set by industry experts for the future of radio, as well as the advances in big data management, have been taken into consideration.

The recent popularity of podcasts and audio-on-demand, as described in Section 1.2, highlights the need for discoverability. To facilitate customization and personalization, listeners have to be able to access radio content based on several criteria, like topic selection, radio producer, guests, music aesthetics, etc. While radio stations may manually provide some tags and descriptions, this direction requires knowledge extraction from the unstructured audio streams.

The main contributors of big audio data are radio stations, producers and amateur users. The aforementioned categories can all benefit from the adaptation of web radio to emerging listening habits. In this paper, a framework is proposed that links knowledge extraction to the production process. A web-application for live radio production is presented that integrates real-time semantic annotation and logging. By providing the application publically, radio producers contribute without stressing their common workflows. A semi-structured XML scheme is described to enhance access and management of stored content. Additionally, deep learning techniques are evaluated for unsupervised knowledge extraction.

In Section 2, the concept and the functionality of the web application are presented. The knowledge structure is also explained within the general framework of radio content on demand access. An approach for speaker recognition and diarization based on deep convolutional neural network modalities is introduced. The implementation and experimental procedures are explained thoroughly. In Section 3, the evaluation results are presented. In Section 4, the conclusions of the research are discussed, and future research goals are set.

2. Materials and Methods

2.1. A Framework for Knowledge Extraction from Radio Content

Radio shows are usually recorded to be stored by the station for archiving reasons or to be accessible to users online. In the most common scenario, the producer is responsible for writing a small description that accompanies the provided podcast, while the audio file is unstructured. The vision of personalized and customized radio content access requires a structured information extraction from audio files that can be discovered by listeners. Users have to be able to browse podcasts based on search queries regarding the content. Moreover, listeners should be able to access the parts of interest in a specific audio file, based on the content.

As it has been described in Section 1.2, one of the main visions for the new radio concerns the enrichment of live broadcast and on-demand content with visual information that can be accessed across multiple platforms and social media. For this reason, along with segmentation and intelligent information retrieval, textual information concerning the broadcast content can be provided to the audience.

The proposed knowledge extraction scheme is demonstrated in Figure 1. The delivered content is segmented into music and speech segments [48]. Music metadata is extracted from every music segment, providing information concerning the title, artist, genre, etc. For the analysis of the speech segments, speaker transition detection, speaker diarization, or speaker recognition in the case of

known speakers/radio producers is applied, to divide discussions into specific-speaker speech excerpts. Speech-to-text allows the extraction of transcripts for every excerpt. The structure of the segmented audio file is shown in Figure 1.

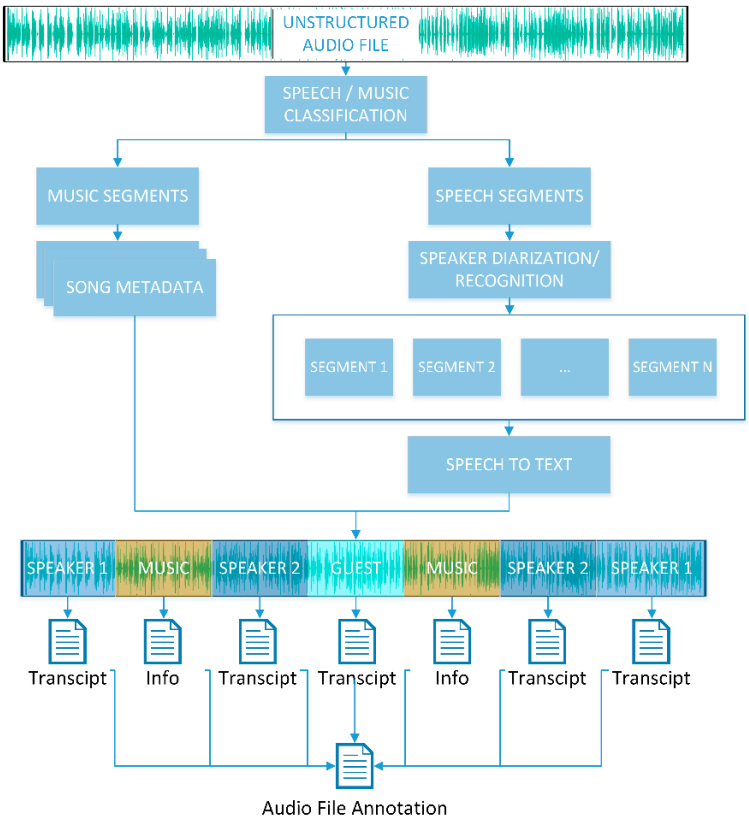

Figure 1. The proposed framework for podcast segmentation and knowledge extraction.

Given this information, the interactive transcripts of podcasts and radio shows can be provided so that the users can navigate through the document, search for keywords, read the generated text, acquire information and jump to the desired audio segment. For example, a user interested in commentary on a specific topic by several analysts can search using suitable keywords, and listen to the specific parts of conversations. This is also applicable for hearing-impaired audiences, who are usually excluded from anything that happens on the radio. An implementation of such an interactive transcript in HTML and Javascript is shown in Figure 2. Every annotation of a different speaker with the respective starting point in time is a functional button that plays the audio at the specified time of the selected utterance.

Furthermore, when real-time segmentation and annotation is available, textual data streaming containing semantic information of the broadcast audio is possible, on the web, on social media and other platforms, or even as dynamic label segment (DLS) in digital audio broadcasting (DAB) [49].

