SPEECH OBFUSCATION IN MEL SPECTRA THAT ALLOWS FOR CENTRALISED ANNOTATION AND CLASSIFICATION OF SOUND EVENTS

Michiel Jacobs[1,2,3], *Lode Vuegen*[1,2,3], *Suraj Khan*[1,2,3], *Peter Karsmakers*[1,2,3*]

[1] KU Leuven, Dept. of Computer Science, Kleinhoefstraat 4, B-2440 Geel, Belgium
[2] Flanders Make @ KU Leuven
[3] Leuven.AI - KU Leuven Institute for AI
Corresponding authors: {michiel.jacobs, peter.karsmakers}@kuleuven.be

ABSTRACT

Nowadays, computerised Sound Event Classification (SEC) aids in several applications, e.g. monitoring domestic events in smart homes. SEC model development typically requires data collected from a diverse set of remote locations. However, this data could disclose sensitive information about uttered speech that might have been present during the acquisition.
In this work, three data preprocessing techniques are investigated that obstruct recognising semantics in speech, but retain the required information in the data for annotating sound events and SEC model development. At the remote location, the data are first preprocessed before transferring to a central place. At the central location, speech should not be interpretable anymore, while still having the opportunity to annotate data with relevant sound event labels. For this purpose, starting from a log-mel representation of the sound signals, three speech obfuscation techniques are assessed: 1) calculating a moving average of the log-mel spectra, 2) sampling a few of the most energetic log-mel spectra and 3) shredding the log-mel spectra. Both intelligibility and SEC experiments were carried out. All considered techniques proved effective in obfuscating speech, while still allowing SEC. For stationary sound events, calculating the moving average of the log-mel spectra is recommended, as well as shredding the log-mel spectra. For impulsive sound events, sampling a few of the most energetic log-mel spectra is recommended.

Index Terms— Speech obfuscation, sound event classification, log-mel spectra

1. INTRODUCTION

Since the past couple of years, there is a growing trend in utilising sound to monitor certain processes. Sound monitoring allows to continuously perceive environments in an automated manner using machine or deep learning (resp. ML, DL) models, e.g. to predict when machine components are about to fail [1], to classify home activities [2] or to automatically interpret auscultation sounds for disease monitoring [3]. Next to the signals of interest, the microphones used are prone to picking up speech signals as well. These speech signals could potentially contain private and/or sensitive information. Therefore, speech obfuscating techniques have to be applied to the recorded sound.

In this work, three preprocessing techniques are evaluated that make the speech present in the sound data unintelligible. Each of these techniques can be run on an extreme edge device with minimal computational overhead, prior to communicating the sound data to the central location. This way, speech is obfuscated while still retaining enough information to centrally perform the tasks of Sound Event Classification (SEC). Moreover, although no words can be recognised, the data still allow human annotators to add class labels to the restored sound data at the central location for refining the ML models.

The main contributions of this work are the following: a) study the effects of three different preprocessing techniques for speech obfuscation, b) perform speech and event intelligibility experiments to study the effect of these speech obfuscation techniques on sound intelligibilty (speech should be obfuscated, while sound events should still be identifiable), and c) perform domestic SEC experiments to study the effect of these speech obfuscation techniques on SEC model classification performance.

The remainder of this paper is organised as follows: Section 2 discusses the related work. Section 3 explains the obfuscation techniques studied. Next, Section 4 describes the intelligibility and classification experiments. Section 5 then covers the results of the intelligibility and the classification experiments. Next, Section 6 discusses the results and findings. Finally, Section 7 summarises this work and forms the conclusion.

2. RELATED WORK

Kumar, Nguyen, Zeng et al. [4] presented subsampling and sound shredding techniques, both applied on mel-frequency cepstral coefficients (MFCC). With subsampling, MFCC feature vectors get thrown out of the sequence in a nonrandom manner. When applying subsampling, the event duration plays an important factor. The time duration of the shortest event (class) has to exceed the subsampling period, since otherwise this event can completely get lost when the corresponding frames are thrown away. Therefore, subsampling MFCCs at regular time intervals is considered not relevant in this research, since it could remove events with short duration. With sound shredding, blocks of MFCCs (units) get randomly shuffled inside of a so-called snippet. The authors indicated that subsampling and sound shredding are valid speech obfuscation techniques that still allow context, gender and speaker recognition using both a k-nearest neighbour (kNN) and a support vector machine (SVM).

