Selective Hearing: A Machine Listening Perspective

Estefanía Cano
Semantic Music Technologies
Fraunhofer IDMT
Ilmenau, Germany
cano@idmt.fraunhofer.de

Hanna Lukashevich
Semantic Music Technologies
Fraunhofer IDMT
Ilmenau, Germany
lkh@idmt.fraunhofer.de

***Abstract*—Selective hearing (SH) refers to the listeners' capability to focus their attention on a specific sound source or a group of sound sources in their auditory scene. This in turn implies that the listeners' focus is minimized for sources that are of no interest. This paper describes the current landscape of machine listening research, and outlines ways in which these technologies can be leveraged to achieve SH with computational means. To do so, a brief overview of the state-of-the-art in the fields of detection, classification, separation, localization and enhancement of sound sources is presented, highlighting recent advances in each field, and drawing connections between them. Two main challenges lie ahead in the development of SH applications: (1) Unified methods that can jointly detect/classify/localize and separate/enhance sound sources are required to provide both the flexibility and robustness required for real-life SH. (2) Low-latency methods suitable for real-time performance are critical when dealing with the dynamic nature of real-life auditory scenes.**

***Index Terms*—Selective Hearing, Machine Listening, Audio Event Detection**

I. Introduction

With the remarkable advances in deep learning, machine listening, and smart hearables in the last years, the development of devices that could enable listeners to selectively modify their auditory scene is a step closer to reality. In this paper, we address the concept of *selective hearing (SH)* from a computational perspective, and more formally define it as the possibility of a listener to selectively enhance, attenuate, suppress or modify sound sources in the auditory scene by means of a hearing device such as headphones, earbuds, etc. Figure 1 presents a SH scenario as considered in this article: The user is the center of their auditory scene. In this case, four external sound sources (S1-S4) are active around the user. A user interface allows the listener to manipulate the auditory scene. Sources S1-S4 can be attenuated, enhanced or suppressed with their corresponding sliders. As seen in Figure 1, the listener can define sound sources or events that should be retained or suppressed from the auditory scene. In Figure 1, the background noise from the city should be suppressed, whereas alarms or telephones ringing should be retained. At all times, the user has the possibility to play an additional audio stream such as music or the radio through the hearing device.

978-1-7281-1817-8/19$31.00 ©2019 European Union

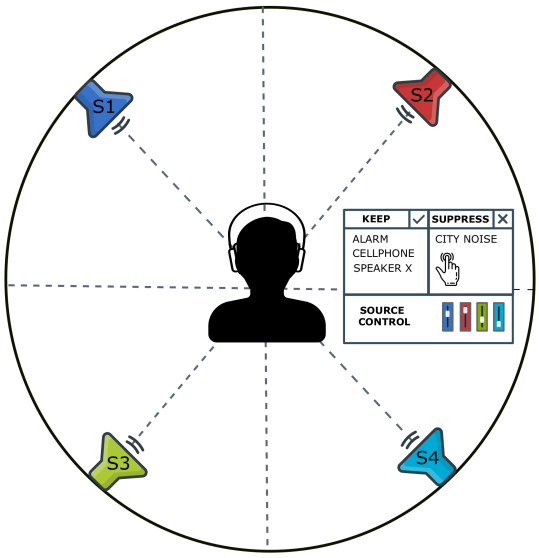

Fig. 1. Selective hearing scenario

The concept of selective hearing is related to other terms in the literature such as assisted listening [1], virtual, and augmented auditory environments [2]. Assisted listening is an overarching term that includes virtual, augmented, and SH applications. We differentiate selective hearing from virtual and augmented auditory environments by constraining selective hearing to those applications where only real audio sources in the auditory scene are modified, without attempting to add any virtual sources to the scene.

