Postprint of: Kąkol K., Korvel G., Kostek B., Noise profiling for speech enhancement employing machine learning models, Journal of the Acoustical Society of America Vol. 152, iss. 6 (2022), pp. 3595-3605, DOI: 10.1121/10.0016495

Noise profiling for speech enhancement employing machine learning models

Krzysztof Kąkol,[1] Grazina Korvel,[2] and Bożena Kostek[3,a]

[1] *PGS Software, Wrocław, 50-086, Poland*

[2] *Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, 08412, Lithuania*

[3] *Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdańsk, 80-233, Poland*

This paper aims to propose a noise profiling method that can be performed in near real-time based on machine learning (ML). To address challenges related to noise profiling effectively, we start with a critical review of the literature background. Then, we outline the experiment performed consisting of two parts. The first part concerns the noise recognition model built upon several baseline classifiers and noise signal features derived from the Aurora noise dataset. This is to select the best-performing classifier in the context of noise profiling. Therefore, a comparison of all classifier outcomes is shown based on effectiveness metrics. Also, confusion matrices prepared for all tested models are presented. The second part of the experiment consists of selecting the algorithm that scored the best, i.e., Naïve Bayes, resulting in an accuracy of 96.76%, and using it in a noise-type recognition model to demonstrate that it can perform in a stable way. Classification results are derived from the real-life recordings performed in momentary and averaging modes. The key contribution is discussed regarding speech intelligibility improvements in the presence of noise, where identifying the type of noise is crucial. Finally, conclusions deliver the overall findings and future work directions.

[a] bokostek@audioacoustics.org

I. INTRODUCTION

Research in speech signal processing and enhancement has attracted considerable interest over the past decades. Major progress has been achieved in various applications, including automatic speech recognition (Li, 2021; Korvel et al., 2021; Michalopoulou et al., 2021), speaker recognition (Krcadinac et al., 2021), and emotion recognition from speech (Gosztolya, 2019; Liu et al., 2021; Morgan et al., 2021). However, when referring to robust speech processing, i.e., in noisy conditions, the progress in this field is below expectations (Li et al., 2015; Srinivasan et al., 2019). Environmental or ambient noise decreases the quality and intelligibility of the speech signal (Trujillo et al., 2021). Therefore, it is vital need to improve the assessment of speech intelligibility in the presence of interference noise. Various noise-robust approaches are adopted for this purpose. Typically, signal processing techniques are employed to reduce noise and enhance voice quality. There is a rich body of work focused on speech enhancement algorithms that use sparse Bayesian learning to solve the sound source localization problem of speech mixtures in noise (Xenaki et al., 2018) and improve speech enhancement by considering power spectral density (PSD) characteristics (Kavalekalam et al., 2018; Kim and Shin, 2022), or aim to improve the quality and intelligibility of noise-corrupted speech through spectral or temporal modifications (Cooke et al., 2019; Kąkol et al., 2020). The limitation of speech enhancement algorithms is that they are based on additive background noise or statistical properties of the speech and noise signal. However, the performance of speech enhancement in a real noisy environment, such as traffic, wind, or a cocktail-party effect when people talk simultaneously (i.e., babble speech), is often unsatisfactory. That is why the challenge of increasing real-world noise recognition robustness is still a significant problem, especially in cases where noise profiling is a necessary step for correct speech signal processing and quality and intelligibility enhancement is the primary goal.

In the literature, there exist several definitions of noise profiling that are related to the task needed, e.g., automatic annotation of noise data (Lin and Tsao, 2021) or attenuation of the noise to certain predefined target levels (Zou et al., 2011). It may also be defined by the automatic threshold selection within lower and higher limit values (Dias et al., 2022), by clustering classification sound types (Kong et al., 2019), or by a noise profile observation in detected silent intervals (Xu et al., 2020).

The present study goes beyond the state-of-the-art methodology of speech enhancement as it incorporates noise inference profiling. In this work, noise profiling is understood similarly to noise type recognition but with a slightly different focus. While for the sound recognition models, it is crucial to obtain correct sound classification (e.g., whether it is a train sound or speech), for profiling task, it is critical to identify the sound characteristics (e.g., spectral features) which are specific to a given type of sound (i.e., noise in our case). In the latter case, precise noise identification is of less importance (Zou et al., 2011). Our previous research (Korvel et al., 2020) demonstrated that using the Lombard effect might improve speech intelligibility in the presence of noise. However, it is crucial to know the noise type to apply the best possible speech modifications. That is the context of our research.

To some extent, our research fits the paradigm of gathering experience based on interactions with the environment through some actions, as the process of noise recognition is sequential, and a decision on enhancing the speech signal should be taken based on satisfying the reward hypothesis (Mahmud et al., 2018).

This work aims to prepare the machine-based model recognizing the noise type and correctly classifying it in near real-time. Based on noise classification, it may then be possible to modify the speech signal appropriately to increase the probability of improving its quality and intelligibility. The study is conducted with a new perspective, focusing not on assigning a disturbance to a given class only but rather on investigating the stability of this assignment − understood as a classification consistency over a longer time, i.e., at least 5 seconds. This allows for stabilizing the decision rules, which might be placed in the system after the profiling block. This adds a new quality to noise profiling that is time-dependent. This research area requires a thorough analysis of speech and noise elements based on a microscopic scale. Therefore, we left the large-scale deep learning analysis outside of this research, disregarding that noise recognition robustness is well served by deep learning methods (Roch et al., 2021; Watanabe et al., 2017). However, state-of-the-art baseline techniques that incorporate the extraction of features and machine learning, such as Naïve Bayes (Zhang, 2014; Barber, 2012), linear SVM (Cortes and Vapnik, 1995; Platt, 1999), SVM with the polynomial kernel (Wu et al., 2004), Gaussian process classifiers (Rasmussen and Williams, 2006; Byrd et al., 1995; Zhu et al., 1997), Decision tree (DT) (Kamiński et al., 2017), Random forest (RF) (Ho, 1995), Multilayer Perceptron (MLP) (Pedregosa et al., 2011), AdaBoost classifier (Rojas, 2009), and Quadratic Discriminant Analysis (Ghojogh and Crowley, (2019) that arose from different families and areas of knowledge (Fernández-Delgado, 2014) are used. It is worth noting that the methodology based on feature extraction and baseline classifiers shows its superiority in many speech signal processing tasks such as speech emotion recognition (Bhavan et al., 2019; Tuncer et al., 2021) or allophones classification (Piotrowska et al., 2019). These studies focused on preparing an optimized feature vector and utilizing this vector in the classification process. In the case of speech emotion recognition, the SVM classifier is used for classification in the mentioned above works. According to Bhavan et al. (2010), SVMs provide reasonably good estimates with lesser effort. In contrast, hidden Markov models and deep neural networks are more challenging to build and train and require higher computational power and time. In the work of Piotrowska et al. (2019), automatic classification methods, such as artificial neural networks (ANNs), the k-nearest neighbor (kNN), and self-organizing maps (SOMs), are applied to lateral allophone analysis and returned satisfactory results.