2.2. A Web Application for Live Radio Production and Annotation

The technologies involved in the extraction of the aforementioned knowledge scheme include speech/music classification, song identification, speaker diarization, speaker recognition, and speech-to-text. These problems are mostly addressed with supervised and unsupervised machine learning. While all of these fields have grown significantly in the past decade, data-driven predictions always involve a certain error rate. Additionally, some of them, like speaker recognition, cannot be treated globally, but a dedicated model has to be trained for a specified group of known speakers, e.g., the employees of a radio station. In this case, vast amounts of labeled data are required. Manual

data labeling is a time-consuming process, vulnerable to human error. This often sets a bottleneck to the viability of deep learning approaches, since small organizations cannot afford this cost.

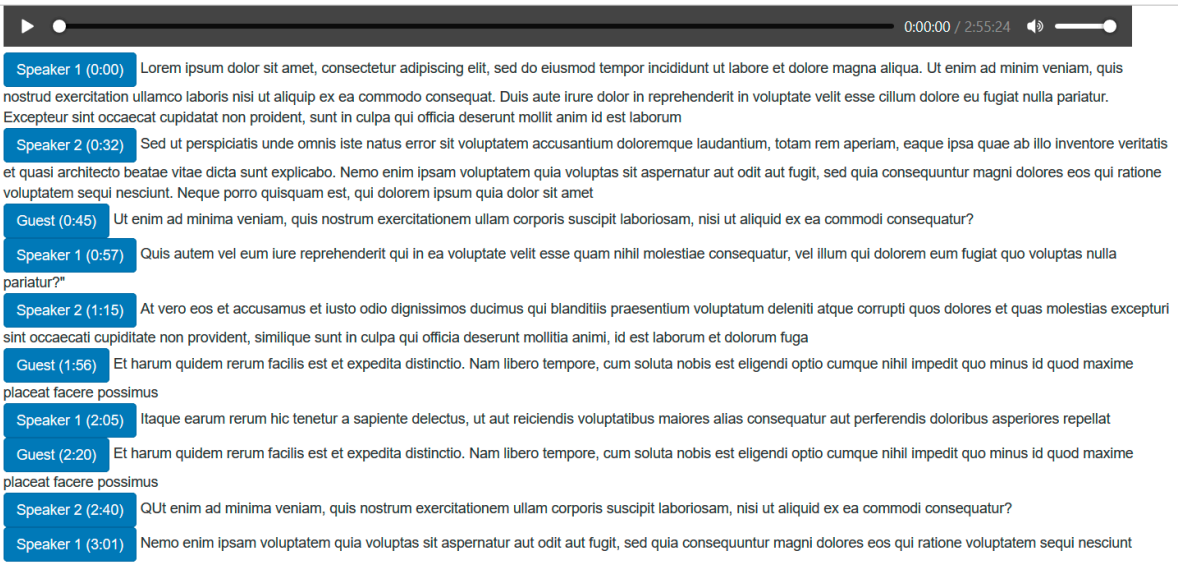

Figure 2. Interactive navigation through an annotated podcast.

To surpass this obstacle, a web application has been developed. The application covers the common workflows and functionality of most live mixing and radio production software, while, at the same time, integrating semantic enhancement of the produced files on-the-fly. The functionality is available to the user through an HTML5-based graphical user interface, which is demonstrated in Figure 3. The producer can load audio files to create playlists, adjust the audio settings, apply cross-fade for transitions, and turn on and off speaker microphones. For every change of sound source (microphone inputs) or songs from the playlist, a transition event is registered in the log file. This log file serves as the annotation file, containing the time segmentation and information accompanying the audio file. Depending on the recording and streaming setup, the inputs of the computer can match the different channels of a multichannel audio interface or the mixed output of the external audio mixing console. In the first case, every speaker is logged separately. In the second, only the transition from music to speech is annotated. The metadata of the songs are also stored in the log file, as shown in Figure 3.

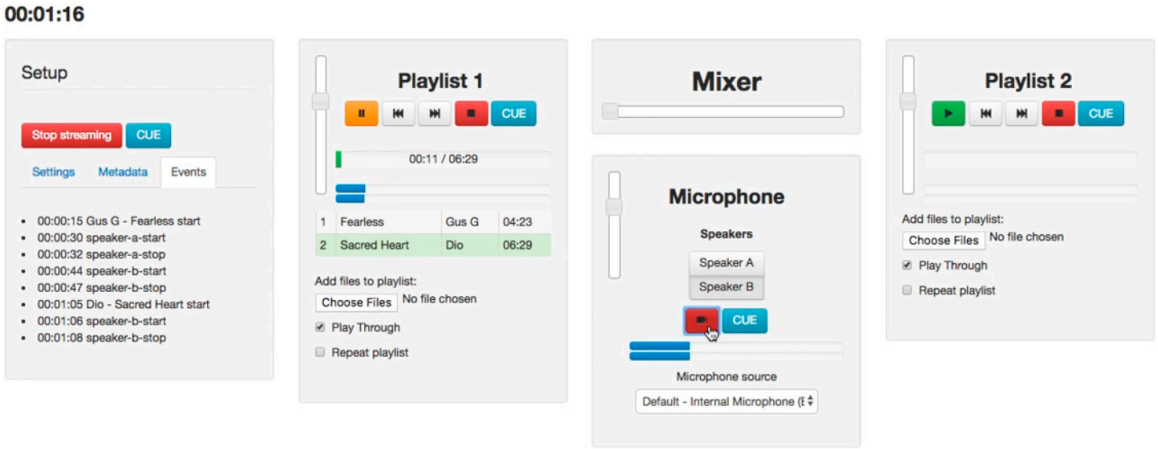

Figure 3. Graphical user interface of the web application for live mixing, annotation, and broadcasting.

