Intelligent Audio Signal Processing − Do We Still Need Annotated Datasets?

Bozena Kostek[0000-0001-6288-2908]

[1] Gdansk University of Technology, 81-230 Gdansk, g. Narutowicza 11/12, Poland
bokostek@audioakustyka.org

Abstract. In this paper, intelligent audio signal processing examples are shortly described. The focus is, however, on the machine learning approach and datasets needed, especially for deep learning models. Years of intense research produced many important results in this area; however, the goal of fully intelligent signal processing, characterized by its autonomous acting, is not yet achieved. Therefore, a review of state-of-the-art concerning this area is given. The aspect of showing the importance of acquiring an appropriate dataset containing audio samples dedicated to the task is also shown. The paper starts with samples of audio-related datasets resulting from the search engine inquiry. Then, examples of research studies along with results are given. Also, several works carried out by the author and her collaborators are presented. Some thoughts on future work are included with answering a question of whether annotated datasets are still needed.

Keywords: Intelligent signal processing, machine learning, datasets.

1. Introduction

The International Year of Sound, a global initiative envisioned to highlight the importance of sound and related sciences and technologies for all in society, was expected to be a year full of activities at regional, national, and international levels. However, the COVID-19 pandemic stopped these festivities even before they were due. Still, understanding sound, whatever its nature and role, i.e., communication, art, bridging the semantic gap in music, speech, and audio, i.e., looking for a strong correlation between human knowledge and audio object description that utilizes signal features and labels, audio-based warning systems, environmental control, etc., is of great importance, especially when an automated approach is sought to be applied. There is a lot of progress in all the mentioned topics, often called auditory scene recognition, as machine learning is no longer a frequented visitor but is interrelated with all the aspects of audio-related tasks. Overall, this concerns the audio ecosystem, defined as an inventory controlled by computer-based technology, which nowadays includes not only audio but also its users. Thus, whatever we listen to audio, talk to a machine, download or stream audio is duly noted and annotated, creating another record in a specific dataset.

In this paper, examples of research are cited, along with machine learning algorithms and datasets available for the effective training of machine learning algorithms and particularly deep models that solve audio signal-related problems.

2. Audio-Related Research

2.1. Machine Learning Applied to Audio-Related Topics

Overall, a variety of machine learning (ML) techniques are used in audio-related topics. They concern the so-called baseline algorithms, such as k-NN (*k-nearest neighbor*) decision trees, Random Forest (RF), SVM (Support Vector Machine*)* [1], Self Sequential Minimal Optimization (SMO), Organizing Maps (SOM), Naïve Bayes, XGBoost (regression models) [2], etc., or deep model architectures. Such models − to name a few − can be given as Convolutional Neural Networks (CNN), Multilayer Perceptrons, Generative Adversarial Networks (GANs), Recurrent NN (RNN), Long Short Term Memory Networks (LSTMs), autoencoders, etc.

These algorithms are well-known, so only some of them are recalled here as they cover either working principles of statistical models or the selection of the hyperparameters of deep models. The Naïve Bayesian classifier is a simple model that uses the assumption of independent variables, which is often not true [3]. The a posteriori maximum likelihood method is often used in such a case. A Random Forest concerns a set of independent random trees. Each random tree makes a decision based on given zero-one parameters placed at successive levels. The final decision of a random forest is the decision that is made more often among its constituent trees. Often, individual trees are trained on only a subset of the parameters to reach a solution of greater generality. Thus, a random forest consists of a large number of decision trees, each of which solves the problem individually in a binary fashion, and the final decision is the sum of the scores of the individual trees (the decision corresponds to the highest number of votes). Critical in this case is the low correlation between the trees. In this way, the trees protect each other from individual errors [4]. The algorithm is based on regression trees, which differ from decision trees in that the leaves contain an actual value instead of a binary decision. Classification is performed based on the values predicted by the corresponding leaves. Model learning is conducted in an additive manner, and the selection of the next splits considered is based on the gradient of the given loss function, with its second-order approximation used instead of the exact value. Weight regularization based on L1 and L2 norms is used. As in a random forest, multiple parallel trees can be used. If multiple trees are used, each tree is constructed based on some subset of samples from the learning set.