From a machine listening perspective, selective hearing applications require technologies to automatically detect, locate, classify, separate, and enhance sound sources. To further clarify the terminology around selective hearing, we define the following terms, highlighting their differences and relationships:

Sound Source Localization refers to the ability to detect the position of a sound source in the auditory scene. In the context of audio processing, source location usually refers to the direction-of-arrival (DOA) of a given source, which can be given either as a 2-D coordinate (azimuth), or as a 3-D coordinate when it includes elevation. Some systems also estimate the distance from the source to the microphone as

location information [3]. In the context of music processing, location often refers to the panning of the source in the final mixture, and is usually given as an angle in degrees [4].

Sound Source Detection refers to the ability to determine whether any instance of a given sound source type is present in the auditory scene. An example of a detection task is to determine whether any speaker is present in the scene. In this context, determining the number of speakers in the scene or the identity of the speakers is beyond the scope of sound source detection. Detection can be understood as a binary classification task where the classes correspond to "source present" and "source absent".

Sound Source Classification assigns a class label from a set of predefined classes to a given sound source or event. An example of a classification task is to determine whether a given sound source corresponds to speech, music, or environmental noise. Sound source classification and detection are closely related concepts. In some cases, classification systems encapsulate a detection stage by considering "no class" as one the possible labels. In these cases, the system implicitly learns to detect the presence or not of a sound source, and is not forced to assign a class label when there is not enough evidence of any of the sources being active.

Sound Source Separation refers to the extraction of a given sound source from an audio mixture or an auditory scene. An example of sound source separation is the extraction of the singing voice from an audio mixture, where besides the singer, other musical instruments are playing simultaneously [5]. Sound source separation becomes relevant in a selective hearing scenario as it allows to suppress sound sources that are of no interest to the listener. Some sound separation systems implicitly perform a detection task before extracting the sound source from the mixture. However, this is not necessarily the rule and hence, we highlight the distinction between these tasks. Additionally, separation often serves as a pre-processing stage for other types of analysis such as source enhancement [6] or classification [7].

Sound Source Identification goes a step further and aims to identify specific instances of a sound source in an audio signal. Speaker identification is perhaps the most common use of source identification today. The goal in this task is to identify whether a specific speaker is present in the scene. In the example in Figure 1, the user has chosen "speaker X" as one of the sources to be retained in the auditory scene. This requires technologies beyond speech detection and classification, and calls for speaker-specific models that allow this fine-grained identification.

Sound Source Enhancement refers to the process of increasing the saliency of a given sound source in the auditory scene [8]. In the case of speech signals, the goal is often to increase their perceptual quality and intelligibility. A common scenario for speech enhancement is the de-noising of speech corrupted by noise [9]. In the context of music processing, source enhancement relates to the concept of remixing, and is often performed in order to make one musical instrument (sound source) more salient in the mix. Remixing applications often use sound separation front-ends to gain access to the individual sound sources and change the characteristic of the mix [10]. Even though source enhancement can be preceded by a sound source separation stage, this is not always the case and hence, we also highlight the distinction between these terms.

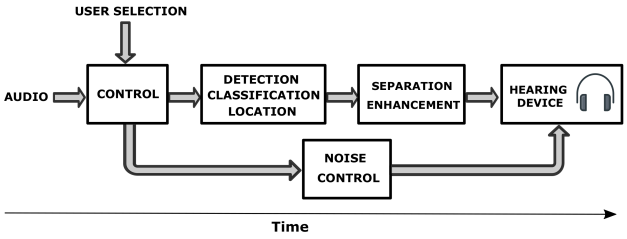

Fig. 2. Processing workflow of a selective hearing application.

In order to understand the technical challenges of selective hearing technologies, this paper presents a processing workflow for selective hearing systems in section II. Current capabilities and challenges in the different relevant fields of research are highlighted through an overview of the state-of-the-art in section III. Finally, current and future research perspectives are presented in section IV.

II. Selective Hearing in Practice

Figure 2 presents a processing workflow of a SH application as defined in section I. The user is always the center of the system, and controls the auditory scene by means of a control unit. The user can modify the auditory scene with a user interface as the one depicted in Figure 1, or with any type of interaction such as speech control, gestures, sight direction, etc. Once the user has provided feedback to the system, the next step consists of a detection/classification/location stage. In some cases, only detection is necessary, e.g., the user wishes to keep any speech occurring in the auditory scene. In other cases, classification might be necessary, e.g., the user wishes to keep fire alarms in the auditory scene but not telephone rings or office noise. In some cases, only the location of the source is relevant for the system. This is the case, for example, of the four sources in Figure 1: The user can decide to remove or attenuate the sound source coming from a certain direction, regardless of the type or characteristics of the source.