For broadcasting, two open-source tools have been integrated into the application. Liquidsoap [50] is a popular and reliable solution for web radio and television stations for multimedia streaming. On top of liquidsoap, Icecast [51] is used as a streaming server for audio/visual content, to create the web radio channel. It supports all popular formats and allows the addition of more. The audience can receive the broadcast stream through an internet browser or a dedicated application capable of receiving and playing audio data streams, like the open-source VLC media player.

In Figure 4, the architecture of the developed application is demonstrated [52]. During live broadcasting, the annotation log file can be enriched with information concerning audience engagement and ratings (e.g., Google Analytics), as well as live interaction, comments, emoticons, etc. This provides further semantic metadata that concern quality-of-experience evaluation, emotional public reaction, etc. [53,54]. Such analytics, in correlation with the delivered content, provide insight for future planning. The baseline metadata scheme can also be extended to involve speech [55,56] and music [57,58] emotional cues. As it is depicted in Figure 4, the functionality that concerns different groups of interest is unified in a common framework. The radio stations and producers want to efficiently annotate and manage their produced content, enrich their content with live information and metadata, make their content discoverable and more appealing, gather and analyze analytics from public interaction during the shows. In addition, they can achieve the above by using free and publically available software. The audience can address their need for accessing radio content on-demand, have a richer augmented experience, and personalize and customize the consumed radio content according to their interests and listening habits. Meanwhile, the automated creation of labeled big audio data provides to the engineering and academic community huge and high-quality publically available datasets to develop multi-purpose models.

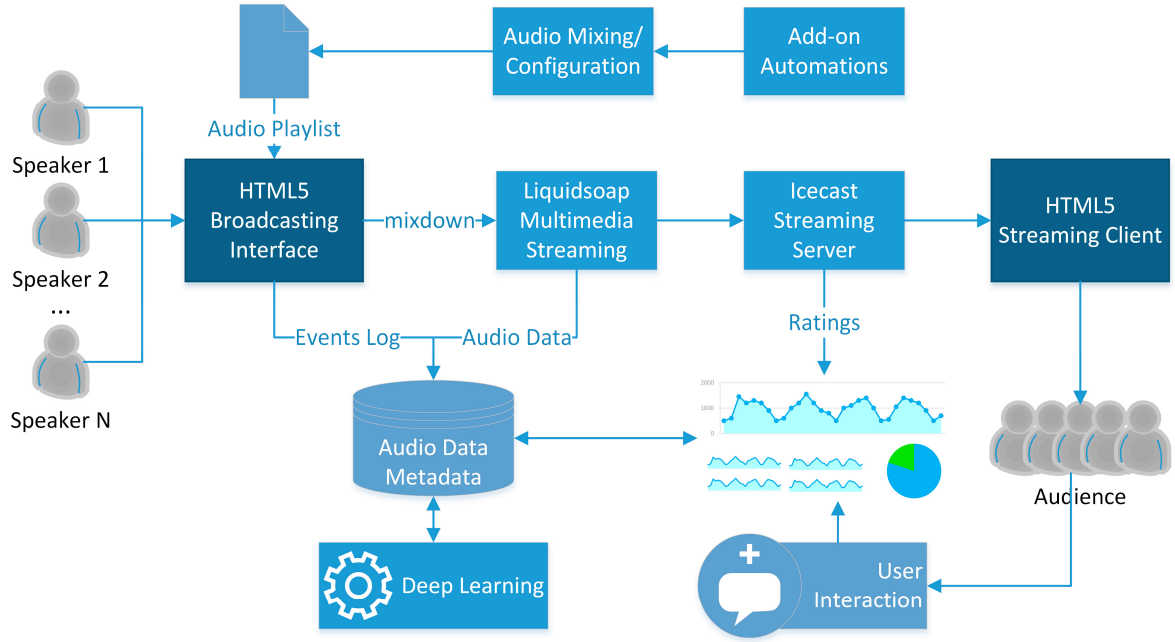

Figure 4. The architecture of the web application for web radio production, streaming, and annotation.

2.3. *Speaker Recognition with Convolutional Neural Networks*

Speaker recognition is the supervised machine learning task of assigning speaker classes to audio segments. For this task, a fixed group of speakers/classes and a respective annotated dataset are required. Speaker recognition applies to the needs of radio stations, where a certain number of known speakers/producers are present in podcasts. However, it is not possible to have a universal model for all interested stations. A dedicated model has to be trained for every defined speaker group. Radio

stations can make use of the web application presented in Section 2.4 for an amount of time to initialize the annotated dataset needed for their speaker recognition purposes.

For modeling and experimentation purposes, we have used the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [59]. The RAVDNESS database contains audiovisual speech and song files. For speaker recognition training and evaluation, we have used the audio-only speech subset of the database. As explained in the accompanying documentation, the database contains 1440 recordings of 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. The dataset covers the case of a radio station with 24 regular producers with gender balance.

