

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

1 I. INTRODUCTION

2 Research in speech signal processing and enhancement has attracted considerable interest over the
3 past decades. Major progress has been achieved in various applications, including automatic speech
4 recognition (Li, 2021; Korvel et al., 2021; Michalopoulou et al., 2021), speaker recognition (Krcadinac
5 et al., 2021), and emotion recognition from speech (Gosztolya, 2019; Liu et al., 2021; Morgan et al.,
6 2021). However, when referring to robust speech processing, i.e., in noisy conditions, the progress in
7 this field is below expectations (Li et al., 2015; Srinivasan et al., 2019). Environmental or ambient noise
8 decreases the quality and intelligibility of the speech signal (Trujillo et al., 2021). Therefore, it is vital
9 need to improve the assessment of speech intelligibility in the presence of interference noise. Various
10 noise-robust approaches are adopted for this purpose. Typically, signal processing techniques are
11 employed to reduce noise and enhance voice quality. There is a rich body of work focused on speech
12 enhancement algorithms that use sparse Bayesian learning to solve the sound source localization
13 problem of speech mixtures in noise (Xenaki et al., 2018) and improve speech enhancement by
14 considering power spectral density (PSD) characteristics (Kavalekalam et al., 2018; Kim and Shin,
15 2022), or aim to improve the quality and intelligibility of noise-corrupted speech through spectral or
16 temporal modifications (Cooke et al., 2019; Kaçkol et al., 2020). The limitation of speech enhancement
17 algorithms is that they are based on additive background noise or statistical properties of the speech
18 and noise signal. However, the performance of speech enhancement in a real noisy environment, such
19 as traffic, wind, or a cocktail-party effect when people talk simultaneously (i.e., babble speech), is often
20 unsatisfactory. That is why the challenge of increasing real-world noise recognition robustness is still
21 a significant problem, especially in cases where noise profiling is a necessary step for correct speech
22 signal processing and quality and intelligibility enhancement is the primary goal.

23 In the literature, there exist several definitions of noise profiling that are related to the task needed,
24 e.g., automatic annotation of noise data (Lin and Tsao, 2021) or attenuation of the noise to certain

25 predefined target levels (Zou et al., 2011). It may also be defined by the automatic threshold selection
26 within lower and higher limit values (Dias et al., 2022), by clustering classification sound types (Kong
27 et al., 2019), or by a noise profile observation in detected silent intervals (Xu et al., 2020).

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

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

41 This work aims to prepare the machine-based model recognizing the noise type and correctly
42 classifying it in near real-time. Based on noise classification, it may then be possible to modify the
43 speech signal appropriately to increase the probability of improving its quality and intelligibility. The
44 study is conducted with a new perspective, focusing not on assigning a disturbance to a given class
45 only but rather on investigating the stability of this assignment – understood as a classification
46 consistency over a longer time, i.e., at least 5 seconds. This allows for stabilizing the decision rules,
47 which might be placed in the system after the profiling block. This adds a new quality to noise profiling
48 that is time-dependent. This research area requires a thorough analysis of speech and noise elements

49 based on a microscopic scale. Therefore, we left the large-scale deep learning analysis outside of this
50 research, disregarding that noise recognition robustness is well served by deep learning methods (Roch
51 et al., 2021; Watanabe et al., 2017). However, state-of-the-art baseline techniques that incorporate the
52 extraction of features and machine learning, such as Naïve Bayes (Zhang, 2014; Barber, 2012), linear
53 SVM (Cortes and Vapnik, 1995; Platt, 1999), SVM with the polynomial kernel (Wu et al., 2004),
54 Gaussian process classifiers (Rasmussen and Williams, 2006; Byrd et al., 1995; Zhu et al., 1997),
55 Decision tree (DT) (Kamiński et al., 2017), Random forest (RF) (Ho, 1995), Multilayer Perceptron
56 (MLP) (Pedregosa et al., 2011), AdaBoost classifier (Rojas, 2009), and Quadratic Discriminant Analysis
57 (Ghojogh and Crowley, (2019) that arose from different families and areas of knowledge (Fernández-
58 Delgado, 2014) are used. It is worth noting that the methodology based on feature extraction and
59 baseline classifiers shows its superiority in many speech signal processing tasks such as speech emotion
60 recognition (Bhavan et al., 2019; Tuncer et al., 2021) or allophones classification (Piotrowska et al.,
61 2019). These studies focused on preparing an optimized feature vector and utilizing this vector in the
62 classification process. In the case of speech emotion recognition, the SVM classifier is used for
63 classification in the mentioned above works. According to Bhavan et al. (2010), SVMs provide
64 reasonably good estimates with lesser effort. In contrast, hidden Markov models and deep neural
65 networks are more challenging to build and train and require higher computational power and time.
66 In the work of Piotrowska et al. (2019), automatic classification methods, such as artificial neural
67 networks (ANNs), the k-nearest neighbor (kNN), and self-organizing maps (SOMs), are applied to
68 lateral allophone analysis and returned satisfactory results.

