Neural Networks-Based Approach for Detecting and Filtering Misleading Audio Segments for Classroom Automatic Transcription

Jorge Hewstone and Roberto Araya *

Posted Date: 26 October 2023

doi: 10.20944/preprints202310.1690.v1

Keywords: n/a; Neural Networks; Noise Detection; Noise Filtering; Classroom Recording; Classroom Analysis

Article

Neural Networks-Based Approach for Detecting and Filtering Misleading Audio Segments for Classroom Automatic Transcription

Jorge Hewstone [1,†] and Roberto Araya [2,*,†]

[1] Department of Mathematical Engineering, Faculty of Physical and Mathematical Sciences, Universidad de Chile, 8370458 Santiago,Chile; jhewstone@dim.uchile.cl
[2] Institute of Education, University of Chile, Periodista Jose Carrasco Tapia 75, 8370458 Santiago, Chile; roberto.araya.schulz@gmail.com
* Correspondence: roberto.araya.schulz@gmail.com; Tel.: +56 99 599 0251
† These authors contributed equally to this work.

Featured Application: Authors are encouraged to provide a concise description of the specific application or a potential application of the work. This section is not mandatory.

Abstract: Audio recording in classrooms is a common practice in educational research, with applications ranging from detecting classroom activities to analysing student behaviour. Previous research has employed neural networks for classroom activity detection and speaker role identification. However, these recordings are often affected by background noise that can hinder further analysis, and the literature has only sought to identify noise with general filters and not specifically designed for classrooms. Although the use of high-end microphones and environmental monitoring can mitigate this problem, these solutions can be costly and potentially disruptive to the natural classroom environment. In this context, we propose the development of a neural network model that can specifically detect and filter out background noise in classroom recordings. This model would allow the use of lower quality recordings without compromising analysis capability, thus facilitating data collection in natural educational environments and reducing the costs associated with high-end recording equipment.

Keywords: neural networks; noise detection; noise filtering; classroom recording; classroom analysis

1. Introduction

Noise in audio recordings is a persistent problem that can significantly hinder the analysis and interpretation of collected data. This issue is particularly pronounced in noisy environments like classrooms, where a variety of noise sources can interfere with the recording quality [1]. These noises can come from a variety of sources, including background student chatter, classroom ambient noise, and heating and cooling system noise. The unpredictable nature of these noise sources makes their removal from recordings a challenge [2].

To address these challenges, we have implemented the use of a lavalier microphone carried by the teacher, specifically designed to capture the teacher's speech. This microphone transmits via a UHF signal to a cell phone placed at the back of the classroom. Employing this method ensures the teacher's talk is predominantly recorded, mitigating privacy concerns as the cell phone mic could inadvertently record both audio and video of students. Furthermore, this setup synchronizes all recordings on a single device. This contrasts with previous studies that recorded video and audio separately and then had to synchronize them, a non-trivial task due to inadvertent interruptions. We use the UHF signal to avoid consuming wifi bandwidth, which is scarce in classrooms. However, the signal can occasionally be interrupted when the teacher moves around the classroom, and it can also occasionally introduce noise. In addition to this, the teacher could carry the necessary equipment (cell phone, microphone and small tripod as detailed in section 3) in a carry-on bag and could take them to other classrooms

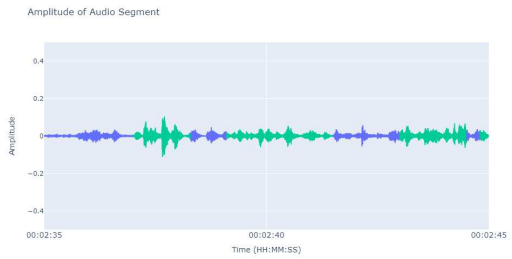
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1690.v1

and even to other schools. This practical and inexpensive solution was key to achieving the scale of recordings we were able to obtain.

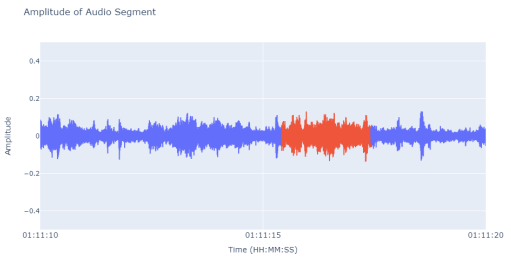
Classrooms are inherently complex and noisy environments due to the nature of the activities carried out in them. Students may be talking amongst themselves, teachers may be giving instructions, and there may be background noise from class materials and other equipment. The presence of multiple speakers and poor acoustics further complicates the process of noise reduction and speaker identification [1].

Collaborative learning environments can be particularly challenging when it comes to speaker identification in noisy audio recordings [2]. In these contexts, the task of identifying individual students speaking in small clusters amidst other simultaneous conversations becomes exceptionally complex. This difficulty underscores the demand for efficient strategies in detecting and filtering noise in classroom recordings. Despite these challenges, the development of advanced techniques such as deep learning and neural networks has shown promise in improving noise detection and filtering in classroom recordings [3–5].

A significant issue in classrooms arises when background noises are too similar to the teacher's voice, making it challenging for a simple audio enhancement to filter out the noise from the voice. When attempting to use a state-of-the-art automatic transcriber like Whisper [10], we encountered not only regular transcription issues but also hallucinations that generated entirely fabricated sentences unrelated to the context of the recording. (Figure 1)



(a) Audio segment where the transcriber had no problem identifying what the professor said and when he said it (The professor's speech in green).



(b) Segment with significant noise where the teacher barely spoke, almost indistinguishable from the room's noise. The transcription associated with the timestamp in red in the image is actually a hallucination that said "Ella es trabajecita, díganme cuánto pueden llegar a casa." (She's a little worker, tell me how much they can get home.)

Figure 1. Sound amplitude graph of audio segments belonging to the same lesson both are 10 seconds long. It can be seen that the noise that produces hallucinations reaches the same volume level as the teacher's speech level, which makes it difficult to separate it from the teacher's speech.

 doi:10.20944/preprints202310.1690.v1

In other instances, phrases like "Voy a estar en la iglesia!" ("I'll be at the church!") or "Se pide al del cielo y le ayuda a la luz de la luz." ("He asks the one from the sky and he helps with the light of the light.") appeared out of nowhere(in non talking segments) in a mathematics class, contaminating any analysis of the content of those audio segments.

While some studies have explored multimodal methods for speaker identification using both visual and acoustic features [6], the focus of this work is solely on acoustic features. The goal is to leverage the power of neural networks to detect and filter the best quality audio segments from classroom recordings, without the need for visual cues or enhancements to the audio during recording.

Moreover, while speech enhancement and inverse filtering methods have shown promise in improving speech intelligibility in noisy and reverberant environments [9], the aim of this work is not to improve the audio during recording. Instead, the focus is on post-recording, using a simple, low-cost recording setup such as a teacher's mobile phone.The proposed algorithm does not enhance the audio but selects the best quality audio segments for transcription. This approach not only facilitates the creation of high-quality datasets for speech enhancement models but also improves the transcriptions. By avoiding the transcription of intervals with higher noise levels, it reduces hallucinations and achieves improvements that cannot be obtained by merely adjusting the parameters of transcription tools like Whisper.

Finally, this work employs a specific classifier to select high-quality segments for transcription. This approach has the potential to improve the results of transcription systems like Whisper, which are trained on a broad and diverse distribution of audio and evaluated in a zero-shot setting, thereby moving closer to human behavior [10].

In this context, the proposal of a neural network model that can specifically detect and filter background noises in classroom recordings presents a promising solution to this persistent problem. The development of these efficient strategies is crucial for improving the quality of data collected and facilitating its analysis and interpretation[2,3,5].

