aerospace

MDPI

Article

An Automatic Speaker Clustering Pipeline for the Air Traffic Communication Domain

Driss Khalil [1,*], Amrutha Prasad [1,2], Petr Motlicek [1,2], Juan Zuluaga-Gomez [1,3], Iuliia Nigmatulina [1,4], Srikanth Madikeri [1] and Christof Schuepbach [5]

[1] Idiap Research Institute, 9120 Martigny, Switzerland; amrutha.prasad@idiap.ch (A.P.); petr.motlicek@idiap.ch (P.M.); juan-pablo.zuluaga@idiap.ch (J.Z.-G.); iuliia.nigmatulina@idiap.ch (I.N.); srikanth.madikeri@idiap.ch (S.M.)
[2] Faculty of Information Technology, Brno University of Technology, 60190 Brno, Czech Republic
[3] LIDIAP, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[4] Institute of Computational Linguistics, University of Zurich, 8006 Zurich, Switzerland
[5] Armasuisse Science and Technology, 3602 Thun, Switzerland; christof.schuepbach@armasuisse.ch
* Correspondence: dkhalil@idiap.ch

Abstract: In air traffic management (ATM), voice communications are critical for ensuring the safe and efficient operation of aircraft. The pertinent voice communications—air traffic controller (ATCo) and pilot—are usually transmitted in a single channel, which poses a challenge when developing automatic systems for air traffic management. Speaker clustering is one of the challenges when applying speech processing algorithms to identify and group the same speaker among different speakers. We propose a pipeline that deploys (i) speech activity detection (SAD) to identify speech segments, (ii) an automatic speech recognition system to generate the text for audio segments, (iii) text-based speaker role classification to detect the role of the speaker—ATCo or pilot in our case—and (iv) unsupervised speaker clustering to create a cluster of each individual pilot speaker from the obtained speech utterances. The speech segments obtained by SAD are input into an automatic speech recognition (ASR) engine to generate the automatic English transcripts. The speaker role classification system takes the transcript as input and uses it to determine whether the speech was from the ATCo or the pilot. As the main goal of this project is to group the speakers in pilot communication, only pilot data acquired from the classification system is employed. We present a method for separating the speech parts of pilots into different clusters based on the speaker's voice using agglomerative hierarchical clustering (AHC). The performance of the speaker role classification and speaker clustering is evaluated on two publicly available datasets: the ATCO2 corpus and the Linguistic Data Consortium Air Traffic Control Corpus (LDC-ATCC). Since the pilots' real identities are unknown, the ground truth is generated based on logical hypotheses regarding the creation of each dataset, timing information, and the information extracted from associated callsigns. In the case of speaker clustering, the proposed algorithm achieves an accuracy of 70% on the LDC-ATCC dataset and 50% on the more noisy ATCO2 dataset.

Keywords: speaker clustering; speaker role detection

check for updates

Citation: Khalil, D.; Prasad, A.; Motlicek, P.; Zuluaga-Gomez, J.; Nigmatulina, I.; Madikeri, S.; Schuepbach, C. An Automatic Speaker Clustering Pipeline for the Air Traffic Communication Domain. *Aerospace* **2023**, *10*, 876. https://doi.org/10.3390/aerospace10100876

Academic Editor: Joost Ellerbroek

Received: 30 April 2023
Revised: 25 September 2023
Accepted: 26 September 2023
Published: 10 October 2023

1. Introduction

Air traffic control (ATC) communication ensures safe and efficient flight operations [1]. The rise of new artificial intelligence/machine learning technologies provides opportunities for a fundamental change in automation and it becomes a central enabler for future air traffic management concepts. Machine learning technologies are typically data-driven and require a large amount of data for training and development. In the case of voice communication, these data are available through Air Traffic Navigation Service Providers (ANSPs). However, obtaining such data through ANSPs is a legally very complex task, as it typically requires access to the operational control rooms of the ANSPs. A cheap and easy alternative (if allowed by national data privacy laws) is the use of data collected by various initiatives worldwide, such as LiveATC1 (https://www.liveatc.net, accessed on 29 April 2023) in the U.S. and ATCO2 (https://www.atco2.org, accessed on 29 April 2023) in Europe, which collect and store freely available voice communications from Very-High-Frequency (VHF) radio channels. In the case of ATCO2, a large set of volunteers collect the voice as well as the supporting contextual data using relatively cheap VHF radio receivers, and the data are then collected through a centralized server. This approach can easily deliver thousands of unlabeled transmissions. Although such data are typically noisier [2], it has been shown that they can be valuable for training machine learning technologies, including the ATC domain [3].

The average length of each utterance in the collected data is around 3.3 s. However, this presents a unique challenge in the ATC domain, where rapid exchanges between pilots and air traffic controllers occur in communication scenarios, and where utterances are often brief. Accurately identifying speaker roles and clustering them can therefore be challenging, especially when multiple speakers communicate simultaneously on the same channel. The task becomes further complicated due to variations in speech patterns, accents, and communication styles. These challenges underscore the need for advanced machine learning techniques that can handle noisy, short-duration audio data while accurately distinguishing between different speakers and their roles in the communication process.

Besides collecting free data for ATM-oriented machine learning technologies, there are also other use cases that are of serious interest to governmental agencies such as prescreening the VHF radio channels and detecting their potential abuse by anonymous private persons. This use case is a principal motivator for this paper. VHF radio channels carry the utterances of both pilots and ATCos as one single-channel recording (i.e., a huge wave file that is not segmented). Even if the segmentation algorithm is applied (i.e., typically, voice/speech activity detection can separate communication into short chunks), there is no other information about whether the utterance comes from the ground (ATCo) or the cockpit (pilot).

This paper focuses on clustering speakers appearing either in the same VHF radio channel or across many channels over a given period of time. This is a principal question of security officers when dealing with the abuse of non-encrypted radio communications. The solution given in this paper is tested by analyzing ATCo–pilot communication captured by ATCO2 data. More specifically, our paper is partially built on the concept of separating ATCos and pilots, as investigated in [4]. As ATCos typically appear in the same VHF radio channel over a relatively long period (up to several hours, depending on the length of their shift), the appearance of a new pilot in the analyzed VHF radio channel is very probable. This paper, therefore, focuses on clustering pilot audio recordings to emulate the reality of automatically clustering random speakers in VHF radio channels.