Also, we justify why the process of improving speech quality and intelligibility should be adaptive and specific modifications may depend on the noise characteristics and be reinforced by them. Based on the rate of change in intensity, noise can be classified into continuous, periodic, impulsive, and low-frequency noise (Tsalera et al., 2020). Therefore, a stable noise profiling method is needed – stable in terms of being consistent over a longer period of time (Yang and Ritzwoller, 2008). Possible speech modifications must fit the disruption to provide the best results in terms of potential loss in intelligibility because of the noise presence. It is because every disturbance has different characteristics and impacts speech differently. However, it is more important to have the noise recognition process repetitive and stable rather than classify a given type of noise as a babble speech or airport noise. Also, noise signals with similar frequency characteristics should always be analogously classified to ensure that the speech signal modification is appropriate and durable.

II. MATERIAL AND METHODS

A. Extraction of signal features

In the learning process, the Aurora noise dataset was employed (Hirsch and Pearce, 2000). The Aurora database contains various speech recordings prepared mainly for speech recognition systems, especially for distributed speech recognition (Kshirsagar and Falk, 2021; Bandela et al., 2021). The noise database within the Aurora dataset has been prepared directly for speech processing, and it is, therefore, appropriate for our research. The noise signals contained in the Aurora dataset are as follows: airport, babble speech, car noise, exhibition, restaurant, street noise, subway, and train. Some noises are reasonably stationary, e.g., the car noise and the recording in the exhibition hall. Others contain non-stationary segments, e.g., recordings on the street and at the airport (Hirsch and Pearce, 2000). In addition, pink noise was generated as this noise type was not present in the Aurora database. To be noted, pink noise is a signal with a frequency spectrum such that the power spectral density is inversely proportional to the signal's frequency, i.e., the power per Hertz in pink noise decreases as the frequency increases (https://www.livescience.com/38464-what-is-pink-noise.html). In pink noise, each octave interval carries an equal amount of noise energy, so the sound of pink noise is perceived as being even.

The following frequency characteristics were chosen and extracted to classify noise types (Klapuri and Davy, 2007; McFee et al., 2015; Das et al., 2021), i.e., spectral centroid, spectral bandwidth, spectral flatness.

The most important factor in evaluating the usefulness of the given feature is the separation of the calculated values in the context of the considered noise type. Three frequency characteristics, calculated in real-time, were considered to increase the separation of different types of noise. What is more, for each of the characteristics, the following short-term statistical parameters are calculated: maximum value, minimum value, average value, amplitude, standard deviation, variance, and median. The given statistic values should provide great noise parameters separation. The frequency characteristics are calculated from the Fourier spectrum computed with a Hamming window of 2048 samples (25% overlap). Below the analyses performed have been described.

1. *Spectral centroid*

Spectral centroid is a metric used in digital signal processing that characterizes the spectrum of the signal. It allows calculating where the center of mass of the spectrum is located. This measure is related perceptually to the impression of the sound brightness. In this study, the spectral centroid is calculated as the weighted mean of the frequencies present in the signal with their magnitudes as the weights:

$$SC = \frac{\sum_{n=0}^{N} f(n)X(n)}{\sum_{n=0}^{N} X(n)} \quad (1)$$

where $X(n)$ is the weighted magnitude of the Fourier transform at frequency bin n, and $f(n)$ represents the center frequency of that bin.

2. *Spectral bandwidth*

The spectral bandwidth (SBW) is used to define the bandwidth of the signal spectrum. This measure shows the concentration of spectrum around the centroid and is computed by:

$$SBW = \left(\sum_{n=0}^{N} X(n)(f(n) - SC)^p\right)^{1/p} \quad (2)$$

where $X(n)$ is the weighted magnitude of the Fourier transform at bin n, $f(n)$ represents the center frequency of that bin, SC is the spectral centroid (see Eq. (1)). Variable p is equal to 2 − this corresponds to a weighted standard deviation around the centroid.

Spectral bandwidth values are calculated for all analyzed noise types and frames within the signal.

3. Spectral flatness

Spectral flatness is a measure of an audio sound spectrum that provides a way to quantify how tone-like a sound is, as opposed to being noise-like. High spectral flatness - approaching 1.0 for white noise - means that the spectrum has a similar amount of power in all spectral bands. Low spectral flatness values (approaching 0.0) convey that the power is concentrated in a small number of bands − typically, it is a mixture of sine waves.

The spectral flatness is calculated by dividing the geometric mean of the power spectrum by the arithmetic mean of the power spectrum, i.e.:

$$SF = \frac{\left[\prod_{n=0}^{N-1} PX(n)\right]^{1/N}}{\frac{1}{N}\sum_{n=0}^{N-1} PX(n)} \tag{3}$$

The power spectrum $PX(n)$ at bin number n is given by the following formula:

$$PX(n) = {}^{1}/_{N} \sqrt{X(n)_{re}^{2} + X(n)_{im}^{2}} \tag{4}$$

where $X(n)$ is Fourier transform coefficient at bin n, re means a real part, and im – an imaginary part.

B. Noise type recognition model

Based on the previously described frequency characteristics, the recognition models were built. For that purpose − as already mentioned − several baseline algorithms were employed, i.e., Naïve Bayes (NB), linear SVM (Support Vector Machines), SVM with the polynomial kernel, Gaussian process classifiers, Decision tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), AdaBoost classifier, and Quadratic Discriminant Analysis (QDA). For both learning and evaluation, the scikit-learn modules from the Python environment were used (https://scikit-learn.org/stable/).

Every recording containing noise was processed in the following way:

- each frame was 2 seconds in length - to retrieve the statistical features for the training process,
- a 2-second window was moved by 0.1 seconds in each analysis step.

The classification models built use relatively long recording fragments because the measured parameter values change in time to a great extent. To clarify, the duration of the Aurora noise recordings is 10 seconds, and the generated pink noise recording is 5 seconds. Since the training is performed on the 2-second long frames, moved by 0.1 seconds, every Aurora noise recording resulted in 81 equally long 2-second frames, while the pink noise resulted in 31 frames of the same length. All frames were represented in the learning process by their calculated parameters – spectral centroid, spectral bandwidth, and spectral flatness. It means that in total, we had 679 samples (frames) – 81 for all 8 Aurora noise recordings and 31 for pink noise recordings.

The above dataset was divided into two almost equal parts: training (consisting of 339 samples) and testing used in generating predictions and calculating scores (composed of 340 samples). The training process was performed on the training set, while calculating scores and generating confusion matrices were performed on the testing set. In other words, the model evaluation process used data that were not seen by the learning process at all.

All classification models employed in the noise profiling task are briefly described below.