2. Related work

Sound detection, and particularly the identification of noises from various sources, has as many applications as varied as the places where these noises can be found. In this section, we summarize several works found in the literature that pertain to the diverse applications, approaches, and systems devised for this task across a multitude of contexts. This will provide perspective on the contribution and positioning of our work in relation to the studies showcased in this section.

Of notable mention is the work by Sabri et al. [11], who in 2003 developed an audio noise detection system using the Hidden Markov Model (HMM). They aimed to classify aircraft noise with an 83% accuracy, using 15 training signals and 28 testing signals. The system combined linear prediction coefficients with Cepstrum coefficients and employed vector quantization based on fuzzy C-mean clustering. This pioneering work, developed before the significant expansion of neural networks in noise detection, has influenced many environments and could even extend to settings such as classrooms.

Another significant work around noise detection was developed by Rangachari and Loizou[12], who proposed a noise-estimation algorithm for highly non-stationary environments, such as a classroom or a busy street, where noise characteristics change rapidly. The algorithm updates the noise estimate using a time-frequency dependent smoothing factor, computed based on the speech-presence probability. This method adapts quickly to changes in the noise environment, making it suitable for real-time applications. The algorithm was found to be effective when integrated into speech enhancement, outperforming other noise-estimation algorithms, and has potential applications in improving the quality of speech communication in noisy settings.

In the domain of Voice Activity Detection (VAD), the influential study by Zhang and Wu [4] stands out for its pioneering approach in addressing the challenges of real-world noisy data. Their research introduces a methodology based on Denoising Deep Neural Networks (DDNN) for VAD,

 doi:10.20944/preprints202310.1690.v1

capitalizing on the network's ability to learn robust features amidst noise. Unlike many traditional systems that lean on clean datasets, this method emphasizes unsupervised denoising pre-training, followed by supervised fine-tuning. This design ensures the model adeptly adapts to real-world noisy data. Once the network undergoes pre-training, it is fine-tuned to mimic the characteristics of clean voice signals. The essence of this approach is its adaptability and proficiency in real-world scenarios, showcasing significant enhancement in voice activity detection amidst noise.

In the realm of polyphonic sound event detection, the work by Çakır et al. [13] stands out for its innovative approach using multi-label deep neural networks (DNNs). The study was aimed at detecting multiple sound events simultaneously in various everyday contexts, such as basketball matches and busy streets, where noise characteristics are complex and varied. The proposed system, trained on 103 recordings from 10 different contexts, totaling 1133 minutes, employed DNNs with 2 hidden layers and log Mel-band energy features. It outperformed the baseline method by a significant margin, offering an average increase in accuracy of 19%. The adaptability and effectiveness of this method in handling polyphonic sound detection make it a valuable contribution to the field of noise detection. It emphasizes the potential of DNNs in accurately identifying noises from various sources and sets a precedent for future research in enhancing real-time audio communication in noisy environments.

The paper by Dinkel et al. [14] proposes a cutting-edge method that overcomes the difficulties of actual noisy data in the rapidly changing environment of Voice Activity Detection (VAD). Their study provides a teacher-student model for VAD that is data-driven and uses vast, unrestricted audio data for training. This approach only requires weak labels, which are a form of supervision where only coarse-grained or ambiguous information about the data is available (e.g., a label for an entire audio clip rather than precise frame-by-frame annotations), during the teacher training phase. In contrast, many conventional systems rely extensively on clean or artificially noised datasets. A student model on an unlabeled target dataset is given frame-level advice by the teacher model after it has been trained on a source dataset. This method's importance rests in its capacity to extrapolate to real-world situations, showing significant performance increases in both artificially noisy and real-world settings. This study sets a new standard for future studies in this field by highlighting the potential of data-driven techniques in improving VAD systems, particularly in settings with unexpected and diverse noise characteristics.

An interesting approach can be found in the work of Rashmi et al. [15], where the focus is on removing noise from speech signals for Speech-to-Text conversion. Utilizing PRAAT, a phonetic tool, the study introduces a Training Based Noise Removal Technique (TBNRT). The method involves creating a noise class by collecting around 500 different types of environmental noises and manually storing them in a database as a noise tag set. This tag set serves as the training data set, and when an input audio is given for de-noising, the corresponding type of noise is matched from the tag set and removed from the input data without tampering with the original speech signal. The study emphasizes the challenges of handling hybrid noise and the dependency on the size of the noise class. The proposed TBNRT has been tested with various noise types and has shown promising results in removing noise robustly, although it has limitations in identifying noise containing background music. The approach opens up possibilities for future enhancements, including applications in noise removal stages in End Point Detection of continuous speech signals and the development of speech synthesis with emotion identifiers.

A significant contribution to this field is the work by Kartik and Jeyakumar [16], who developed a deep learning system to predict noise disturbances within audio files. The system was trained on 1282239 training files, and during prediction, it generates a series of 0s and 1s in intervals of 10 ms, denoting the 'Audio' or 'Disturbance' classes (Total of 3 hours and 33 minutes). The model employs dense neural network layers for binary classification and is trained with a batch size of 32. The performance metrics, including training accuracy of 90.50% and validation accuracy of 83.29%,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1690.v1

demonstrate its effectiveness. Such a system has substantial implications in preserving confidential audio files and enhancing real-time audio communication by eliminating disturbances.

While the research cited above have achieved substantial advances in noise detection and classification, there is still a large gap in tackling the unique issues provided by primary school classroom situations. These environments have distinct noise profiles that are frequently modified by student interactions, classroom activities, and the natural acoustics of the space. Unlike previous work, our research also focuses on the classification aimed at predicting the quality of audio-to-text transcriptions in the classroom environment. Furthermore, the feasibility of applying these methods in real-world applications has not been thoroughly investigated, particularly given budget limits that necessitate the use of low-cost UHF microphones and the requirement for distant detection (e.g., a smartphone put at a distance). Our research aims to close this gap by concentrating on the intricacies of noise interference in primary schools, particularly when employing low-cost equipment.

Given this context, our research questions are:

1. "To what extent can we enhance information acquisition from transcriptions in elementary school classes by eliminating interferences and noises inherent to such settings, using an affordable UHF microphone transmitting to a smartphone located 8 meters away at the back of the classroom?"
2. "To what extent can we ensure that, in our pursuit to enhance transcription quality by eliminating noises and interferences, valuable and accurate information is not inadvertently filtered out or omitted in the process?"

3. Materials and Methods

3.1. Data Collection

Video and audio recordings were carried out by trained teachers using low-cost equipment. The devices used for video recordings were Redmi 9A mobile phones with 2GB of RAM and 32GB of ROM. For audio recordings, Lavalier UHF lapel microphones were used. In addition, small tripods were used to hold the mobile phones during the recordings.

The costs of the equipment used are detailed in Table 1.

Table 1. Costs of the equipment used for video and audio recordings.