Recent advances in machine learning, particularly deep neural networks (DNNs), have shown promise in addressing these issues by modeling the complex relationships between acoustic features and speakers' identities. DNNs can be trained on large amounts of speech data and can learn to extract high-level features from the speech audio, which can be used for speaker clustering [5]. Despite these advances, the task of speaker clustering in ATC communications remains challenging due to the presence of multiple speakers in the same channel, in addition to the lack of ground-truth information. This absence of labeled data poses significant problems in developing such systems. Nevertheless, the development of such automatic pipelines presents great value in both operational and forensic contexts.

The proposed pipeline presented in this paper comprises four stages: speech segment separation, automatic speech recognition (ASR), speaker role classification, and speaker clustering. The first stage separates the speech segments from a single channel, followed by ASR to transcribe into English. The transcribed text is then fed into a speaker classification model, which detects the pilot segments. The speaker role classification is used only to filter the pilot audio required for speaker clustering. Finally, a speaker clustering method separates and groups the pilot speaker from the audio segments. The proposed pipeline aims to improve the accuracy of speaker clustering in the ATC domain and facilitate effective communication between controllers and pilots.

In air traffic control (ATC) communication analysis, most similar works tend to address only a portion of the complete pipeline outlined in this research. Notably, speaker clustering, a critical part of this pipeline, has received limited attention due to the absence of reliable ground truths on publicly available datasets. It is in this context that the importance of our proposed integrated framework becomes evident. By including speech activity detection, automatic speech recognition, text-based speaker role classification, and unsupervised speaker clustering, this pipeline offers a comprehensive solution. This approach not only addresses the limitations of existing methods but also has the potential to significantly improve the analysis of ATC communication. It bridges the crucial gap between individual components, enabling a deeper understanding of speaker roles and ultimately enhancing safety and efficiency in air traffic control. The rest of this paper is organized as follows. In Section 2, we present the different steps of the pipeline through a discussion of related works, as well as the method used in each step in our work. In Section 3, we describe the different datasets used for training and evaluation in each of the two main components of our pipeline. In Section 4, we present the experiments, the method of evaluation, and the results of each experiment. Finally, we conclude the paper in Section 5 and discuss potential future directions for research in this area.

2. Automatic Speaker Clustering Pipeline

In the ATC domain, the communication of the pilots is of particular interest compared to that of ATC controllers (ATCos) because pilots are responsible for executing flight plans and maneuvering the aircraft, making their communication critical for ensuring safe and efficient flights. Therefore, it is vital to separate pilots' communications from those of ATCos to train automatic systems for each group. Separating the communications of individual pilots is essential for post-flight analysis, incident investigations, and pilot training tasks. The proposed method starts by extracting the speech segments using the SAD system and then using ASR to transcribe those extracted segments. The transcripts obtained are used as input to classify the pilot's speech segments, which are used in the final step as input to the speaker clustering model, as shown in Figure 1. The following subsections describe each step of this pipeline.

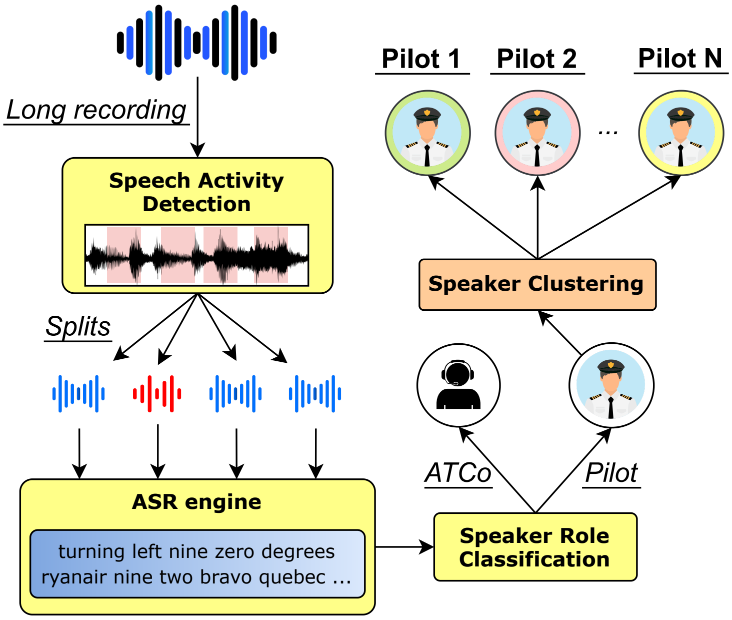
Long recording
Speech Activity Detection
Splits
ASR engine
turning left nine zero degrees
ryanair nine two bravo quebec ...
Speaker Role Classification
ATCo
Pilot
Speaker Clustering
Pilot 1
Pilot 2
...
Pilot N

Figure 1. Overview of the automatic speaker clustering pipeline.

2.1. Speech Activity Detection

Speech activity detection (SAD) is a crucial process in speech processing that involves identifying speech segments within an audio utterance. This system splits the audio based on long-silence regions to generate a subset of audio files without silence. It plays a vital role in many speech-based applications such as automatic speech recognition (ASR), speaker recognition, and speaker diarization. Researchers are actively working on developing a SAD system that can accurately operate in noisy environments. The approach is based on [6], which leverages multilingual ASR to improve speech activity detection. The acoustic model (AM) was trained using a lattice-free maximum mutual information loss to extract contextual information from acoustic frames. Multilingual training enhances robustness to noise and language variability. The proposed multilingual acoustic model was trained on 18 languages from the BABEL datasets (https://catalog.ldc.upenn.edu/byyear, accessed on 29 April 2023), including LDC2018S07, LDC2018S13, LDC2018S02, LDC2017S03, LDC2017S22, LDC2017S08, LDC2017S05, LDC2017S13, LDC2017S01, LDC2017S19, LDC2016S06, LDC2016S08, LDC2016 S02, LDC2016S12, LDC2016S09, LDC2016S13, and LDC2016S10. The primary objective of using this dataset was to develop a SAD system that can operate accurately in noisy environments and is robust to language variability. Within each language-dependent part of the acoustic model, speech and non-speech acoustic frames were mapped to a different set of output context-dependent phones or posteriors. For each language, the index of the maximum output posterior was used as a frame-level speech/non-speech decision function. Conventional logistic regression [7] and majority voting were employed to combine decisions from different languages.