The auditory scene is first modified in the *separation/enhancement* stage in Figure 2. This happens either by suppressing, attenuating, or enhancing a certain sound source(s). As shown in Figure 2, an additional processing alternative in SH is noise control, where the goal is to remove or minimize the background noise in the auditory scene. Perhaps the most popular and wide-spread technology for noise control today is Active Noise Control (ANC) [11].

One of the biggest challenges in selective hearing applications, relates to the strict requirements with respect to processing time. The full processing workflow needs to be carried out with minimal delay in order to maintain the naturalness and perceptual quality for the user. The maximum acceptable latency of a system highly depends on the

application and on the complexity of the auditory scene. For example, McPherson et al. propose 10 ms as an acceptable latency reference for interactive music interfaces [12]. For music performances over a network, the authors in [13] report that delays become perceivable in the range between 20-25 and 50-60 ms. However, active noise control/cancellation (ANC) technologies call for ultra-low latency processing for better performance. In these systems, the amount of acceptable latency is both frequency- and attenuation-dependent, but can be as low as 1 ms for an approximately 5dB attenuation of frequencies below 200Hz [14]. A final consideration in SH applications refers to the perceptual quality of the modified auditory scene. Considerable amount of work has been devoted to methodologies for reliable assessment of audio quality in different applications [15]–[17]. However, the challenge for SH is managing the clear trade-off between processing complexity and perceptual quality.

III. Related Work

This section focuses on presenting a concise overview of the state-of-the-art in machine listening relevant to SH.

A. Sound Source Detection, Classification and Identification

Perhaps some of the most relevant works for SH come from the field of detection and classification of acoustic scenes and events [18]. In this context, methods for audio event detection (AED) in domestic environments have been proposed, where the goal is to detect the time boundaries of a given sound event within 10 sec recordings [19], [20]. In this particular case, 10 sound event classes were considered including cat, dog, speech, alarm, and running water. Methods for polyphonic sound event (several simultaneous events) detection have also been proposed in the literature [21], [22]. In [21], a method for polyphonic sound event detection is proposed where a total of 61 sound events from real-life contexts are detected using binary activity detectors based on a bi-directional long short-term memory (BLSTM) recurrent neural network (RNN).

To deal with weakly-labeled data, some systems incorporate temporal attention mechanisms to focus on certain regions of the signal for classification [23]. The problem of noisy labels in classification is particularly relevant for selective hearing applications where the class labels can be so diverse that having high-quality annotations are very costly [24]. Noisy labels in sound event classification tasks were addressed in [25], where noise-robust loss functions based on the categorical cross-entropy, as well as ways of exploiting both noisy and manually labeled data are presented. Similarly, [26] presents a system for audio event classification based on a convolutional neural network (CNN) that incorporates a verification step for noisy labels based on prediction consensus of the CNN on multiple segments of the training example.

Some works attempt to simultaneously detect and locate sound events. In [27], detection is performed as a multi-label classification task, and location is given as the 3-D coordinates of the direction-of-arrival (DOA) for each sound event.

In the context of speech, work on voice activity detection and on speaker recognition/identification are of great relevance for SH. Voice activity detection has been addressed in noisy environments using denoising auto-encoders [28], recurrent neural networks [29], or as an end-to-end system using raw waveforms [30]. For speaker recognition applications, a great number of system have been proposed in the literature [31], the great majority focusing on increasing robustness to different conditions, for example with data augmentation or with improved embeddings that facilitate recognition [32]–[34].