Equipment	Cost (Chilean Pesos)	Cost (USD)
Redmi 9A Mobile Phone	70,000	90
Lavalier UHF Lapel Microphone	17,000	20
Mobile Phone Tripod	5,000	6
Total per teacher	**92,000**	**116**

For a proper understanding of the cost, consider that each teacher taught around 20 hours of classes weekly. In a typical year, classes are held for 30 weeks, which amounts to 600 hours annually. Considering that the costs in the table are fixed for a specific year, we can estimate the cost per class hour for the materials as:

$$\frac{\text{Annual investment}}{\text{Annual class hours}} = \frac{116}{600} \approx 0.1933 \text{ (USD per hour)}$$

Given an average of 29 students per class, the cost per student is:

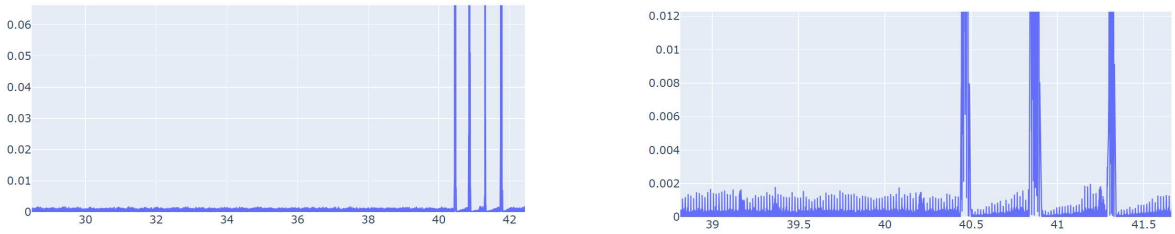
$$\frac{0.1933}{29} \approx 0.0067 \text{ (USD per student per hour)}$$

In total, 14 mobile phones were used for the recordings, representing a total investment of approximately $1,624$ USD.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1690.v1

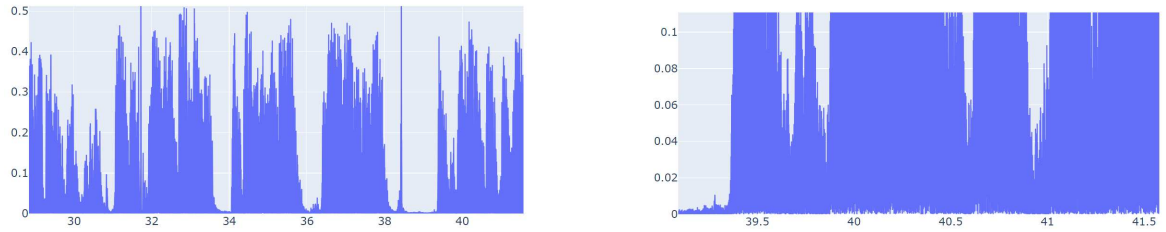
3.2. Preprocessing and Quality Control

We obtained the recordings and then conducted a quality diagnosis to filter out low-quality recordings. The filtering algorithm took uniformly distributed samples from the audio of each class's recordings. For each sample, we calculated the 50th percentile of the volume (or amplitude) of all the frames in the audio segment. If this percentile fell below an established empirical value (<0.005 this level can be deduce from the Figures 2 and 3), which is lower than common background noise, we considered that sample to have audio quality problems. If over 30% of the samples from a specific class recording had audio issues, we deemed the entire recording as low quality and discarded it.



(a) 10 seconds of the sample **(b)** zooming of the same sample

Figure 2. Sample with audio issues(silences).



(a) 10 seconds of the sample **(b)** zooming of the same sample

Figure 3. Sample with good audio(no issues).

3.3. Data labeling and Augmentation

We filtered the recordings and then took on the task of labeling randomly selected 10-second segments as "acceptable" or "not acceptable" for transcription. This task gave us a total of 1040 labeled segments, which amounts to approximately 173 minutes.

For the labeling, the human annotator's confidence in understanding the professor's speech served as the criteria. If the professor's speech was clear and understandable, the segment got an "acceptable" label. If the segment included external noises like white noise, dialogues from children, shouts, or chair movements, or if understanding the professor was difficult, the segment got a "not acceptable" label.

Recognizing the need to increase the quantity of data and mitigate any bias where volume might disproportionately influence noise detection, we implemented a data augmentation strategy. It's important to note that the clarity of an audio segment for transcription isn't determined by its volume but rather by its content and interference. This is because the person labeling the audio segments had the flexibility to adjust the volume or use headphones as needed, making the volume of the segment non-determinative for labeling. However, there exists a range of volume levels within which automatic transcription tools can effectively operate.

To increase the quantity of data and mitigate biases related to volume levels, we implemented a data augmentation strategy. Our algorithm first measures the current volume of each segment in decibels full scale (dBFS). It then calculates the difference between this current volume and a set of predefined optimal volume levels, ranging from -20 dBFS to -28 dBFS. The algorithm selects the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1690.v1

smallest difference, corresponding to the closest predefined volume level that is still greater than the current volume of the segment.

For example, if an original audio segment had a volume of -18 dBFS, the algorithm would lower it to -20 dBFS. Conversely, if a segment had a volume of -32 dBFS, the algorithm would raise it to -28 dBFS.

Moreover, to ensure that the volume of the augmented segment does not match the volume of the original segment, an element of randomness is introduced into the process. After adjusting the volume to the nearest optimal level, a random value is generated and added to the volume of the segment. This random value is either between -1.5 and -0.5 or between 0.5 and 1.5, ensuring a minimum alteration of 0.5 dBFS from the original adjusted volume level.

We then adjust the volume of each segment by this calculated difference. This effectively raises the volume of segments below the target range and lowers the volume of those above it. By applying this algorithm to each of the original 945 segments, we created an additional set of 945 segments with adjusted volumes. This doubled our dataset to a total of 1890 segments, each labeled as "Acceptable" or "Not Acceptable" depending on whether or not it was possible to recognize what was being said in the transcript. The sum of the two classes gave a total time of 5 hours and 15 minutes, their distribution is detailed in Figure 4.

This approach not only made our dataset more robust but also ensured it was unbiased, setting the stage for more reliable subsequent analysis.

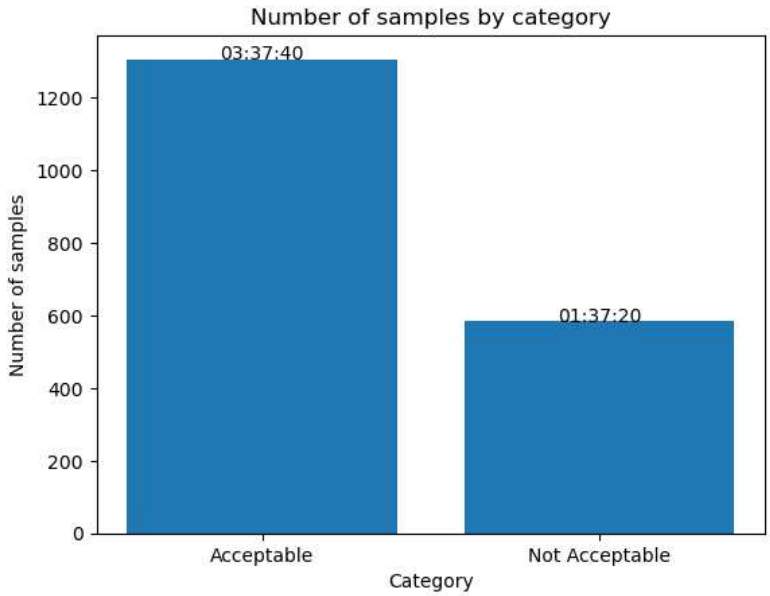


Figure 4. Distribution of the segments that we used to train our model. The numbers above the bars represent the amount of total time that each group of samples represent in format HH:MM:SS.

3.4. Audio Features

To train the neural network, we extracted a series of audio features associated with audio quality. These features include the deciles of the audio volume level, which divide the data into intervals containing equal proportions of the data. We also considered the average and standard deviation of these deciles, which measure the central tendency and dispersion of the audio volume level.