69 Also, we justify why the process of improving speech quality and intelligibility should be adaptive and
70 specific modifications may depend on the noise characteristics and be reinforced by them. Based on
71 the rate of change in intensity, noise can be classified into continuous, periodic, impulsive, and low-
72 frequency noise (Tsalera et al., 2020). Therefore, a stable noise profiling method is needed – stable in

73 terms of being consistent over a longer period of time (Yang and Ritzwoller, 2008). Possible speech
74 modifications must fit the disruption to provide the best results in terms of potential loss in
75 intelligibility because of the noise presence. It is because every disturbance has different characteristics
76 and impacts speech differently. However, it is more important to have the noise recognition process
77 repetitive and stable rather than classify a given type of noise as a babble speech or airport noise. Also,
78 noise signals with similar frequency characteristics should always be analogously classified to ensure
79 that the speech signal modification is appropriate and durable.

80 **II. MATERIAL AND METHODS**

81 **A. Extraction of signal features**

82 In the learning process, the Aurora noise dataset was employed (Hirsch and Pearce, 2000). The Aurora
83 database contains various speech recordings prepared mainly for speech recognition systems,
84 especially for distributed speech recognition (Kshirsagar and Falk, 2021; Bandela et al., 2021). The
85 noise database within the Aurora dataset has been prepared directly for speech processing, and it is,
86 therefore, appropriate for our research. The noise signals contained in the Aurora dataset are as
87 follows: airport, babble speech, car noise, exhibition, restaurant, street noise, subway, and train. Some
88 noises are reasonably stationary, e.g., the car noise and the recording in the exhibition hall. Others
89 contain non-stationary segments, e.g., recordings on the street and at the airport (Hirsch and Pearce,
90 2000). In addition, pink noise was generated as this noise type was not present in the Aurora database.
91 To be noted, pink noise is a signal with a frequency spectrum such that the power spectral density is
92 inversely proportional to the signal's frequency, i.e., the power per Hertz in pink noise decreases as
93 the frequency increases (<https://www.livescience.com/38464-what-is-pink-noise.html>). In pink
94 noise, each octave interval carries an equal amount of noise energy, so the sound of pink noise is
95 perceived as being even.



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

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

107 **1. Spectral centroid**

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

$$112 \quad SC = \frac{\sum_{n=0}^N f(n)X(n)}{\sum_{n=0}^N X(n)} \quad (1)$$

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

115 **2. Spectral bandwidth**

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

$$118 \quad SBW = (\sum_{n=0}^N X(n)(f(n) - SC)^p)^{1/p} \quad (2)$$

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

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

123 **3. Spectral flatness**

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

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

$$131 \quad SF = \frac{[\prod_{n=0}^{N-1} PX(n)]^{1/N}}{\frac{1}{N} \sum_{n=0}^{N-1} PX(n)} \quad (3)$$

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

$$133 \quad PX(n) = \frac{1}{N} \sqrt{X(n)_{re}^2 + X(n)_{im}^2} \quad (4)$$

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

136 **B. Noise type recognition model**

137 Based on the previously described frequency characteristics, the recognition models were built. For
138 that purpose – as already mentioned – several baseline algorithms were employed, i.e., Naïve Bayes
139 (NB), linear SVM (Support Vector Machines), SVM with the polynomial kernel, Gaussian process
140 classifiers, Decision tree (DT), Random Forest (RF), Multilayer Perceptron (MLP), AdaBoost

141 classifier, and Quadratic Discriminant Analysis (QDA). For both learning and evaluation, the scikit-
142 learn modules from the Python environment were used (<https://scikit-learn.org/stable/>).

143 Every recording containing noise was processed in the following way:

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

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

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

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

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

162 A posteriori probability was calculated using the following formula:

$$163 \quad P(C_k | \mathbf{X}) = \frac{P(C_k)P(\mathbf{X}|C_k)}{P(\mathbf{X})} \quad (5)$$

164 where \mathbf{X} represents the vector with n conditionally independent features X_1, X_2, \dots, X_n , and C_k is a
165 possible outcome class.

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

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

$$168 \quad K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) \quad (6)$$

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

172 ***SVM with polynomial kernel (sklearn.svm.SVC)***

173 The following parameters of the polynomial SVM were implemented: regularization parameter $C =$
174 1, gamma coefficient (γ) set to auto (which means that it uses $1/\text{number_features}$), probability
175 estimates were enabled, independent term in kernel function equals 0, tolerance for stopping criterion
176 is equal to 0.001.