2.2. Automatic Speech Recognition (ASR)

Automatic speech recognition (ASR) is a sub-field of speech processing that involves converting speech to text, typically in one language. Hence, this is also termed speech-to-text. A typical ASR system employs an AM and a language model (LM) for converting a speech signal to text. The former is trained on speech recordings with corresponding (ideally manually corrected) text, also referred to as transcripts. The AM represents the relationship between a speech signal and phonemes or other linguistic units that make up the speech. The latter is trained on a large corpus of text data. A probability distribution over sequences of words usually represents the LM. The LM provides context to distinguish between words and phrases that sound similar. Using the knowledge of the AM and LM, a decoding graph is usually built as a weighted Finite State Transducer (FST) [8–10], which generates text output given an observation sequence.

To build a robust speech recognition engine, the artificial intelligence behind it has to be adept at handling challenges such as different acoustic conditions, background noise, model size, and performance. Development in natural language processing and neural network technology has improved speech and voice technology. Past research projects in ATM have provided a platform to develop and improve ASR systems for ATCos and pilots. In [11,12], the authors developed ASR for ATCos to help increase their efficiency and reduce workloads. The authors of [1] provided a benchmark on ASR for different ATC databases. An approach for leveraging non-transcribed audio data to improve ASR was investigated in [13]. A semi-supervised learning approach for enhancing ASR in the ATM domain was employed in [11,12,14]. In [15–17], the authors aimed to improve the recognition of the callsigns in ASR by integrating surveillance data. Finally, the authors of [18] investigated the effect of fine-tuning large pre-trained models, trained using a Transformer architecture, for application in the ATC domain.

This work presents ASR systems that employ two approaches: (i) a hybrid system and (ii) end-to-end training. The hybrid system for ASR uses deep neural network (DNN)-based AMs trained with the lattice-free maximum mutual information (LF-MMI) [19] criterion and n-gram models for the LM. Current state-of-the-art systems use a Transformer architecture, which uses unsupervised [20] or self-supervised [21] learning for speech representations.

2.3. Speaker Role Classification

The task of sequence labeling (SL) assigns labels to words that share a specific role and meaning within the grammatical structure of a sentence. In [22], these groups of words/sentences had similar grammatical properties, and the work focused on two subtasks of SL: named entity recognition (NER) [23,24] and sequence classification (SC) [22,25]. Early work on NER and SC was based on handcrafted ontology, dictionaries, and lexicons, which made them prone to human error. Nowadays, deep learning-based systems are cataloged as state-of-the-art on NER [24] and SC. These models are primarily based on convolutional neural networks [26], recurrent neural networks [22], and Transformers [27].

ATC communications are a rich source of information and follow explicit grammar and ontology. Additionally, ATC communications are built on a well-defined lexicon and dictionary that speakers' errors can sometimes disrupt. One example is the order in which the named entities (e.g., callsign) are uttered in the communication. ATCos utter the callsign (lufthansa seven eight two) at the beginning, whereas pilots invariably do so at the end:

ATCo: *"lufthansa seven eight two descend flight level seven zero"* and,
PILOT: *"descend flight level seven zero lufthansa seven eight two"*.

Following the pros and cons described in Section 2.3, we demonstrate that state-of-the-art NER and SC can be leveraged to automatically identify speaker roles. For instance, one can apply NER to identify ATC-related named entities such as *callsigns*, *command types*, or *units*. Similarly, the structure and type of these 'entities' used in a given communication can be leveraged to identify speaker roles. Our previous research on identifying speaker roles [4] mainly focused on a grammar-based bag-of-words system that was capable of performing speaker role identification with precision/recall values of 0.82/0.81 for ATCos and 0.84/0.85 for pilots, respectively. Also, in [28–30], we explored speaker change detection for ATC text. In [31], the authors mentioned that manually annotating pilot recordings was more challenging than annotating ATCo recordings due to their quality, speech rate, speaker accent, etc. Another reason is that the audio of ATCos is obtained directly from the source, whereas the pilot audio is recorded through the radio receiver. This is one of the reasons why speech processing systems (ASR, diarization, and speaker role identification) perform considerably worse for pilots' recordings compared to ATCos' recordings.

2.4. Speaker Clustering

Over the past few years, there has been growing interest in applying speech processing techniques to the air traffic control (ATC) domain. Specifically, researchers have explored various methods for automatically analyzing and classifying speech in ATC conversations. Although speaker clustering is an essential task in the ATC domain, only a few research studies have focused on it due to the need for ground truths for speaker identity. However, speaker clustering is essential for improving safety and efficiency, especially for pilots, by accurately tracking and managing communication flow, identifying instances of miscommunication and errors, and enabling timely interventions and corrective actions. In [32], the author proposed a method based on graph neural networks (GNNs) to enhance clustering procedures in speaker diarization. The approach aims to purify the similarity matrix used in spectral clustering and assumes a sequence of speaker embeddings that the GNN processes. The GNN outputs a distance metric between the reference and estimated affinity matrices and is trained using a combination of a histogram loss and nuclear norm. Another approach for speaker diarization was proposed in [33], using deep neural networks to learn representations and scoring functions for speaker diarization without relying on i-vector clustering. The proposed method aims to reduce the computational cost and improve the efficiency of speaker diarization in the presence of multiple speakers.