Frequencies play a crucial role in audio quality, so we included several frequency-related parameters. These parameters encompass the Mel-frequency cepstral coefficients (MFCCs), which capture the power spectrum of the audio signal; the spectral contrast, which measures the difference in amplitude between peaks and valleys in the sound spectrum; and the pitch, which orders sounds on a frequency-related scale based on their perceptual property.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1690.v1

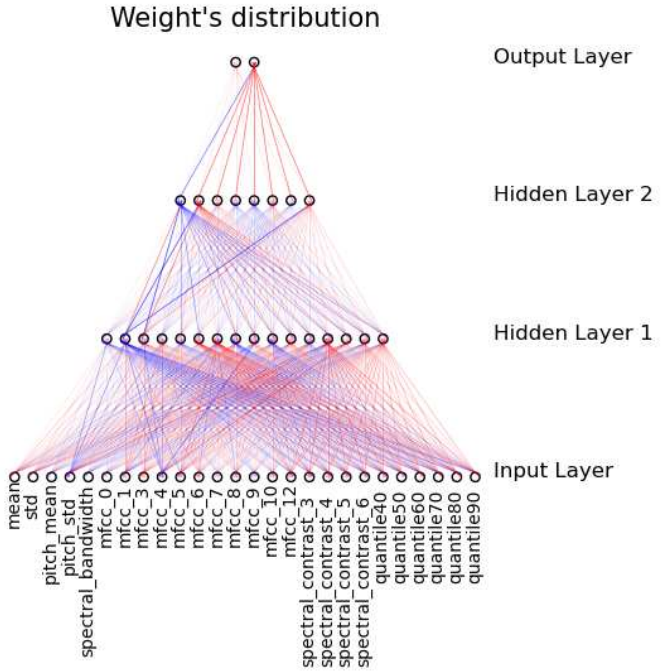
We also incorporated parameters related to the shape of the power spectrum, such as the spectral centroid and bandwidth, which measure the "center of mass" and spread of the power spectrum, respectively; the spectral rolloff, which indicates the frequency below which a specified percentage of the total spectral energy resides; and the zero crossing rate, which tracks the rate a signal changes from positive to zero to negative, or from negative to zero to positive.

From all these parameters, we discarded those with a correlation to the classification of interest that was too close to 0 (an absolute distance of less than 0.1). In the end, we selected only 26 from the original 38 input parameters for the final model.

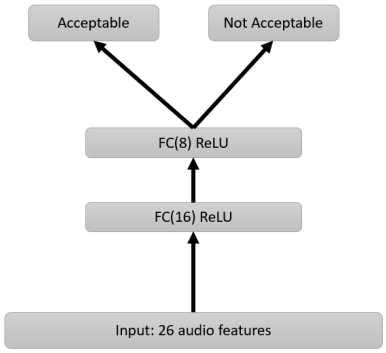
The list of parameters and their correlations can be found in Appendix A.

3.5. Model architecture and training

The proposed model is a classification network designed with the complexity of the task at hand in mind. It consists of two hidden layers with 16, and 8 neurons respectively, summing up to a total of 24 neurons(without counting the input layer of 26 and the output of 2) . The final layer has 2 neurons corresponding to the classification categories, reflecting the nature of the output parameters (Figure 5b).



(a) Distribution of model weights. Negative weights are shown in red and positive weights in blue. Lines with higher opacity have a higher weight (absolute value).



(b) Model architecture

Figure 5. Architecture details.

While the choice of the number of hidden layers and the nodes within each is often a subject of debate, the architecture of a classification model is largely dictated by the inherent complexity of the patterns one seeks to discern within the dataset. In our context, the objective is for our neural network to make informed decisions based on the characteristics of an audio sample. This necessitates, at a minimum, the ability to understand patterns from the audio's waveform plot. Two hidden layers have the capability to represent an arbitrary decision boundary with arbitrary accuracy using rational activation functions and can approximate any smooth mapping to any desired precision. [18] The decision to employ only 2 hidden layers was made after noting that adding more layers did not yield

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1690.v1

significant performance improvements during our tests. This observation aligns with findings reported for neural networks of a similar size in prior research, such as that by [19].

For the node count in each hidden layer, we adhered to the general "2/3" rule, suggesting that one layer should have approximately 2/3 the number of nodes as its preceding layer. [18] For computational reasons, these counts needed to be multiples of 8. This choice is grounded in the recommendation to work with matrices of this dimensionality to optimize processor efficiency, as advised in [17]. Given 26 input parameters and adhering to these guidelines, we arrived at the described architecture in Figure 5a.

We trained the model using the cross-entropy loss function, which is particularly suited for classification problems. The Adaptive Moment Estimation (Adam) optimizer was chosen for its efficiency and ability to handle large datasets[7], with a learning rate of 0.002. The training was carried out over 300 epochs (Figures 6 and 7) to ensure a robust model that can handle the complexity of the 100< input parameters and 10000< data points.

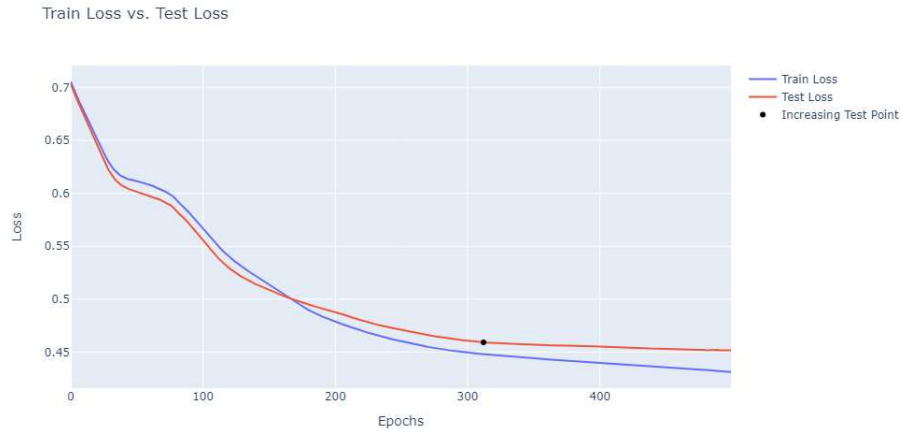


Figure 6. Loss of the model trained with the automatic split.

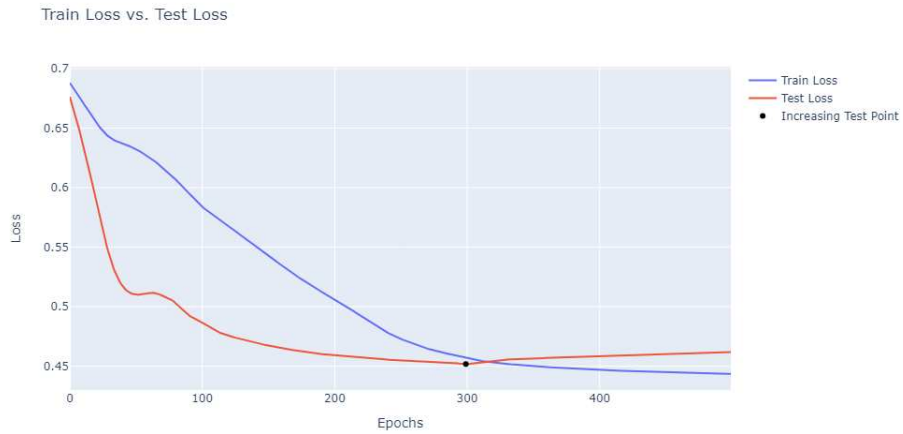


Figure 7. Loss of the model trained with the manual split.

We carefully chose the architecture of this model to address the specific challenges of this task. The use of ReLU activation functions allows the model to learn non-linear relationships, which are expected given the complex nature of audio signals [8]. The multiple layers enable the model to learn hierarchical representations of the input features, which is crucial given the large number of input parameters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1690.v1

The selection of the Adam optimizer was primarily influenced by its efficiency, its popularity in state-of-the-art machine learning applications, and its robustness in handling local convergence. While the dataset size is not exceptionally large(< 5000), the task is complex due to the 26 input parameters and the challenge of recognizing patterns in 10-second audio segments, which are considerably longer than those typically used in other noise detectors. For example, as cited in the related works section [16], the focus is often on detecting noises that last for fractions of a second. In contrast, our goal is to predict the quality of audio-to-text transcriptions, for which a 10-second audio segment is actually quite short. Therefore, the architecture of this model is designed to strike a balance between computational efficiency and the specific complexities associated with audio quality classification.