177 ***Gaussian process classifiers (GPCs) (sklearn.gaussian_process.GaussianProcessClassifier*** 178 ***module)***

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

$$181 \quad k_{exp}(\mathbf{X}, \mathbf{Y}) = k(\mathbf{X}, \mathbf{Y})^p \quad (8)$$

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

$$184 \quad K(\mathbf{x}_i, \mathbf{x}_j) = \left(1 + \frac{(\mathbf{x}_i - \mathbf{x}_j)^2}{2\alpha l^2}\right)^{-\alpha} \quad (9)$$

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

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

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

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

192 ***Random Forest (sklearn.ensemble.RandomForestClassifier module)***

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

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

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

200 ***AdaBoost classifier (sklearn.ensemble.AdaBoostClassifier module)***

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

204 ***Quadratic Discriminant Analysis***

205 ***(sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis module)***

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

210 III. COMPARISON OF THE CLASSIFIER RESULTS

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

- 214 – overall accuracy for the whole prediction process,
- 215 – precision, recall, and F1-score for every class.

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

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

- 221 – TP_n – the number of true positive recognitions for distortion type n (e.g., subway),
- 222 – TN_n – the number of true negative recognitions for distortion type n ,
- 223 – FP_n – the number of false positive recognitions for distortion type n (in other words – the number
224 of samples recognized incorrectly as type n),
- 225 – FN_n – the number of false negative recognitions for distortion type n (in other words – the number
226 of n distortion samples recognized as something different than type n).

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

$$229 \text{Acc} = \sum_n \frac{TP_n}{N} \quad (10)$$

230 In other words – it is a sum of true positives for all distortion types divided by the number of samples
231 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:

$$Precision_n = \frac{TP_n}{TP_n + FP_n} \quad (11)$$

Recall for type n is defined as follows:

$$Recall_n = \frac{TP_n}{TP_n + FN_n} \quad (12)$$

F1-score for type n is defined as follows:

$$F1score_n = 2 \cdot \frac{Precision_n * Recall_n}{Precision_n + Recall_n} \quad (13)$$

Tables I-III show the comparison of the above-described classification models. Also, metrics such as P – precision, R – recall, $F1$ – 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, $F1$ – 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

247
248 TABLE II. Results of classification using Gaussian process, Decision tree, and Random forest
249 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

250

251 TABLE III. Results of the classification using MLP, AdaBoost, and QDA classification models. All

252 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

253

254 One can notice that most tested algorithms give sufficiently good results with an accuracy of over
 255 90%; however, only three have better accuracy than 96%, i.e., Naïve Bayer, Linear SVM, and Decision
 256 Tree. For all three algorithms, all other metrics (averages of precision, recall, and F1 for all noise types)
 257 are similar; however, Naïve Bayer is a little better than Linear SVM and Decision Tree. The

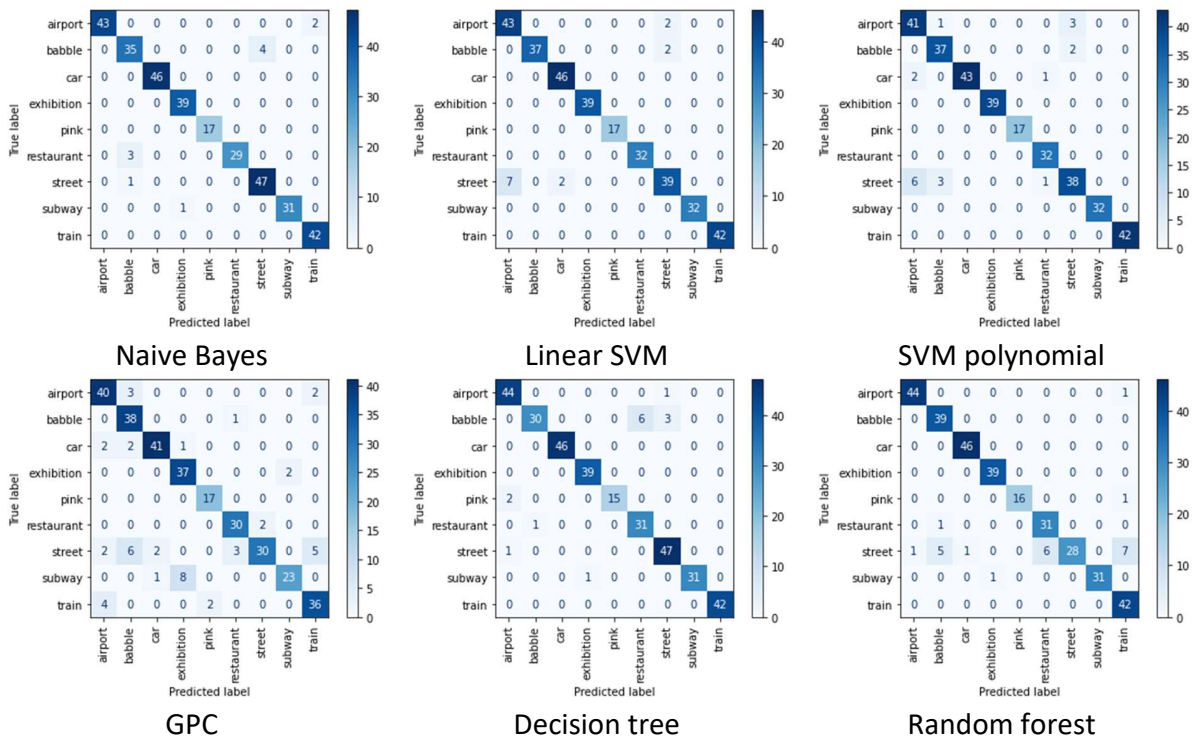
258 computational complexity for inference for all these methods is also similar and linearly dependent on
 259 the number of dimensions (for Linear SVM and Naive Bayer) or the number of tree depths for the
 260 Decision Tree.