It is essential to mention that, regardless of the algorithm used, quality metrics should be used for the algorithm's efficiency. So, accuracy, precision, recall, F1-score, the Area under the Curve (AUC), Receiver Operating Characteristics (AUC-ROC) method [5], [6], or other measures are used to evaluate the classification quality. The AUC-ROC is a performance measurement for the classification problem at different threshold settings. The ROC represents the probability curve, while the AUC

represents the degree or measure of separability. The ROC curve is plotted using TPR (True Positive Rate) versus FPR (False Positive Rate).

Since the rise of GPU-based computations, audio-based studies are easier to be carried out, as the GPU-accelerated calculation can be employed [7]. Machine learning (ML) methods using deep structures are a rapidly growing field of knowledge, with multiple applications in areas of expertise such as speaker and source localization [8], audio rendering [9], analysis of acoustic signals originating from urban-related sources [10], [11], [12], [13], acoustic-based terrain classification [14], or music-related tasks [15], or even for obtaining results associated with "creative" tasks [16]. Reinforcement learning algorithms are a group of ML methods for tackling problems involving interaction with the environment and managing getting knowledge from such an interaction, so the agent interacts with the environment by taking actions. In return, it obtains feedback through the reward signal [17].

When using machine learning algorithms, critical is the hyperparameter selection as they impact the results obtained. Hyperparameter selection in reinforcement learning algorithms, especially ones based on a deterministic way of determining future actions, is the concept of greedy policies. It enhances the ability of the algorithm to explore the state space instead of exploiting known trajectories [18]. It should also be noted that much of the research is carried out in the Python environment, and deep models are implemented using the Keras framework and functional API [19]. So, the proposed architecture, the number of filters, kernels, activation function, etc., are easily available to build one's model.

In a recent paper by Lerch and Knees, a review of audio-related papers belonging to the special issue on Machine Learning Applied to Music/Audio Signal Processing [20,] is to be found. This is a subset of subjects and machine learning approaches covering the area; however, these papers are state-of-the-art [21]−[35], so some of the techniques are to be recalled in the context of their practical application.

Recurrent Convolutional Network and HRNet are used in vocal/singing voice separation [21][23]. U-net architecture is applied to jazz bass transcription [24]. Some other deep models, such as Convolutional Neural Networks (CNN) [25] – investigate polyphony; Convolutional NMF − an integration of Non-Negative Matrix Factorization (NMF) with Convolutional Neural Network [27] – solves the problem of drum mixture decomposition. Generative Adversarial Networks [30] is a basis for the restoration of compressed musical audio. Also, automatic melody harmonization is carried out using reinforcement learning [35].

Additional examples of such will be given further on when referring to research carried out by the author and her collaborators and students.

2.2. Audio Datasets

This Section should start with whether we still need datasets annotated and how large they should be. For the purpose of audio-related classification tasks, when baseline algorithms were employed, individuals added tags and labels to music files manually. Such a method of dataset creation is not only arduous but has a subjective bias and is also time-consuming. Another approach is called *social tagging*, when a statistically

significant number of people participate in the process, resulting in *collaborative filtering*. Still, the process is time-taking and biased by a person's performance and choices; however, this is used in music social services. In contrast, based on low-level descriptors, labeling may be automatized to some extent. This method searches for similarities within the audio signals to carry out automatic tagging. However, parameterization quality depends on the algorithm used in automatic classification, so the process is not without problems. Nevertheless, many datasets were created in music for challenges, e.g., GZTAN, ISMIR2004, MIREX2005, Million Song Dataset (Magnatune), ISMIS [36], SYNAT, containing 1,000 to approx. 50,000 music excerpts [37]. In parallel, manually annotated datasets for speech (e.g., MODALITY [38], SAVEE (Surrey Audio-Visual Expressed Emotion) [39], RAVDESS [40], TESS [41], and environmental sound [42] processing were created, often containing more than one modality. Deep learning brought new, often huge datasets, especially when compared to the existing ones, as these models are greedy and need a lot of data. It should be noted that many corporations created technology and devices to gather data when communicating with them. However, for many years, the datasets were still manually labeled, and only recently has the process been revolutionized.

