The effects on speech detection of low sample frequency audio data

Taichi Uno
Supervisor(s): Hayley Hung, Jose Vargas Quiros
EEMCS, Delft University of Technology, The Netherlands

June 19, 2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

The interactions between human and machines are now common in our daily life. The audio data of human communication is a rich source of information, but it is considered privacy-invasive for machines to listen to it. By reducing sampling frequency, it is possible to preserve privacy by making conversation unclear while still being possible to detect if someone is speaking or not. The topic of this paper is to investigate how low sampled frequency audio data hinders the detection of speech. To detect speaking, voice activity detection has been applied, which is a technology in the signal processing field that identifies which short segments of audio contain speakings. Two types of state-of-art voice activity detector(VAD) were used for this experiment including a supervised (pyannote) and two unsupervised (rVAD pitch and flatness mode) methods. As a result, the unsupervised methods outperformed the supervised model, where rVAD pitch mode has resulted in the best performance out of all three. More specifically, the unsupervised VAD's performance became lower as the sample rates decreased while the supervised VAD did not work well at higher sample frequency. rVAD pitch mode at sample rates of 8000Hz or higher was possible to perform at the almost same level as a state-of-art supervised VAD that is trained in a similar data set. Furthermore, it was able to perform as well as a modern unsupervised VAD at 2000Hz or higher sample frequencies. At the sample rate of 1250Hz or lower, any VAD was not able to perform at the same level as a state-of-art VAD. Regarding the privacy aspect, it is observed that human ears detect speaking better than computers, where humans can understand parts or all of the contents of speaking at 2000Hz or higher, which infers that current technology is not enough to detect speech from downsampled privacy-preserving audio. However, there is still a need for further research to verify the effects of the training set and its sample frequencies for the supervised method and also proper scientific social experiments to test the ability of humans of speech detection for reduced sampled audio.

1 Introduction

In our daily life, the interactions between machines and humans are getting more normal and important. One way of such interactions is based on audio data of conversations and many state-of-art technology processes this communication in an automatic way using different social audio signals. However, it is considered privacy-invasive for machines to listen to and record what people are saying. Luckily, verbal conversation is not the only important element. In fact, it is known that non-verbal social signals play important role in communication [1]. Therefore, it is important to investigate what and how we can approach sensing non-verbal social signals from privacy-sensitive audio data.

An example of privacy-sensitive data collection is ConfLab[2]. This is a social experimental event organized by the Socially Perceptive Computing Lab at Delft University of Technology. This event is aimed to collect a variety of data, including acceleration, proximity, video and audio. Its audio data is recorded at a low sample frequency of 1250Hz so that it is enough to detect if someone is speaking but the contents are kept private. They claimed that the contents are not recognizable by just listening to the audio but any scientific justification were not given.

The reason why low sampled audio has such property of preserving privacy is that the sampling frequency needs to be higher than a certain threshold value called Nyquist frequency, otherwise, it is not possible to reproduce the original audio signal and causes aliasing[3]. The ConfLab audio data is all sampled at 1250Hz and thus aliasing is occuring. This is because the maximum frequency of human hearing is at around 20kHz[4] thus it at least needs 40kHz of sampling rate to be able to reconstruct the original continuous wave according to the Shannon Sampling Theorem[3].

In order to detect the speaking by machines, the easiest approach is to use Voice Activity Detector (VAD). There are many applications of VAD because the information of whether someone is talking or not can be useful in many different scenes. For example, iPhone from Apple has functionalities of triggering Apple's personal assistant "Siri" with a voice command. It keeps running a VAD at the backend to detect if someone is saying "Hey Siri" [5]. Because it is a very important topic in the fields of social signal or audio processing, there is a lot of research done to increase the efficiency and performance of voice activity detection using different approaches including a supervised and an unsupervised model.

Even though researchers looked into speech detection techniques and there is a widely known theory about sample frequency, there is not much research done in the past about the topic of speech detection of low sample frequency audio data. Therefore, this research aims to investigate how the reduction in sample frequency hinders the detection of speaking. It is such an important topic to research now as society has strict eyes on privacy concerns while social signal processing is now used on daily basis. By making this question clear, it is possible to know at which the lowest sample frequency should be used to detect speech automatically and what types of VADs should be used. As one of the essential aspects is privacy, this research also looks into privacy concerns, not just technical perspectives. More specifically, the difference between human and machines detecting speech are also discussed.

The hypothesis to the research question is that it gets harder and harder as the sample rates are reduced because it gets more difficult to reconstruct the original sound wave because of the loss of waves' information at lower sample frequencies. In addition, the model of VAD will affect the results of speaking detection as different implementations focus on different parts of sound waves to detect speech. Lastly, it can be expected that an unsupervised model outperforms a pre-trained supervised model since the training set is taken at normal sample rates, which can make it harder to adapt to an unusual situation like downsampled audio.

The structure of this paper is as followings. After this introductory chapter, the background information about this topic is provided in chapter 2 and then there will be a description of what methods are used to test the hypothesis in chapter 3. In chapter 4, the detailed experiment setups and results are shown. Next, the reflection of the ethical perspectives of this research is discussed in chapter 5. Chapter 6 provides the discussion of the results from both technical and privacy aspects and finally chapter 7 concludes this research.