As described above, the purpose of speaker clustering is to classify segmented speech into clusters so that each group only contains speech from one specific speaker. Our approach is based on the methodology described in [34], in which speech segments were preprocessed using the Kaldi FBank features with 40 dimensions, a 16k Hz sampling rate, and 40 filter-bank channels. These features were used as input to the RESNET34 neural network, which processed them using 2-dimensional CNN layers to generate fixed-size embeddings for each speaker. To train the model, we used 500,000 utterances by thousands of speakers from the publicly available VOXCeleb 2 dataset. We applied Probabilistic Linear Discriminant Analysis (PLDA) to the embeddings, which were trained on the VOXCeleb 2 data. The x-vector features generated by the neural network were centered using the training data mean, and Linear Discriminant Analysis (LDA) was applied to further improve the system's performance. For speaker clustering, we used the unweighted pair group method with arithmetic mean (UPGMA), which is a variant of agglomerative hierarchical clustering (AHC). The method consists of grouping similar objects or data points based on their pairwise distances. The algorithm follows the following steps:

1. Calculate the distance matrix D:

$$D_{ij} = \begin{cases} 0, & i = j \\ d_{ij}, & i \neq j, \end{cases} \tag{1}$$

where d_{ij} is the distance between objects (i, j).

2. Calculate the minimum distance pair (i^*, j^*) in the distance matrix D:

$$(i^*, j^*) = \arg\min_{i,j} D_{ij} \tag{2}$$

3. Calculate the new cluster k by averaging the distances between i^* and all objects in the cluster containing i^* and j^*:

$$k = \frac{1}{|C_i| + |C_j|} \sum_{l \in C_i \cup C_j} d_{il} \tag{3}$$

where C_i and C_j are the clusters containing objects i* and j*, and k represents the distance value associated with the newly formed cluster.

4. Update the distance matrix D by removing rows and columns i^* and j^* and adding a new row and column for the newly formed cluster k:

$$D_{ik} = D_{ki} = \frac{d_{ik} + d_{jk}}{2}, \quad \forall i \neq i, j \tag{4}$$

$$D_{kj} = D_{jk} = \infty \tag{5}$$

$$D_{kl} = D_{lk} = \frac{|C_i| d_{il} + |C_j| d_{jl}}{|C_i| + |C_j|}, \quad \forall l \neq i, j, k \tag{6}$$

In this step, k is an index representing the newly formed cluster, whereas i' and j' represent the indices of the objects selected for merging.

5. Repeat steps 2–4 until all objects are in a single cluster or the process is stopped based on a fixed threshold.

In our case, we obtain a pairwise log-likelihood ratio scores matrix using our PLDA model. We represent the distance between clusters by subtracting this matrix from zero, which is then fed into our clustering algorithm. The output generated by the clustering algorithm groups the audio files into clusters, with files that are potentially spoken by the same speaker being assigned to the same cluster.

3. Datasets

In this section, we describe the datasets used to train and evaluate our speaker role classification and speaker clustering components. For training, we employed different datasets for each of the two components. However, for evaluation, we utilized identical testing datasets to evaluate the performance of both components.

3.1. Training

The following sub-section describes the datasets used in the training of speaker role classification and speaker clustering.

3.1.1. Speaker Role Classification

LDC-ATCC corpus: The Air Traffic Control Corpus (LDC-ATCC: https://catalog.ldc.upenn.edu/LDC94S14A, accessed on 29 April 2023). (ATCC) consists of recorded speech for use in ATC research in the area of ASR and NLP. The audio data contains voice communication traffic between various ATCos and pilots. The audio files are sampled at 8 kHz, 16-bit linear, representing continuous monitoring without squelch or silence elimination. Each file captures a single radio frequency channel over one to two hours of audio. The corpus contains gold annotations and metadata (metadata covers voice activity segmentation details, speaker role information (who is talking), and callsigns in ICAO format). The corpus consists of approximately 25 h of ATCo and pilot transmissions (after SAD).

UWB-ATCC corpus: This corpus (released by the University of West Bohemia: https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0001-CCA1-0, accessed on 29 April 2023) is a free and public resource for research on ATC. The communication between ATCos and pilots is manually transcribed and labeled with the speaker information, i.e., pilot/controller. The total duration of speech after removing silences is 13 h. The audio data are single-channel and sampled at 8kHz and 16-bit PCM.

3.1.2. Speaker Clustering

VoxCeleb 1 [35]: This is a dataset comprising 100,000 utterances by 1251 celebrities, which were extracted from videos uploaded to YouTube. The dataset ensures a gender balance, with 55% of the speakers being male, and it features a wide range of speakers of different ages, accents, and ethnicities. The dataset includes speech audio taken in different environments. The diversity of the speakers and environments makes the dataset valuable for training and evaluating different systems for speaker and speech recognition.

VoxCeleb 2 [36]: This is the second version of the dataset, and it builds upon the success of its predecessor, VoxCeleb 1. Like VoxCeleb 1, it contains a large number of utterances by celebrities extracted from YouTube videos, and it features a diverse set of speakers in terms of age, ethnicity, and accent. However, VoxCeleb 2 has more than 1,000,000 utterances by 6000 celebrities, and it has no overlap with the identities of VoxCeleb 1. As a result, it provides an even larger and more diverse dataset for training and evaluating speech processing systems.

Librispeech [37]: This is a corpus of approximately 1000 h of read English speech sampled at 16 kHz from the LibriVox project. The LibriVox project is responsible for the creation of approximately 8000 public domain audio books, the majority of which are in English. Most of the recordings are based on texts from Project Gutenberg2, also in the public domain.

3.2. Evaluation

We employed the same testing datasets for evaluating the performance of automatic speech recognition, speaker role classification, and speaker clustering components.

LDC-ATCC dataset: The test set of LDC-ATCC was used for evaluation. This set consists of 2961 utterances, featuring dialogues between ATCos and pilots, and uses the same speech audio data described in the previous section.

ATCO2 corpus: This dataset was built for the development and evaluation of ASR and NLP technologies for English ATC communications from several airports worldwide (e.g., LKTB, LKPR, LZIB, LSGS, LSZH, LSZB, and YSSY). We used the *ATCO2 test set corpus*, which comprises ∼4 h of annotated data. The corpus is available for purchase through ELDA (http://catalog.elra.info/en-us/repository/browse/ELRA-S0484, accessed on 29 April 2023). The recordings are mono-channel sampled at 16 kHz and 16-bit PCM [38].