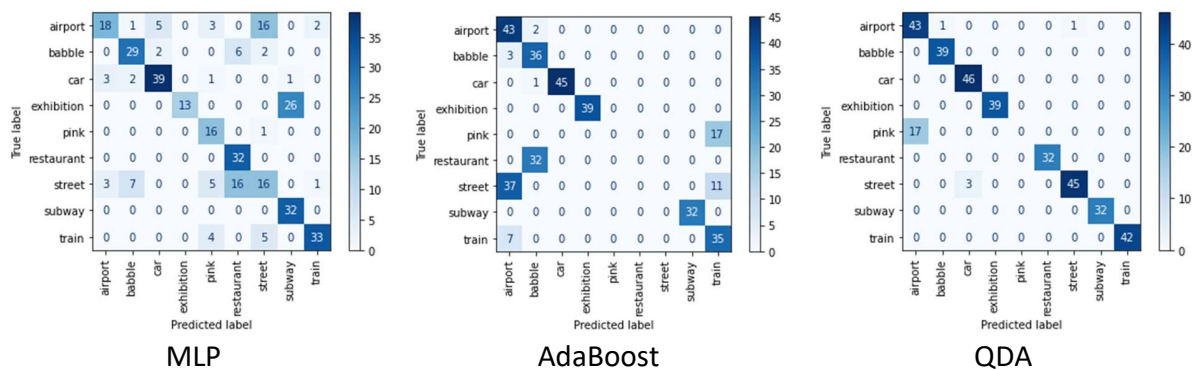
261 The other algorithms are not as accurate as the three mentioned above. Some of them have no true
 262 positives for some noise types, which results in zeroing the basic metrics for these types. This can be
 263 observed in Figure 1 (e.g., pink noise recognition for the AdaBoost classifier). That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