3.6. Training results

We employed two different data splits for training and testing. In the first split, we randomly divided the data without specific separation criteria, using 80% (1664 segments) for training and the remaining 20% (416 segments) for testing. We didn't train the model on this test data. For the second, more manual split, we ensured that the 416 testing segments came from entirely different teaching sessions than the 1664 training segments. We chose this deliberate separation for two main reasons: first, to avoid testing on "duplicated" segments that the model had already seen during training, and second, to evaluate the model's performance with new teaching sessions, distinct voices, varied dynamics, and other unique characteristics.

To further ensure the accuracy and reliability of our model's predictions, we employed a human classifier to test on teaching sessions not included in either the training or testing sets. We will present and discuss the results of both the automated and human classifications in the subsequent sections.

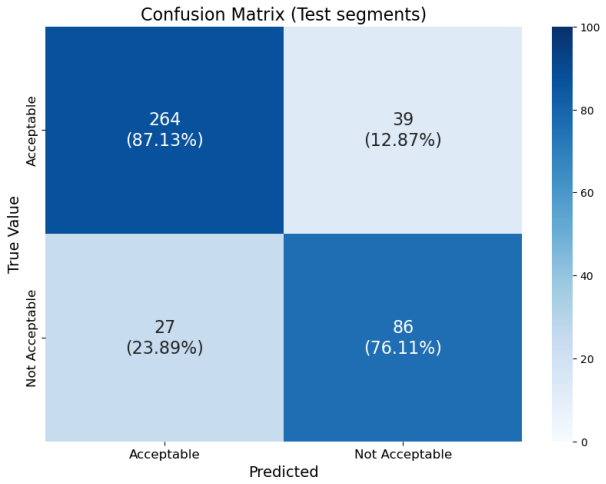


Figure 8. Confusion matrixs showcasing the model trained with automatic random split.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1690.v1

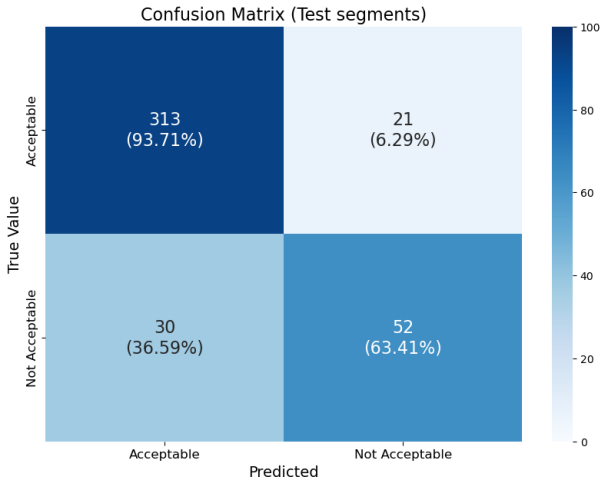


Figure 9. Confusion matrixs showcasing the model trained with manual split.

Metrics of the training results

The Area Under the Curve (AUC) is a metric that provides an aggregate measure of a model's performance across all possible classification thresholds. It is particularly useful for binary classification problems, as it evaluates the model's ability to distinguish between the "Acceptable" and "Non Acceptable" classes. In the context of the provided code, the AUC was calculated using the `roc_auc_score` function from the `sklearn.metrics` module.

To compute the AUC, the true labels and the probability scores for the "Non Acceptable" class were used. Specifically, the probability scores for the "Non Acceptable" class were extracted from the `y_pred_train` and `y_pred_test` tensors. These extracted scores, along with the true labels, were then passed to the `roc_auc_score` function to obtain the AUC values for both the training and test sets.

In the evaluation process of the classification model, several performance metrics were computed using the probability scores and the true labels. These probability scores, generated by the model, represent the likelihood of an instance belonging to a particular class. These scores were then converted into predicted classes using a threshold of 0.5. From the predicted classes and the true labels, True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) were computed for both the training and test datasets.

The performance metrics were defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The results of the model metrics in the test sample are shown in the Table 2.

Table 2. We take the metrics with respect to the "Not Acceptable" class from the randomly separated training set since the main purpose of the filter is to predict this class.

Datos	Precision	Accuracy	Recall	F1	AUC
train	0.703	0.823	0.751	0.726	0.835
test	0.631	0.782	0.646	0.639	0.810

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1690.v1

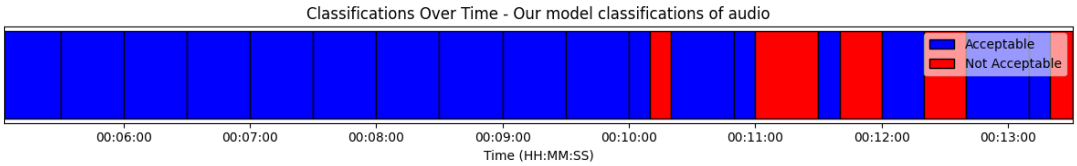
4. Results

To evaluate the model's effectiveness in enhancing transcription quality, three classes were transcribed using both our tool and a dummy model. Subsequently, a human evaluator assessed the transcription quality at various intervals throughout each class. We instructed the evaluator to label the transcription as "Good" if the automatic transcriber's output was accurate, "Regular" if there were minor errors, and "Bad" if the transcription bore little to no resemblance to the audio segment. Only transcribed segments underwent this classification.

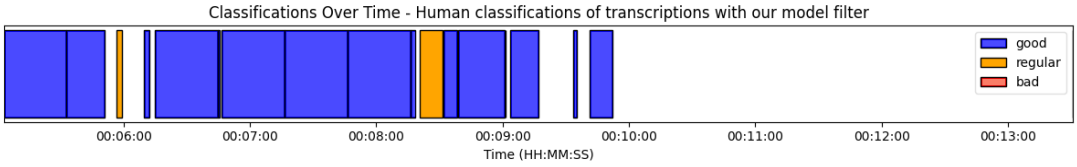
The human evaluator labeled the transcriptions at consistent time intervals for both models. The dummy model predicts every transcription as "acceptable", akin to using the Whisper transcriber without any filtering. In contrast, our model filters out the "Not acceptable" labels based on the trained neural network's predictions.

We tallied the number of "Good", "Regular", and "Bad" transcriptions, and the results are presented in the subsequent table.

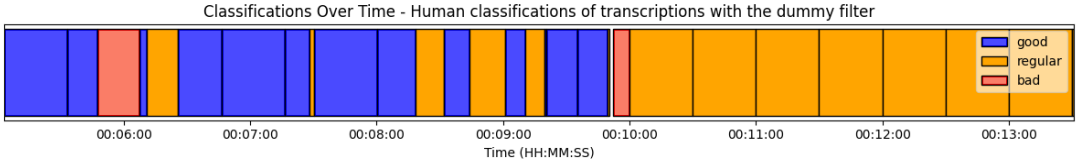
There are two primary reasons some audio segments lack transcription. Firstly, the Whisper transcriber sometimes determines that certain audio portions don't contain transcribable content. Secondly, our model's filter instructs the Whisper transcriber not to transcribe specific segments. The following image (Figure 10) illustrates how our model's filter impacts transcriptions and human labels. For comparison, we also provide an image showcasing the transcription without our filter, termed the "dummy model".



(a) Our Model Classifications. The "Not Acceptable" segments are those that our model decides to filter out.



(b) Human Classifications (Filter). The "Not Acceptable" segments were filtered out and therefore are not transcribed. There are some untranscribed portions that are given by the transcriber himself and not by our filter.



(c) Human Classifications (Dummy). No segment was filtered by our model. It can be seen that the filtered parts are generally rated worse in terms of their resulting transcripts.