4. Experiments and Results

In this section, we present the experimental setup and the results obtained for the different modules of the automatic speaker clustering pipeline described in Section 2. The results include ASR, speaker role classification, speaker clustering modules, as well as the overall performance of the pipeline when all modules are combined. The results are discussed in detail in the following subsections.

4.1. Automatic Speech Recognition

As mentioned in Section 2.2, we adopted two approaches for training an ASR engine: (i) a hybrid-based approach and (ii) an end-to-end training approach. The automatic transcripts were generated using automatic speech recognition systems trained with ∼190 h of annotated ATC data.

Baseline: The main blocks of the hybrid ASR system are the acoustic model (AM) and language model (LM). In our experiments, conventional biphone convolutional neural network (CNN) [26] + TDNN-F [39]-based acoustic models trained with the Kaldi [40] toolkit (i.e., nnet3 model architecture) were used. AMs were trained with the LF-MMI training framework, which is considered to achieve state-of-the-art performance for hybrid ASR. Threefold speed perturbation with MFCC features was used, and i-vectors were used for speaker representation. The 3-gram LM was trained on all the manual transcripts available in the ATC datasets.

XLSR-KALDI: As mentioned earlier, the self-supervised learning approaches using the wav2vec framework facilitated the state-of-the-art performance in ASR. These models were pre-trained with 50k h of speech data. One such model is the XLSR [41], which can then be fine-tuned to ATC data. The authors of [42] proposed to use the LF-MMI criterion (similar to hybrid-based ASR) for the supervised adaptation of the self-supervised pretrained XLSR model [41]. We employed this technique to fine-tune the pre-trained model on our annotated ATC data.

The performance of our ASR system is presented in Table 1 using the Word Error Rate (WER) metric. The system that achieved the lowest WER on the test data was used as the input for the speaker role classification system.

Table 1. WER (%) of our ASR used in our experiments for speaker role classification evaluated on the ATCO2 and LDC-ATCC test sets described in Section 3.2.

Model	ATCO2	LDC-ATCC
Baseline	36.6	**13.5**
XLSR-KALDI	**25.7**	18.7

4.2. Speaker Role Classification

A BERT-based speaker role identification module was implemented that allowed us to attribute a speaker role (i.e., ATCo or pilot) to a given ATC communication. We fetched a BERT (BERT-base-uncased model: 110 M parameters) model [27] from Huggingface [43,44]. We then used ground-truth speaker labels to fine-tune the model on the sequence classification task with the data defined in Section 3.1.1.

Fine-tuning: the BERT model was fine-tuned for 3k steps (∼5 epochs), with a 500-step warm-up phase. The learning rate was increased linearly up to 5×10^{-5} during warm-up, and then it decayed linearly. We fine-tuned each model using the Adam optimizer, a batch size of 32, and a gradient accumulation of 2. After the training, we simply performed inference on either the manual transcripts or automatic transcripts generated through ASR.

Results: Table 2 shows the performance of the data-driven model trained for speaker role classification. The performance is shown for the test sets—ATCO2 and LDC ATCC—trained with all combinations of the training data sets mentioned in Section 3.1.1. We also report the F1 score of the system when (i) manual transcripts and (ii) automatic transcripts are used for classification.

Table 2. Averaged F1 score [0–1] for speaker role classification using different training (column 1) and test sets. All the experiments used the same model (BERT-base-uncased) and the same hyperparameters. We report the mean of five runs with different seeds (the standard deviation was less than 0.01 for all cases, thus we omit it). **Bold** refers to the best performance in each column. Metrics reported on ground-truth transcripts and automatic transcripts generated using the speech recognition system.

Model	ATCO2	LDC-ATCC
Manual Transcripts		
LDC-ATCC	0.83	**0.94**
UWB	0.85	0.87
LDC-ATCC + UWB	**0.87**	0.93
Automatic Transcripts		
LDC-ATCC	0.5	0.9
UWB	0.51	0.8
LDC-ATCC + UWB	0.53	0.9

4.3. Speaker Clustering

For all the experiments in this study, hypotheses concerning the ground truth of pilot identities were generated based on information about the creation of the datasets. Two datasets, ATCO2 and LDC-ATCC, were used to evaluate the performance of our model. In ATCO2, the ground truth was generated using both the callsign and flight date as the pilot identity information. In LDC-ATCC, only the callsign was used as the ground truth for the pilot identity. To determine the optimal threshold for hierarchical clustering, we randomly selected a representative subset of the LDC-ATCC training set consisting of three files per callsign from 259 different callsigns. We fine-tuned the threshold on this selected set as a whole, extracting the value that resulted in the highest accuracy, as shown in Figure 2. The resulting threshold was then used for evaluation on both datasets.

Upon evaluating the test set using this ground-truth generation approach, we observed a total of 929 distinct speakers in the ATCO2 dataset. The ATCO2 dataset covers a span of 7 months from October 2020 to May 2021. Additionally, in the LDC-ATCC dataset, we identified 189 distinct speakers. This indicates the number of unique speakers identified within each respective dataset. The output generated by the clustering algorithm represents the different clusters.

To evaluate the accuracy of our system, we proposed the following evaluation approach: Using the ground truth, we assigned to each cluster the speaker that was assigned to it the most. The utterances that were not assigned to that specific cluster but had this speaker as their label are considered errors. The idea was to map each speaker with one cluster, while all the remaining clusters would be considered errors. Using the same approach, when evaluating the performance of the entire pipeline, we added a constraint to our evaluation method. The pipeline first extracts the speech segments of the pilots. All utterances that are incorrectly classified as belonging to a pilot are also considered errors in our speaker clustering accuracy.