To make the most of the available data and improve generalization, some common audio data augmentation techniques [56,60] have been applied:

- Background Noise: Additive Gaussian White Noise (AGWN) was added to the existing audio files. This technique doubles the number of files used for training and helps generalization in noisy conditions.
- Dynamic Range: The audio files provided in the RAVDNESS database are not normalized. Machine learning models are vulnerable to overfitting to energy characteristics. In the training session, we used the existing audio files as well as normalized versions of them. This aims at making recognition performance robust to energy fluctuations, e.g., with the speaker moving farther and closer to the mic.
- Time Shift: Time shifting of audio segments is achieved by extracting features from heavily overlapping windows to increase the instances. This is similar to the different image cropping data augmentation technique that has been popularized in visual object detection. In our approach, 90% overlapping between successive observation windows was chosen.

A convolutional neural network architecture was used for classification [48,56]. The CNN is much more lightweight than the LSTM architectures, while it also models spectro-temporal information when it is fed with spectrograms as input [48,56]. The architecture of the network along with the selected hyperparameter values are presented in Table 1 and Figure 5. CNNs are vulnerable to overfitting the training data. Dropout layers randomly discard a portion of calculated weights. To estimate overfitting, the accuracy in the training and the validation set are compared. In the initial experiment, training accuracy was found to be much higher, even with the dropout layers. Thus, an L2 regularization was added in Layer 12. The model was compiled using categorical cross-entropy loss function, for the estimation of multi-class probabilities. The Adamax optimizer, an alternative to the popular optimizer Adam [61], was chosen after experimentation.

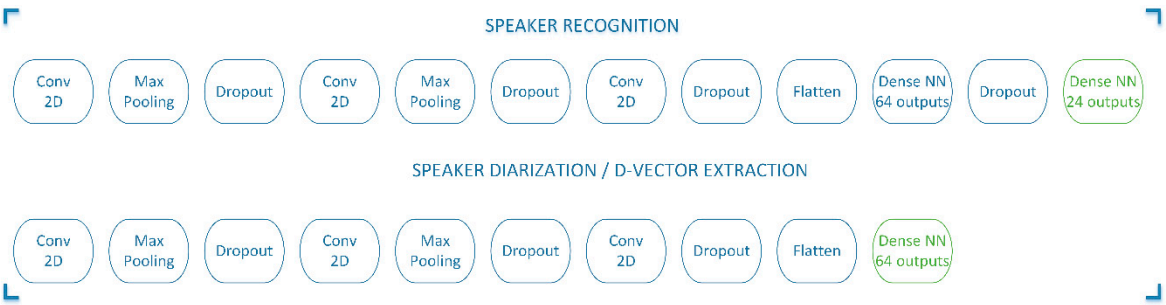

Figure 5. The architecture of the convolutional neural network used for speaker recognition. The last two layers of the trained model are discarded to be used for d-vector extraction for speaker diarization.

Implementation of the described architecture and training was held using the Keras toolbox for the Python programming language [62]. The librosa toolkit for Python [63] was used to extract Mel-scale spectrograms with a dimension of 128 Mel-coefficients from the audio files with a sampling frequency

of fs = 44,100 samples/s for windows of 1 s with 90% overlap [48,56]. The extracted spectrograms were used as input to the 2D convolutional neural network.

Table 1. Architecture and hyperparameters of the CNN model for speaker recognition.

Layer	Type	Hyperparameters
1	Convolutional 2D Layer	16 filters Kernel size = (3,3) Strides = (1,1)
2	Max Pooling 2D Layer	Pool size = (2,2)
3	Dropout	Rate = 0.25
4	Convolutional 2D Layer	32 filters Kernel size = (3,3) Strides = (1,1)
5	Max Pooling 2D Layer	Pool size = (2,2)
6	Dropout	Rate = 0.25
7	Convolutional 2D Layer	64 filters Kernel size = (3,3) Strides = (1,1)
8	Dropout	Rate = 0.25
9	Convolutional 2D Layer	128 filters Kernel size = (3,3) Strides = (1,1)
10	Convolutional 2D Layer	256 filters Kernel size = (3,3) Strides = (1,1)
11	Flatten Layer	
12	Dense Neural Network	Output weights = 64 Activation = ReLU L2 regularizer
13	Dense Neural Network	Output weights = 64 Activation = ReLU
14	Dropout	Rate = 0.25
15	Dense Neural Network	Output weights = 24 Activation = Softmax

2.4. Speaker Diarization

While speaker recognition is sufficient for the case of known speakers, this is not always the case in radio podcasts. Many producers have guests in their shows. Speaker diarization is needed to segment audio files were unknown speakers are present. As described in the literature review of Section 1.3, speaker diarization is an unsupervised clustering task. Identity vectors (i-vectors) are extracted from unstructured audio files to be used as input to clustering algorithms. In the past few years, state-of-the-art approaches have used deep learning models to extract i-vectors, which are in this case called d-vectors.

The trained model described in Section 2.3 which is fit for multi-class speaker recognition is used for the formulation of the d-vectors. The two last layers of the network, the dropout Layer 11 and the dense neural network Layer 12, which are used for the classification task, are discarded. The resulting architecture is shown in Figure 5.

Having as input Mel-scale spectrograms of the same shape as the ones used for training (fs = 44,100, window = 1 s, 128 Mel-coefficients), the output of the model is the vector of the 64 weights of the dense

neural network Layer 10, which is the final layer of the modified network. This fix sized vector of 64 values is used as the d-vector for clustering and speaker diarization.

The d-vector should be efficient for the clustering of audio files with unknown speakers who were not included in the training dataset. This is why cross-corpus evaluation was performed. An audio file of 20 min in length was used, containing speech from three speakers, two male and one female. The three speakers were not included in the group of 24 speakers of the RAVDNESS dataset, and they were recorded in the same studio, using the same equipment. The d-vector is extracted from windows of 1 s with a 90% overlap, as described.