Naive Bayes (NB) (sklearn.naive_bayes.GaussianNB module)

A posteriori probability was calculated using the following formula:

$$P(C_k|\boldsymbol{X}) = \frac{P(C_k)P(\boldsymbol{X}|C_k)}{P(\boldsymbol{X})} \quad (5)$$

where $\boldsymbol{X}$ represents the vector with n conditionally independent features $X_1, X_2, \ldots, X_n$, and C_k is a possible outcome class.

Linear Support Vector Machines (SVM) (sklearn.svm.SVC module)

A kernel used to train linear SVM takes the following form:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) \tag{6}$$

where $\boldsymbol{\phi}$ is a function that maps training data into higher dimensional space, $\boldsymbol{x}_i, \boldsymbol{x}_j \in R^n$. The following parameters of linear SVM were implemented: regularization C=0.025, probability estimates have been enabled, and tolerance for stopping criterion is equal to 0.001.

SVM with polynomial kernel (sklearn.svm.SVC)

The following parameters of the polynomial SVM were implemented: regularization parameter $C = 1$, gamma coefficient (γ) set to auto (which means that it uses 1/number_features), probability estimates were enabled, independent term in kernel function equals 0, tolerance for stopping criterion is equal to 0.001.

Gaussian process classifiers (GPCs) (sklearn.gaussian_process.GaussianProcessClassifier module)

In our test, the exponential kernel was used − it takes one base kernel and a scalar parameter and combines them via:

$$k_{exp}(\boldsymbol{X}, \boldsymbol{Y}) = k(\boldsymbol{X}, \boldsymbol{Y})^p \tag{8}$$

In this study, the exponent is equal to 2. As a source kernel, a Rational Quadratic kernel was used. It is parameterized by the length scale parameter and a scale mixture parameter. The kernel is given by:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(1 + \frac{(x_i - x_j)^2}{2\alpha l^2}\right)^{-\alpha} \tag{9}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are vectors of features computed from training or test samples, $\alpha > 0$ is the scale mixture parameter, $l > 0$ is the length scale of the kernel.

The L-BFGS-B (a limited memory Broyden–Fletcher–Goldfarb–Shanno) algorithm is used in the context of finding a (local) minimum of an objective function.

Decision Tree (DT) (sklearn.tree.DecisionTreeClassifier module)

The parameters used in this test are as follows: the quality of the split is Gini impurity, maximum depth of the tree is 5.

Random Forest (sklearn.ensemble.RandomForestClassifier module)

Parameters used in this research: the quality of the split is Gini impurity, the maximum depth of the tree is 5, number of estimators (trees in the forest) is set to 10.

Multilayer Perceptron (MLP) Classifier (sklearn.neural_network.MLPClassifier module)

The following parameters of the MLP classifier were used: L2 regularization parameter (alpha) is set to 1, and the maximum number of iterations equals 1000. The hidden layer contains 100 neurons, and the activation function is ReLU. The optimizer used for weight is Adam optimization, which refers to the stochastic gradient descent optimizer (Pedregosa et al., 2011).

AdaBoost classifier (sklearn.ensemble.AdaBoostClassifier module)

In this study, the following parameters were used: the maximum number of estimates at which boosting is stopped equals 50, the learning rate equals 1, and SAMME.R is used as the boosting algorithm.

Quadratic Discriminant Analysis (sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis module)

The quadratic Discriminant Analysis classifier is based on the Bayes rule presented above in the description of the Naïve Bayes classifier (see Eq. 5). If there is an assumption that the covariance matrices are diagonal, then the input features are assumed independent - the resulting classifier is then equivalent to Naïve Bayes. For our test, the regularization parameter is set to 0.

III. COMPARISON OF THE CLASSIFIER RESULTS

The classification results are provided in the form of overall accuracy and a confusion matrix, allowing for a straightforward interpretation of the results. For the multiclass classification problems, the following metrics have been used (Grandini et al., 2020):

− overall accuracy for the whole prediction process,

− precision, recall, and F1-score for every class.

The F1 metric was used because, in our classification procedure, both false positives and false negatives are equally undesirable, which is best reflected by F1 (Lipton et al., 2014). The dataset used in our study is well-balanced; therefore, AUC ROC has been chosen as it suits balanced datasets (Huang and Ling, 2005).

To calculate these metrics, the following prediction results need to be obtained:

− TP_n – the number of true positive recognitions for distortion type n (e.g., subway),

− TN_n – the number of true negative recognitions for distortion type n,

− FP_n – the number of false positive recognitions for distortion type n (in other words – the number of samples recognized incorrectly as type n),

− FN_n – the number of false negative recognitions for distortion type n (in other words – the number of *n* distortion samples recognized as something different than type n).

The overall accuracy can be measured only using the full recognition results. For the multiclass classification problem, the formula is as follows:

$$Acc = \sum_n \frac{TP_n}{N} \quad (10)$$

In other words – it is a sum of true positives for all distortion types divided by the number of samples being recognized. The typical definition of two-class accuracy has the sum of true positives and true negatives in the denominator of the equation. Still, it is the same as the sum of all true positives if both classes are treated as being detected.

Precision for type n is defined as follows:

$$Precision_n = \frac{TP_n}{TP_n + FP_n} \tag{11}$$

Recall for type n is defined as follows:

$$Recall_n = \frac{TP_n}{TP_n + F\ \ _n} \tag{12}$$

F1-score for type n is defined as follows:

$$F1score_n = 2 \cdot \frac{Precision_n * Recall_n}{Precision_n + Recall_n} \tag{13}$$

Tables I-III show the comparison of the above-described classification models. Also, metrics such as *P* – precision, *R* – recall, *F*1 – F1-score, and *S* – support are included. The pair of the best accuracy and ROC AUC (area under the receiver operating characteristic curve) achieved – is highlighted in bold. Moreover, recognition time for all models is included as well. Values of recognition time for all models are calculated as a time used for classifying all 340 testing samples.

TABLE I. Results of the classification using Naïve Bayes, Linear SVM, and SVM polynomial classification models. *P* – precision, *R* – recall, *F*1 – F1-score, *S* – support.

	Naïve Bayes	Linear SVM	SVM polynomial
Accuracy	96.76%	96.17%	94.41%
ROC AUC	0.99	0.99	0.99
Recognition time	0.67 ms	1.56 ms	1.29 ms

Noise distortions	P	R	$F1$	S	P	R	$F1$	S	P	R	$F1$	S
Airport	1.00	0.96	0.98	45	0.86	0.96	0.91	45	0.84	0.91	0.87	45
Babble speech	0.90	0.90	0.90	39	1.00	0.95	0.97	39	0.90	0.95	0.93	39
Car	1.00	1.00	1.00	46	0.96	1.00	0.98	46	1.00	0.93	0.97	46
Exhibition	0.98	1.00	0.99	39	1.00	1.00	1.00	39	1.00	1.00	1.00	39
Pink noise	1.00	1.00	1.00	17	1.00	1.00	1.00	17	1.00	1.00	1.00	17
Restaurant	1.00	0.91	0.95	32	1.00	1.00	1.00	32	0.94	1.00	0.97	32
Street noise	0.92	0.98	0.95	48	0.91	0.81	0.86	48	0.88	0.79	0.84	48
Subway	1.00	0.97	0.98	32	1.00	1.00	1.00	32	1.00	1.00	1.00	32
Train	0.95	1.00	0.98	42	1.00	1.00	1.00	42	1.00	1.00	1.00	42

TABLE II. Results of classification using Gaussian process, Decision tree, and Random forest classification models. All denotations are as shown in TABLE I.