It is worth looking for available audio datasets through a search machine, one of a wide array of existing search engines. For example, Google opens such lists with "40 Open-Source Audio Datasets for ML," a collection of datasets covering 2 TBs of labeled audio datasets. They are publicly available and parseable on DagsHub [43]. For each sample, additional information such as 2D representation in the form of waveforms, spectrograms, as well as file metadata is available. Moreover, these datasets contain seven languages and refer to various domains and sources. Then comes a description of another source of the labeled 25 Large audio datasets [44]. This source contains over 1.5 TBs of Labeled Audio Datasets. They are music-related, i.e., Free Music Archive, Million Song dataset, speech such as Free Spoken Digit, LibriSpeech, VoxCeleb, The Spoken Wikipedia Corpora, FlickrCaptionCorpus, and many others. Among them, one can find datasets used in various challenges. They cover audio-only, audio in noise, audio-video, emotions contained in audio, a combination of them, or multimodal data. Other categories belong to nature-related datasets, containing bird sounds and environmental or urban audio. The Google list continues with 6,182 machine learning datasets [45], some of them can be found under various links, but some others – are unique. Most of the datasets are open to the public [46], even though examples of commercial usage are also to be seen.

Another example is AudioSet which consists of a collection of 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos that are structured in the form of an ontology of audio event classes [47].

A comprehensive explanation behind the audio event categories is given by the dataset creators. They list several recommendations such as, e.g., the categories should refer to sounds existing in real-world, and the relationship between the sound and its label should be obvious, the labels should be unique in the sense that a listener can easily identify the sound and the assigned label, the hierarchy should also be easily identified and its structure not too deep, etc. On the other hand, it is assumed that the ontology may be expandable. Obviously, this is not the only attempt to categorize

audio datasets. There are many such examples in the literature [42][48] [49][50][51][52][53].

Finally, the characteristics of the datasets are diverse and given in hours, bytes, events, classes, languages, samples, speakers, and – as already mentioned – containing different modalities, e.g., MODALITY [38], etc., and their availability seems to grow exponentially.

3. Examples of Audio-Related Work Performed

Examples of research work carried out by the author and their collaborators and students are plenty, so it would not be possible to recall them all. So, this Section starts with some topics researched, and then an example of such will be shown in more detail. Overall, they concern music, speech, and ambient noise recognition, i.e., the mood in music classification [54]; classifying emotions in film music [55]; discovering the relationship between personality types and preferred music genres [56]; musical instrument identification [57], [58]; detection of lexical stress errors in non-native English speakers [59]; analysis of 2D feature spaces for deep learning-based speech recognition [60]; highlighting interlanguage phoneme differences [61]; cross-linguistic speech emotion recognition [62]; recognition of types of vehicle-associated noise [63]; acoustic sensing analytics applied to speech in reverberation [64].

Among the works performed are such as musical instrument sound separation and identification supported by baseline algorithms as well as deep learning models [57], [58]. For the purpose of the research carried out with the baseline algorithms, a dataset of approximately 50,000 audio excerpts was created [37]. In contrast, for the deep learning approach, data from the Slakh dataset [65] was used, which contains 2100 audio tracks with aligned MIDI files, and separate instrument stems along with tagging. It should be noted that the deep model-based approach brought higher accuracy, though the task was, to some extent – different. The classification effectiveness was close to 100% in some instrument configurations [58]. However, the first study [57] focused on automatic music genre classification while using the original and separated tracks. The instrument separation approach was selected to improve the results of music genre classification and, in particular, decrease the misclassification between selected genres in the context of the influence of the specific instrument on selected genres [57]. Moreover, such studies employing baseline and deep models, even though devoted to the same task, are comparable only to some extent. This is because, along with a difference in ML approach, different feature extraction is applied. At the same time, it should be pointed out that feature extraction, regardless of its representation, i.e., a feature vector, 2D representation, or an audio stream, plays a crucial role in the overall audio recognition process.