When opting to interpret the sound events centrally, another possibility would be to add a speech filter to the edge device. This way, detected speech is simply not sent to the central server. One example would be a so-called Voice Activity Detector (VAD). The drawback of such a VAD is that when an event of interest overlaps a speech signal, it will be discarded and the event gets lost [5]. Also,

*Funding provided by Flanders Innovation & Entrepreneurship Agency (VLAIO) and Flanders Make.

the use of such speech filters is more complex in terms of computation as compared to the techniques considered in this work.

The work of Larson, Lee, Liu et al. [6] focussed on detecting cough, while disguising speech. The authors proved that ten principal components suffice for classifying the cough sounds. When the number of principal components was increased to 25, the quality of the coughing sound was good and 84% of the spoken words was concealed. However, in the context described in this work such preprocessing is likely to hinder the post-hoc labelling of any audio events that are present in the data.

Chen, Adcock and Krishnagiri [7] used a methodology that identified the vocalic regions using a vocalic syllable detector and replaced the local vocalic linear predictive coefficients (LPC) with those of pre-recorded vowels. Speech intelligibility experiments showed that this methodology can reduce the word recognition rate to 7%. Furthermore, Liaqat, Nemati, Rahman et al. [5] applied this methodology to detect coughs. The mean classification accuracy of the raw audio was 75.86%, while the mean classification accuracy of the filtered audio was 75.75%. The t-test p-value equalled 0.985, thus showing no significant difference between the raw and filtered classification accuracies.

3. METHODOLOGY

This section starts by discussing the calculation of the log-mel spectra, which are the features most commonly used for sound event classification using DL techniques. Next, it discusses which speech obfuscation techniques were applied in the experiments: 1) calculating the moving average of the log-mel spectra, 2) sampling the most energetic log-mel spectra and 3) shredding the log-mel spectra. Finally, this section outlines the steps required to transform the processed log-mel spectra back into a time-domain sound signal (sound restoration).

3.1. Feature extraction: log-mel spectra

The most popular choice for acoustical features in combination with DL are the so-called log-mel spectra [8]. To calculate the short-time Fourier transform (STFT), the following parameters were used: 32 milliseconds (ms) window length, 16 ms hop length (50% overlap), and Hamming window. Finally, the STFT frames were converted into mel frames using a 64-dimensional mel filterbank and the logarithm was taken.

3.2. Speech obfuscation techniques

When averaging consecutive log-mel spectra, nonstationary speech signals get diffused when they are represented by an aggregated log-mel spectrum spanning a larger time horizon. The larger the number of frames being averaged, the more difficult it becomes to restore the original speech afterwards. As can be seen in Table 1, three moving average (MA) configurations were tested. First, *MA-light* refers to averaging over a sliding window having a length of 4 frames (80 ms) and a step of 3 frames (48 ms). Second, *MA-medium* refers to averaging over a sliding window having a length of 8 frames (144 ms) and a step of 5 frames (80 ms). Third, *MA-heavy* refers to averaging over a sliding window having a length of 12 frames (208 ms) and a step of 7 frames (112 ms).

When sampling the most energetic windows, a sliding window is moved over the mel frequency domain audio signal within a nonoverlapping larger segment. At each position, the energy (sum of squares inside smaller, overlapping segment) is calculated and only those segments having highest energy are retained. As can be seen in Table 1, again three configurations were tested. First, with *ENERGY-light* a block of 56 contiguous (912 ms) frames is taken and is replaced by the block of 18 contiguous (304 ms) frames having the highest energy. Second, with *ENERGY-medium* a block of 112 contiguous frames (1,808 ms) is taken and is replaced by the block of 18 contiguous frames having the highest energy. Third, with *ENERGY-heavy* a block of 168 contiguous frames (2,704 ms) is taken and is replaced by the block of 18 contiguous frames having the highest energy.