	GPC	Decision tree	Random forest
Accuracy	85.88%	95.59%	92.94%
ROC AUC	0.98	0.98	0.99

Recognition time	45 ms				0.25 ms				1.66 ms			
Noise distortions	P	R	F1	S	P	R	F1	S	P	R	F1	S
Airport	0.83	0.89	0.86	45	0.94	0.98	0.96	45	0.98	0.98	0.98	45
Babble speech	0.78	0.97	0.86	39	0.97	0.77	0.86	39	0.87	1.00	0.93	39
Car	0.93	0.89	0.91	46	1.00	1.00	1.00	46	0.98	1.00	0.99	46
Exhibition	0.80	0.95	0.87	39	0.98	1.00	0.99	39	0.98	1.00	0.99	39
Pink noise	0.89	1.00	0.94	17	1.00	0.88	0.94	17	1.00	0.94	0.97	17
Restaurant	0.88	0.94	0.91	32	0.84	0.97	0.90	32	0.84	0.97	0.90	32
Street noise	0.94	0.63	0.75	48	0.92	0.98	0.95	48	1.00	0.58	0.74	48
Subway	0.92	0.72	0.81	32	1.00	0.97	0.98	32	1.00	0.97	0.98	32
Train	0.84	0.86	0.85	42	1.00	1.00	1.00	42	0.82	1.00	0.90	42

TABLE III. Results of the classification using MLP, AdaBoost, and QDA classification models. All denotations are as shown in TABLE I.

	MLP	AdaBoost	QDA
Accuracy	67.05%	67.64%	93.52%

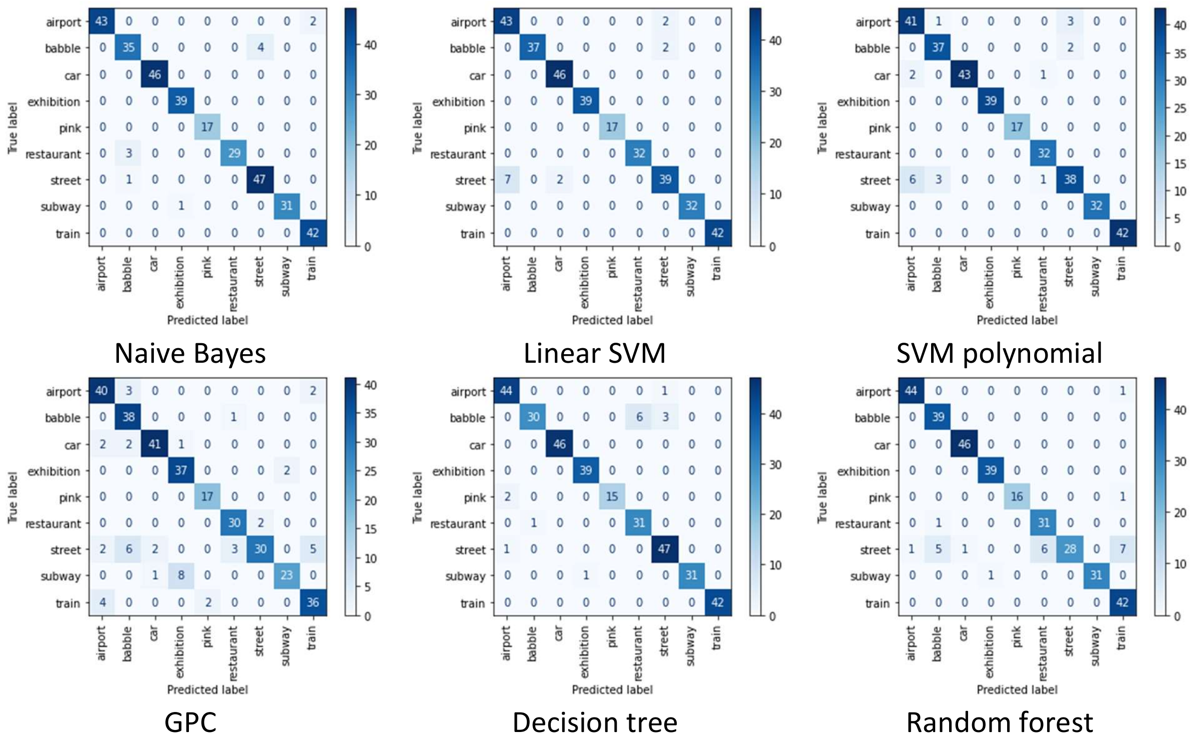
ROC AUC	0.95				0.95				0.94			
Recognition time	0.49 ms				15.66 ms				0.72 ms			
Noise distortions	P	R	F1	S	P	R	F1	S	P	R	F1	S
Airport	0.75	0.40	0.52	45	0.48	0.96	0.64	45	0.72	0.96	0.82	45
Babble speech	0.74	0.74	0.74	39	0.51	0.92	0.65	39	0.98	1.00	0.99	39
Car	0.85	0.85	0.85	46	1.00	0.98	0.99	46	0.94	1.00	0.97	46
Exhibition	1.00	0.33	0.50	39	1.00	1.00	1.00	39	1.00	1.00	1.00	39
Pink noise	0.55	0.94	0.70	17	0.00	0.00	0.00	17	0.00	0.00	0.00	17
Restaurant	0.59	1.00	0.74	32	0.00	0.00	0.00	32	1.00	1.00	1.00	32
Street noise	0.40	0.33	0.36	48	0.00	0.00	0.00	48	0.98	0.94	0.96	48
Subway	0.54	1.00	0.70	32	1.00	1.00	1.00	32	1.00	1.00	1.00	32
Train	0.92	0.79	0.85	42	0.56	0.83	0.67	42	1.00	1.00	1.00	42

One can notice that most tested algorithms give sufficiently good results with an accuracy of over 90%; however, only three have better accuracy than 96%, i.e., Naïve Bayer, Linear SVM, and Decision Tree. For all three algorithms, all other metrics (averages of precision, recall, and F1 for all noise types) are similar; however, Naïve Bayer is a little better than Linear SVM and Decision Tree. The computational complexity for inference for all these methods is also similar and linearly dependent on the number of dimensions (for Linear SVM and Naive Bayer) or the number of tree depths for the Decision Tree.