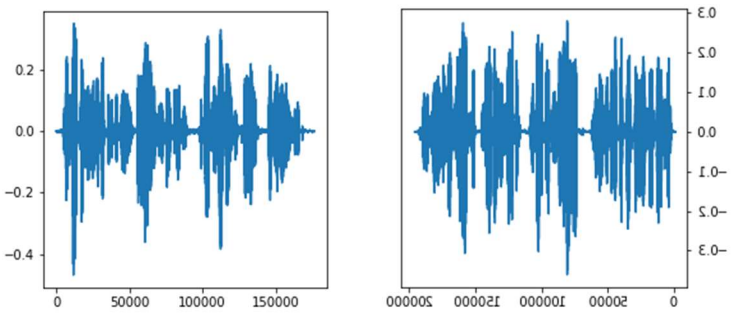
Earlier research used a variety of algorithms and their combination. Notably, a fuzzy logic-based approach was combined with other ML techniques when applied to the automated evaluation of singing voice quality [66] or gesture-based system for sound mixing [67]. In both cases, datasets were created by the authors as the recognition tasks were unique at that time.

Another example focuses on showing the problem of speech analysis in the presence of ambient noise [68].

The environmental noise changes the manner of expression. This concerns the so-called Lombard effect, which involuntary affects speech production. Speech with the Lombard effect is more intelligible in noisy environments than normal/neutral speech. So, the idea behind this study was that the speech synthesis model might retain Lombard effect characteristics. Therefore, the main goal of the experiment was to check how the Lombard-based speech models are recognizable and at what type of noise and threshold a particular model stops working. The investigations were carried out in the context of speech enhancement. So, the ultimate goal is to prepare a system capable of synthetically generating Lombard speech through noise profiling for enhancing speech automatically in the presence of noise.

An illustration of discerning between neutral and Lombard speech by CNN is shown in Fig. 1 (a. neutral speech, b. Lombard speech; the upper part corresponds to the time-domain signals, the middle part – spectral domain – mel spectrogram visualization; the lower part to the recognition by employing CNN). It refers to a sentence uttered by a male speaker in the presence of noise. The system is capable of recognizing this effect in speech. The data include information on F0, and the first two MFCC (Mel Frequency Cepstrum Coefficients) for the entire recordings (Number of samples used for training: 3156, no. of samples for validation: 790; accuracy on validation set: 0.9899; loss on validation set: 0.0370; recognition accuracy: 0.9594; precision: 0.9681; recall: 0.95).

a. b.



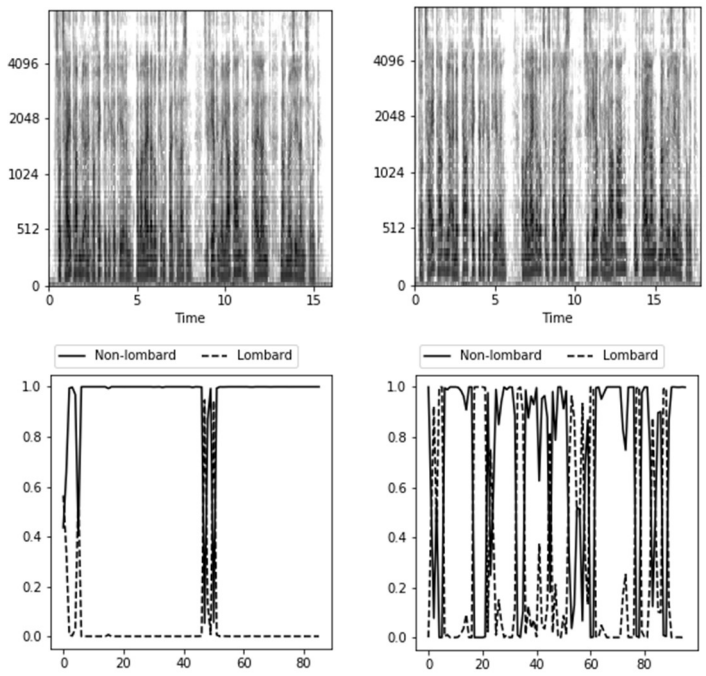


Fig. 1. Discerning between neutral and Lombard speech by CNN; a. neutral speech, b. Lombard speech; the upper part corresponds to the time-domain signals, the middle part – spectral domain – mel spectrogram visualization; the lower part to the recognition by employing CNN).