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

269

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





271 FIG. 1. Confusion matrices for all tested models.

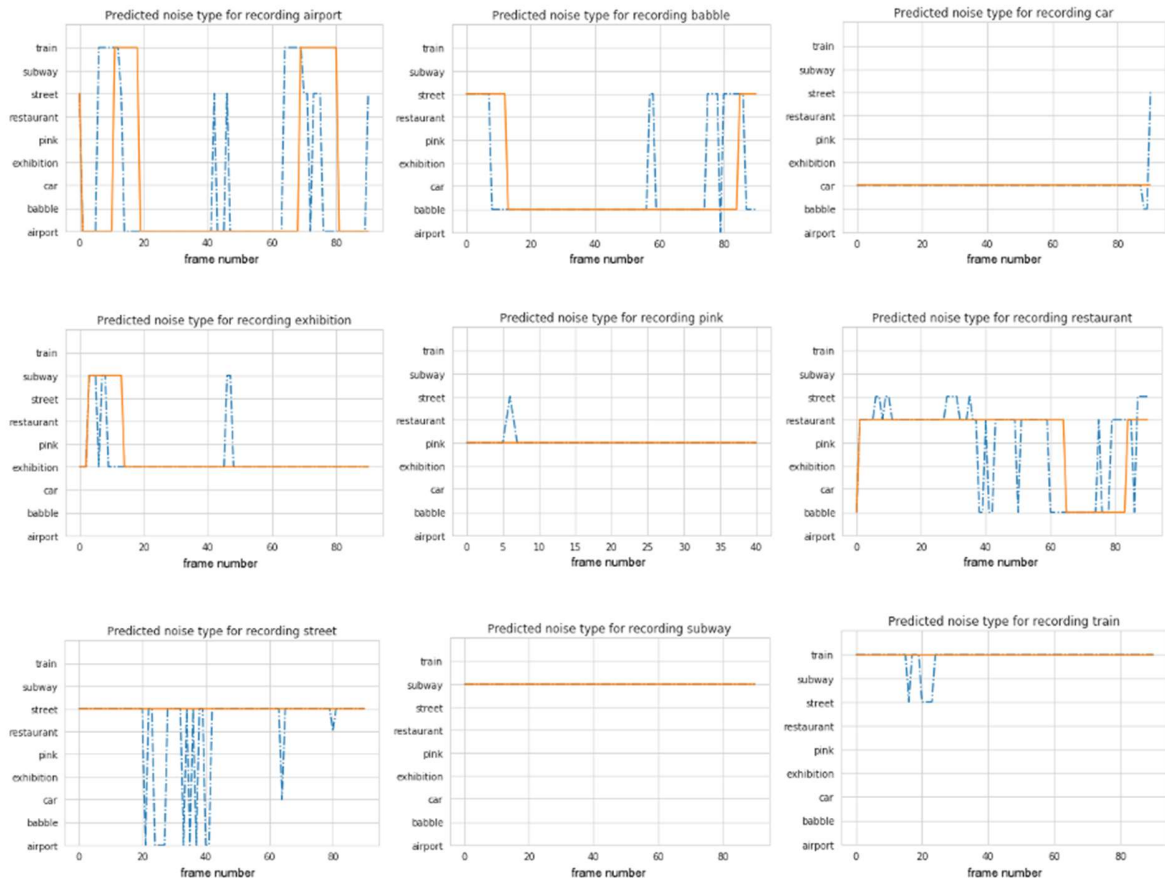
272 **IV. DISCUSSION**

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

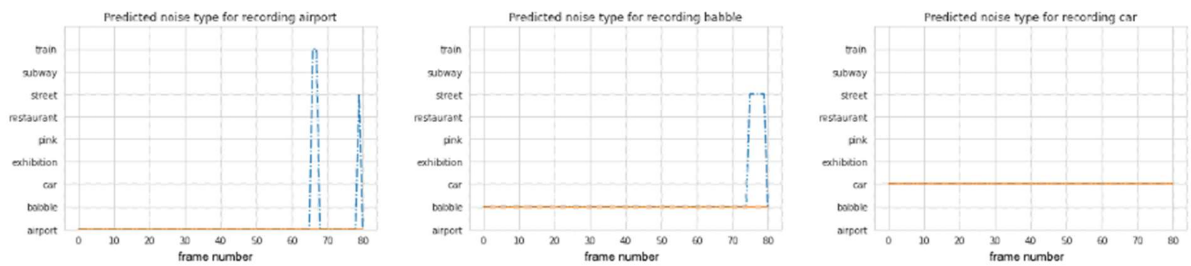
289 The recognition process was carried out in two modes: momentary and averaging. In both modes, the
290 window/frame analyzed was 1 or 2 seconds, and the window was moved by 0.1 seconds with every
291 step. In the momentary mode, classification was performed for every frame. In the averaging mode,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

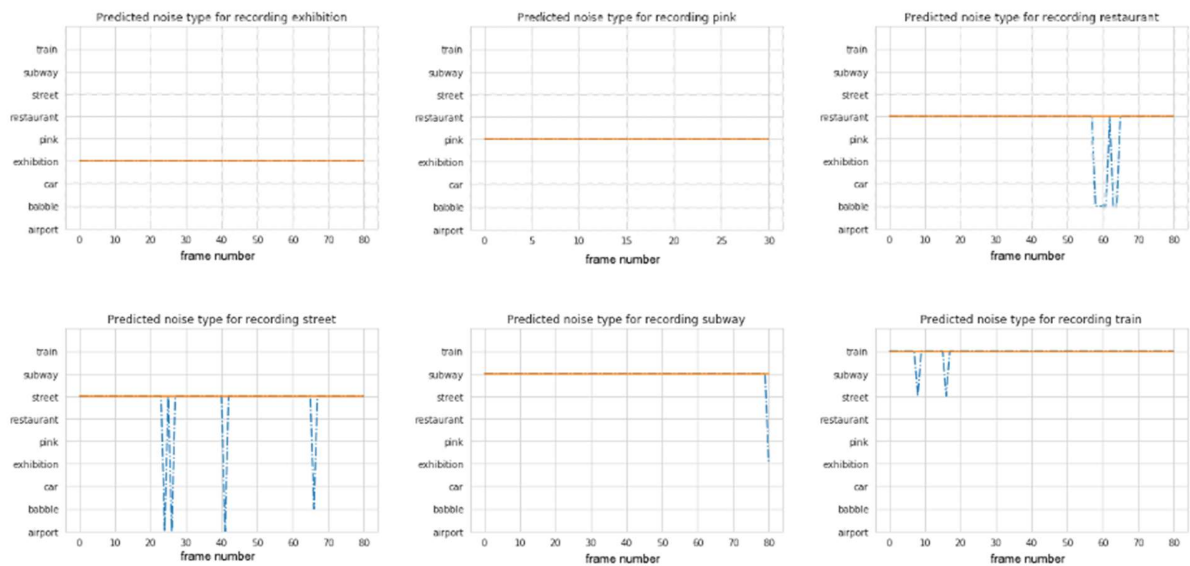
297 Thanks to this procedure, the recognition model avoids a temporary disturbance, usually caused by
298 non-stationary noise.