The other algorithms are not as accurate as the three mentioned above. Some of them have no true positives for some noise types, which results in zeroing the basic metrics for these types. This can be observed in Figure 1 (e.g., pink noise recognition for the AdaBoost classifier). That is why these algorithms have been disqualified, i.e., MLP, AdaBoost, and Quadratic Discriminant Analysis. Moreover, since these times are of a millisecond level, we can assume that near-real-time recognition is possible with the assumption that the initial 1-second recognition has already passed.

Considering the above results, we have selected the Naïve Bayes model as a source model for the subsequent experiments.

In Figure 1, confusion matrices are presented that were prepared for all tested models.



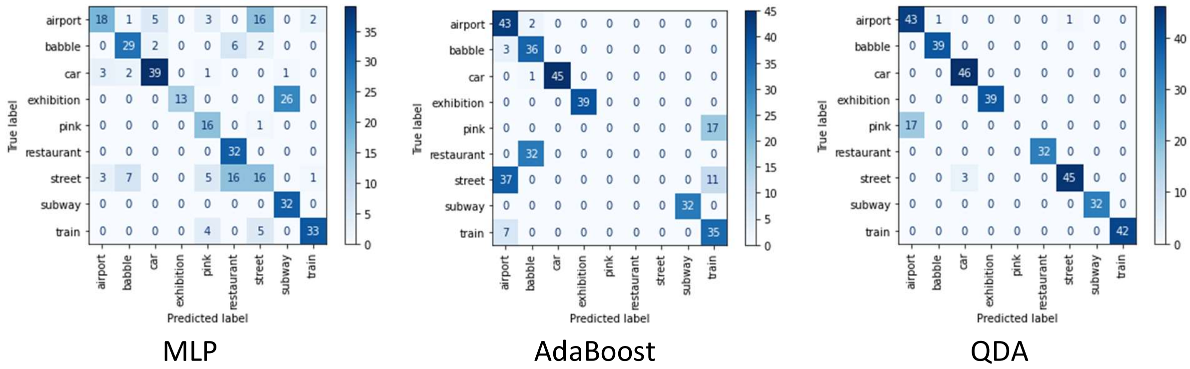


FIG. 1. Confusion matrices for all tested models.

IV. DISCUSSION

The created model using the Naïve Bayes classification was tested on recordings that were used for training (but different parts of these recordings) and on the additional recordings from the multimodal corpus of English speech recordings called MODALITY (Czyzewski et al., 2017). As mentioned before, in the context of noise profiling, the model's usefulness is measured by evaluating its stability, understood as a classification consistency over a longer period of time, not correctness – presumed as class-level accuracy. This is because the recording conditions might be very different - such as the recording method and equipment, source of noise, and its characteristics. Therefore, for instance, the airport recording might be identified as street noise. What matters here is that this recording is always (or almost always) identified as street noise. That is why the correctness of classification is of less importance in general. The value of this model is in recognizing the abstract type of distortion using its frequency parameters – and this is the basis of improving speech intelligibility in the presence of noise. The process of speech quality/intelligibility enhancement requires particular conditioning – and the values of the parameters used should correspond to the type of noise. These values strongly impact the efficiency of speech intelligibility improvement. So, it is crucial to effectively classify the particular types of distortion to an assigned number of classes, enabling to modify the speech in the best way in given noise conditions.

The recognition process was carried out in two modes: momentary and averaging. In both modes, the window/frame analyzed was 1 or 2 seconds, and the window was moved by 0.1 seconds with every step. In the momentary mode, classification was performed for every frame. In the averaging mode, the classification was made with delay - it means that the momentary classification should change across five consecutive frames to calculate the average classification. However, it does not mean that the results should be considered valid if and only if the five consecutive frames will occur. What is more, the 1-second frame does not necessarily have to be an uninterrupted fragment. It only means that the system should wait a little longer for the first recognition.

Thanks to this procedure, the recognition model avoids a temporary disturbance, usually caused by non-stationary noise.

Figures 2 and 3 present the outcomes of classification. The solid line represents the classification in the averaging mode, while the dashed line represents the momentary classification. The classification results for 1- and 2-second frames are different – first of all, it is because the learning process was performed using a 2-second frame; what is more, a longer window allows for better evaluation of the statistical features of the frequency characteristics. When using 2-second windows, the classification results are very good. For a 1-second window, the statistical characteristics might not be clearly visible, but the averaging mode provides satisfying results.

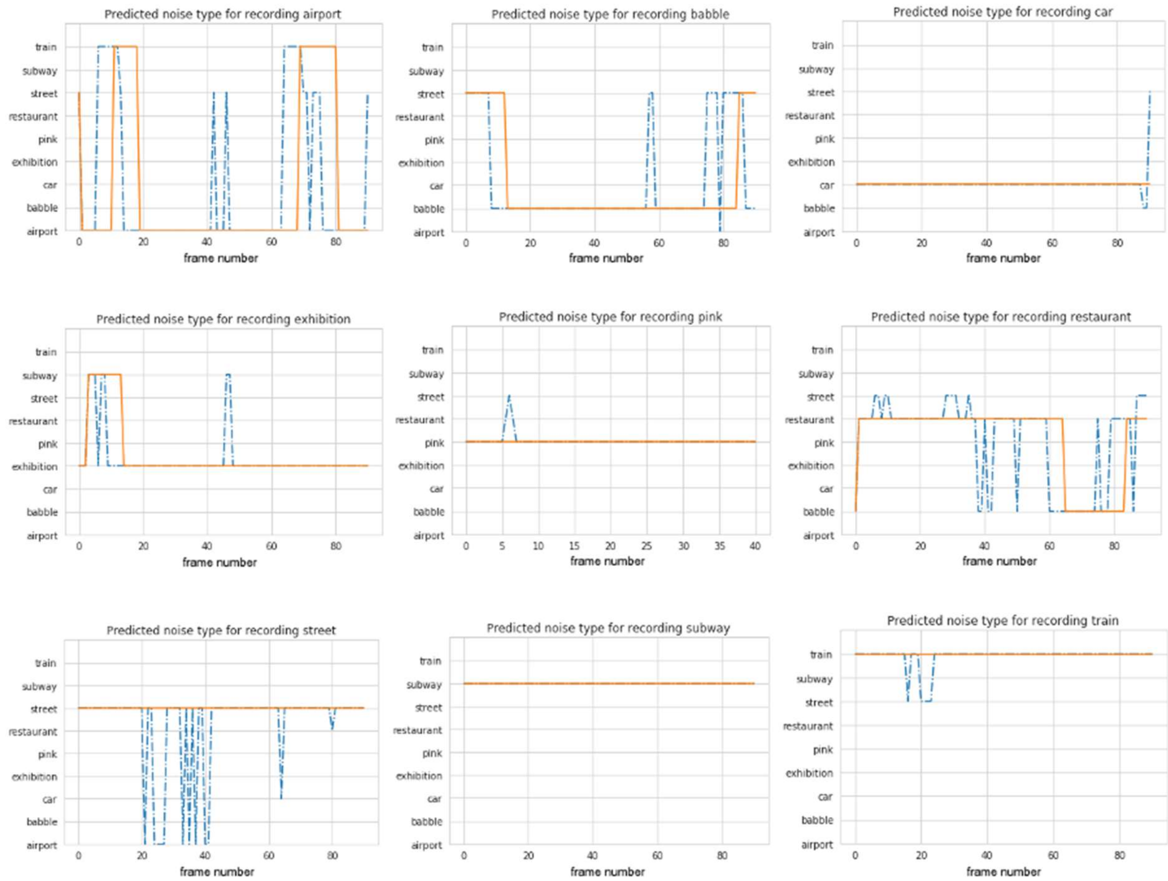
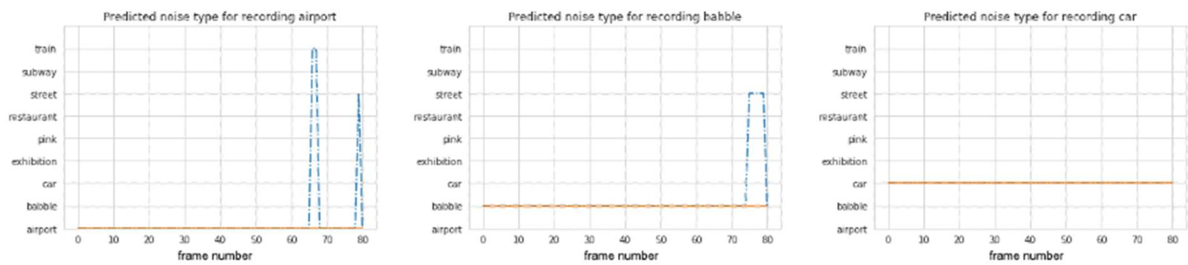