The performance of several clustering algorithms in the task of speaker diarization with d-vector input is evaluated. The results for all algorithms are presented in Section 3. K-means [64] is one of the most popular clustering algorithms in the literature, assigning observations to the centroid with the nearest mean after a number of iterations, minimizing the inertia. The initial values of the centroids are chosen using the kmeans++ algorithm for faster convergence. Agglomerative clustering is a hierarchical unsupervised classficiation approach, starting from an initial cluster and performing successive splits and merges [65]. The ward linkage criterion was selected, which minimizes the variance of the clusters being merged. The linkage is computed using Euclidean distance. The Birch algorithm builds a tree called the clustering feature tree (CFT) [66]. The maximum number of sub-clusters in each node was set to 50 and the threshold for the merging/splitting of neighboring sub-clusters to 0.5.

3. Results

The data augmentation procedures, described in Section 2.3, resulted in a dataset with 144,271 instances, balanced between the 24 classes. For the evaluation of the speaker recognition model, the dataset was split into three sets, following common practice: training (70%), validation (15%) and testing (15%). The validation set was used for hyperparameter tuning, while the test set contained unseen data held out for evaluation. After the fine-tuning of the model, as presented in Section 2.3, classification accuracy in the test set was estimated equal to 88.34%. The learning curves for the training and validation sets are shown in Figure 6. The results in the three sets (training/validation/test) show a balanced performance in known and unseen data and prove that overfitting issues have been addressed efficiently. A graphical representation confusion matrix is depicted in Figure 7, indicating the relation between the predicted and actual speakers.

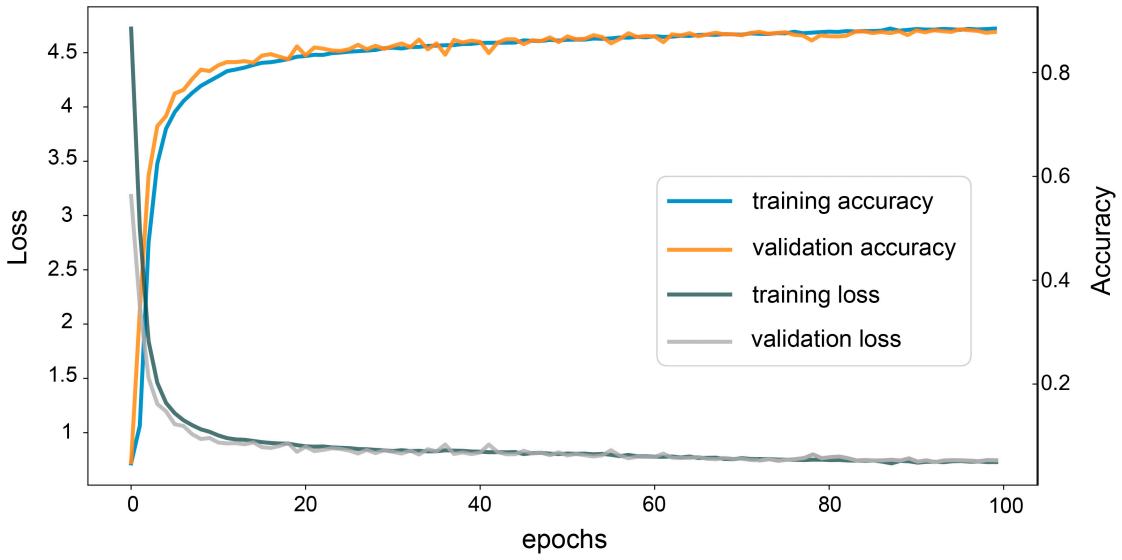

Figure 6. Learning curves for the accuracy and loss of the model for the training and validation set.

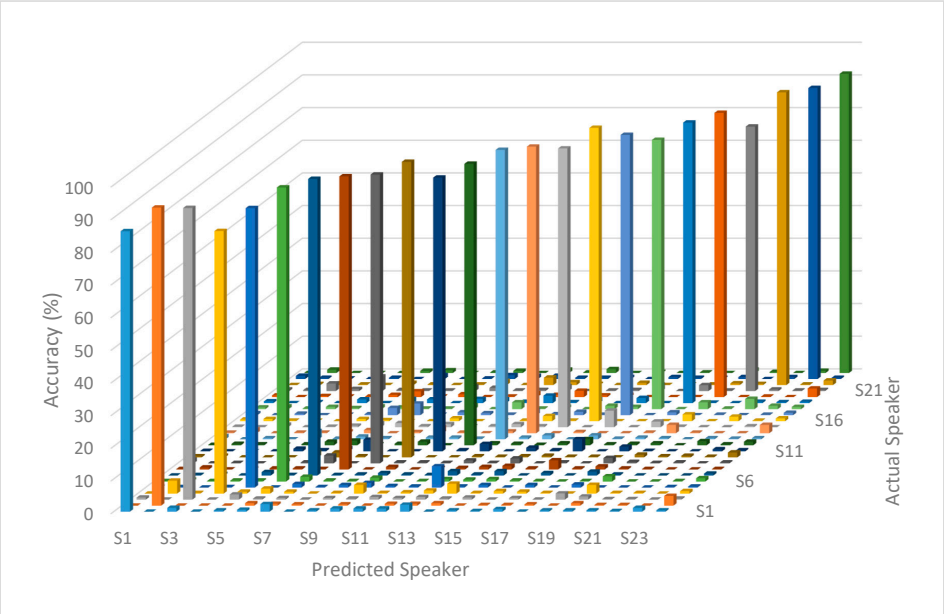

Figure 7. Graphical representation of the speaker recognition confusion matrix.