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



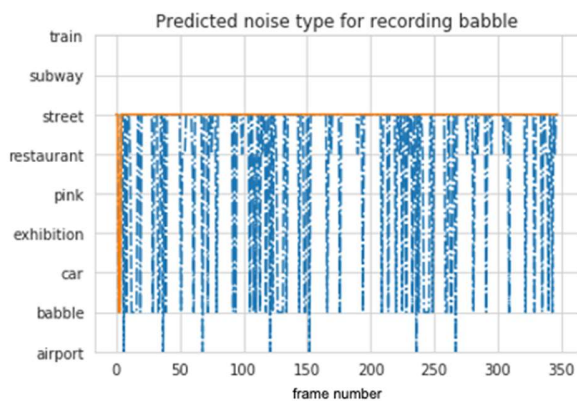
306 FIG. 2. Classification results on the real-life recordings using a 1-second-length frame (dashed line –
 307 result from momentary mode, solid line – result from averaging model).





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

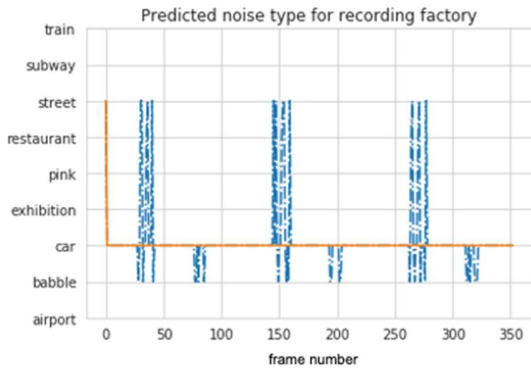
310 The recognition process was also performed on a completely different set of noise recordings
 311 contained in the MODALITY multimodal corpus of English speech recordings (Rasmussen et al.,
 312 2006). The recordings used in this test were very long (between 11 minutes 45 seconds and 14 minutes
 313 54 seconds). The test was performed only for a 2-second frame, and the window was moved by 2
 314 seconds (due to the overall recording length) with every step. The averaging was also used to remove
 315 random fluctuations in the recognition results. Figures 4-6 present recognition results, where dashed
 316 lines represent the single window classification and the solid line depicts the averaged result.



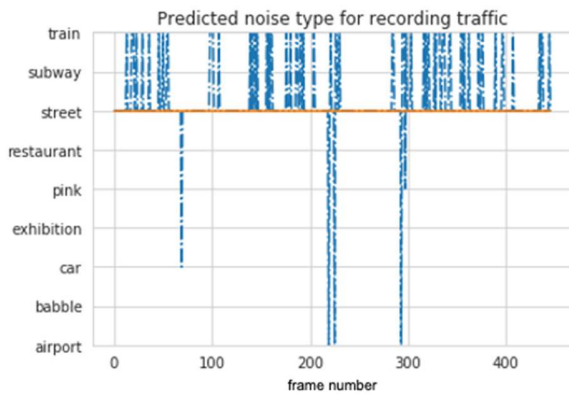
317

20

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



320
321 FIG. 5. An example where the “factory” recording was classified as “car noise” (there was no such
322 class as “factory” in the training set).



323
324 FIG. 6. An example where the recording “traffic” was classified as “street,” which is the correct
325 classification.

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

330 V. CONCLUSIONS

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

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

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

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

357

358 REFERENCES

359 Bandela, S. R., & Kumar, T. K. (2021). Unsupervised feature selection and NMF de-noising for robust
360 Speech Emotion Recognition. *Applied Acoustics*, 172, 107645, doi:
361 10.1016/J.APACOUST.2020.107645.

362 Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press. ISBN 978-
363 0-521-51814-7.

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

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

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

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

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

- 376 Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J., & Szykalski, M. (2017), An audio-visual corpus
377 for multimodal automatic speech recognition, *J. of Intelligent Information Systems*, pp. 1-26,
378 DOI: 10.1007/s10844-016-0438-z.
- 379 Das, N., Chakraborty, S., Chaki, J., Padhy, N., & Dey, N. (2021). Fundamentals, present and future
380 perspectives of speech enhancement. *International Journal of Speech Technology*, 24(4), 883-
381 901.
- 382 Dias, F. F., Ponti, M. A., & Minghim, R. (2022). Implementing simple spectral denoising for
383 environmental audio recordings. *arXiv preprint arXiv:2201.02099*.
- 384 Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of
385 classifiers to solve real world classification problems?. *Journal of Machine Learning Research*
386 15 (2014) 3133-3181
- 387 Ghojogh, B., & Crowley, M. (2019). Linear and quadratic discriminant analysis: Tutorial. *arXiv preprint*
388 *arXiv:1906.02590*.
- 389 Gosztolya, G. (2019). Posterior-thresholding feature extraction for paralinguistic speech classification.
390 *Knowledge-Based Systems*, 186, 104943, <https://doi.org/10.1016/j.knosys.2019.104943>.
- 391 Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv*
392 *preprint arXiv:2008.05756*.
- 393 Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining,*
394 *Inference, and Prediction*. Springer, New York, NY.
- 395 Hirsch, H. G., & Pearce, D. (2000). The Aurora experimental framework for the performance
396 evaluation of speech recognition systems under noisy conditions. In *ASR2000-Automatic*
397 *speech recognition: challenges for the new Millenium ISCA tutorial and research workshop*
398 *(ITRW)*.