FIG. 2. Classification results on the real-life recordings using a 1-second-length frame (dashed line – result from momentary mode, solid line – result from averaging model).



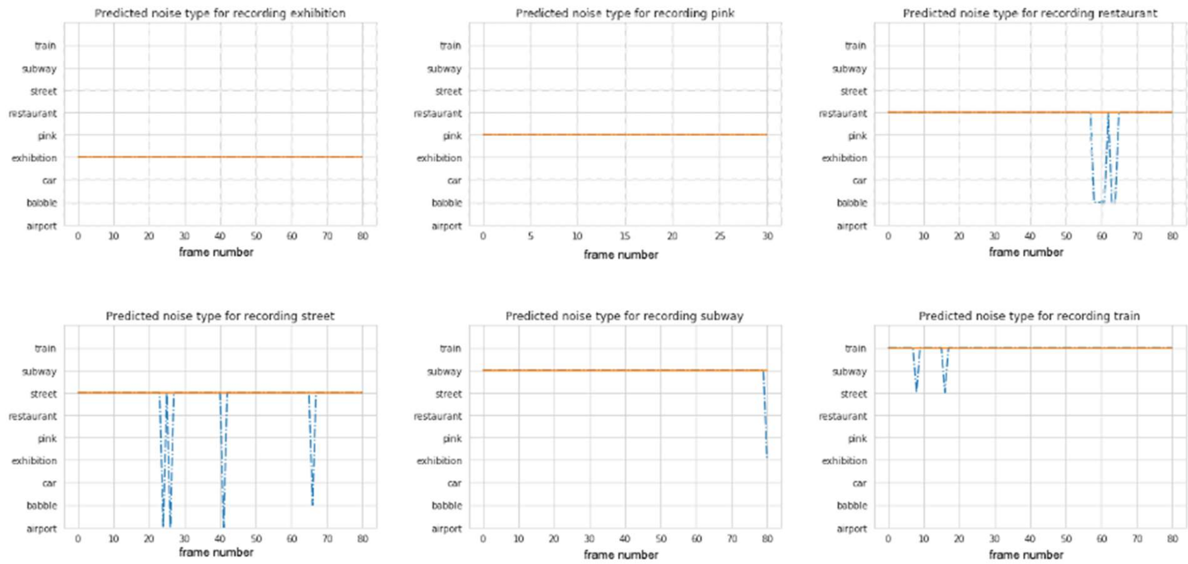


FIG. 3. Classification results on the real-life recordings using a 2-second-length frame (dashed line – result from momentary mode, solid line – result from averaging model).

The recognition process was also performed on a completely different set of noise recordings contained in the MODALITY multimodal corpus of English speech recordings (Rasmussen et al., 2006). The recordings used in this test were very long (between 11 minutes 45 seconds and 14 minutes 54 seconds). The test was performed only for a 2-second frame, and the window was moved by 2 seconds (due to the overall recording length) with every step. The averaging was also used to remove random fluctuations in the recognition results. Figures 4-6 present recognition results, where dashed lines represent the single window classification and the solid line depicts the averaged result.

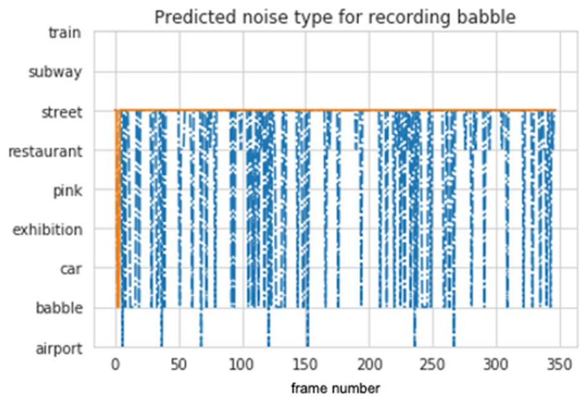


FIG. 4. An example where the classification model has selected both “street” and “babble speech,” but after averaging, the resulting classification was “street.”

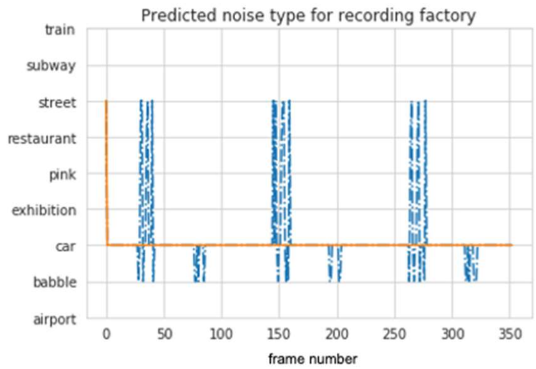


FIG. 5. An example where the “factory” recording was classified as “car noise” (there was no such class as “factory” in the training set).