2 Background

2.1 Overview of Voice Activity Detectors (VAD)

Voice Activity Detection, also known as speech detection, is a technology in the audio processing field that detects if someone is talking or not in a short segment of audio data. It is widely used in practical applications, such as speech enhancement, transcription, estimating signal-to-noise ratio, or speaker diarization[6]. In general, VAD takes digitized audio as an input and first extracts particular features from the processed input. Then, the extracted features will be inputted into a model that represents them in noise and speech. Finally, based on a threshold of the model, it returns the final binary outcome of speech or not. Fourier coefficients, periodicity and zero-crossing rates are often used as features but it depends on each implementation. Likewise, various models are used including statistical ones like Gaussian and Laplacian distribution, or other heuristics methods[7].

2.2 Wearable badges for social signal collections

Before the Socially Perceptive Computing Lab at Delft University of Technology carries out Conflab events, a group of researchers at MIT has invented a wearable badge called "Rhythm"[8]. It measures three kinds of signals in social interactions, namely vocal activity, proximity to other badges and location. To collect vocal activity, it samples the sound signal using a microphone at 700Hz sample rates, averages its amplitude and stores it in a chunk every 50ms. By doing so, it is not possible to reconstruct the original audio, meaning it preserves the speaker's privacy. This data will be processed to a multi-step voice activity detection algorithm, which detects speaking purely based on whether the signal has higher energy than a threshold over a fixed-sized window. It is a good example of an application of speech detection and downsampled privacy-sensitive audio data. However, since speech detection is only dependent on the signal energy, its performance might not be as high as a state-of-art VAD. Also, the explanation and scientific evidence of why they used 700Hz sample rates were not given.

3 Methodology

3.1 Voice Activity Detectors (VADs)

In order to best answer the research question of "how does the reduction of in sample frequency hinders speech detection completely", we will use two types of VAD and measure how the performance of speaking detection changes by reducing sampling frequencies of audio data. Here, we have decided to use a supervised VAD called "pyannote" and an unsupervised VAD called "rVAD". These two are chosen because both of them are available online and free to use, which makes it easier to reproduce the experiments. Also, they have completely different implementations, which makes it possible to see how the implementations and models affect the performance of speech detection. Finally, both VADs represent the performance of state-of-art speech detection technology as claimed in their papers[9, 10].

As briefly mentioned just now, these two VADs have completely different approaches and methods for speech detection. The main difference between these two VADs is that pyannote is supervised whereas rVAD is unsupervised. In general, a supervised model usually works well under similar circumstances as it has been trained on, but it cannot possibly function well in unexpected conditions[9]. Since many other research papers focused on the performance of VADs in a setting where the tested and training audio data are collected at the normal regular sampling frequency, it is possible to also look into whether it is suitable to use such a supervised model for the speech detection at the lower sampling frequencies in comparison to unsupervised model.

Firstly, pyannote provides a variety of multimedia processing tools including a supervised voice activity detector(VAD) [11]. It implements a sequence labelling task of classifying a given array of feature vectors to an array of labels, where $y_t = 0$ indicates no speech while $y_t = 1$ as speech with help of Recurrent Neural Network (RNN), where it is used for calculating the probability of speech by making use of low-level acoustic features and also for a fusion and decision making [10, 12]. It has been trained on a data set called AMI Meeting Corpus[13]. It consists of 100hours of meeting recordings sampled at 16kHz and includes both close and far talking to the microphone. Since it is collected during meetings, there can be a couple of people talking simultaneously or some small noises, but background music or noises of crowds is not present.

Next, rVAD is proposed by Zheng Hua Tan in 2019[9]. While most VADs out there are supervised, he has implemented an unsupervised one that is robust to both stationary and burst-like noise. The rVAD's pitch segment detection algorithm has two modes: pitch and flatness mode. For the pitch mode, its pipeline consists of two denoising filters and detects speech or non-speech based on a posteriori signal-to-noise ratio(SNR) weighted energy difference. There are two roles of this first filter, which are not to estimate noise too much because of the burst-like noise when a noise estimator of the second filter is applied and to detect and remove noises of non-speech segments with high energy. The second denoising filter contributes to speech enhancement based on three different noise estimation approaches. On the other hand, the flatness mode relies on spectral flatness and uses the SFT feature to extract pitch. The difference between these two methods is that the flatness mode has much less computational complexity but is slightly less accurate according to the research[9].

3.2 Dataset

The audio data used for the experiments, named March15LaRedBirthdayParty, were collected at a social experiment event similar to ConfLab by the Socially Perceptive Computing Lab at Delft University of Technology. Its audio contains mainly the following components: silence, talking in Dutch, talking in English, and talking with noises in the background. It consists of combinations of these elements with some noises of other people chatting or music in the background.

3.3 Data preparation

For the experiments, we want to measure the performance of two VADs using audio data which are sampled at different rates. However, by reducing the sampling frequency, the audio data will lose some information. Instead of resampling the audio at different sampling frequencies, we decided to use a low pass filter for practical convenience. Low pass filters remove the high-frequency audio and remain the lower one. This way, it is possible to have the same effect on audio as downsampling but the number of samples per second remains the same. According to Shannon Sampling Theorem, the sampling frequency has to be greater than twice the maximum frequency[3]. It means that a 10kHz low pass filter needs to be applied, for example, if we want to simulate a downsampling of 20kHz.