In the music domain, the classification of music instruments can be seen as an analogous problem to sound event detection. Musical instrument classification in both monophonic and polyphonic settings has been addressed in the literature [35], [36]. In [35], the predominant instrument in 3 sec audio segments are classified between 11 instrument classes, proposing several aggregation techniques. Similarly, [37] proposes a method for musical instrument activity detection able to detect instruments in a finer temporal resolution of 1 sec. A significant amount of research has been done in the topic of singing voice analysis. In particular, methods such as [38] have been proposed for the task of detecting segments in an audio recording where the singing voice is active.

B. Sound Source Localization

Sound source localization is closely related to the problem of source counting, as the number of sound sources in the auditory scene is usually not known in real-life applications. Some systems work under the assumption that the number of sources in the scene is known. That is the case, for example, of the model presented in [39] that uses histograms of active intensity vectors to locate the sources. From a supervised perspective, [40] proposes a CNN-based algorithm to estimate the DOA of multiple speakers in the auditory scene using phase maps as input representations. In contrast, several works in the literature jointly estimate the number of sources in the scene and their location information. This is the case of [41], where a system for multi-speaker localization in noisy and reverberant environments is proposed. The system uses a complex-valued Gaussian Mixture Model (GMM) to estimate both the number of sources and the their localization.

Sound source localization algorithms can be computationally demanding as they often involve scanning a large space around the auditory scene [42]. In order to reduce computational requirements in localization algorithms, some works attempt to reduce the search space by introducing clustering algorithms [43], or by performing multi-resolution searches [42] on well-established methods such as those based on the steered response power phase transform (SRP-PHAT). Other methods impose sparsity constraints and assume only one sound source is predominant in a given time-frequency region [44]. More recently, an end-to-end system for azimuth detection directly from the raw waveforms has been proposed in [45].

C. Sound Source Separation (SSS)

When it comes to source separation of audio signals, the majority of the research output comes from the speech separation and music separation communities.

The most relevant research for SH in the speech community comes from the work on speaker-independent separation. These systems attempt to perform separation without any prior information about the speakers in the scene [46]. Some systems also attempt to exploit the spatial location of the speakers to perform separation [47].

Given the importance of computational performance in selective hearing applications, research conducted with the specific aim of achieving low-latency is of particular relevance. Some works have been proposed to perform low-latency speech separation ($< 10\ ms$) with little training data available [48]. In order to avoid delays caused by framing analysis in the frequency domain, some systems approach the separation problem by carefully designing filters to be applied in the time domain [49]. Other systems achieve low-latency separation by directly modelling the time-domain signal using encoder-decoder framework [50]. In contrast, some systems have attempted to reduce the framing delay in frequency domain separation approaches [51].

Music sound separation (MSS) attempts to extract a music source from an audio mixture [5]. A great number of systems have been proposed to deal with the problem of lead instrument-accompaniment separation [52]. These algorithms take the most salient sound source in the mixture, regardless of its class label, and attempt to separate it from the remaining accompaniment. Considerable amount of research has been devoted to the problem of singing voice separation [53]. In most cases, either specific source models [54] or data-driven models [55] are used to capture the characteristics of the singing voice. Even though systems such as the one proposed in [55], do not explicitly incorporate a classification or a detection stage to achieve separation, the data-driven nature of these approaches, allows these systems to implicitly learn to detect the singing voice with certain accuracy before separation. Another class of algorithms in the music domain attempt to perform separation using only the location of the sources, without attempting to classify or detect the source before separation [4].

D. Active Noise Control (ANC)

A particularly relevant line of research in the context of SH is the work conducted towards the development of active noise control/cancellation (ANC) methods. ANC systems mostly aim at removing background noise for headphone users by introducing an anti-noise signal to cancel it out [11]. ANC can be considered a special case of SH, and faces equally strict performance requirement [14]. Some works have focused on active noise control in specific environments such as automobile cabins [56] or industrial scenarios [57]. The work in [56], analyses the cancellation of different types of noises such as road noise and engine noise, and calls for unified noise control systems capable of dealing with different types of noises. Some work has focused on developing ANC systems to cancel noise over specific spatial regions. In [58], ANC over a spatial region is addressed using spherical harmonics as basis functions to represent the noise field.