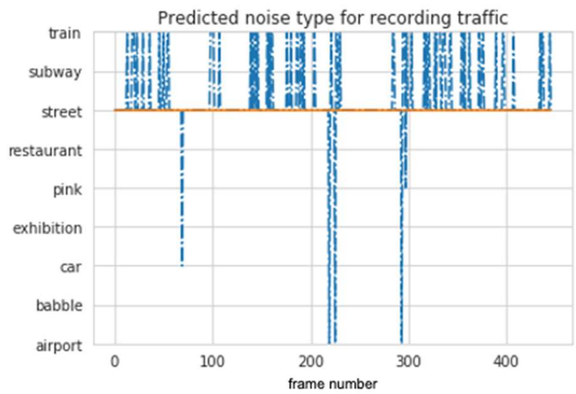


FIG. 6. An example where the recording “traffic” was classified as “street,” which is the correct classification.

As pointed out, it must be underlined that the classification quality is impacted by the stability of the classification, not correctness. That is why the results are generally satisfying, even if the noise recordings are not always correctly classified. As previously mentioned, the classification would strongly be impacted by the recording place, recording equipment, sampling frequency, etc.

V. CONCLUSIONS

In this study, an efficient method of noise profiling was presented, understood as critical to identify the sound characteristics specific to a given type of sound. It was demonstrated that stable and predictable noise profiling is possible using noise spectral characteristics. These characteristics can be calculated almost in real time so that noise profiling can be fast and efficient. The stability, however, depends on the length of the frame and the number of frames used in the averaging process. It may mean that the noise profiling process is delayed up to 2-3 seconds), but it can strongly be decreased after a couple of initial seconds of a signal. This means that the presented method can efficiently be used when trying to improve speech quality and intelligibility when the speech is played back in noisy conditions. The experiments, however, assumed that noise was separated from the speech signal. This can be extended to situations where speech is recorded with noise by separating both signals and processing them in separate flows, which could be the next step in improving the overall speech intelligibility improvement model.

Overall, the proposed method is fast and stable so that it can be used in near real-time systems. The algorithmic simplicity of the machine learning models employed results in low computational complexity while classifying the recorded noise, thus allowing for obtaining low inference times. Even though the classification is not binary, and the number of classes is quite large, a relatively simple model using spectral measures provides high accuracy. This allows for building applications on top of the model proposed.

In future research, we plan to use noise profiling along with the P.563 objective metric ITU-T Recommendation P. 563 (2004) as an input to the feedback system in classical reinforcement learning. We will follow the methodology in which predictors are trained on human quality ratings (Reddy et al., 2021) but use the reward derived from the Reinforcement Learning (RL) paradigm. This is because RL refers to learning by interacting with the environment (Sutton and Barto, 2018).

Indeed, our focus will be on the speed of stable recognition in our future research. Following our experiments, future research should also be directed to reducing the time needed for noise profiling and trying to use this approach in noise suppression systems.

REFERENCES

Bandela, S. R., & Kumar, T. K. (**2021**). Unsupervised feature selection and NMF de-noising for robust Speech Emotion Recognition. Applied Acoustics, 172, 107645, doi: 10.1016/J.APACOUST.2020.107645.

Barber, D. (2012). Bayesian Reasoning and Machine Learning. Cambridge University Press. ISBN 978-0-521-51814-7.

Bhavan, A., Chauhan, P., & Shah, R. R. (**2019**). Bagged support vector machines for emotion recognition from speech. Knowledge-Based Systems, 184, 104886, https://doi.org/10.1016/j.knosys.2019.104886.

Byrd, R. H., Lu, P., & Nocedal, J. (**1995**). A Limited Memory Algorithm for Bound Constrained Optimization, SIAM Journal on Scientific and Statistical Computing, 16, 5, pp. 1190-1208.

Cooke, M., Aubanel, V., & García Lecumberri M. L. (**2019**). Combining spectral and temporal modification techniques for speech intelligibility enhancement. Computer Speech and Language, Elsevier, 55, pp.26-39. 10.1016/j.csl.2018.10.003.

Cortes, C., & Vapnik, V. N. (**1995**). Support-vector networks. Machine Learning. 20 (3): 273–297. CiteSeerX 10.1.1.15.9362. doi:10.1007/BF00994018. S2CID 206787478.

Cortes, C., Haffner, P., & Mohri, M. (**2004**). Rational kernels: Theory and algorithms. Journal of Machine Learning Research, 5(Aug), 1035-1062.

Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J., & Szykulski, M. (**2017**), An audio-visual corpus for multimodal automatic speech recognition, J. of Intelligent Information Systems, pp. 1-26, DOI: 10.1007/s10844-016-0438-z.

Das, N., Chakraborty, S., Chaki, J., Padhy, N., & Dey, N. (**2021**). Fundamentals, present and future perspectives of speech enhancement. International Journal of Speech Technology, 24(4), 883-901.

Dias, F. F., Ponti, M. A., & Minghim, R. (**2022**). Implementing simple spectral denoising for environmental audio recordings. arXiv preprint arXiv:2201.02099.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (**2014**). Do we need hundreds of classifiers to solve real world classification problems?. Journal of Machine Learning Research 15 (2014) 3133-3181

Ghojogh, B., & Crowley, M. (**2019**). Linear and quadratic discriminant analysis: Tutorial. arXiv preprint arXiv:1906.02590.

Gosztolya, G. (**2019**). Posterior-thresholding feature extraction for paralinguistic speech classification. Knowledge-Based Systems, 186, 104943, https://doi.org/10.1016/j.knosys.2019.104943.

Grandini, M., Bagli, E., & Visani, G. (**2020**). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.

Hastie, T., Tibshirani, R., & Friedman, J. (**2009**). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY.

Hirsch, H. G., & Pearce, D. (**2000**). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In ASR2000-Automatic speech recognition: challenges for the new Millenium ISCA tutorial and research workshop (ITRW).

Ho, T. K. (**1995**). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

Huang, Jin, & Charles, X. Ling. (**2005**). Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on knowledge and Data Engineering, 17.3: 299-310.

ITU-T Recommendation P.563. (**2004**), "Single-ended method for objective speech quality assessment in narrow-band telephony applications," ITU-T Recommendation P.563.

Kamiński, B., Jakubczyk, M., & Szufel, P. (**2017**). A framework for sensitivity analysis of decision trees. Central European Journal of Operations Research. 26 (1): 135–159. doi:10.1007/s10100-017-0479-6. PMC 5767274. PMID 29375266.

Kavalekalam, M. S., Nielsen, J. K., Christensen, M. G., & Boldt, J. B. (**2018**). A study of noise PSD estimators for single channel speech enhancement. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5464-5468). IEEE.

Kąkol, K., Korvel, G., & Kostek, B. (2020). Improving Objective Speech Quality Indicators in Noise Conditions. Studies in Computational Intelligence, vol. 869, 199-218. https://doi.org/10.1007/978-3-030-39250-5.

Kim, M., & Shin, J. W. (**2022**). Improved Speech Enhancement Considering Speech PSD Uncertainty. IEEE/ACM Transactions on Audio, Speech, and Language Processing.