3.4 Metrics

In order to evaluate the performance of VADs on different sampling frequencies, false alarm rates (FAR), false rejection rates (FRR) and false error rates (FER) will be calculated using the following definition:

$$FAR = \frac{Total\ number\ of\ nonspeech\ frames\ mislabeled\ as\ speech}{Total\ number\ of\ nonspeech\ frames}$$

$$FRR = \frac{Total\ number\ of\ speech\ frames\ mislabeled\ as\ nonspeech}{Total\ number\ of\ speech\ frames}$$

$$FER = \frac{Total\ number\ of\ mislabeled\ frames}{Total\ number\ of\ frames}$$

FAR, FRR and FER all indicate error rates so the lower the values are, the better they are. They are chosen because some other research papers have used them to evaluate the state-of-art VADs so it is convenient to make a comparison between them. The performance is calculated by comparing it against the ground truth data. Here, we assumed that the result of rVAD pitch mode using 44.1 kHz sampled audio is correct and thus treat as ground truth. In fact, by manually inspecting the result and the original audio, it seems the performance is fairly high with some minor mistakes. These mistakes are usually when someone is talking at a small volume or loud background noise is present. Therefore, it can be predicted that the metrics will result in a better score, meaning lower error rates, as the ground truth will have some mistakes which are likely to be in favour of VADs.

4 Experimental Setup and Results

4.1 Experiment steps

The first step is to process audio. The original audio is shortened to 1-3 hours from 4 hours long by removing the silence parts at the beginning and the end manually. Then, the audio data is downsampled to different frequencies using low pass filters as explained in 3.3. The values are the followings; 300, 350, 500, 800, 1250, 2000, 3150, 5000, 8000, 12000, 20000, 30000, 44100Hz. These values are chosen in such a way that the steps are logarithmic and include 1250Hz and 800Hz. That completes the data process and it will be passed to the two VADs. For rVAD, both modes of pitch and flatness are used.

rVAD returns a list of 1 (speech) and 0 (non-speech) per frame, where the frameshift is 10ms[9] by default. To process results data to be able to compare, the outputs of pyannote are also converted into the same format as it originally returns a list of tuples containing the start time and the end time when a speech is detected.

4.2 Experiment results

The results on table 1 represent average FAR, FFR and FER using 12 different 1-3 hours long audio at different sample frequencies. Figure 1, 2 and 3 visualize how those rates changed over different sample frequencies. Individual results of all 12 audio data are shown in table 2 in Appendix A.

5 Responsible Research

It is important to be reproducible and ethical to ensure scientific integrity so this chapter provides information about reproducibility and ethical aspect of this research.

The reproducibility of this research is assured by systematic use of libraries, such as rVAD and pyannote, where both of them are publicly available online. The experiment method can easily be repeated by following the steps as in 4.1 because it consists of a simple low pass filtering process and python codes for pyannote and Matlab commands for rVAD to execute the operations.

Next, about the ethical aspect, since March15LaRedBirthdayParty audio data contained conversations of people at social events, and it has to be treated with extra attention to the speakers' privacy. Therefore, our research group had to sign the End User License Agreement before having access to the database. It states that the data can only be used for academic

Table 1: Performance of rVAD (unsupervised methods (pitch and flatness mode)) and pyannote (a supervised method) over different sample frequencies

	Unsupervised						Supervised		
	rVAD pitch			rVAD flatness			pyannote		
Sample Frequency (Hz)	FAR (%)	FRR (%)	FER (%)	FAR (%)	FRR (%)	FER (%)	FAR (%)	FRR (%)	FER (%)
300	0.10	99.69	34.91	8.26	94.97	38.34	0.10	99.83	34.90
350	0.17	99.14	34.73	7.85	94.76	38.06	0.34	97.38	33.74
500	0.47	90.51	30.63	7.80	87.07	33.95	2.62	71.70	25.42
800	2.48	69.64	22.62	9.43	67.73	26.11	18.49	19.69	18.64
1250	3.42	42.29	13.68	10.41	40.98	17.24	29.64	14.38	24.54
2000	2.87	21.66	7.48	10.24	22.00	11.65	26.62	15.30	22.87
3150	2.38	12.18	4.65	10.41	15.12	10.05	35.60	11.45	26.38
5000	1.98	6.28	2.89	10.77	10.75	9.25	53.65	6.81	36.16
8000	0.91	2.93	1.39	10.95	9.44	9.06	58.88	6.12	39.38
12000	0.66	2.32	1.05	10.96	8.93	8.94	61.06	6.00	40.50
20000	0.51	1.81	0.81	10.96	8.47	8.83	67.22	4.95	44.30
30000	0.42	1.68	0.70	10.77	8.31	8.64	67.18	4.99	44.31
44100	0.00	0.00	0.00	11.13	8.58	8.94	68.69	4.72	45.34