We conducted experiments using two datasets, LDC-ATCC and ATCO2, and utilized the speaker role classification (SRC) method to extract the speech segments of pilots from the datasets. The number of speech utterances for pilots was initially 1350 for ATCO2 and 1446 for LDC-ATCC using the SRC ground truth. However, 243 and 281 utterances were, respectively, removed from ATCO2 and LDC-ATCC datasets due to their short duration (less than 1 s). We further applied the SRC on the manual transcripts of the same dataset, resulting in 1455 utterances for ATCO2 and 1563 utterances for LDC-ATCC. However, 322 and 300 of these utterances were excluded due to their short duration. Lastly, we used SRC on the ASR transcripts of the datasets, resulting in 1705 and 1563 speech utterances for ATCO2 and LDC-ATCC, respectively. However, 389 and 288 of the ASR transcripts were excluded due to their short duration. These excluded segments were not used in the speaker clustering part of the experiment. The details of the pipeline's output are summarized in Table 3, whereas the performance of our model across all experiments is summarized in Table 4.

In our experiments, we found that the level of noise in the data had a significant impact on the accuracy of the speaker clustering pipeline. We observed that the speaker clustering model performed better on the LDC-ATCC dataset, which contained less noise, compared to the noisier ATCO2 dataset. After analyzing the results, we concluded that the difference in the accuracy of all pipelines was mainly due to the performance of the automatic speech recognition (ASR) and speaker role classification (SRC) components of the pipeline, which exhibited lower performance on the noisier ATCO2 data. However, on the LDC-ATCC dataset, we observed that both ASR and SRC exhibited better performance, resulting in a smaller decrease in accuracy. In addition, we found that the difference in performance between clustering alone and the complete pipeline was 8% on the LDC-ATCC dataset and 16% on the ATCO2 dataset. These findings suggest that more research is necessary to improve the performance of the ASR and SRC components, especially in datasets with higher levels of noise like ATCO2 to achieve optimal results with the speaker clustering pipeline.

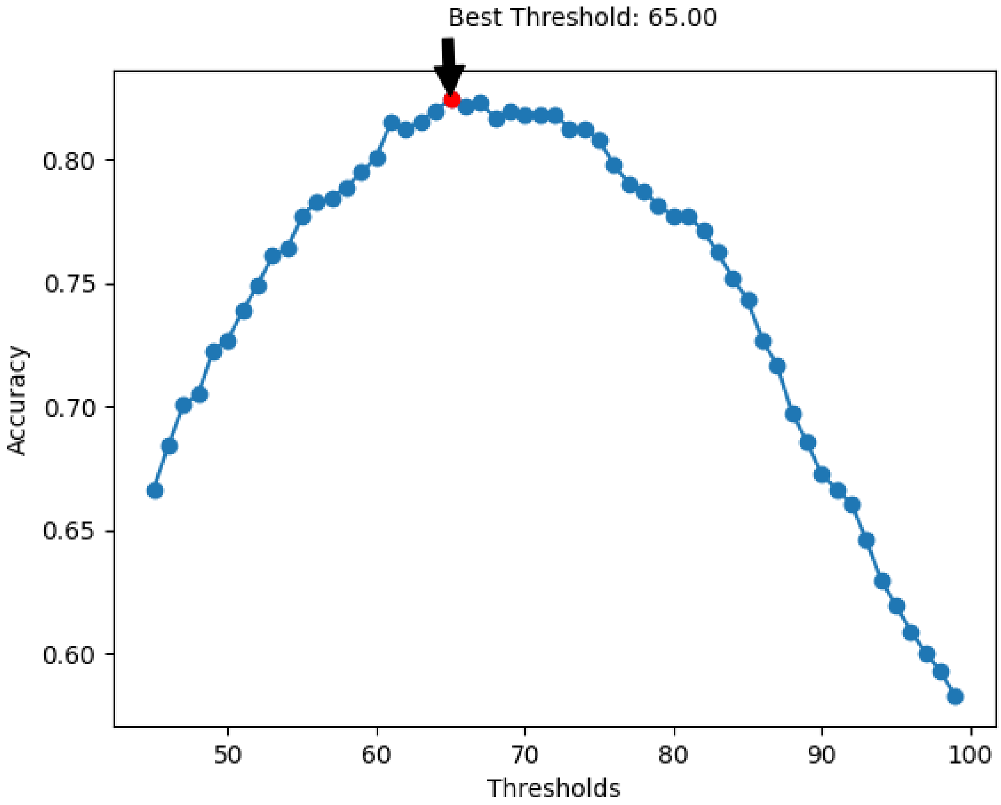

Figure 2. Accuracy vs. thresholds plot used to fine-tune the threshold on a representative subset of the LDC-ATCC dataset for the speaker clustering algorithm. The x-axis shows the threshold values and the y-axis shows the corresponding accuracy values. The red circle indicates the best threshold value (65) with a maximum accuracy of 82%.

Table 3. Automatic speaker clustering performance details for ATCO2 and LDC-ATCC datasets. **Number of segments classified as ATCo**: The segments that were classified as ATCo instead of pilot using the SRC model. **Number of Correct segments**: The segments that were assigned the same label as the ground truth after the evaluation mapping. **Number of Incorrect segments**: The segments that were assigned a different label from the ground truth after the evaluation mapping.

Experiment	Number of Segments Classified as ATCo	Number of Correct Segments	Number of Incorrect Segments
ATCO2			
SRC Ground Truth	-	739	368
Manual transcript	118	663	470
ASR transcript	337	661	655
LDC-ATCC			
SRC Ground Truth	-	908	257
Manual transcript	111	897	366
ASR transcript	128	896	379

Table 4. Accuracy (%) of the speaker clustering on ATCO2 and LDC-ATCC datasets.

Dataset	SRC Ground Truth	Manual Transcript	ASR Transcript
ATCO2	66%	58%	50%
LDC-ATCC	78%	71%	70%

5. Discussion and Conclusions

In conclusion, the presented pipeline offers a viable solution to the speaker clustering problem in ATC communication. By using a combination of speech activity detection, automatic speech recognition, text-based speaker role classification, and unsupervised speaker clustering, the pipeline can accurately identify and group speech segments from the same pilot among different speakers. The reported accuracies of 70% and 50% on the LDC-ATCC and ATCO2 datasets, respectively, signify the pipeline's proficiency in identifying pilot speakers within the ATC domain. It is important to note that these accuracies reflect the overall performance of the entire pipeline. There is an observed variation in accuracy when dealing with datasets of different noise levels, such as the LDC-ATCC and ATCO2 datasets, which show a notable deviation of approximately 20%. This discrepancy can be attributed to the effect of noise appearing in VHF data. Specifically, when noise levels increase, it not only challenges the initial component (SAD) by making it harder to accurately identify speech segments but also significantly impacts the ASR component, leading to transcription errors. These inaccuracies propagate through the pipeline and affect the performance of all the remaining components. Consequently, the decrease in speaker clustering accuracy from 70% to 50% on the LDC-ATCC and ATCO2 datasets illustrates the sensitivity of the entire pipeline to noise interference. Nevertheless, when considering the speaker clustering step alone and utilizing the speaker role classification as ground truth, even higher accuracy rates of 78% and 66% can be achieved on the same LDC-ATCC and ATCO2 datasets. This technology has the potential to improve ATC safety, facilitating post-flight analysis and incident investigation. As such, further research in this area is warranted to refine and improve these automated methods for speaker clustering in ATC communication.