With sound shredding, log-mel frames get randomly shuffled inside of a so-called snippet, which is a region of contiguous log-mel spectra wherein sound shredding is applied [4]. Two parameters have to be defined, i.e. the unit size refers to the number of frames that are seen as a whole (a block of contiguous, adjoining frames), and the number of units inside one snippet. Again, three configurations were tested and can be found in Table 1. First, *SHRED-light* uses a snippet size of 3 units (208 ms). Second, *SHRED-medium* uses a snippet size of 6 units (400 ms). Third, *SHRED-heavy* uses a snippet size of 16 units (1,040 ms). In all three configurations the unit size was kept at 4 contiguous frames (80 ms), since this was required by the convolutional kernels of the DL-based SEC models in the automated classification experiments. One limitation of sound shredding is that all sound information is kept, i.e. given sufficient effort an attacker could still rearrange the units in the correct order again.

3.3. Sound restoration

In order to assess the speech and event intelligibility of the obfuscated sound features, these features have to be restored to the time-domain signal. Recall that prior to obfuscation the sound data were transformed to log-mel spectra. To return to the time domain, the log, mel and STFT operations have to be reversed. The logarithm can be perfectly undone without introducing artefacts. The inversion from mel frequency scale back to regular frequency scale can be achieved by equation (1):

$$|\hat{X}[k]|^2 = \sum_{b=0}^{B-1} M_{bk}^{\dagger} m_b \approx |X[k]|^2 \tag{1}$$

where $M^{\dagger}$ is the Moore-Penrose pseudo-inverse of the mel matrix M, $|X[k]|^2$ the magnitude spectrogram and m_b the mel value of bin b. This equation guarantees that $|\hat{X}[k]|^2$ is the best solution with minimum norm [9]. This inversion of mel frequency might introduce minor artefacts. The inverse STFT operation was performed using NumPy's `numpy.fft.irfft` function and corresponding phase information [10].

4. EXPERIMENTS

4.1. Speech intelligibility experiment

The speech intelligibility experiment aimed to evaluate the level of obfuscation by having participants grade the restored audio. The data used in this experiment were derived from the Mozilla Common Voice Dutch Subset (v10.0) [11]. A subset of 27 sound files was taken from the Dutch (NL) dataset. Messages of varying length (6 to 9 words) and of both male and female speakers were included.

Twelve native Dutch-speaking participants each got one of three sets of 27 sound recordings obfuscated with varying techniques and configurations. The participants had to grade their obfuscated recordings on an ordinal scale from one to three. Herein,

Table 1: The abbreviations used in the experiments, alongside a brief description of the corresponding configuration. "Light" always refers to the least obfuscating technique, while "heavy" refers to the most obfuscating configuration.

Group	Abbreviation	Description
	baseline	32 ms STFT window size and 16 ms STFT step size, 64 log-mel bins.
Moving average of log-mel spectra	MA-light	Moving average over 4 log-mel frames, step size of 3 log-mel frames.
	MA-medium	Moving average over 8 log-mel frames, step size of 5 log-mel frames.
	MA-heavy	Moving average over 12 log-mel frames, step size of 7 log-mel frames.
Sampling log-mel spectra	ENERGY-light	For each block of 56 frames, apply a sliding window with length 18 frames and hop size 1 frame and retain the 18 frames having highest energy.
	ENERGY-medium	For each block of 112 frames, apply a sliding window with length 18 frames and hop size 1 frame and retain the 18 frames having highest energy.
	ENERGY-heavy	For each block of 168 frames, apply a sliding window with length 18 frames and hop size 1 frame and retain the 18 frames having highest energy.
Shredding log-mel spectra	SHRED-light	Sound shredding with unit size: 4 log-mel frames, snippet size: 3 units.
	SHRED-medium	Sound shredding with unit size: 4 log-mel frames, snippet size: 6 units.
	SHRED-heavy	Sound shredding with unit size: 4 log-mel frames, snippet size: 16 units.

a score of '1' represented sound that is completely incomprehensible, a score of '2' represented sound that had a portion of the words comprehensible, while a score of '3' referred to perfectly understandable audio. The mean intelligibility score then represented the mean grade for each preprocessing obfuscation method.

Next to the ordinal score, each participant had to write down the message he/she understood. By comparing the understood message and the true transcription, an objective measure of the obfuscated sound quality could be made. In case a participant noted the sound recording as '2' but none of the words in the message he/she understood were correct, then the score was altered afterwards to '1'. In case the participant assessed the recording as '3' but the sentence understood was different or incomplete, then the score was changed to '2'. If the assigned score equalled '3' and the understood message differed by only a single word as compared to the transcription and the meaning of the sentence did not become very different, only then the score was kept as '3'.