- 399 Ho, T. K. (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference
400 on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.
- 401 Huang, Jin, & Charles, X. Ling. (2005). Using AUC and accuracy in evaluating learning algorithms.
402 IEEE Transactions on knowledge and Data Engineering, 17.3: 299-310.
- 403 ITU-T Recommendation P.563. (2004), “Single-ended method for objective speech quality assessment
404 in narrow-band telephony applications,” ITU-T Recommendation P.563.
- 405 Kamiński, B., Jakubczyk, M., & Szufel, P. (2017). A framework for sensitivity analysis of decision trees.
406 Central European Journal of Operations Research. 26 (1): 135–159. doi:10.1007/s10100-017-
407 0479-6. PMC 5767274. PMID 29375266.
- 408 Kavalekalam, M. S., Nielsen, J. K., Christensen, M. G., & Boldt, J. B. (2018). A study of noise PSD
409 estimators for single channel speech enhancement. In 2018 IEEE International Conference
410 on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5464-5468). IEEE.
- 411 Kąkol, K., Korvel, G., & Kostek, B. (2020). Improving Objective Speech Quality Indicators in Noise
412 Conditions. Studies in Computational Intelligence, vol. 869, 199-218.
413 <https://doi.org/10.1007/978-3-030-39250-5>.
- 414 Kim, M., & Shin, J. W. (2022). Improved Speech Enhancement Considering Speech PSD Uncertainty.
415 IEEE/ACM Transactions on Audio, Speech, and Language Processing.
- 416 Klapuri, A., & Davy, M. (Eds.). (2007). Signal processing methods for music transcription, chapter 5.
417 Springer Science and Business Media LLC.
- 418 Kong, Q., Xu, Y., Sobieraj, I., Wang, W., & Plumbley, M. D. (2019). Sound event detection and time–
419 frequency segmentation from weakly labelled data. IEEE/ACM Transactions on Audio,
420 Speech, and Language Processing, 27(4), 777-787, doi:
421 <https://doi.org/10.1109/TASLP.2019.2895254>.

- 422 Korvel, G., Kałkol, K., Kurasova, O., & Kostek, B. (2020). Evaluation of Lombard speech models in
423 the context of speech in noise enhancement. *IEEE Access*, 8, 155156-155170,
424 <https://doi.org/10.1109/access.2020.3015421>.
- 425 Korvel, G., Treigys, P., & Kostek, B. (2021). Highlighting interlanguage phoneme differences based
426 on similarity matrices and convolutional neural network. *The Journal of the Acoustical Society
427 of America*, 149(1), 508-523.
- 428 Krčadinac, O., Šošević, U., & Starčević, D. (2021). Evaluating the Performance of Speaker
429 Recognition Solutions in E-Commerce Applications. *Sensors*. 21(18):6231.
430 <https://doi.org/10.3390/s21186231>.
- 431 Kshirsagar, S. R., & Falk, T. H. (2022). Quality-Aware Bag of Modulation Spectrum Features for
432 Robust Speech Emotion Recognition. *IEEE Transactions on Affective Computing*, doi:
433 10.1109/TAFFC.2022.3188223.
- 434 Li, J. (2021). Recent Advances in End-to-End Automatic Speech Recognition, invited paper submitted
435 to *APSIPA Transactions on Signal and Information Processing*,
436 <https://arxiv.org/abs/2111.01690>.
- 437 Li J., Deng L., Haeb-Umbach, R., & Gong, Y. (2015). Robust automatic speech recognition: A bridge
438 to practical applications. Academic Press, Elsevier, [https://doi.org/10.1016/C2014-0-02251-](https://doi.org/10.1016/C2014-0-02251-4)
439 [4](https://doi.org/10.1016/C2014-0-02251-4).
- 440 Lin, T. H., & Tsao, Y. (2020). Source separation in ecoacoustics: a roadmap towards versatile
441 soundscape information retrieval. *Remote Sensing in Ecology and Conservation*, 6(3), 236-
442 247. <https://zslpublications.onlinelibrary.wiley.com/doi/epdf/10.1002/rse2.141>.
- 443 Lipton, Z. C., & Elkan, C., & Narayanaswamy, B. (2014). Thresholding classifiers to maximize F1
444 score. arXiv preprint arXiv:1402.1892.