Conclusion

In conclusion, challenges that could be identified within the audio processing technology area are related to the role of human factors such as, for example, the user's personality and experience, emotions in the user's models, the naturalness of the processed sound, and personalized services. This means that machine learning-based approaches should have a built-in evaluation module, mimicking the human way of assessing the quality of audio or judging the match between the audio content and the application envisioned. These elements still need a lot of attention. In contrast, in the era of statistical and deep models, it occurred that manually annotated audio sets may no longer be required as two domains, i.e., analysis and synthesis, become one. So, instead of recording and tagging audio, synthesized audio signals are employed. Their quality is not any more questionable; thus, speech can be reconstructed from the TTS (Text-to-Speech) model outputs [69]. Moreover, deep models help to transfer speech style, so voice conversion across different speakers is now achievable−this time, it is based on the speech-to-speech approach [70]. The same notion refers to music–music style transfer between musical pieces. The family of such tasks is called

one-shot music style transfer [71]. This may sound like a simple task, but it has at least several elements to be dealt with, such as overall composition and performance translating into accompaniment, harmonic structure, timbre, etc., transfer [72].

Finally, a concept presented in a recent paper entitled "Computer-assisted pronunciation training—Speech synthesis is almost all you need" may be of interest in this context as it answers the posed question [73].

References

1. Candel, D., Nanculef, R., Concha, C., Allende, H.: A Sequential Minimal Optimization Algorithm for the All-Distances Support Vector Machine, CIARP 2010, LNCS 6419, Springer Verlag, Berlin, 484-491 (2010).
2. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System, KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794 (2016). https://doi.org/10.1145/2939672.2939785 v
3. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers, Machine Learning, 29, 139-164 (1997).
4. Yiu, T., Understanding Random Forest. How the Algorithm Works and Why it Is So Effective, Towards Data Science, https://towardsdatascience.com/understanding-random-forest-58381e0602d2 /last accessed 2022/06/21.
5. Classification: ROC Curve and AUC Machine Learning Crash Course Google Developers. Available online: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc /last accessed 2022/06/21.
6. Narkhede, S., Understanding AUC – ROC Curve, Towards Data Science, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 /last accessed 2022/06/21.
7. Cao, X., Cai. Y., Cui, X.: A parallel numerical acoustic simulation on a GPU using an edge-based smoothed finite element method, Adv. Eng. Softw., 148, 2020, doi: 10.1016/j.advengsoft.2020.102835 /last accessed 2022/06/21.
8. Bianco, M., Gerstoft, P., Traer, J., Ozanich, E., Roch, M., Gannot, S., Deledalle C.: Machine learning in acoustics: Theory and applications, J. Acoust. Soc. Am. 146(5) 3590 (2019). doi: 10.1121/1.5133944
9. Tang, Z., Bryan, N., Li, D., Langlois, T., Manocha, D.: Scene-Aware Audio Rendering via Deep Acoustic Analysis, IEEE Transactions on Visualization and Computer Graphics, 26 (5), 1991-2001, doi: 10.1109/TVCG.2020.2973058.
10. Huang, P., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (12), 2136-2147 (2015), doi:10.1109/TASLP.2015.2468583.
11. Kurowski, A., Zaporowski, S., Czyżewski, A.: Automatic labeling of traffic sound recordings using autoencoder-derived features, 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), Poznan, Poland, 38-43 (2019). doi: 10.23919/SPA.2019.8936709.
12. Naranjo-Alcazar, J., Perez-Castanos, S., Zuccarello, P., Cobos, M.: Acoustic Scene Classification with Squeeze-Excitation Residual Networks, IEEE Access 8, 112287–112296, 2020, doi: 10.1109/ACCESS.2020.3002761