For speaker diarization evaluation, the classifiers described in Section 2.4 were evaluated for their performance in clustering unknown speakers. We have used a cross-corpus, vocabulary-independent and language-independent evaluation strategy. The unknown speakers have been selected from the AESDD dataset described in [55]. The unsupervised diarization results are shown in Table 2. The BIRCH clustering algorithm scores the best results in every experiment.

Table 2. Speaker diarization accuracy for unknown speakers.

Clustering Algorithm	**Accuracy (%)**		
	2 speakers	3 speakers	4 speakers
Agglomerative	95.98	81.7	78.12
K-means	90.4	85.83	81.81
BIRCH	**96.72**	**87.22**	**82.33**

4. Discussion

4.1. Conclusions

We have presented a framework that addresses the need for evolution in radio production practice in the evolving digital media ecosystem and the challenges for efficient management of publically available big audio data. Radio production has to adapt in the direction of providing richer content, and more discoverable audio-on-demand, to maintain a broader audience and be more appealing to younger audiences. We have proposed a knowledge extraction scheme for podcast segmentation and annotation, involving speech/music detection, speaker recognition and diarization, speech-to-text for transcription extraction, and music metadata extraction. While many approaches to supervised and unsupervised machine learning models for audio information retrieval appear in the literature, we estimate that the most effective, robust and inexpensive in terms of working hours is the annotation of radio shows during the production stage. For this reason, an application has been developed, covering live production functionality, while providing real-time event logging. The annotated audio files can be browsed through a dedicated interactive GUI which integrates the extracted information. This methodology may also be applied for live metadata streaming to enrich the provided content.

The proposed web radio application can also be used for the creation of a dataset to train models for recognition of the regular speakers on audio stations. A CNN model was trained and evaluated for

24 different speakers, a number that corresponds to the upper limit of a typical radio station needs. The recognition accuracy of the model is 88.34% for windows of 1 sec. Since it is quite common for radio shows to have one or more guests along with the regular speakers, an unsupervised clustering module was also considered necessary. Following a state-of-the-art approach to speaker diarization, the CNN model was used to extract identity vectors. Experimentation with two, three and four unknown speakers was conducted, resulting in unsupervised clustering with a maximum of 82.33%–96.72% accuracy for the BIRCH algorithm, which corresponds to a DER of 17.67%–3.28%, since voice activity detection errors are not taken into consideration. This score is very close to the results for supervised speaker recognition. Prior knowledge of the number of guests/clusters was proved very important for performance. The resulting diarization error rate (DER) outperforms common reported results of i-vector approaches [41] (DER = 20.54%–42.63%), and is comparable to state-of-the-art d-vector approaches including: ANN [40] (DER = 25.9%–32%), CNN [47] (DER = 15.3%–24.6%), and LSTM [41] (DER = 12.3%–27.3%). The results from papers [40,41,47] are mentioned to provide further insight to the reader concerning the state-of-the-art. However, a direct comparison between the results is not applicable due to different problem definitions in every study (different datasets, number of speakers, experimental parameters, etc.).

4.2. Limitations and Future Work

The main limitations of the proposed system are set by the diarization error rates. A more lightweight diarization methodology has been proposed, which produces results comparable to the state-of-the-art and clearly outperforms older i-vector systems. Moreover, the clustering performance for an undefined number of unknown speakers is satisfactory only for the birch algorithm. Since the novelty of the present research lies mostly in the convergence of state-of-the-art technologies in a modular approach, the individual modules (speaker diarization, speaker recognition, speech-to-text) can be updated over the use of the framework, based on developments of the aforementioned fields. Furthermore, in the presented approach, speaker recognition is performed for individual windows of 1 s. With an aggregation over longer complete speech excerpts, which is the most common case in real-life scenarios, robustness is expected to improve by use of majority voting across successive frames and discarding of outlier values, thus, the proposed methodology is considered applicable.

The proposed automation is expected to create value through its use. The benefits concern radio organizations, which can organize and document their archive more efficiently, the audience, who can enjoy a more personalized and customized experience, as well as the scientific community. Every annotated podcast available online can serve as a valuable big dataset. The manual labeling of this data corresponds to many working hours and an unavoidable annotation error rate. The future plans of the project include the collaboration with radio stations for testing and feedback collection. The application will be available as open-source to make it more appealing to radio producers. However, in the initial state, it is estimated that the most efficient kick-start would be the collaboration with some of the many academic web radio stations.

Author Contributions: Conceptualization, C.D.; methodology, N.V., N.T. and C.D.; software, N.V. and N.T.; validation, N.V. and N.T.; formal analysis, N.V.; investigation, N.V. and N.T.; resources, N.V. and N.T.; data curation, N.V.; writing—original draft preparation, N.V., N.T. and C.D.; writing—review and editing, N.V., N.T. and C.D.; visualization, N.V., N.T. and C.D.; supervision, C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The Titan V GPU used for this research was donated by the NVIDIA Corporation.