Figure 1: FAR (False Alarm rates) over different sample frequencies

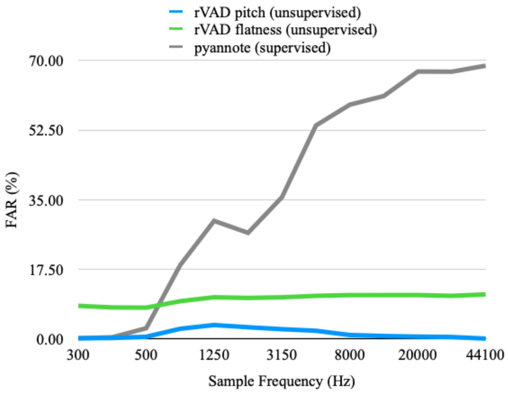
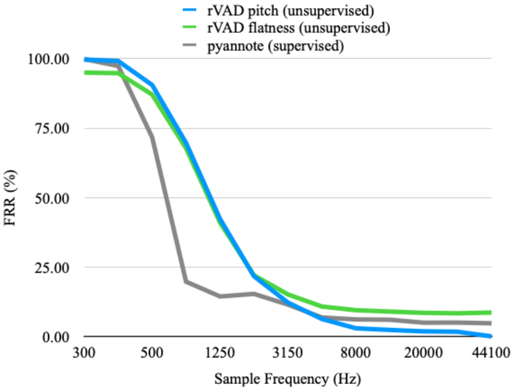
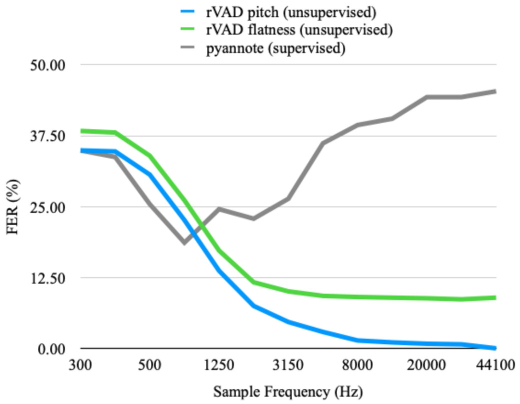


Figure 2: FRR (False Rejection rates) over different sample frequencies



non-governmental research with non-commercial purposes. By signing, we had to agree to several conditions including that the data set should not be distributed to the third person and should not be used to identify persons. Additionally, the data should not be uploaded online as there are risks of data leakage, therefore all experiments are run locally on our personal computers. The results are only showing the performance of VADs in terms of metrics as mentioned in 3.4 so there is no risk of data exposure. By obeying the agreement and also being mindful of the sensitivity of the data, our research group has avoided unethical use of resources.

Figure 3: FER (False Error rates) over different sample frequencies



6 Discussion

6.1 Discussion of the results

As the figures in 4.2 show, there is a noticeable difference in performance between two rVAD modes and Pyannote. For rVAD methods, they produced significantly smaller FAR in comparison to pyannote at rates of 800Hz or higher. In addition, FAR of pyannote kept increasing ranging from 0.10% to 68.69%, while rVAD flatness and pitch mode remained almost steady and only a small fluctuation of around ± 2%. In other words, pyannote tends to misclassify the non-speech segments as speech as sample frequencies increases, while it does not change much of rVAD methods as shown in figure 1. In contrast to FAR, FRR over different sample frequencies for rVAD and Pyannote had almost similar trends of decreasing as the sample frequencies are lower but pyannote had a huge drop between 500 and 800Hz whereas both rVAD modes had it between 500 and 2000Hz, which can be seen graphically on 2.

FER is useful to compare the overall performance of VADs as it represents the total error rate. As shown in figure 3, pyannote performed the best under 800Hz, but its FER increased as sample frequencies increased from 800Hz or higher. On the other hand, both rVAD modes resulted in decreasing trend as sample frequency increased. However, the pitch mode outperformed the flatness mode, which was as expected as in what has been claimed in 3.1.

Given the fact that the false alarm rate is 3.5% and missed detection, which is equivalent to the false rejection rate, is 2.7%[10], it can be said that rVAD pitch mode with 8000Hz or higher sample frequency works at a similar or higher level of supervised state-of-art speech detection model. Also, rVAD pitch mode works at clean background noise with 6.90% of FER[9], it can be also said that it is possible to detect speaking at 2000Hz sample rates or higher at the almost similar level of the unsupervised speech detection model. In other words, with the audio sampled at the lowest of 1250Hz or lower sample frequencies, the detection of speaking is hindered using a state-of-art voice activity detector. It implies that the ConfLab audio of 1250Hz sample rates is not enough to detect speaking.

As it has been discussed above and shown in table 1, pyannote has resulted in very poor results compared to what has been shown in its research paper[10]. The main reason might be that the training set was too different from the experimental data set. The training set was recorded at meetings so the background noises like music in our test set might have affected the performance. Given the fact that pyannote has been trained on 16000Hz sampled audio but the performance at that sampling frequency did not stand out, it seems that the sampling frequency of training data does not have much effect on the final performance, but it has to be experimentally verified to conclude that.

In summary, the unsupervised methods outperformed a supervised model. More specifically, the unsupervised VAD's performance became lower as the sample rates decrease while the supervised VAD did not work well at higher sample frequency as well. This is most likely because of the mismatch between the supervised one's training set and the test audio. These results are the same as in the hypothesis as in 1. However, such a huge gap between supervised and unsupervised methods was unexpected.

There are some aspects that this experiment could not cover. Firstly, the influence of training set in the supervised method on the performance is unclear yet. Even though pyannote resulted in a completely different trend for FRR and FER, the information from the results and the experiment is not enough to fully explain the causes. To clarify that, an experiment can be performed with a training set from similar data as the test set. At the same time, the sample rates of the training set can be changed so that it is possible to find out the effect of the sample frequency of the training set on performance. Also, it is worth studying with other VADs with different models to be able to generalize more and find the best approach for low sampled data.