E. Sound Source Enhancement

In the context of speech enhancement, one of the most common applications is the enhancement of speech that has been corrupted by noise. A great deal of work has focused on phase processing of single-channel speech enhancement [8]. From a deep neural network perspective, the problem of speech denoising has been addressed with denoising autoencoders in [59], as an non-linear regression problem between clean and noisy speech using a deep neural networks (DNN) in [60], and as an end-to-end system using Generative Adversarial Networks (GANs) in [61]. In many cases, speech enhancement is applied as a front-end for automatic speech recognition (ASR) systems, as is the case of [62], where speech enhancement is approached with an LSTM RNN. Speech enhancement is also often done in conjunction with sound source separation approaches where the idea is to first extract the speech, to then apply enhancement techniques on the isolated speech signal [6].

As mentioned in the introduction, source enhancement in the context of music most often refers to applications of music remixing. In contrast to speech enhancement where often the assumption is that the speech is only corrupted by noisy, music applications most often assume that other sound sources (music instruments) are simultaneously playing with the source to be enhanced. For this reason, music remixing applications always come preceded by a source separation stage. In [10] for example, early jazz recordings were remixed by applying lead-accompaniment and harmonic-percussive separation techniques in order to achieve better sound balance in the mixture. Similarly, [63] studied the use of different singing voice separation algorithms in order to change the relative loudness of the singing voice and the backing track, showing that a 6 dB increase is possible by introducing minor but audible distortions into the final mixture. In [64], the authors study ways of enhancing music perception for cochlear implant users by applying sound source separation techniques to achieve new mixes.

IV. Current and Future Perspectives

After analyzing the state-of-the-art presented in section III, three general trends can be observed. First, even though earlier methods are focused on a single task, a clear trend to develop methods to jointly address two or more machine listening tasks can be observed. That is the case, for example, of the method for counting and localization in [41], the method for localization and detection in [27], the method for separation and classification in [65], and the method for separation and counting in [66].

Second, a general effort to increase robustness of current machine listening methods is evident in the community [25],

[26], [32], [34], where new emerging directions include domain adaptation [67] and training on data sets recorded with multiple devices [68].

Finally, clear efforts to increase computational efficiency of machine listening methods can be observed [48], with a clear tendency to move into end-to-end systems capable of dealing with raw waveforms [30], [45], [50], [61].

While important advances in all the relevant fields can be identified, two main challenges define the future of SH: (1) Methods that can jointly detect/classify/locate and separate/enhance under a unified optimization scheme are required to be able to selectively modify sound sources in the scene. While independent detection, separation, localization, classification, and enhancement methods can be reliable under constrained conditions, they do not provide the robustness and flexibility required for SH. (2) Realistic SH applications require methods suitable for real-time processing. While real-time processing is already achievable in certain tasks, there is in general a clear trade-off between algorithmic complexity and performance.

As a final remark, we would like to highlight the potential of exploring joint models for ANC and machine listening. Even though ANC has not traditionally been developed within the machine listening community, the potential of methods that can for example, first classify the acoustic scene and then selectively apply ANC, is to the authors' knowledge, still open for exploration.

REFERENCES

[1] V. Valimaki, A. Franck, J. Ramo, H. Gamper, and L. Savioja, "Assisted listening using a headset: Enhancing audio perception in real, augmented, and virtual environments," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 92–99, March 2015.

[2] K. Brandenburg, E. Cano, F. Klein, T. Köllmer, H. Lukashevich, A. Neidhardt, U. Sloma, and S. Werner, "Plausible augmentation of auditory scenes using dynamic binaural synthesis for personalized auditory realities," in *Proc. of AES International Conference on Audio for Virtual and Augmented Reality*, Aug 2018.

[3] S. Argentieri, P. Dans, and P. Soures, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech Language*, vol. 34, no. 1, pp. 87–112, 2015.

[4] D. FitzGerald, A. Liutkus, and R. Badeau, "Projection-based demixing of spatial audio," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1560–1572, 2016.

[5] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, and F. Stöter, "Musical source separation: An introduction," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 31–40, Jan 2019.