In our speech intelligibility experiment, phase information was not discarded when calculating the STFT to simulate the best reconstruction possible (worst-case scenario from the point of speech intelligibility). Therefore, this information could be used during reconstruction. The baseline had the same transformation and reconstruction applied.

4.2. Sound event intelligibility experiment

The sound event intelligibility experiment is similar to the speech intelligibility experiment and differs only in the type of sound to label, i.e. the same participants had to recognise varying domestic sound events in 18 obfuscated recordings. The event classes are summarised in Table 2 and originate from the same dataset as used in the classification experiment (Section 4.3). The labelled events were graded in a binary true/false manner, i.e. a correct label received score '1', while an incorrect label received '0'. The participants did not receive any prior knowledge about the recording procedure (e.g. microphone location) that could help them.

In our event intelligibility experiment, the phase information was discarded after calculating the STFT and was replaced by a random Gaussian noise phase (worst-case scenario from the point of event intelligibility). The baseline had the same transformation and reconstruction applied.

4.3. Sound event classification experiment

For the domestic event classification task, the data and classifier model from Vuegen and Karsmakers [12] were used. The considered dataset contains domestic sound events collected from 72 home environments. In total, data for eight different domestic sound events are available. The recordings were made using a sampling frequency equal to 32 kHz and each sample had a 16-bit resolution. In total, 47.7 hours of data were recorded, spread out over 1519 recordings. Table 2 gives an overview of the dataset distribution.

Table 2: Overview of the dataset used in the SEC task and the event intelligibility experiment. "*Background*" refers to silence and sounds that do not belong to any of the other classes. [12].

Class	Hours	Recordings
Background	10.5	205
Door & window	5.3	141
Faucet & shower	9.3	386
Footstep	4.2	220
Kitchen hood	4.0	140
Speech	4.9	217
Toilet	5.5	136
Radio & television	4.0	74

As a classifier model a convolutional neural network (CNN) is used. Its performance is evaluated in a 4-fold cross-validation scheme using the previously discussed log-mel features. The model consists of three convolutional layers having 32 filters and ReLU activation (no pooling), followed by one fully-connected layer having 64 neurons with ReLU activation and finally one fully-connected output layer of 8 neurons with softmax activation (8 classes). The dimensions of the convolutional kernel were 4×4, with a stride of 1×4. Note that in case sound shredding is used as an speech obfuscation technique, the horizontal stride of the convolutional kernel was modified to have a value of 4. As such the kernel always spanned a single shredding unit (4 log-mel frames). This way, a kernel never covered a mix of two neighbouring shredded units which are expected to have an unnatural transient from one unit to the other. Zero padding was added to keep the correct dimensions.

The input dimension of the CNN models can be found in Table 3 and was set to one second for the baselines of MA and SHRED techniques. For the ENERGY techniques, the input of the baseline CNN was thrice the segment length.

5. RESULTS

Two types of experiments were carried out to test both the comprehensibility of speech and events, and the SEC model performance on the obfuscated log-mel spectra. The first set of experiments tried to assess the level of speech obfuscation through intelligibility experiments, while the second experiment assessed SEC performance. Table 1 lists the abbreviations used, alongside a brief description of each of the nine tested configurations.

5.1. Speech intelligibility experiment

The results of the speech intelligibility experiment can be found in Table 4a. As was mentioned in Section 4.1, a mean intelligibility

score equal to 1 represents a perfect obfuscation, while a mean intelligibility score equal to 3 represents perfectly comprehensible audio. It can be seen that both MA-medium and MA-heavy are able to achieve the best obfuscation in this preliminary speech intelligibility experiment. Furthermore, ENERGY-heavy and SHRED-heavy can be recommended as well, since both have a mean opinion score below 1.20. SHRED-light performs worst in obfuscating speech.

5.2. Event intelligibility experiment

In the event intelligibility experiment, participants had to label obfuscated domestic events. As can be seen in Table 4b, all of the MA and SHRED obfuscation techniques (with the exception of MA-medium and SHRED-medium) score above 0.70, while all ENERGY obfuscating techniques score less.

5.3. Classification

The results of the domestic event classification experiment are presented in Table 3. It can be seen that all models have comparable classification results.

Table 3: Results of the CNN classification experiment. The baseline always spanned the same time horizon at the network's input.