13. Shen, Y., Cao, J., Wang, J., Yang, Z.: Urban acoustic classification based on deep feature transfer learning, J. Franklin Inst. 357 (1), 667–686 (2020). doi: 10.1016/j.jfranklin.2019.10.014.
14. Valada, A., Spinello, L. Burgard, W.: Deep Feature Learning for Acoustics-Based Terrain Classification, 21–37 (2018). doi: 10.1007/978-3-319-60916-4_2.
15. Avramidis, K., Kratimenos, A., Garoufis, C., Zlatintsi, A., Maragos, P.: Deep Convolutional and Recurrent Networks for Polyphonic Instrument Classification from Monophonic Raw Audio Waveforms. In Proceedings of the 46th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021), Toronto, Canada, 6–11 June 2021; pp. 3010–3014,https://doi.org/10.48550/arXiv.2102.06930.
16. Thoma, M.: Creativity in Machine Learning, ArXiv preprint no. 1601.03642, 2016, https://arxiv.org/abs/1601.03642 /last accessed 2022/06/21.
17. Kurowski, A., Kostek, B.: Reinforcement Learning Algorithm and FDTD-based Simulation Applied to Schroeder Diffuser Design Optimization. IEEE Access, 9, 136004-136017 (2021). https://doi.org/10.1109/access.2021.311462.
18. Buduma, N., Locasio, N.: Fundamentals of Deep Learning. Designing next-generation machine intelligence algorithms, O'Reilly Media, Inc., (2017).
19. The Functional API. Available online: https://keras.io/guides/functional_api/ /last accessed 2022/06/21.
20. Lerch, A., Knees P.: Machine Learning Applied to Music/Audio Signal Processing. Electronics 10, no. 24: 3077 (2021). https://doi.org/10.3390/electronics10243077.
21. Zhang, X., Yu, Y., Gao, Y., Chen, X., Li, W.: Research on Singing Voice Detection Based on a Long-Term Recurrent Convolutional Network with Vocal Separation and Temporal Smoothing. Electronics, 9, 1458 (2020).
22. Krause, M., Müller, M., Weiß, C.: Singing Voice Detection in Opera Recordings: A Case Study on Robustness and Generalization. Electronics, 10, 1214 (2021).
23. Gao, Y., Zhang, X., Li, W.: Vocal Melody Extraction via HRNet-Based Singing Voice Separation and Encoder-Decoder-Based F0 Estimation. Electronics 2021, 10, 298.
24. Abeßer, J., Müller, M.: Jazz Bass Transcription Using a U-Net Architecture. Electronics, 10, 670 (2021).
25. Taenzer, M., Mimilakis, S.I., Abeßer, J.: Informing Piano Multi-Pitch Estimation with Inferred Local Polyphony Based on Convolutional Neural Networks. Electronics, 10, 851 (2021).
26. Hernandez-Olivan, C., Zay Pinilla, I., Hernandez-Lopez, C., Beltran, J.R.: A Comparison of Deep Learning Methods for Timbre Analysis in Polyphonic Automatic Music Transcription. Electronics, 10, 810 (2021).
27. Vande Veire, L., De Boom, C., De Bie, T.: Sigmoidal NMFD: Convolutional NMF with Saturating Activations for Drum Mixture Decomposition. Electronics, 10, 284 (2021).
28. Pinto, A.S.; Böck, S.; Cardoso, J.S.; Davies, M.E.P. User-Driven Fine-Tuning for Beat Tracking. Electronics, 10, 1518 (2021).
29. Carsault, T., Nika, J., Esling, P., Assayag, G.: Combining Real-Time Extraction and Prediction of Musical Chord Progressions for Creative Applications. Electronics, 10, 2634 (2021).
30. Lattner, S., Nistal, J.: Stochastic Restoration of Heavily Compressed Musical Audio Using Generative Adversarial Networks. Electronics, 10, 1349 (2021).
31. Venkatesh, S., Moffat, D., Miranda, E.R.: Investigating the Effects of Training Set Synthesis for Audio Segmentation of Radio Broadcast. Electronics, 10, 827 (2021).
32. Grollmisch, S., Cano, E.: Improving Semi-Supervised Learning for Audio Classification with FixMatch. Electronics, 10, 1807 (2021).