[6] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, April 2017.

[7] E. Cano, J. Nowak, and S. Grollmisch, "Exploring sound source separation for acoustic condition monitoring in industrial scenarios," in *Proc. of 25th European Signal Processing Conference (EUSIPCO)*, Aug 2017, pp. 2264–2268.

[8] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, March 2015.

[9] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. Wiley, 2018.

[10] D. Matz, E. Cano, and J. Abeßer, "New sonorities for early jazz recordings using sound source separation and automatic mixing tools," in *Proc. of the 16th International Society for Music Information Retrieval Conference*. Malaga, Spain: ISMIR, Oct. 2015, pp. 749–755.

[11] S. M. Kuo and D. R. Morgan, "Active noise control: a tutorial review," *Proceedings of the IEEE*, vol. 87, no. 6, pp. 943–973, June 1999.

[12] A. McPherson, R. Jack, and G. Moro, "Action-sound latency: Are our tools fast enough?" in *Proceedings of the International Conference on New Interfaces for Musical Expression*, July 2016.

[13] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An overview on networked music performance technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.

[14] S. Liebich, J. Fabry, P. Jax, and P. Vary, "Signal processing challenges for active noise cancellation headphones," in *Speech Communication; 13th ITG-Symposium*, Oct 2018, pp. 1–5.

[15] E. Cano, J. Liebetrau, D. Fitzgerald, and K. Brandenburg, "The dimensions of perceptual quality of sound source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 601–605.

[16] P. M. Delgado and J. Herre, "Objective assessment of spatial audio quality using directional loudness maps," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 621–625.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[18] M. D. Plumbley, C. Kroos, J. P. Bello, G. Richard, D. P. Ellis, and A. Mesaros, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*. Tampere University of Technology. Laboratory of Signal Processing, 2018.

[19] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 19–23.

[20] L. JiaKai, "Mean teacher convolution system for dcase 2018 task 4," DCASE2018 Challenge, Tech. Rep., September 2018.

[21] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 6440–6444.

[22] E. Çakir and T. Virtanen, "End-to-end polyphonic sound event detection using convolutional recurrent neural networks with learned time-frequency representation input," in *Proc. of International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–7.

[23] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-Scale Weakly Supervised Audio Classification Using Gated Convolutional Neural Network," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 121–125.

[24] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 845–869, May 2014.

[25] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[26] M. Dorfer and G. Widmer, "Training general-purpose audio tagging networks with noisy labels and iterative self-verification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, Surrey, UK, 2018.

[27] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2018.

[28] Y. Jung, Y. Kim, Y. Choi, and H. Kim, "Joint learning using denoising variational autoencoders for voice activity detection," in *Proc. of Interspeech*, September 2018, pp. 1210–1214.

[29] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 483–487.

[30] R. Zazo-Candil, T. N. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform CLDNNs for voice activity detection," in *Proc. of Interspeech*, 2016.

[31] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," in *Proc. of IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2015, pp. 4814–4818.

[32] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5329–5333.

[33] M. McLaren, D. Castán, M. K. Nandwana, L. Ferrer, and E. Yilmaz, "How to train your speaker embeddings extractor," in *Odyssey*, 2018.

[34] S. O. Sadjadi, J. W. Pelecanos, and S. Ganapathy, "The IBM speaker recognition system: Recent advances and error analysis," in *Proc. of Interspeech*, 2016, pp. 3633–3637.

[35] Y. Han, J. Kim, and K. Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 208–221, Jan 2017.

[36] V. Lonstanlen and C.-E. Cella, "Deep convolutional networks on the pitch spiral for musical instrument recognition," in *Proceedings of the 17th International Society for Music Information Retrieval Conference*. New York, USA: ISMIR, 2016, pp. 612–618.

[37] S. Gururani, C. Summers, and A. Lerch, "Instrument activity detection in polyphonic music using deep neural networks," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*. Paris, France: ISMIR, Sep. 2018, pp. 569–576.