	Macro average recall $\pm$ SD (4 folds; in %)		Macro F1 $\pm$ SD (4 folds; in %)		Nr. of input frames	
	Obfuscated	**Baseline**	**Obfuscated**	**Baseline**	**Obfuscated**	**Baseline**
MA-light	84 ± 0.0	85 ± 0.0	83 ± 0.0	84 ± 0.0	20	61
MA-medium	83 ± 0.0	85 ± 0.0	82 ± 0.5	84 ± 0.5	12	63
MA-heavy	82 ± 0.5	85 ± 0.0	81 ± 0.5	84 ± 0.0	8	61
ENERGY-light	83 ± 0.5	87 ± 0.5	82 ± 0.5	87 ± 0.0	54	168
ENERGY-medium	82 ± 0.0	89 ± 0.0	82 ± 0.6	89 ± 0.5	54	336
ENERGY-heavy	81 ± 0.8	90 ± 0.6	81 ± 1.0	89 ± 0.5	54	504
SHRED-light	84 ± 0.0	84 ± 0.5	83 ± 0.5	83 ± 0.5	60	60
SHRED-medium	85 ± 0.5	85 ± 0.5	85 ± 0.6	85 ± 0.6	72	72
SHRED-heavy	84 ± 0.5	85 ± 0.5	83 ± 0.0	84 ± 0.6	64	64

6. DISCUSSION

The results of the speech intelligibility experiment (Table 4a) are as expected, with the exception of MA-medium and MA-heavy. The MA-medium speech appeared to be less intelligible as compared to MA-heavy, but this could be explained by the limited number of participants and assessments. For all ENERGY techniques, the impulsive events all had perfect classifications and can therefore be recommended for this kind of events.

The results of the event intelligibility experiment (Table 4b) are as expected as well, with the exception of MA-medium and SHRED-medium. Possible reasons for these inconsistencies are the possibility for the participants to choose "I don't know", the limited number of participants and the limited diversity in combinations of obfuscation techniques and event types. In practice, the annotators could also have access to additional information, e.g. the microphone location and the spectrogram representation. This would aid them in annotating the events. Furthermore, when looking at the participants' annotations it can be noted that most mistakes were between "speech" and "radio & television", and "footsteps" and "door & window" (impulsive sounds), and between "faucet & shower", "toilet", "background" and "kitchen hood" (stationary sounds). More experienced annotators would be better at distinguishing these different types of events. Note that our participants were not trained beforehand, which could also explain why the baseline is lower than MA-light and SHRED-light.

In the results of the CNN classification experiment (Table 3), a decrease in performance can be seen with all three ENERGY techniques as compared to their corresponding baselines. This decrease

Table 4: Results of the intelligibility experiments.

(a) Speech, lower is better, range $[1, 3]$.

Obfuscation technique	**Mean opinion score $\pm$ SD**
Baseline	2.96 ± 0.09
MA-light	1.69 ± 0.43
MA-medium	1.03 ± 0.08
MA-heavy	1.06 ± 0.17
ENERGY-light	1.64 ± 0.33
ENERGY-medium	1.22 ± 0.23
ENERGY-heavy	1.19 ± 0.27
SHRED-light	1.97 ± 0.29
SHRED-medium	1.36 ± 0.42
SHRED-heavy	1.17 ± 0.22

(b) Events, higher is better, range $[0, 1]$.

Obfuscation technique	**Mean score $\pm$ SD**
Baseline	0.77 ± 0.20
MA-light	0.83 ± 0.20
MA-medium	0.58 ± 0.20
MA-heavy	0.71 ± 0.19
ENERGY-light	0.67 ± 0.26
ENERGY-medium	0.67 ± 0.30
ENERGY-heavy	0.50 ± 0.35
SHRED-light	0.88 ± 0.14
SHRED-medium	0.58 ± 0.20
SHRED-heavy	0.75 ± 0.22

could be explained by the fact that the CNN model has less information at its input. For ENERGY-light, the time at the model's input is reduced by 66.7% as compared to its baseline. For ENERGY-medium this reduction is equal to 83.2%, and for ENERGY-heavy this reduction is equal to 88.8%. A smaller decrease in performance is also noticeable with MA, due to the reduced resolution at the network's input. SHRED does not suffer from a decrease in performance, because the same information is still present at the network's input.