6.2 Difference between a human and a computer

As explained in chapter 1, having computers listening to what people are saying is considered privacy-invasive so it is important to discuss whether there is a gap between humans and machines in terms of speech detection.

For human ears, it is possible to easily detect if someone is speaking or not until 800Hz sample rates. This is why the ConfLab website claimed that 1250Hz sampled audio is privacy-preserving but is enough to detect speech. It is partially possible to detect the content of the speech at 2000Hz but it becomes impossible for lower frequencies. For all frequencies higher than 2000Hz, it is possible to understand what people are saying to a large extent even if they are speaking in the background. However, from this experiment, we found that at least a 2000Hz sample rate is necessary to detect speech by a computer. This implies that the technology of the current state-of-art VAD seems not up to the level of being used with privacy-preserving downsampled audio, which is most likely collected under 2000Hz.

As a limitation of this discussion, this observation is not a result of proper scientific or social experiments, so it is very subjective and it is not possible to generalize to the entire population. Therefore, to examine this topic scientifically, human experiments with a sufficient number of people need to be done to properly test speech detection abilities of low sample frequency.

7 Conclusion

In this research, how the low sampled audio data hinders the detection of speech was made clear. Two approaches of unsupervised (rVAD pitch and flatness mode) and another supervised method (pyannote) were used for the experiments. As a result, the unsupervised methods outperformed the supervised model, where rVAD pitch mode has resulted in the best performance out of all three. More specifically, the unsupervised VAD's performance became lower as the sample rates decreased while the supervised VAD did not work well at higher sample frequency. rVAD pitch mode at sample rates of 8000Hz or higher was possible to perform at the almost same level as a state-of-art supervised VAD that is trained in a similar data set. Furthermore, it was able to perform as well as a modern unsupervised VAD at 2000Hz or higher sample frequencies. At the sample rate of 1250Hz or lower, any VAD was not able to perform at the same level as a state-of-art VAD. This implies that it is not possible to detect speech from ConfLab audio which is sampled at 1250Hz. There is still room for further research to verify the effects of the training set and its sample frequencies to make the reasons clear why the supervised method ended up with such unexpected results.

Additionally, human ears detect speaking better than computers, where humans can understand parts or all of the contents of speaking at 2000Hz or higher, which implies that current technology is not up to the level of detecting speech with computers by using downsampled privacy-preserving audio. However, this is just an observation from listening to the audio so scientific social experiments to test the ability of humans of speech detection for reduced sampled audio is worth investigating.

Appendix A Results of individual audio data

Table 2: Results (FAR, FRR, FER) of rVAD and pyannote for all individual audio files over different sample frequencies

		Unsupervised						**Supervised**		
Audio File		**rVAD pitch mode**			**rVAD flatness mode**			**Pyannote**		
Sample rates (Hz)	**File index**	**FAR (Hz)**	**FRR (Hz)**	**FER (Hz)**	**FAR (Hz)**	**FRR (Hz)**	**FER (Hz)**	**FAR (Hz)**	**FRR (Hz)**	**FER (Hz)**
300	1	0.01	99.99	46.76	11.17	94.26	50.03	0.00	99.99	46.76
300	2	0.05	99.87	51.15	10.57	93.80	53.18	0.10	99.29	50.88
300	3	0.02	99.95	45.00	10.04	94.15	47.89	0.14	99.99	45.07
300	4	0.01	99.98	52.71	7.82	95.45	54.02	0.20	99.30	52.45
300	5	0.11	99.57	31.78	9.34	96.48	37.08	0.07	99.53	31.74
300	6	0.02	99.94	37.41	10.64	94.92	42.17	0.00	100.00	37.42
300	7	0.00	99.99	33.48	4.05	92.88	33.79	0.00	100.00	33.48
300	8	0.37	99.35	25.91	7.75	95.36	30.36	0.18	99.89	25.91
300	9	0.03	99.97	16.75	5.25	95.71	20.39	0.00	100.00	16.73
300	10	0.08	99.87	36.27	10.04	96.17	41.27	0.00	100.00	36.27
300	11	0.16	99.12	30.29	5.84	96.44	33.42	0.38	100.00	30.71
300	12	0.29	98.65	11.38	6.59	94.05	16.45	0.09	99.98	11.36
350	1	0.01	99.88	46.71	10.80	94.69	50.03	0.19	97.65	45.76
350	2	0.18	97.07	49.78	10.44	92.82	52.61	1.12	84.24	43.67
350	3	0.10	99.74	44.94	9.74	94.53	47.90	0.02	98.95	44.54
350	4	0.02	99.81	52.63	8.26	95.26	54.13	2.68	88.21	47.77
350	5	0.21	98.58	31.53	6.73	97.00	35.47	0.00	100.00	31.84
350	6	0.08	99.46	37.26	10.38	95.03	42.06	0.00	100.00	37.42