[38] J. Schlütter and B. Lehner, "Zero mean convolutions for level-invariant singing voice detection," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*. Paris, France: ISMIR, Sep. 2018, pp. 321–326.

[39] S. Delikaris-Manias, D. Pavlidi, A. Mouchtaris, and V. Pulkki, "DOA estimation with histogram analysis of spatially constrained active intensity vectors," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 526–530.

[40] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, March 2019.

[41] X. Li, L. Girin, R. Horaud, and S. Gannot, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1997–2012, Oct 2017.

[42] F. Grondin and F. Michaud, "Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations," *Robotics and Autonomous Systems*, vol. 113, pp. 63 – 80, 2019.

[43] D. Yook, T. Lee, and Y. Cho, "Fast sound source localization using two-level search space clustering," *IEEE Transactions on Cybernetics*, vol. 46, no. 1, pp. 20–26, Jan 2016.

[44] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, Oct 2013.

[45] P. Vecchiotti, N. Ma, S. Squartini, and G. J. Brown, "End-to-end binaural sound localisation from the raw waveform," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 451–455.

[46] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, April 2018.

[47] Z. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 1–5.

[48] G. Naithani, T. Barker, G. Parascandolo, L. BramslŁw, N. H. Pontoppidan, and T. Virtanen, "Low latency sound source separation using convolutional recurrent neural networks," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2017, pp. 71–75.

[49] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 216–220.

[50] Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 696–700.

[51] J. Chua, G. Wang, and W. B. Kleijn, "Convolutive blind source separation with low latency," in *Proc. of IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sep. 2016, pp. 1–5.

[52] Z. Rafii, A. Liutkus, F. Stöter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1307–1335, Aug 2018.

[53] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation*, Y. Deville, S. Gannot, R. Mason, M. D. Plumbley, and D. Ward, Eds. Cham: Springer International Publishing, 2018, pp. 293–305.

[54] J.-L. Durrieu, B. David, and G. Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 6, pp. 1180 –1191, Oct. 2011.

[55] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[56] P. N. Samarasinghe, W. Zhang, and T. D. Abhayapala, "Recent advances in active noise control inside automobile cabins: Toward quieter cars," *IEEE Signal Processing Magazine*, vol. 33, no. 6, pp. 61–73, Nov 2016.

[57] G. S. Papini, R. L. Pinto, E. B. Medeiros, and F. B. Coelho, "Hybrid approach to noise control of industrial exhaust systems," *Applied Acoustics*, vol. 125, pp. 102 – 112, 2017.

[58] J. Zhang, T. D. Abhayapala, W. Zhang, P. N. Samarasinghe, and S. Jiang, "Active noise control over space: A wave domain approach," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 774–786, April 2018.

[59] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. of Interspeech*, 2013.

[60] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan 2015.

[61] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: speech enhancement generative adversarial network," in *Proc. of Interspeech*, August 2017, pp. 3642–3646.

[62] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Latent Variable Analysis and Signal Separation*, E. Vincent, A. Yeredor, Z. Koldovský, and P. Tichavský, Eds. Cham: Springer International Publishing, 2015, pp. 91–99.

[63] H. Wierstorf, D. Ward, R. Mason, E. M. Grais, C. Hummersone, and M. D. Plumbley, "Perceptual evaluation of source separation for remixing music," in *Proc. of Audio Engineering Society Convention 143*, Oct 2017.

[64] J. Pons, J. Janer, T. Rode, and W. Nogueira, "Remixing music using source separation algorithms to improve the musical experience of cochlear implant users," *The Journal of the Acoustical Society of America*, vol. 140, no. 6, pp. 4338–4349, 2016.

[65] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "A joint separation-classification model for sound event detection of weakly labelled data," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2018.

[66] T. v. Neumann, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, and R. Haeb-Umbach, "All-neural online source separation, counting, and diarization for meeting analysis," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 91–95.

[67] S. Gharib, K. Drossos, E. Cakir, D. Serdyuk, and T. Virtanen, "Unsupervised adversarial domain adaptation for acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, November 2018, pp. 138–142.

[68] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop*, Surrey, UK, 2018.