7. CONCLUSION

In this work, three techniques based on the log-mel spectra were investigated for the purpose of speech obfuscation. A requisite was that sound data could be labelled by human raters at a later point in time, without having intelligible speech in the recordings. The first technique was calculating the moving average of 4 (MA-light), 8 (MA-medium) or 12 (MA-heavy) log-mel frames. The second technique was sampling those windows of log-mel frames having the highest energy, where 18 out of 56 frames (ENERGY-light), 18 out of 112 frames (ENERGY-medium) or 18 out of 168 frames (ENERGY-heavy) were kept. The final technique was sound shredding, where 4 contiguous log-mel frames were kept in a so-called unit. These units were then randomised inside of a snippet of length 3 units (SHRED-light), 6 units (SHRED-medium) or 16 units (SHRED-heavy).

Both a speech and event intelligibility experiment (12 participants) and a SEC classification experiment were carried out. The intelligibility experiment demonstrated that both MA-heavy and SHRED-heavy achieved good speech obfuscation levels, while still having the possibility to label the data. Furthermore, the proposed techniques only had minor impact on the classification performance when evaluating on a dataset with sounds from domestic events, except for the ENERGY techniques.

All considered techniques proved effective in obfuscating speech, while still allowing SEC. For stationary sound events, calculating the moving average or shredding the log-mel spectra is recommended. For impulsive sound events, sampling a few of the most energetic log-mel spectra is recommended.

8. ACKNOWLEDGEMENT

The authors would like to thank Flanders Innovation & Entrepreneurship Agency (VLAIO) and Flanders Make for providing research funding.

9. REFERENCES

[1] B. Boons, M. Verhelst, and P. Karsmakers, "Low power online machine monitoring at the edge," in *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*, 2021, pp. 1–8.

[2] L. Vuegen, P. Karsmakers, B. Vanrumste, and H. Van hamme, "Acoustic event classification using low-resolution multi-label non-negative matrix deconvolution," *Journal of the Audio Engineering Society*, vol. 66, no. 5, pp. 369–384, May 2018.

[3] S. H. Lee, Y.-S. Kim, M.-K. Yeo, M. Mahmood, N. Zavanelli, C. Chung, J. Y. Heo, Y. Kim, S.-S. Jung, and W.-H. Yeo, "Fully portable continuous real-time auscultation with a soft wearable stethoscope designed for automated disease diagnosis," *Science Advances*, vol. 8, no. 21, p. eabo5867, 2022. [Online]. Available: https://www.science.org/doi/abs/10.1126/sciadv.abo5867

[4] S. Kumar, L. T. Nguyen, M. Zeng, K. Liu, and J. Zhang, "Sound shredding: Privacy preserved audio sensing," in *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, ser. HotMobile '15. New York, NY, USA: Association for Computing Machinery, 2015, pp. 135–140.

[5] D. Liaqat, E. Nemati, M. Rahman, and J. Kuang, "A method for preserving privacy during audio recordings by filtering speech," in *2017 IEEE Life Sciences Conference (LSC)*, 2017, pp. 79–82.

[6] E. C. Larson, T. Lee, S. Liu, M. Rosenfeld, and S. N. Patel, "Accurate and privacy preserving cough sensing using a low-cost microphone," in *Proceedings of the 13th International Conference on Ubiquitous Computing*, ser. UbiComp '11. New York, NY, USA: Association for Computing Machinery, 2011, pp. 375–384.

[7] F. Chen, J. Adcock, and S. Krishnagiri, "Audio privacy: Reducing speech intelligibility while preserving environmental sounds," in *Proceedings of the 16th ACM International Conference on Multimedia*, ser. MM '08. New York, NY, USA: Association for Computing Machinery, 2008, pp. 733–736.

[8] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2021.

[9] L. E. Boucheron and P. L. De Leon, "On the inversion of mel-frequency cepstral coefficients for speech enhancement applications," in *2008 International Conference on Signals and Electronic Systems*, 2008, pp. 485–488.

[10] NumPy. Numpy v1.23 `numpy.fft.irfft` function. [Online]. Available: https://numpy.org/doc/1.23/reference/generated/numpy.fft.irfft.html

[11] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[12] L. Vuegen and P. Karsmakers, "Real-time on-edge classification: an application to domestic acoustic event recognition," in *Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2021.