- 445 Liu, S., Zhang, M., Fang, M., Zhao, J., Hou, K., & Hung, C. C. (2021). Speech emotion recognition
446 based on transfer learning from the FaceNet framework. *The Journal of the Acoustical Society*
447 *of America*, 149(2), 1338-1345.
- 448 McFee, B., Colin, R., Liang, D., Ellis, D. P. W., McVicar M., Battenberg E., & Nieto O. (2015). librosa:
449 Audio and music signal analysis in python. In *Proceedings of the 14th python in science*
450 *conference*, pp. 18-25.
- 451 Michalopoulou, Z. H., Gerstoft, P., Kostek, B., & Roch, M. A. (2021). Introduction to the special issue
452 on machine learning in acoustics. *The Journal of the Acoustical Society of America*, 150(4),
453 3204-3210, <https://doi.org/10.1121/10.0006783>
- 454 Morgan, M. M., Bhattacharya, I., Radke, R. J., & Braasch, J. (2021). Classifying the emotional speech
455 content of participants in group meetings using convolutional long short-term memory
456 network. *The Journal of the Acoustical Society of America*, 149(2), 885-894.
- 457 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,
458 Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher,
459 M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python;
460 12(85):2825–2830.
- 461 Piotrowska, M., Korvel, G., Kostek, B., Ciszewski, T., & Czyżewski, A. (2019). Machine learning-
462 based analysis of English lateral allophones. *International Journal of Applied Mathematics and*
463 *Computer Science*, 29(2).
- 464 Platt, J. (1999). Probabilistic outputs for SVMs and comparisons to regularized likelihood methods,
465 *Advances in Large Margin Classifiers*. In: *Advances in Large Margin Classifiers*, MIT Press.
- 466 Rasmussen, C.E., & Williams C.K.I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.
467 ISBN 978-0-262-18253-9.

- 468 Roch, M. A., Gerstoft, P., Kostek, B., & Michalopoulou, Z. H. (2021). How machine learning
469 contributes to solve acoustical problems. *Journal of the Acoustical Society of America*, 17(4),
470 17, 48-57, <https://doi.org/10.1121/at.2021.17.4.48>.
- 471 Rojas, R. (2009). *AdaBoost and the super bowl of classifiers a tutorial introduction to adaptive
472 boosting* (Tech. Rep.). Freie University, Berlin.
- 473 Srinivasan, T., Sanabria, R., & Metze, F. (2019). Analyzing utility of visual context in multimodal speech
474 recognition under noisy conditions. In arXiv preprint arXiv:1907.00477.
475 <https://scikit-learn.org/stable/> (last accessed November 2022).
- 476 Tuncer, T., Dogan, S., & Acharya, U. R. (2021). Automated accurate speech emotion recognition
477 system using twine shuffle pattern and iterative neighborhood component analysis techniques.
478 *Knowledge-Based Systems*, 211, 106547, <https://doi.org/10.1016/j.knosys.2020.106547>.
- 479 Tsalera, E., Papadakis, A., & Samarakou, M. (2020). Monitoring, profiling and classification of urban
480 environmental noise using sound characteristics and the KNN algorithm. *Energy Reports*, 6,
481 223-230. <https://doi.org/10.1016/j.egy.2020.08.045>.
- 482 Trujillo, J., Özyürek, A., Holler, J., Drijvers, L. (2021). Speakers exhibit a multimodal Lombard effect
483 in noise. *Sci Rep* 11, 16721. <https://doi.org/10.1038/s41598-021-95791-0>.
- 484 Watanabe, S., Delcroix, M., Metze, F., & Hershey, J. R. (Eds.) (2017). *New Era for Robust Speech
485 Recognition*. Springer International Publishing, doi:10.1007/978-3-319-64680-0.
- 486 Wu, T-F., Lin, C.-J., & Weng, R. C.-H. (2004). Probability estimates for multi-class classification by
487 pairwise coupling, *Journal of Machine Learning Research*, 5:975-1005.
- 488 Xenaki, A., & Bünsow Boldt, J., & Græsbøll Christensen, M. (2018). Sound source localization and
489 speech enhancement with sparse Bayesian learning beamforming. *The Journal of the
490 Acoustical Society of America*, 143(6), 3912-3921.

- 491 Xu, R., Wu, R., Ishiwaka, Y., Vondrick, C., & Zheng, C. (2020). Listening to sounds of silence for
492 speech denoising. *Advances in Neural Information Processing Systems*, 33, 9633-9648.
- 493 Yang, Y., & Ritzwoller, M. H. (2008). Characteristics of ambient seismic noise as a source for surface
494 wave tomography. *Geochemistry, Geophysics, Geosystems*, 9(2).
- 495 Zhang, H. (2004). The Optimality of Naïve Bayes, FLAIRS Conference, AAAI Press.
- 496 Zhu, H., Byrd, R. H., & Nocedal, J., (1997). Algorithm 778: L-BFGS-B, FORTRAN routines for large
497 scale bound constrained optimization, *ACM Transactions on Mathematical Software*, 23, 550-
498 560.
- 499 Zou, G., Antila, M., & Kataja, J. (2011). Practical active noise profiling in a passenger car. *Proc.*
500 *Akustiikkapäivät*, 11-12.