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 Audio Engineering Society Convention 132, Budapest, Hungary, 26–29 April 2012.
3. Tsipas, N.; Vrysis, L.; Dimoulas, C; Papanikolaou, G. Efficient audio-driven multimedia indexing through similarity-based speech /music discrimination. *Multimed. Tools Appl.* **2017**, *76*, 25603–25621. [CrossRef]
4. Dimoulas, C.A.; Symeonidis, A.L. Syncing Shared Multimedia through Audiovisual Bimodal Segmentation. *IEEE Multimed.* **2015**, *22*, 26–42. [CrossRef]
5. 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]
6. 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.
7. 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]
8. Dimoulas, C. Machine Learning. In *The SAGE Encyclopedia of Surveillance, Security, and Privacy*; Arrigo, B.A., Ed.; Sage Publications Inc.: California, CA, USA, 2018; pp. 591–592. [CrossRef]
9. 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]
10. Tsipas, N.; Vrysis, L.; Dimoulas, C.; Papanikolaou, G. Content-Based Music Structure Analysis using Vector Quantization. In Proceedings of the 138th AES Convention, Warsow, Poland, 7–10 May 2015.
11. Tsipas, N.; Dimoulas, C.; Kalliris, G.; Papanikolaou, G. Collaborative annotation platform for audio semantics. In Proceedings of the 134th AES Convention, Rome, Italy, 4–7 May 2013; pp. 218–222.
12. Baume, C.; Plumbley, M.; Frohlich, D.; Calic, J. PaperClip: A digital pen interface for semantic speech editing in radio production. *J. Audio Eng. Soc.* **2018**, *66*, 241–252. [CrossRef]
13. Baume, C.; Plumbley, M.D.; Ćalić, J.; Frohlich, D. A Contextual Study of Semantic Speech Editing in Radio Production. *Int. J. Hum.-Comput. Stud.* **2018**, *115*, 67–80. [CrossRef]
14. Jauert, P.; Ala-Fossi, M.; Föllmer, G.; Lax, S.; Murphy, K. The future of radio revisited: Expert perspectives and future scenarios for radio media in 2025. *J. Radio Audio Media* **2017**, *24*, 7–27. [CrossRef]
15. Berry, R. Part of the establishment: Reflecting on 10 years of podcasting as an audio medium. *Convergence* **2016**, *22*, 661–671. [CrossRef]
16. Mensing, D. Public radio at a crossroads: Emerging trends in US public media. *J. Radio Audio Media* **2017**, *24*, 238–250. [CrossRef]
17. Edmond, M. All platforms considered: Contemporary radio and transmedia engagement. *New Media Soc.* **2015**, *17*, 1566–1582. [CrossRef]
18. McHugh, S. How podcasting is changing the audio storytelling genre. *Radio J.: Int. Stud. Broadcast Audio Media* **2016**, *14*, 65–82. [CrossRef]
19. Berry, R. Podcasting: Considering the evolution of the medium and its association with the word 'radio'. *Radio J.: Int. Stud. Broadcast Audio Media* **2016**, *14*, 7–22. [CrossRef]
20. Jędrzejewski, S. Radio in the new media environment. In *Radio: Resilient Medium*; Oliveira, M., Stachyra, G., Starkey, G., Eds.; Sunderland: Centre for Research in Media and Cultural Studies: London, UK, 2014; Volume 1, pp. 17–26.
21. Uddin, M.F.; Gupta, N. Seven V's of Big Data understanding Big Data to extract value. In Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education, Bridgeport, CT, USA, 3–5 April 2014; pp. 1–5.
22. O'Leary, D.E. Artificial intelligence and big data. *IEEE Intell. Syst.* **2013**, *28*, 96–99. [CrossRef]
23. Sagiroglu, S.; Sinanc, D. Big data: A review. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, USA, 20–24 May 2013; pp. 42–47.
24. Gandomi, A.; Haider, M. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manag.* **2015**, *35*, 137–144. [CrossRef]
25. Barnaghi, P.; Sheth, A.; Henson, C. From data to actionable knowledge: Big data challenges in the web of things [Guest Editors' Introduction]. *IEEE Intell. Syst.* **2013**, *28*, 6–11. [CrossRef]
26. Han, Q.; Liang, S.; Zhang, H. Mobile cloud sensing, big data, and 5G networks make an intelligent and smart world. *IEEE Netw.* **2015**, *29*, 40–45. [CrossRef]