Figure 10. Comparison of Classifications

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1690.v1

The following tables (Tables 3–5) shows the results of the human ratings on all audio segments of each teaching session comparing our filter with the dummy filter, the **Time** column displays the total sum of segment durations grouped according to the label assigned by the human classifier based on how well the transcription matched the audio. The *No transcription* row represents the sum of segments that were not labeled due to the absence of a transcription. The **Percentage of Transcription** column considers 100% as the sum of the times of segments that have an assigned transcription. Meanwhile, the **Total Percentage** column considers the entire audio duration, including those segments without an assigned transcription, as 100%.:

Table 3. Comparison between Our Model and the Dummy Model for Class 1 .

Teaching session 1			
Label	**Time(HH:MM:SS)**	**Percentage of Transcription**	**Total Percentage**
the dummy			
Good	00:04:08.23	51.61%	47.14%
Regular	00:01:12.71	15.12%	13.81%
Bad	00:02:40.04	33.27%	30.39%
No transcription	00:00:45.63		8.66%
Total	00:08:46.62		100.0%
our model			
Good	00:04:15.62	93.12%	48.54%
Regular	00:00:14.07	5.13%	2.67%
Bad	00:00:04.82	1.76%	0.92%
No transcription	00:04:12.12		47.87%
Total	00:08:46.63		100.0%

Table 4. Comparison between Our Model and the Dummy Model for Class 2 .

Teaching session 2			
Label	**Time(HH:MM:SS)**	**Percentage of Transcription**	**Total Percentage**
the dummy			
Good	00:07:07.85	54.34%	53.92%
Regular	00:05:31.97	42.16%	41.84%
Bad	00:00:27.55	3.5%	3.47%
No transcription	00:00:06.04		0.76%
Total	00:13:13.44		100.0%
our model			
Good	00:05:00.35	91.16%	37.85%
Regular	00:00:29.12	8.84%	3.67%
Bad	00:00:00.00	0.0%	0.0%
No transcription	00:07:43.95		58.47%
Total	00:13:13.43		100.0%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1690.v1

Table 5. Comparison between Our Model and the Dummy Model for Class 3 .

Teaching session 3			
Label	**Time(HH:MM:SS)**	**Percentage of Transcription**	**Total Percentage**
the dummy			
Good	00:13:16.33	91.16%	86.83%
Regular	00:00:25.01	2.86%	2.73%
Bad	00:00:52.17	5.97%	5.69%
No transcription	00:00:43.55		4.75%
Total	00:15:17.06		100.0%
our model			
Good	00:12:21.01	99.73%	80.8%
Regular	00:00:02.00	0.27%	0.22%
Bad	00:00:00.00	0.0%	0.0%
No transcription	00:02:54.06		18.98%
Total	00:15:17.08		100.0%

First and foremost, we aim to gauge whether our filter is inadvertently causing any "harm" - that is, filtering out segments that should rightfully be categorized as "Good". In the ideal scenario, the total percentage of "Good" segments from the filtered data should align closely with the "Good" percentage from the dummy model, which underwent no filtration. To ascertain this, we conducted a Z-test for each teaching session, and the outcomes of these tests are presented below (Table 6):

Table 6. The hypothesis is that the total percentage of the dummy model is equivalent to the total percentage from our model.

	(Z-tests)	
Lecture	Z-value	*p*
1	-0.15	0.5309
2	7.23	5.01e-13
3	3.64	2.71e-04

From the table, it's evident that for Teaching session 1, the difference between our model and the dummy model is not statistically significant, with a p-value of 0.5309 This implies that our filter performs comparably to the dummy model for this class, not omitting a notable number of "Good" segments. However, for Teaching session 2 and Teaching session 3, the differences are statistically significant with p-values of 5.01e-13 and 2.71e-04, respectively. This indicates that our model is likely excluding a significant number of segments that should have been categorized as "Good" for these lectures. This provides evidence that, at least for lectures 2 and 3, the filter in our model might be causing inadvertent "harm" by misclassifying or omitting segments that in the dummy model are identified by the human classifier as "Good."

Having previously assessed the level of "harm" our filter may have caused by potentially omitting segments that should have been categorized as "Good", we now shift our focus to measure its efficacy. Specifically, we aim to determine how well our filter managed to exclude segments that indeed should have been filtered out. Assuming each measurement as independent, we conducted a two-sided Z-test for proportions to individual categories, and we performed a chi-squared test for overall model comparison, yielding the following results (Table 7)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1690.v1

Table 7. Hypothesis test results. For the Z-tests, a positive value suggests that the dummy model has a higher percentage in that category, whereas a negative value indicates the opposite. A p value less than 0.05 indicates a statistically significant difference between the two models.

	(Z-tests Good)		(Z-tests Regular)		(Z-tests Bad)		(χ^2 Overall)
Lecture	Z-value	p	Z-value	p	Z-value	p	p
1	-6.5624	5.29e-11	2.3417	0.0192	5.8619	4.58e-09	1.55e-10
2	-5.8475	4.99e-09	5.4056	6.46e-08	1.8874	0.0591	3.09e-08
3	-2.9063	3.66e-03	1.4755	0.1401	2.4807	0.0131	0.0143

For Teaching session 1, the negative Z-value in the 'Good' category indicates that our model had a better performance, achieving a higher "Good" percentage than the dummy model. Conversely, the positive Z-values in both the 'Regular' and 'Bad' categories suggest that our model had a lower percentage in these categories, indicating a better performance compared to the dummy model.

For Teaching session 2, the negative Z-value in the 'Good' category indicates that our model had a better performance. The positive Z-values in both the 'Regular' and 'Bad' categories suggest that our model had a lower percentage in these categories, indicating a better performance compared to the dummy model. However, the 'Bad' category's p-value indicates that this difference is not statistically significant.

For Teaching session 3, the negative Z-value in the 'Good' category indicates that our model had a better performance. The positive Z-value in the 'Bad' category suggests that our model had a lower "Bad" percentage, indicating a better performance. However, the 'Regular' category did not show a statistically significant difference between the two models.

Overall, the chi-squared tests further support these findings, indicating a significant difference in performance between our model and the dummy model across all three lectures.

4.1. Suppression of hallucinations in transcripts

In the introductory chapter, we discussed the issue of hallucinations generated by automatic transcribers and how they could emerge in unexpected situations, contaminating any data obtained from the text. We employed our filter specifically to gauge the impact on hallucinations.

From the classroom recordings exposed in the results we proceeded to evaluate the number of hallucinations, measured in terms of words obtained from these same recordings transcribed with our filter. The results of this evaluation can be seen in Tables 8 and 9.

Table 8. Number of words representing automatic transcriber hallucinations in class sessions with different audio quality.

Lecture	Duration	Word Count (Segment)	Word Count (Hallucination)	Percentage
1	01:18:40	3668	119	3.24%
2	01:19:26	4309	3	0.07%
3	01:26:10	8989	38	0.42%

Table 9. Number of words representing automatic transcriber hallucinations in class sessions with different audio quality, post-application of our model's filter.

Lecture	Duration	Word Count (Segment)	Word Count (Hallucination)	Percentage
1	01:18:40	2175	22	1.01%
2	01:19:26	4281	0	0.00%
3	01:26:10	7191	0	0.00%

5. Discussion

The obtained results provide a detailed insight into the performance of our model compared to a dummy model across different scenarios. We will discuss these findings in detail below.

 doi:10.20944/preprints202310.1690.v1

5.1. General Analysis of the Results

Upon examining the three classes, it's evident that each presents distinct characteristics in terms of transcription quality. In Teaching session 1, the audio quality is at extremes, with a significant proportion of both "Good" and "Bad" transcriptions, and a smaller proportion of "Regular". Teaching session 2, on the other hand, displays a high proportion of "Good" and "Regular" transcriptions, with almost none being "Bad". Finally, Teaching session 3 stands out for having exceptionally high audio quality, with a vast majority of "Good" transcriptions and very low proportions of "Regular" and "Bad". These differences, confirmed qualitatively by a human classifier, allow us to delve deep into the behavior of our model across different contexts.