Potential future work could focus on enhancing the performance of speaker clustering models with noisy data such as the ATCO2 dataset. We aim to adapt the embedding used for the speaker clustering model on ATC data to improve its performance with such types of noisy data. Another approach is to investigate some speech processing methods to reduce noise and improve the quality of the input data. We also plan to incorporate language identification (LID) as prior information for the speaker clustering in our proposed pipeline. This could potentially improve the accuracy of the clustering by providing additional information about the language and dialect being spoken. Another approach could be to expand the pipeline to support a variety of languages and accents, which would make it more suitable for use in actual ATM systems. By making these modifications, we believe that we can enhance the performance of the speaker clustering model and make it more appropriate for use in real-world scenarios.

Author Contributions: Writing—original draft, D.K. and A.P.; Writing—review & editing, P.M., J.Z.-G., I.N., S.M. and C.S.; Supervision, P.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Armasuisse Science and Technology.

Data Availability Statement: Private and public databases are used in this paper. They are covered in detail in Section 3.

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

References

1. Zuluaga-Gomez, J.; Motlicek, P.; Zhan, Q.; Veselý, K.; Braun, R. Automatic Speech Recognition Benchmark for Air-Traffic Communications. In *Proceedings of the Interspeech*; ISCA: Singapore, 2020; pp. 2297–2301. [CrossRef]
2. Szöke, I.; Kesiraju, S.; Novotný, O.; Kocour, M.; Veselý, K.; Černocký, J. Detecting English Speech in the Air Traffic Control Voice Communication. In *Proceedings of the Interspeech*; ISCA: Singapore, 2021; pp. 3286–3290. [CrossRef]
3. Zuluaga-Gomez, J.; Veselỳ, K.; Blatt, A.; Motlicek, P.; Klakow, D.; Tart, A.; Szöke, I.; Prasad, A.; Sarfjoo, S.; Kolčárek, P.; et al. Automatic call sign detection: Matching air surveillance data with air traffic spoken communications. *Proc. Multidiscip. Digit. Publ. Inst.* **2020**, *59*, 14.
4. Prasad, A.; Juan, Z.G.; Motlicek, P.; Sarfjoo, S.S.; Iuliia, N.; Ohneiser, O.; Helmke, H. Grammar Based Speaker Role Identification for Air Traffic Control Speech Recognition. *arXiv* **2022**, arXiv:2108.12175.
5. Lukic, Y.X.; Vogt, C.; Dürr, O.; Stadelmann, T. Learning embeddings for speaker clustering based on voice equality. In Proceedings of the 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP), Tokyo, Japan, 25–28 September 2017; pp. 1–6. [CrossRef]
6. Sarfjoo, S.S.; Madikeri, S.; Motlicek, P. Speech Activity Detection Based on Multilingual Speech Recognition System. *arXiv* **2020**, arXiv:2010.12277.
7. Kleinbaum, D.G.; Dietz, K.; Gail, M.; Klein, M.; Klein, M. *Logistic Regression*; Springer: Berlin/Heidelberg, Germany, 2002.
8. Mohri, M.; Pereira, F.; Riley, M. Weighted finite-state transducers in speech recognition. *Comput. Speech Lang.* **2002**, *16*, 69–88. [CrossRef]
9. Mohri, M.; Pereira, F.; Riley, M. Speech recognition with weighted finite-state transducers. In *Springer Handbook of Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 559–584.
10. Riley, M.; Allauzen, C.; Jansche, M. OpenFst: An Open-Source, Weighted Finite-State Transducer Library and its Applications to Speech and Language. In Proceedings of the Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts, Boulder, Colorado, 31 May–5 June 2009; Association for Computational Linguistics: Boulder, CO, USA, 2009; pp. 9–10.
11. Srinivasamurthy, A.; Motlicek, P.; Himawan, I.; Szaszak, G.; Oualil, Y.; Helmke, H. Semi-supervised learning with semantic knowledge extraction for improved speech recognition in air traffic control. In Proceedings of the 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 20–24 August 2017.
12. Kleinert, M.; Helmke, H.; Siol, G.; Ehr, H.; Cerna, A.; Kern, C.; Klakow, D.; Motlicek, P.; Oualil, Y.; Singh, M.; et al. Semi-supervised adaptation of assistant based speech recognition models for different approach areas. In Proceedings of the 37th Digital Avionics Systems Conference (DASC), London, UK, 23–27 September 2018; IEEE: Piscataway, NJ, USA, 2018, pp. 1–10.
13. Khonglah, B.; Madikeri, S.; Dey, S.; Bourlard, H.; Motlicek, P.; Billa, J. Incremental semi-supervised learning for multi-genre speech recognition. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020, pp. 7419–7423.
14. Zuluaga-Gomez, J.; Nigmatulina, I.; Prasad, A.; Motlicek, P.; Veselỳ, K.; Kocour, M.; Szöke, I. Contextual Semi-Supervised Learning: An Approach to Leverage Air-Surveillance and Untranscribed ATC Data in ASR Systems. In *Proceedings of the Interspeech*; ISCA: Singapore, 2021; pp. 3296–3300. [CrossRef]
15. Kocour, M.; Veselý, K.; Blatt, A.; Gomez, J.Z.; Szöke, I.; Cernocky, J.; Klakow, D.; Motlicek, P. Boosting of Contextual Information in ASR for Air-Traffic Call-Sign Recognition. In *Proceedings of the Interspeech*; ISCA: Singapore, 2021; pp. 3301–3305. [CrossRef]
16. Nigmatulina, I.; Braun, R.; Zuluaga-Gomez, J.; Motlicek, P. Improving callsign recognition with air-surveillance data in air-traffic communication. *arXiv* **2021**, arXiv:2108.12156.
17. Nigmatulina, I.; Zuluaga-Gomez, J.; Prasad, A.; Sarfjoo, S.S.; Motlicek, P. A two-step approach to leverage contextual data: Speech recognition in air-traffic communications. In *Proceedings of the ICASSP*; ISCA: Singapore, 2022.