33. Zinemanas, P.; Rocamora, M.; Miron, M.; Font, F.; Serra, X. An Interpretable Deep Learning Model for Automatic Sound Classification. Electronics, 10, 850 (2021).
34. Krug, A., Ebrahimzadeh, M., Alemann, J., Johannsmeier, J.;, Stober, S.: Analyzing and Visualizing Deep Neural Networks for Speech Recognition with Saliency-Adjusted Neuron Activation Profiles. Electronics, 10, 1350 (2021).
35. Zeng, T., Lau, F.C.M.: Automatic Melody Harmonization via Reinforcement Learning by Exploring Structured Representations for Melody Sequences. Electronics, 10, 2469 (2021).
36. Kostek, B., Kupryjanow, A., Zwan, P., Jiang, W., Ras, Z., Wojnarski, M., Swietlicka, J.: Report of the ISMIS 2011 Contest: Music Information Retrieval, Foundations of Intelligent Systems. ISMIS 2011, LNAI 6804, 715-724, M. Kryszkiewicz et al. (eds.), Springer Verlag, Berlin, Heidelberg (2011).
37. Kostek, B.; Music Information Retrieval in Music Repositories. Rough Sets and Intelligent Systems - Professor Zdzisław Pawlak in Memoriam. Vol. 1, 464-489 (2013). https://doi.org/10.1007/978-3-642-30344-9_17.
38. Czyżewski, A., Kostek, B., Bratoszewski, P., Kotus, J., & Szykulski, M.: An audio-visual corpus for multimodal automatic speech recognition. Journal of Intelligent Information Systems, 49(2), 167-192 (2017). https://doi.org/10.1007/s10844-016-0438-z.
39. Haq, P. Jackson, J.E.: Speaker-dependent audio-visual emotion recognition, in AVSP, Norwich, UK, pp. 53–58, Sept. 2009.
40. Livingstone, S.R., Russo, F.A.: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE, vol. 13, no. 5, 2018: e0196391. https://doi.org/10.1371/journal.pone.0196391.
41. Dupuis M.K.P.K.: Toronto emotional speech set (TESS), (2010). https://tspace.library.utoronto.ca/handle/1807/24487 / last accessed 2022/05/21.
42. Piczak, K.J.: ESC: Dataset for environmental sound classification. In: Proceedings of the ACM International Conference on Multimedia. ACM, 2015, pp. 1015–1018 (2015).
43. https://towardsdatascience.com/40-open-source-audio-datasets-for-ml-59dc39d48f06 / last accessed 2022/05/21.
44. https://towardsdatascience.com/a-data-lakes-worth-of-audio-datasets-b45b88cd4ad / last accessed 2022/05/21.
45. https://paperswithcode.com/datasets?mod=audio, last accessed 2022/05/21
46. https://www.twine.net/blog/100-audio-and-video-datasets/ last accessed 2022/05/21
47. Gemmeke J. F. et al.: Audio Set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776-780 (2017), doi: 10.1109/ICASSP.2017.7952261.
48. Salamon J., Jacoby C., Bello J. P.: A dataset and taxonomy for urban sound research. In: Proceedings of the ACM International Conference on Multimedia. ACM, pp. 1041–1044 (2014).
49. Mesaros, A., Heittola, T., Virtanen, T.: TUT database for acoustic scene classification and sound event detection. In: 24th European Signal Processing Conference (EUSIPCO), pp. 1128-1132 (2016), doi: 10.1109/EUSIPCO.2016.7760424.
50. Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., Plumbley, M. D.: Detection and Classification of Acoustic Scenes and Events. IEEE Transactions on Multimedia, 17, 10, 1733-1746 (Oct. 2015), doi: 10.1109/TMM.2015.2428998.
51. Fonseca, E., Favory X., Pons J., Font F., Serra X.: FSD50K: An Open Dataset of Human-Labeled Sound Events. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30 (2022), https://arxiv.org/pdf/2010.00475.pdf.