5.2. Damage Control

One of the primary goals in developing our filter was to ensure it didn't cause "harm" by filtering out segments that should be transcribed correctly. The Z-test results indicate that, in general, our filter tends to perform better in classes where there's a clear contrast between "Good" and "Bad" transcriptions. However, in classes like Teaching session 2, where there's a mix of "Regular" and "Good" transcriptions, the filter tends to make more mistakes in filtering out segments that could have been transcribed correctly.

This suggests that the model struggles to distinguish between "Good" and "Regular" transcriptions in scenarios where the errors are minor and more subtle. In such cases, the model tends to be more conservative and filters out segments that might have been transcribed correctly.

5.3. Benefits of Using the Filter

Despite the aforementioned limitations, it's evident that our filter offers significant benefits in terms of improving the overall quality of transcriptions. In classes where there's a high proportion of "Bad" transcriptions, like Teaching session 1 and Teaching session 2, the filter is capable of significantly improving the proportion of "Good" transcriptions. However, in classes like Teaching session 3, where there are very few "Bad" segments to filter out, the benefit is less pronounced.

It's crucial to consider how the Whisper transcriber works and how the segments were labeled. Given that Whisper uses 30-second chunks for transcription and can base itself on the text from the previous chunk, erroneous transcriptions in one chunk can negatively impact subsequent transcriptions. This cumulative effect can have a significant impact on the overall quality of transcriptions.[10]

6. Conclusions

In addressing our first research question regarding the enhancement of information acquisition by eliminating interferences and noises, we have found that it is indeed feasible to improve transcription quality in elementary school classrooms. Importantly, this can be achieved without a significant budget. Our study demonstrates that modern technology, specifically neural network-based filters applied to small audio samples (3-5 hours), can effectively enhance the quality of noisy transcriptions. This is particularly significant as it negates the need for high-end sound equipment and does not disrupt the normal operation of a classroom. The use of an affordable UHF microphone transmitting to a smartphone located 8 meters away has proven to be adequate for this purpose.

Addressing our second research question about ensuring valuable information is not inadvertently filtered out, we recognize that one of the primary challenges of our model is the non-negligible number of potential good-quality transcriptions that are lost. This is due to the misprediction of good-quality audio being labeled as poor quality.

Regarding areas of improvement, our filter model performs commendably but could benefit from an expanded dataset. This suggests that the extent to which we can ensure valuable information is not inadvertently filtered out could be further optimized by increasing the variety and quantity of data used for training.

One of the biggest challenges of the model is the non-negligible number of potential good-quality transcriptions that are lost due to the poor prediction of good-quality audio being labeled as poor quality.

In the future, we anticipate improved versions of this model, which will not only allow us to enhance the transcriptions but also perform other related tasks, such as improving speaker diarization, for research purposes like detecting the activity of teachers and/or students in the classroom.

Additionally, there's a growing interest in applying our filter to other automatic transcription systems specifically tailored for classroom environments, such as the one developed by [21]. This transcriber has shown to be particularly suited for classroom settings, demonstrating robustness against ambient sounds and the specific discourse style of teaching. By enhancing the quality of transcriptions produced by such systems, we could facilitate more accurate analyses of teacher discourse. This, in turn, would bolster efforts to classify and understand teaching practices, like those detailed in [20], which leverage acoustic features to identify pedagogical activities in the classroom as per protocols like COPUS [22] , a protocol that's uniquely designed to depend minimally (if at all) on the classifier's subjective interpretations of ongoing classroom activities. As such, it holds significant potential as a protocol for automated classification. In summary, merging efficient filtering technologies with specialized automatic transcriptions could revolutionize the way we analyze and comprehend classroom dynamics, providing potent tools for both researchers and educators alike.

Author Contributions: "Conceptualization, R.A. and J.H.; methodology, R.A. and J.H..; software, J.H.; validation, R.A. and J.H.; formal analysis, J.H..; investigation, R.A. and J.H..; resources, R.A..; data curation, J.H..; writing—original draft preparation, J.H..; writing—review and editing, R.A. and J.H..; visualization, J.H..; supervision, R.A..; project administration, R.A..; funding acquisition, R.A. All authors have read and agreed to the published version of the manuscript."

Funding: This work was supported by the Chilean National Agency for Research and Development (ANID), grant number ANID/PIA/Basal Funds for Centers of Excellence FB0003.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Comité de Ética de la Investigación en Ciencias Sociales y Humanidades, of the Universidad de Chile (date of approval 11-23-2015).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data from the study can be shared with interested parties by personal communication with the corresponding author.

Acknowledgments: We acknowledge support from the Chilean National Agency for Research and Development (ANID), grant number ANID/PIA/Basal Funds for Centers of Excellence FB0003. In addition, we acknowledge technical support to Paulina Jaure and Carlos Aguirre.

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

Sample Availability: Samples can be shared with interested parties by personal communication with the corresponding author.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1690.v1

Appendix A

Appendix A.1. List of parameters and their correlations

Table A1. Feature Descriptions

Feature	Brief Description	Detailed Description
mean	Mean of the absolute magnitude.	The mean of the absolute value of all samples in the audio segment was computed using `np.abs(segment_samples).mean()`.
std	Standard deviation of the absolute magnitude.	The standard deviation of the absolute value of the samples was computed using `np.abs(segment_samples).std()`.
pitch_mean	Mean of pitch frequencies.	Pitch frequencies were detected in the Mel spectrogram and their mean was obtained with `np.mean(pitches)`.
pitch_std	Standard deviation of pitch frequencies.	The standard deviation of the detected pitch frequencies was computed with `np.std(pitches)`.
pitch_confidence	Confidence in pitch detection.	The mean of the magnitudes associated with pitch frequencies was computed using `np.mean(magnitudes)`.

Table A2. Feature Descriptions

spectral_centroid	"Gravity" center of the spectrum.	`librosa.feature.spectral_centroid` was used to compute the spectral centroid of the signal, giving a measure of the perceived "brightness".
spectral_bandwidth	Spectrum width.	`librosa.feature.spectral_bandwidth` was used to obtain the width of the frequency in which most of the energy is concentrated.
spectral_rolloff	Spectrum cutoff frequency.	`librosa.feature.spectral_rolloff` was used to determine the frequency below which a specified percentage of the total energy is located.
zero_crossing_rate	Zero crossing rate.	`librosa.feature.zero_crossing_rate` was used to compute how many times the signal changes sign within a period.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 October 2023 doi:10.20944/preprints202310.1690.v1

Table A3. Feature Descriptions

mfcc_x	Mel frequency cepstral coefficients for component x.	This was computed using `librosa.feature.mfcc`. The 'x' index refers to the x-th component of the MFCC coefficients. It describes the shape of the power spectrum of an audio signal.
spectral_contrast_x	Amplitude difference between peaks and valleys of the spectrum for band x.	`librosa.feature.spectral_contrast` was used to measure the amplitude difference between peaks and valleys in the spectrum. The 'x' index indicates that it is the average spectral contrast for band x.
quantile_x	Deciles of the absolute magnitude.	Deciles of the absolute magnitude of audio samples were computed using `np.quantile(np.abs(segment_samples), q)`, for various q values.