350	7	0.04	99.86	33.46	4.88	91.35	33.83	0.00	100.00	33.48
350	8	0.63	98.58	25.91	5.32	96.31	28.81	0.00	100.00	25.81
350	9	0.16	99.77	16.83	8.20	93.10	22.41	0.00	100.00	16.73
350	10	0.20	99.02	36.04	7.66	96.93	40.04	0.00	100.00	36.27
350	11	0.17	99.12	30.30	5.31	96.79	33.16	0.09	99.47	30.34
350	12	0.20	98.81	11.32	6.49	93.27	16.28	0.00	100.00	11.28
500	1	0.49	90.72	42.68	8.70	92.96	48.10	1.75	72.47	34.82
500	2	1.91	22.45	12.42	11.06	6.76	8.86	7.59	41.52	24.96
500	3	0.38	94.81	42.88	9.85	94.54	47.97	2.75	65.40	30.94
500	4	0.31	91.25	48.25	11.52	85.40	50.47	2.36	69.08	37.54
500	5	0.20	98.42	31.47	7.28	96.99	35.84	2.87	69.49	24.08
500	6	0.24	98.27	36.92	7.30	97.03	40.88	1.13	71.59	27.49
500	7	0.36	98.43	33.19	6.61	92.81	35.47	1.05	82.15	28.20
500	8	0.46	98.36	25.72	5.39	96.38	28.87	4.09	80.47	23.80
500	9	0.50	98.61	16.91	7.42	94.02	21.91	1.77	91.15	16.73
500	10	0.23	97.75	35.59	5.91	96.74	38.85	2.62	73.41	28.29
500	11	0.30	98.08	30.07	5.53	97.04	33.39	2.60	48.63	16.62
500	12	0.32	98.95	11.44	7.01	94.17	16.85	0.92	95.04	11.53
800	1	4.51	24.03	13.64	13.73	18.13	15.78	14.27	17.75	15.90
800	2	4.74	8.73	6.78	13.37	2.32	7.71	30.36	7.12	18.46
800	3	3.61	29.13	15.10	12.04	23.91	17.38	17.32	15.85	16.66
800	4	3.98	41.60	23.81	14.59	35.70	25.72	15.33	31.48	23.84
800	5	1.68	96.57	31.89	7.89	96.98	36.26	15.94	15.10	15.68
800	6	1.87	91.89	35.55	8.81	94.26	40.78	16.12	9.88	13.78
800	7	1.60	91.94	31.85	7.85	89.99	35.35	7.47	19.45	11.48
800	8	1.79	95.87	26.07	6.19	95.89	29.34	36.98	14.65	31.22
800	9	1.65	96.40	17.50	7.96	94.70	22.47	23.40	40.90	26.33
800	10	1.45	94.76	35.29	7.35	94.81	39.07	10.64	13.58	11.71
800	11	1.46	68.31	21.81	6.45	73.74	26.93	3.67	18.13	8.07
800	12	1.46	96.41	12.17	6.88	92.37	16.52	30.34	32.39	30.57
1250	1	6.92	11.91	9.25	17.16	7.20	12.50	23.17	15.75	19.70
1250	2	4.21	5.94	5.09	11.78	1.62	6.58	38.77	5.86	21.92
1250	3	4.91	14.68	9.31	13.12	8.58	11.08	31.88	11.73	22.81
1250	4	7.48	18.22	13.14	19.33	12.95	15.96	21.00	28.70	25.06
1250	5	1.91	60.57	20.59	8.57	66.64	27.06	28.63	10.27	22.78
1250	6	2.08	22.53	9.73	7.66	17.40	11.30	36.27	6.70	25.21
1250	7	2.12	54.62	19.70	9.32	55.14	24.66	14.40	14.38	14.39
1250	8	2.22	75.04	21.01	6.85	81.99	26.24	46.66	10.82	37.41
1250	9	2.58	92.85	17.69	8.23	92.76	22.38	39.61	20.14	36.35
1250	10	2.57	51.97	20.48	8.12	51.71	23.93	25.51	8.03	19.17
1250	11	2.04	19.76	7.44	7.94	14.40	9.91	10.10	15.36	11.70
1250	12	2.05	79.41	10.78	6.86	81.40	15.27	39.69	24.79	38.01
2000	1	3.92	7.12	5.42	15.22	4.65	10.28	19.22	15.73	17.59
2000	2	2.49	2.85	2.67	11.13	1.36	6.13	39.72	5.94	22.43
2000	3	3.14	8.23	5.43	12.47	4.77	9.01	26.19	13.57	20.51
2000	4	6.51	8.78	7.71	19.32	5.54	12.05	22.63	28.61	25.78