18. Zuluaga-Gomez, J.; Prasad, A.; Nigmatulina, I.; Sarfjoo, S.; Motlicek, P.; Kleinert, M.; Helmke, H.; Ohneiser, O.; Zhan, Q. How Does Pre-trained Wav2Vec2.0 Perform on Domain Shifted ASR? An Extensive Benchmark on Air Traffic Control Communications. *arXiv* **2023**, arXiv:2203.16822.
19. Povey, D.; Peddinti, V.; Galvez, D.; Ghahremani, P.; Manohar, V.; Na, X.; Wang, Y.; Khudanpur, S. Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *Proceedings of the Interspeech*; ISCA: Singapore, 2016; pp. 2751–2755.
20. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised Pre-Training for Speech Recognition. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; ISCA: Singapore, 2019; pp. 3465–3469. [CrossRef]
21. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020.
22. He, Z.; Wang, Z.; Wei, W.; Feng, S.; Mao, X.; Jiang, S. A Survey on Recent Advances in Sequence Labeling from Deep Learning Models. *arXiv* **2020**, arXiv:2011.06727.
23. Grishman, R.; Sundheim, B. Message Understanding Conference-6: A Brief History. In Proceedings of the COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 5–9 August 1996.
24. Yadav, V.; Bethard, S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; Association for Computational Linguistics: Santa Fe, NM, USA, 2018; pp. 2145–2158.
25. Zhou, C.; Cule, B.; Goethals, B. Pattern based sequence classification. *IEEE Trans. Knowl. Data Eng.* **2015**, *28*, 1285–1298. [CrossRef]
26. LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. *Handb. Brain Theory Neural Netw.* **1995**, *3361*, 1995.
27. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
28. Zuluaga-Gomez, J.; Sarfjoo, S.S.; Prasad, A.; Nigmatulina, I.; Motlicek, P.; Ondre, K.; Ohneiser, O.; Helmke, H. BERTraffic: BERT-based Joint Speaker Role and Speaker Change Detection for Air Traffic Control Communications. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 9–12 January 2023.
29. Prasad, A.; Zuluaga-Gomez, J.; Motlicek, P.; Sarfjoo, S.; Nigmatulina, I.; Veselý, K. Speech and Natural Language Processing Technologies for Pseudo-Pilot Simulator. *arXiv* **2022**, arXiv:2212.07164.
30. Zuluaga-Gomez, J.; Prasad, A.; Nigmatulina, I.; Motlicek, P.; Kleinert, M. A Virtual Simulation-Pilot Agent for Training of Air Traffic Controllers. *Aerospace* **2023**, *10*, 490. [CrossRef]
31. Pellegrini, T.; Farinas, J.; Delpech, E.; Lancelot, F. The Airbus Air Traffic Control speech recognition 2018 challenge: Towards ATC automatic transcription and call sign detection. *arXiv* **2018**, arXiv:1810.12614.
32. Wang, J.; Xiao, X.; Wu, J.; Ramamurthy, R.; Rudzicz, F.; Brudno, M. Speaker diarization with session-level speaker embedding refinement using graph neural networks. *arXiv* **2020**, arXiv.2005.11371.
33. 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. [CrossRef]
34. Zeinali, H.; Wang, S.; Silnova, A.; Matějka, P.; Plchot, O. BUT System Description to VoxCeleb Speaker Recognition Challenge *arXiv* **2019**, arXiv.1910.12592.
35. Nagrani, A.; Chung, J.S.; Zisserman, A. VoxCeleb: A Large-Scale Speaker Identification Dataset. *arXiv* **2017**, arXiv:1706.08612.
36. Chung, J.S.; Nagrani, A.; Zisserman, A. VoxCeleb2: Deep Speaker Recognition. *arXiv* **2018**, arXiv:1806.05622.
37. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 5206–5210.
38. Zuluaga-Gomez, J.; Veselỳ, K.; Szöke, I.; Motlicek, P.; Kocour, M.; Rigault, M.; Choukri, K.; Prasad, A.; Sarfjoo, S.S.; Nigmatulina, I.; et al. ATCO2 corpus: A Large-Scale Dataset for Research on Automatic Speech Recognition and Natural Language Understanding of Air Traffic Control Communications. *arXiv* **2022**, arXiv:2211.04054.
39. Povey, D.; Cheng, G.; Wang, Y.; Li, K.; Xu, H.; Yarmohammadi, M.; Khudanpur, S. Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. In *Proceedings of the Interspeech*; ISCA: Singapore, 2018; pp. 3743–3747.
40. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011; IEEE Signal Processing Society: Piscataway, NJ, USA, 2011.
41. Conneau, A.; Baevski, A.; Collobert, R.; Mohamed, A.; Auli, M. Unsupervised cross-lingual representation learning for speech recognition. *arXiv* **2020**, arXiv:2006.13979.
42. Vyas, A.; Madikeri, S.; Bourlard, H. Lattice-Free Mmi Adaptation of Self-Supervised Pretrained Acoustic Models. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6219–6223. [CrossRef]

43. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
44. Lhoest, Q.; del Moral, A.V.; Jernite, Y.; Thakur, A.; von Platen, P.; Patil, S.; Chaumond, J.; Drame, M.; Plu, J.; Tunstall, L.; et al. Datasets: A Community Library for Natural Language Processing. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 7–11 November 2021; pp. 175–184.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.