27. Chandarana, P.; Vijayalakshmi, M. Big data analytics frameworks. In Proceedings of the 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), Mumbai, India, 4–5 April 2014; pp. 430–434.
28. Qiu, J.; Wu, Q.; Ding, G.; Xu, Y.; Feng, S. A survey of machine learning for big data processing. *EURASIP J. Adv. Signal Process.* **2016**, *2016*, 67. [CrossRef]
29. Najafabadi, M.M.; Villanustre, F.; Khoshgoftaar, T. M.; Seliya, N.; Wald, R.; Muharemagic, E. Deep learning applications and challenges in big data analytics. *J. Big Data* **2015**, *2*, 1. [CrossRef]
30. Anguera, X.; Bozonnet, S.; Evans, N.; Fredouille, C.; Friedland, G.; Vinyals, O. Speaker diarization: A review of recent research. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 356–370. [CrossRef]
31. Moattar, M.H.; Homayounpour, M.M. A review on speaker diarization systems and approaches. *Speech Commun.* **2012**, *54*, 1065–1103. [CrossRef]
32. Rouvier, M.; Dupuy, G.; Gay, P.; Khoury, E.; Merlin, T.; Meignier, S. An open-source state-of-the-art toolbox for broadcast news diarization. In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013.
33. Shum, S.; Dehak, N.; Glass, J. On the use of spectral and iterative methods for speaker diarization. In Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012.
34. Sell, G.; Garcia-Romero, D. Speaker diarization with PLDA i-vector scoring and unsupervised calibration. In Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, CA, USA, 7 December 2014; pp. 413–417.
35. Shum, S.H.; Dehak, N.; Dehak, R.; Glass, J.R. Unsupervised methods for speaker diarization: An integrated and iterative approach. *IEEE Trans. Audio, Speech Lang. Process.* **2013**, *21*, 2015–2028. [CrossRef]
36. Rouvier, M.; Meignier, S. A global optimization framework for speaker diarization. In Proceedings of the Odyssey, Singapore, 25–28 June 2012.
37. Ferras, M.; Boudard, H. Speaker diarization and linking of large corpora. In Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, 2–5 December 2012; pp. 280–285.
38. Garcia-Romero, D.; Snyder, D.; Sell, G.; Povey, D.; McCree, A. Speaker diarization using deep neural network embeddings. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 4930–4934.
39. Snyder, D.; Garcia-Romero, D.; Sell, G.; McCree, A.; Povey, D.; Khudanpur, S. Speaker Recognition for Multi-speaker Conversations Using X-vectors. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5796–5800.
40. Yella, S. H.; Stolcke, A.; Slaney, M. Artificial neural network features for speaker diarization. In Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, CA, USA, 7–10 December 2014; pp. 402–406.
41. Wang, Q.; Downey, C.; Wan, L.; Mansfield, P.A.; Moreno, I.L. Speaker diarization with lstm. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5239–5243.
42. Lin, Q.; Yin, R.; Li, M.; Bredin, H.; Barras, C. LSTM based similarity measurement with spectral clustering for speaker diarization. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019.
43. Sun, L.; Du, J.; Gao, T.; Lu, Y. D.; Tsao, Y.; Lee, C. H.; Ryant, N. A novel LSTM-based speech preprocessor for speaker diarization in realistic mismatch conditions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5234–5238.
44. Zhang, A.; Wang, Q.; Zhu, Z.; Paisley, J.; Wang, C. Fully supervised speaker diarization. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6301–6305.
45. Hagerer, G.; Pandit, V.; Eyben, F.; Schuller, B. Enhancing lstm rnn-based speech overlap detection by artificially mixed data. In Proceedings of the 2017 AES International Conference on Semantic Audio, Audio Engineering Society, Erlangen, Germany, 22–24 June 2017.
46. Hrúz, M.; Zajíc, Z. Convolutional neural network for speaker change detection in telephone speaker diarization system. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 4945–4949.

47. Cyrta, P.; Trzciński, T.; Stokowiec, W. Speaker diarization using deep recurrent convolutional neural networks for speaker embeddings. In Proceedings of the International Conference on Information Systems Architecture and Technology, Szklarska Poręba, Poland, 17–19 September 2017; pp. 107–117.
48. 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]
49. Hoeg, W.; Lauterbach, T. *Digital Audio Broadcasting*; Wiley: Hoboken, NJ, USA, 2003.
50. Baelde, D.; Beauxis, R.; Mimram, S. Liquidsoap: A high-level programming language for multimedia streaming. In Proceedings of the International Conference on Current Trends in Theory and Practice of Computer Science, Nový Smokovec, Slovakia, 22–28 January 2011; pp. 99–110.
51. Dengler, B. Creating an internet radio station with icecast and liquidsoap. *Linux J.* **2017**, *280*, 1.
52. Dalakas, A.; Tsipas, N.; Dimoulas, C.; Kalliris, G. Web radio automation and "big data" of audio (and audiovisual) semantic annotation. In Proceedings of the HELINA Conference on Acoustics, Patras, Greece, 8–9 October 2018.
53. 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]
54. Kalliris, G.; Matsiola, M.; Dimoulas, C.; Veglis, A. Emotional aspects and quality of experience for multifactor evaluation of audiovisual content. *Int. J. Monit. Surveill. Technol. Res. (IJMSTR)* **2014**, *2*, 40–61. [CrossRef]
55. Vryzas, N.; Kotsakis, R.; Liatsou, A.; Dimoulas, C. A.; Kalliris, G. Speech emotion recognition for performance interaction. *J. Audio Eng. Soc.* **2018**, *66*, 457–467. [CrossRef]
56. 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]
57. Oxholm, E.; Hansen, E. K.; Triantafyllidis, G. Auditory and Visual based Intelligent Lighting Design for Music Concerts. *EAI Endorsed Trans. Creat. Technol.* **2018**, *5*, e2. [CrossRef]
58. Barthet, M.; Fazekas, G.; Sandler, M. Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models. In Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR), London, UK, 19–22 June 2012; pp. 492–507.
59. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef] [PubMed]
60. Salamon, J.; Bello, J.P. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process. Lett.* **2017**, *24*, 279–283. [CrossRef]
61. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015.
62. Chollet, F. *Keras: The Python Deep Learning Library*; Astrophysics Source Code Library (ASCL): Leicester, UK, 2018.
63. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. Librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference, Austin, TX, USA, 6–12 July 2015.
64. Steinley, D.; Brusco, M.J. Initializing k-means batch clustering: A critical evaluation of several techniques. *J. Classif.* **2007**, *24*, 99–121. [CrossRef]
65. Murtagh, F.; Legendre, P. Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *J. Classif.* **2014**, *31*, 274–295. [CrossRef]
66. Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: An efficient data clustering method for very large databases. *ACM Sigmod Rec.* **1996**, *25*, 103–114. [CrossRef]