Klapuri, A., & Davy, M. (Eds.). (**2007**). Signal processing methods for music transcription, chapter 5. Springer Science and Business Media LLC.

Kong, Q., Xu, Y., Sobieraj, I., Wang, W., & Plumbley, M. D. (**2019**). Sound event detection and time–frequency segmentation from weakly labelled data. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(4), 777-787, doi: https://doi.org/10.1109/TASLP.2019.2895254.

Korvel, G., Kąkol, K., Kurasova, O., & Kostek, B. (**2020**). Evaluation of Lombard speech models in the context of speech in noise enhancement. IEEE Access, 8, 155156-155170, https://doi.org/10.1109/access.2020.3015421.

Korvel, G., Treigys, P., & Kostek, B. (**2021**). Highlighting interlanguage phoneme differences based on similarity matrices and convolutional neural network. The Journal of the Acoustical Society of America, 149(1), 508-523.

Krčadinac, O., Šošević, U., & Starčević, D. (**2021**). Evaluating the Performance of Speaker Recognition Solutions in E-Commerce Applications. Sensors. 21(18):6231. https://doi.org/10.3390/s21186231.

Kshirsagar, S. R., & Falk, T. H. (**2022**). Quality-Aware Bag of Modulation Spectrum Features for Robust Speech Emotion Recognition. IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2022.3188223.

Li, J. (**2021**). Recent Advances in End-to-End Automatic Speech Recognition, invited paper submitted to APSIPA Transactions on Signal and Information Processing, https://arxiv.org/abs/2111.01690.

Li J., Deng L., Haeb-Umbach, R., & Gong, Y. (**2015**). Robust automatic speech recognition: A bridge to practical applications. Academic Press, Elsevier, https://doi.org/10.1016/C2014-0-02251-4.

Lin, T. H., & Tsao, Y. (**2020**). Source separation in ecoacoustics: a roadmap towards versatile soundscape information retrieval. Remote Sensing in Ecology and Conservation, 6(3), 236-247. https://zslpublications.onlinelibrary.wiley.com/doi/epdf/10.1002/rse2.141.

Lipton, Z. C., & Elkan, C., & Narayanaswamy, B. (**2014**). Thresholding classifiers to maximize F1 score. arXiv preprint arXiv:1402.1892.

Liu, S., Zhang, M., Fang, M., Zhao, J., Hou, K., & Hung, C. C. (**2021**). Speech emotion recognition based on transfer learning from the FaceNet framework. The Journal of the Acoustical Society of America, 149(2), 1338-1345.

McFee, B., Colin, R., Liang, D., Ellis, D. P. W., McVicar M., Battenberg E., & Nieto O. (**2015**). librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference, pp. 18-25.

Michalopoulou, Z. H., Gerstoft, P., Kostek, B., & Roch, M. A. (**2021**). Introduction to the special issue on machine learning in acoustics. The Journal of the Acoustical Society of America, 150(4), 3204-3210, https://doi.org/10.1121/10.0006783

Morgan, M. M., Bhattacharya, I., Radke, R. J., & Braasch, J. (**2021**). Classifying the emotional speech content of participants in group meetings using convolutional long short-term memory network. The Journal of the Acoustical Society of America, 149(2), 885-894.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (**2011**). Scikit-learn: Machine Learning in Python; 12(85):2825−2830.

Piotrowska, M., Korvel, G., Kostek, B., Ciszewski, T., & Czyżewski, A. (**2019**). Machine learning-based analysis of English lateral allophones. International Journal of Applied Mathematics and Computer Science, 29(2).

Platt, J. (**1999**). Probabilistic outputs for SVMs and comparisons to regularized likelihood methods, Advances in Large Margin Classifiers. In: Advances in Large Margin Classifiers, MIT Press.

Rasmussen, C.E., & Williams C.K.I. (**2006**). Gaussian Processes for Machine Learning. MIT Press. ISBN 978-0-262-18253-9.

Roch, M. A., Gerstoft, P., Kostek, B., & Michalopoulou, Z. H. (**2021**). How machine learning contributes to solve acoustical problems. Journal of the Acoustical Society of America,17(4), 17, 48-57, https://doi.org/10.1121/at.2021.17.4.48.

Rojas, R. (**2009**). AdaBoost and the super bowl of classifiers a tutorial introduction to adaptive boosting" (Tech. Rep.). Freie University, Berlin.

Srinivasan,T., Sanabria, R., & Metze, F. (**2019**). Analyzing utility of visual context in multimodal speech recognition under noisy conditions. In arXiv preprint arXiv:1907.00477.

https://scikit-learn.org/stable/ (last accessed November 2022).

Tuncer, T., Dogan, S., & Acharya, U. R. (**2021**). Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques. Knowledge-Based Systems, 211, 106547, https://doi.org/10.1016/j.knosys.2020.106547.

Tsalera, E., Papadakis, A., & Samarakou, M. (**2020**). Monitoring, profiling and classification of urban environmental noise using sound characteristics and the KNN algorithm. Energy Reports, 6, 223-230. https://doi.org/10.1016/j.egyr.2020.08.045.

Trujillo, J., Özyürek, A., Holler, J., Drijvers, L. (2021). Speakers exhibit a multimodal Lombard effect in noise. Sci Rep 11, 16721. https://doi.org/10.1038/s41598-021-95791-0.

Watanabe, S., Delcroix, M., Metze, F., & Hershey, J. R. (Eds.) (**2017**). New Era for Robust Speech Recognition. Springer International Publishing, doi:10.1007/978-3-319-64680-0.

Wu, T-F., Lin, C.-J., & Weng, R. C.-H. (**2004**). Probability estimates for multi-class classification by pairwise coupling, Journal of Machine Learning Research, 5:975-1005.

Xenaki, A., & Bünsow Boldt, J., & Græsbøll Christensen, M. (**2018**). Sound source localization and speech enhancement with sparse Bayesian learning beamforming. The Journal of the Acoustical Society of America, 143(6), 3912-3921.

Xu, R., Wu, R., Ishiwaka, Y., Vondrick, C., & Zheng, C. (**2020**). Listening to sounds of silence for speech denoising. Advances in Neural Information Processing Systems, 33, 9633-9648.

Yang, Y., & Ritzwoller, M. H. (2008). Characteristics of ambient seismic noise as a source for surface wave tomography. Geochemistry, Geophysics, Geosystems, 9(2).

Zhang, H. (**2004**). The Optimality of Naïve Bayes, FLAIRS Conference, AAAI Press.

Zhu, H., Byrd, R. H., & Nocedal, J., (**1997**). Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization, ACM Transactions on Mathematical Software, 23, 550-560.

Zou, G., Antila, M., & Kataja, J. (**2011**). Practical active noise profiling in a passenger car. Proc. Akustiikkapäivät, 11-12.