Table A4. Sorted correlations in absolut value of included parameters

Feature	Value	Status
spectral_contrast_6	0.397998	Included
mfcc_1	0.381041	Included
std	0.380255	Included
mfcc_3	0.373912	Included
mfcc_10	0.350538	Included
mfcc_8	0.336277	Included
quantile90	0.324888	Included
mean	0.274777	Included
quantile80	0.271496	Included
mfcc_0	0.271207	Included
spectral_contrast_3	0.267938	Included
mfcc_9	0.237584	Included
spectral_bandwidth	0.233203	Included
quantile70	0.224501	Included
mfcc_7	0.203561	Included
spectral_contrast_4	0.197853	Included
quantile60	0.183169	Included
mfcc_4	0.169382	Included
mfcc_5	0.164113	Included
pitch_mean	0.159703	Included
quantile50	0.146121	Included
pitch_std	0.128296	Included
spectral_contrast_5	0.122807	Included
mfcc_6	0.117541	Included
mfcc_12	0.113921	Included
quantile40	0.107676	Included

 doi:10.20944/preprints202310.1690.v1

Table A5. Sorted correlations in absolut value of not included parameters

Feature	Value	Status
quantile30	0.067156	Not Included
mfcc_11	0.061327	Not Included
mfcc_2	0.044224	Not Included
zero_crossing_rate	0.040466	Not Included
pitch_confidence	0.035399	Not Included
quantile20	0.028609	Not Included
Segmento	0.015662	Not Included
time	0.015662	Not Included
spectral_contrast_2	0.009942	Not Included
spectral_centroid	0.006812	Not Included
quantile10	0.002807	Not Included
spectral_rolloff	0.001632	Not Included
spectral_contrast_0	0	Not Included
spectral_contrast_1	0	Not Included

References

1. Li, H., Wang, Z., Tang, J., Ding, W., & Liu, Z. *Siamese Neural Networks for Class Activity Detection*; In: Bittencourt, I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds) *Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science*, vol 12164. Springer, Cham, 2020; DOI: 10.1007/978-3-030-52240-7_3.
2. Li, H., Kang, Y., Ding, W., Yang, S., Yang, S., Huang, G. Y., & Liu, Z. *Multimodal Learning for Classroom Activity Detection*; In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9234-9238, 2020; DOI: 10.1109/ICASSP40776.2020.9054407.
3. Cosbey, R., Wusterbarth, A., Hutchinson, B. Deep Learning for Classroom Activity Detection from Audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2019; pp. 3727–3731. doi:10.1109/ICASSP.2019.8683365.
4. Zhang, X.-L., Wu, J. Denoising Deep Neural Networks Based Voice Activity Detection. In *arXiv preprint arXiv:1303.0663*; 2013. https://arxiv.org/pdf/1303.0663.pdf.
5. Thomas, S., Ganapathy, S., Saon, G., Soltau, H. Analyzing Convolutional Neural Networks for Speech Activity Detection in Mismatched Acoustic Conditions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: Piscataway, NJ, USA, 2014; pp. 2519–2523. https://api.semanticscholar.org/CorpusID:1646846.
6. Ma, Y., Wiggins, J.B., Celepkolu, M., Boyer, K.E., Lynch, C., Wiebe, E. The Challenge of Noisy Classrooms: Speaker Detection During Elementary Students' Collaborative Dialogue. *Lecture Notes in Artificial Intelligence* **2021**, *12748*, 268–281.
7. Kingma, D. P., & Ba, J. *Adam: A Method for Stochastic Optimization*. 2017; arXiv preprint arXiv:1412.6980. https://arxiv.org/abs/1412.6980.
8. Abreu Araujo, F., Riou, M., Torrejon, J., et al. *Role of non-linear data processing on speech recognition task in the framework of reservoir computing*. Sci Rep 10, 328 (2020); DOI: https://doi.org/10.1038/s41598-019-56991-x.
9. Dong, H.-Y. *Speech intelligibility improvement in noisy reverberant environments based on speech enhancement and inverse filtering*; EURASIP Journal on Audio, Speech, and Music Processing, 2018.
10. Radford, A., Kim, J., Xu, T., Brockman, G., McLeavey, C., Sutskever, I. *Robust Speech Recognition via Large-Scale Weak Supervision*; 2022.
11. Sabri, M., Alirezaie, J., Krishnan, S. Audio noise detection using hidden Markov model. In *Statistical Signal Processing, 2003 IEEE Workshop on*; IEEE: 2003; pp. 637–640.
12. Rangachari, S., Loizou, P.C. A noise estimation algorithm with rapid adaptation for highly nonstationary environments. In *Speech Communication*; Elsevier: Amsterdam, Netherlands, 2006; pp. 220–231.
13. Çakır, E.; Heittola, T.; Huttunen, H.; Virtanen, T. Polyphonic sound event detection using multi label deep neural networks. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*; IEEE: Killarney, Ireland, 2015; pp. 1–7. doi:10.1109/IJCNN.2015.7280624.
14. Dinkel, H., Wang, S., Xu, X., Wu, M., Yu, K. Voice Activity Detection in the Wild: A Data-Driven Approach Using Teacher-Student Training. In *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE*

PROCESSING; IEEE: Piscataway, NJ, USA, 2021; Vol. 29. https://myw19.github.io/AboutPageAssets/papers/hedi7-dinkel-taslp2021-2.pdf.

15. Rashmi, S.; Hanumanthappa, M.; Gopala, B. Training Based Noise Removal Technique for a Speech-to-Text Representation Model. In *Journal of Physics: Conference Series*, Volume 1142, Second National Conference on Computational Intelligence (NCCI 2018); IOP Publishing: Bangalore, India, 2018; pp. 1–2, doi:10.1088/1742-6596/1142/1/012019.
16. Kartik, P., & Jeyakumar, G. *A Deep Learning Based System to Predict the Noise (Disturbance) in Audio Files*; In *Advances in Parallel Computing*, Volume 37; ISBN 9781643681023, 2020; DOI: 10.3233/APC200135.
17. NVIDIA Corporation. *Matrix Multiplication Background User's Guide*; NVIDIA Docs; © 2020-2023 NVIDIA Corporation & affiliates. All rights reserved. Last updated on Feb 1, 2023. Available at: https://docs.nvidia.com/deeplearning/performance/pdf/Matrix-Multiplication-Background-User-Guide.pdf.
18. Heaton, J. *Artificial Intelligence for Humans: Deep Learning and Neural Networks*, Volume 3 of Artificial Intelligence for Humans Series; Heaton Research, Incorporated., 2015; ISBN 1505714346.
19. Adil, M., Ullah, R., Noor, S. et al. *Effect of number of neurons and layers in an artificial neural network for generalized concrete mix design*. Neural Comput & Applic, Volume 34, 8355–8363, 2022; DOI: https://doi.org/10.1007/s00521-020-05305-8.
20. Schlotterbeck, D., Uribe, P., Araya, R., Jimenez, A., Caballero, D. *What Classroom Audio Tells About Teaching: A Cost-effective Approach for Detection of Teaching Practices Using Spectral Audio Features*. Pages 132-140, 2021; DOI: https://doi.org/10.1145/3448139.3448152.
21. Schlotterbeck, D., Jiménez, A., Araya, R., Caballero, D., Uribe, P., & Van der Molen Moris, J. *"Teacher, Can You Say It Again?" Improving Automatic Speech Recognition Performance over Classroom Environments with Limited Data*. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Volume 13355 LNCS, Pages 269-280, 23rd International Conference on Artificial Intelligence in Education, AIED 2022; Durham; United Kingdom; 27 July 2022 through 31 July 2022.
22. Smith, M. K., Jones, F. H. M., Gilbert, S. L., & Wieman, C. E. *The Classroom Observation Protocol for Undergraduate STEM (COPUS): A New Instrument to Characterize University STEM Classroom Practices*. Published Online: 13 Oct 2017; DOI: https://doi.org/10.1187/cbe.13-08-0154.