52. Hershey, S., Ellis, D. P. W., Fonseca, E., Jansen, A., Liu, C., Moore, R. C., Plakal, M.: The benefit of temporally-strong labels in audio event classification. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2021).
53. Foster, P., Sigtia, S., Krstulovic, S., Barker, J., Plumbley, M. D.: CHiME-home: A dataset for sound source recognition in a domestic environment. In: 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE (2015).
54. Kostek B., Plewa M.: Parametrisation and correlation analysis applied to music mood classification. Int. J. Comput. Intell. Stud. 2(1): 4-25 (2013).
55. Ciborowski, T., Reginis, S., Kurowski, A., Weber, D., Kostek, B.: Classifying Emotions in Film Music - A Deep Learning Approach. Electronics, 10, 2955v(2021). https://doi.org/10.3390/electronics10232955
56. Dorochowicz, A., Kurowski, A., & Kostek, B.: Employing Subjective Tests and Deep Learning for Discovering the Relationship between Personality Types and Preferred Music Genres. Electronics, 9, 2016 (2020). https://doi.org/10.3390/electronics9122016.
57. Rosner, A., Kostek, B.: Automatic music genre classification based on musical instrument track separation. J. Intell. Inf. Syst., 50(2), 363-384 (2018). https://doi.org/10.1007/s10844-017-0464-5.
58. Blaszke, M., Kostek, B.: Musical Instrument Identification Using Deep Learning Approach. Sensors, 22, 3033. (2022). https://doi.org/10.3390/s22083033.
59. Korzekwa, D., Barra-Chicote, R., Zaporowski, S., Beringer, G., Lorenzo-Trueba, J., Serafinowicz, A., Droppo, J., Drugman, T., Kostek, B.: Detection of Lexical Stress Errors in Non-Native (L2) English with Data Augmentation and Attention. (2021). https://doi.org/10.21437/interspeech.2021-86.
60. Korvel, G., Treigys, P., Tamulevicus, G., Bernataviciene, J., Kostek, B.: Analysis of 2D Feature Spaces for Deep Learning-based Speech Recognition. Journal of the Audio Engineering Society, 66(12), 1072-1081 (2018). https://doi.org/10.17743/jaes.2018.0066.
61. Korvel, G., Treigys, P., Kostek, B.: Highlighting interlanguage phoneme differences based on similarity matrices and convolutional neural network. Journal of the Acoustical Society of America, 149, 508-523 (2021). https://doi.org/10.1121/10.0003339.
62. Tamulevicius, G., Korvel, G., Yayak, A. B., Treigys, P., Bernataviciene, J., Kostek, B.: A Study of Cross-Linguistic Speech Emotion Recognition Based on 2D Feature Spaces. Electronics, 9, 1725 (2020). https://doi.org/10.3390/electronics9101725.
63. Kurowski, A., Marciniuk, K., B.: Separability Assessment of Selected Types of Vehicle-Associated Noise. MISSI 2016: 113-121 (2016).
64. Odya, P., Kotus, J., Kurowski, A., Kostek, B.: Acoustic Sensing Analytics Applied to Speech in Reverberation Conditions. Sensors, 21, 6320 (2021). https://doi.org/10.3390/s21186320
65. Slakh Demo Site for the Synthesized Lakh Dataset (Slakh). Available online: http://www.slakh.com/ last accessed 2022/06/20.
66. Żwan, P., Kostek, B.: System for automatic singing voice recognition. Journal of the Audio Engineering Society, 56(no. 9), 710-723 (2008).
67. Lech, M., Kostek, B., Czyzewski, A.: Examining Classifiers Applied to Static Hand Gesture Recognition in Novel Sound Mixing System. MISSI 2012: 77-86 (2012).
68. Korvel, G., Kąkol, K., Kurasova, O., Kostek, B.: Evaluation of Lombard Speech Models in the Context of Speech in Noise Enhancement. IEEE Access, 8, 155156-155170 (2020) https://doi.org/10.1109/access.2020.3015421
69. Ezzerg, A., Gabrys, A., Putrycz, B., Korzekwa, D., Saez-Trigueros, D., McHardy, D., Pokora, K., Lachowicz, J., Lorenzo-Trueba, J., Klimkov, V.: Enhancing audio quality for

expressive Neural Text-to-Speech. Proc. 11th ISCA Speech Synthesis Workshop (SSW 11), 78-83, (2021). doi: 10.21437/SSW.2021-14.
70. AlBadawy, E.A., Lyu, S.: Voice Conversion Using Speech-to-Speech Neuro-Style Transfer. Proc. Interspeech 2020, 4726-4730 (2020). doi: 10.21437/Interspeech.2020-3056.
71. Cífka, O., Şimşekli, U., G. Richard, G.: Groove2Groove: One-Shot Music Style Transfer With Supervision From Synthetic Data. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 2638-2650 (2020). doi: 10.1109/TASLP.2020.3019642.
72. Mukherjee, S., Mulimani, M.: ComposeInStyle: Music composition with and without Style Transfer. Expert Systems with Applications, 191, 116195 (2022). https://doi.org/10.1016/j.eswa.2021.116195.
73. Korzekwa, D., Lorenzo-Trueba, J., Drugman, T., Kostek, B.: Computer-assisted pronunciation training—Speech synthesis is almost all you need. Speech Communication, 142, 22-33 (2022) https://doi.org/10.1016/j.specom.2022.06.003.