2000	5	2.10	22.69	8.65	9.39	25.37	14.48	26.23	10.48	21.22
2000	6	2.10	8.41	4.46	6.35	3.62	5.33	29.14	8.01	21.23
2000	7	2.28	18.98	7.87	9.91	13.48	11.10	10.38	17.43	12.74
2000	8	2.66	39.50	12.17	7.78	48.94	18.40	42.59	11.99	34.69
2000	9	1.95	64.81	12.47	7.74	74.11	18.84	36.88	23.32	34.61
2000	10	2.85	25.58	11.10	8.88	25.46	14.89	19.62	9.41	15.92
2000	11	2.51	12.35	5.51	8.07	7.07	7.76	7.71	16.52	10.39
2000	12	1.91	40.61	6.28	6.66	49.61	11.51	39.14	22.61	37.28
3150	1	2.88	5.52	4.11	13.26	4.51	9.17	38.03	8.59	24.26
3150	2	2.75	1.58	2.15	11.26	1.27	6.14	50.53	4.53	26.98
3150	3	2.39	5.07	3.60	12.52	4.10	8.73	46.37	8.12	29.15
3150	4	4.17	4.73	4.47	20.23	3.39	11.35	41.80	13.35	26.80
3150	5	2.41	10.58	5.02	10.49	14.05	11.62	33.70	8.44	25.65
3150	6	1.20	4.65	2.49	6.44	3.00	5.15	44.58	5.54	29.97
3150	7	2.10	7.98	4.07	10.44	8.10	9.66	14.32	15.08	14.57
3150	8	2.40	23.91	7.95	7.90	33.38	14.48	45.19	10.00	36.10
3150	9	1.85	40.41	8.30	7.78	53.30	15.40	41.93	18.23	37.97
3150	10	2.53	11.74	5.87	9.41	13.44	10.87	24.40	8.58	18.67
3150	11	1.73	6.45	3.17	7.95	6.20	7.42	8.77	14.31	10.46
3150	12	2.18	23.49	4.58	7.25	36.70	10.58	37.62	22.69	35.94
5000	1	1.47	2.96	2.17	12.77	4.25	8.79	64.89	3.71	36.28
5000	2	2.12	1.07	1.58	11.46	1.20	6.20	59.59	3.13	30.69
5000	3	2.01	2.50	2.23	13.58	3.55	9.06	76.15	3.51	43.46
5000	4	3.20	2.66	2.91	20.71	2.51	11.12	61.64	4.97	31.77
5000	5	2.36	5.82	3.46	11.41	11.00	11.28	55.28	4.29	39.05
5000	6	1.06	2.15	1.47	6.78	2.41	5.15	59.50	3.88	38.69
5000	7	2.02	4.57	2.88	10.75	6.38	9.29	33.28	11.12	25.86
5000	8	1.71	9.53	3.73	8.43	21.66	11.85	60.82	5.51	46.54
5000	9	1.52	18.80	4.41	7.73	34.60	12.22	55.32	9.21	47.61
5000	10	2.76	7.28	4.40	9.65	6.22	8.40	49.46	4.52	33.16
5000	11	1.97	3.90	2.56	8.28	5.51	7.44	22.92	9.77	18.91
5000	12	1.53	14.10	2.94	7.70	29.78	10.19	44.89	18.09	41.87
8000	1	0.61	1.93	1.23	13.48	3.85	8.98	67.39	4.00	37.75
8000	2	0.97	0.65	0.81	11.53	1.17	6.23	63.99	3.18	32.86
8000	3	0.63	1.13	0.85	13.95	3.39	9.20	81.49	2.80	46.07
8000	4	1.26	1.55	1.41	20.65	2.15	10.90	64.28	4.41	32.72
8000	5	0.95	2.87	1.56	11.52	9.97	11.02	61.76	3.56	43.23
8000	6	0.57	1.24	0.82	6.82	2.30	5.13	61.87	4.33	40.34
8000	7	1.18	2.87	1.75	10.88	5.35	9.03	37.23	10.01	28.12
8000	8	1.12	5.23	2.18	8.74	18.15	11.17	67.00	4.55	50.88
8000	9	0.89	6.90	1.89	8.04	26.36	11.11	58.62	8.83	50.29
8000	10	0.83	2.06	1.28	9.99	5.89	8.50	65.15	2.96	42.60
8000	11	0.87	1.70	1.12	8.19	5.26	7.29	29.79	8.25	23.23
8000	12	1.06	6.97	1.73	7.66	29.42	10.11	48.03	16.53	44.47
12000	1	0.39	0.80	0.58	13.74	3.61	9.01	72.77	2.62	39.97
12000	2	0.20	0.37	0.29	11.55	1.18	6.24	68.80	2.65	34.94

12000	3	0.39	0.66	0.51	14.07	3.24	9.20	86.79	1.56	48.43
12000	4	0.71	1.03	0.88	20.82	1.97	10.88	69.86	3.30	34.77
12000	5	0.92	3.73	1.82	11.65	9.89	11.09	62.12	3.75	43.54
12000	6	0.37	1.01	0.61	6.80	2.24	5.09	63.61	3.71	41.20
12000	7	1.13	2.38	1.55	10.70	5.12	8.83	35.84	11.14	27.57
12000	8	0.73	4.15	1.61	8.57	17.45	10.86	66.21	5.07	50.43
12000	9	0.71	4.36	1.32	7.90	24.21	10.63	59.88	9.55	51.46
12000	10	0.49	1.24	0.76	10.21	5.56	8.52	65.32	3.35	42.84
12000	11	0.77	1.43	0.97	8.13	5.30	7.27	32.18	9.16	25.17
12000	12	1.08	6.69	1.71	7.35	27.39	9.61	49.40	16.08	45.64
20000	1	0.21	0.37	0.28	13.88	3.37	8.96	77.63	2.51	42.50
20000	2	0.13	0.28	0.21	11.51	1.20	6.23	73.85	2.09	37.12
20000	3	0.16	0.27	0.21	14.10	3.21	9.20	89.70	1.54	50.02
20000	4	0.47	0.45	0.46	21.05	1.88	10.94	75.42	2.65	37.06
20000	5	1.15	3.37	1.85	11.59	9.40	10.90	69.41	2.72	48.18
20000	6	0.27	0.73	0.45	6.87	2.23	5.14	68.26	3.31	43.95
20000	7	1.00	2.06	1.35	10.73	5.10	8.84	42.34	8.97	31.16
20000	8	0.49	3.18	1.19	8.76	16.23	10.69	71.31	4.64	54.11
20000	9	0.61	3.59	1.11	7.72	23.95	10.44	64.73	7.40	55.14
20000	10	0.27	0.93	0.51	10.34	5.19	8.47	76.62	2.80	49.85
20000	11	0.61	1.19	0.79	7.99	5.20	7.14	40.47	8.24	30.66
20000	12	0.81	5.36	1.32	6.99	24.74	8.99	56.92	12.48	51.91
30000	1	0.14	0.23	0.18	13.86	3.29	8.92	77.96	2.38	42.62
30000	2	0.07	0.24	0.16	11.48	1.20	6.22	73.26	2.28	36.92
30000	3	0.12	0.21	0.16	14.08	3.24	9.20	89.54	1.59	49.96
30000	4	0.24	0.33	0.29	20.98	1.88	10.91	75.36	2.69	37.05
30000	5	0.99	3.24	1.71	11.49	8.98	10.69	69.50	2.66	48.22
30000	6	0.29	0.59	0.40	6.81	2.23	5.10	68.08	3.31	43.85
30000	7	0.66	1.75	1.02	10.49	5.11	8.69	42.00	9.16	31.01
30000	8	0.42	3.03	1.09	8.48	15.87	10.39	71.06	4.64	53.92
30000	9	0.52	3.39	1.00	7.51	23.04	10.11	65.26	7.43	55.59
30000	10	0.19	0.89	0.45	10.16	5.17	8.35	76.92	2.79	50.04
30000	11	0.57	1.12	0.73	7.39	4.84	6.61	40.02	8.11	30.31
30000	12	0.78	5.11	1.27	6.46	24.86	8.53	57.22	12.79	52.21
44100	1	0.00	0.00	0.00	13.99	3.23	8.96	77.74	2.45	42.53
44100	2	0.00	0.00	0.00	11.51	1.19	6.23	74.08	2.20	37.28
44100	3	0.00	0.00	0.00	14.16	3.33	9.28	90.15	1.42	50.22
44100	4	0.00	0.00	0.00	21.34	1.87	11.08	75.54	2.63	37.10
44100	5	0.00	0.00	0.00	11.68	9.52	10.99	71.30	2.71	49.46
44100	6	0.00	0.00	0.00	7.12	2.14	5.25	69.84	3.18	44.90
44100	7	0.00	0.00	0.00	10.78	5.22	8.92	43.27	8.73	31.71
44100	8	0.00	0.00	0.00	8.81	14.29	10.22	72.71	4.49	55.10
44100	9	0.00	0.00	0.00	7.85	23.49	10.47	66.48	6.87	56.51
44100	10	0.00	0.00	0.00	10.51	5.58	8.73	77.78	2.72	50.56
44100	11	0.00	0.00	0.00	8.39	5.39	7.47	44.19	7.65	33.07
44100	12	0.00	0.00	0.00	7.42	27.76	9.71	61.21	11.55	55.61

References

[1] P. Deepika, "The importance of non-verbal communication," *IUP Journal of Soft Skills; Hyderabad*, vol. 9, no. 4, pp. 43–49, 2015. [Online]. Available: https://www-proquest-com.tudelft.idm.oclc.org/scholarly-journals/importance-non-verbal-communication/docview/1759007009/se-2?accountid=27026

[2] D. U. o. T. The Socially Perceptive Computing Lab. (2019) Conflab - acm mm 2019. [Online]. Available: https://conflab.ewi.tudelft.nl

[3] R. E.Isufi, "Cse2220 signal processing sampling iir filters," 2020.

[4] W. contributors. (2022) Hearing range. [Online]. Available: https://en.wikipedia.org/wiki/Hearing_range

[5] S. Team, "Hey siri: An on-device dnn-powered voice trigger for appleâs personal assistant," *research area Speech and Natural Language Processing*, 2017. [Online]. Available: https://machinelearning.apple.com/research/hey-siri

[6] S. S. Meduri and R. Ananth, "A survey and evaluation of voice activity detection algorithms," 2012.

[7] J. Kola, C. Espy-Wilson, and T. Pruthi, "Voice activity detection," *Merit Bien*, pp. 1–6, 2011.

[8] O. Lederman, A. Mohan, D. Calacci, and A. S. Pentland, "Rhythm: A unified measurement platform for human organizations," *IEEE MultiMedia*, vol. 25, no. 1, pp. 26–38, 2018.

[9] Z.-H. Tan, A. kr. Sarkar, and N. Dehak, "rvad: An unsupervised segment-based robust voice activity detection method," *Computer Speech Language*, vol. 59, pp. 1–21, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230819300920

[10] H. B. et al, "pyannote.audio: neural building blocks for speaker diarization," 2019. [Online]. Available: https://arxiv.org/abs/1911.01255

[11] H. Bredin. (2022) pyannote. [Online]. Available: https://pyannote.github.io

[12] G. Gelly and J.-L. Gauvain, "Optimization of rnn-based speech activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 646–656, 2018.

[13] H. B. et al. (2019) ami. datasets at hugging face. [Online]. Available: https://huggingface.co/datasets/ami#citation-information