Multimedia Tools and Applications (2021) 80:28365–28389
https://doi.org/10.1007/s11042-021-11080-y

N-HANS: A neural network-based toolkit for in-the-wild audio enhancement

Shuo Liu[1] · Gil Keren[1] · Emilia Parada-Cabaleiro[1] · Björn Schuller[1,2]

Received: 1 September 2020 / Revised: 6 May 2021 / Accepted: 11 May 2021 /
Published online: 3 June 2021

Abstract

The unprecedented growth of noise pollution over the last decades has raised an always increasing need for developing efficient audio enhancement technologies. Yet, the variety of difficulties related to processing audio sources in-the-wild, such as handling unseen noises or suppressing specific interferences, makes audio enhancement a still open challenge. In this regard, we present N-HANS (the Neuro-Holistic Audio-eNhancement System), a Python toolkit for in-the-wild audio enhancement that includes functionalities for audio denoising, source separation, and —for the first time in such a toolkit—selective noise suppression. The N-HANS architecture is specially developed to automatically adapt to different environmental backgrounds and speakers. This is achieved by the use of two identical neural networks comprised of stacks of residual blocks, each conditioned on additional speech- and noise-based recordings through auxiliary sub-networks. Along to a Python API, a command line interface is provided to researchers and developers, both of them carefully documented. Experimental results indicate that N-HANS achieves great performance w. r. t. existing methods, preserving also the audio quality at a high level; thus, ensuring a reliable usage in real-life application, e. g., for in-the-wild speech processing, which encourages the development of speech-based intelligent technology.

Keywords Residual block · Auxiliary network · Selective noise suppression · Conditioned recording · Audio context · Source separation · Speech denoising

✉ Shuo Liu
shuo.liu@informatik.uni-augsburg.de

Gil Keren
gil.keren@informatik.uni-augsburg.de

Emilia Parada-Cabaleiro
emilia.parada-cabaleiro@informatik.uni-augsburg.de

Björn Schuller
schuller@ieee.org

[1] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany

[2] GLAM – Group on Language, Audio, Music, Imperial College London, London, UK

1 Introduction

Noise pollution has become an indiscernible limitation of today's society. Through a constant increment in magnitude and severity [9], environmental noise impairs human's health and well-being more than ever before [15]. However, negative background auditory interferences, such as those produces by transportation noise [38], industrial noise [1], or urban noise [79], not only impair human's cognitive [62, 73] and communicative [44, 45] skills, but also limit the performance of general audio and specific speech-driven applications, such as, automatic speech recognition [40], speech emotion recognition [2, 61], and speaker verification [24, 28, 37, 56]. Hence, audio enhancement, which generally aims at extracting targeted signals, is broadly exploited to improve audio and speech quality for real-life applications [16, 26, 55]. Two of the main procedures for enhancing audio and speech are source separation and denoising: The former aims to extract a target audio from a mixture of multiple overlapping signals [66]; the latter attempts to suppress the background noise [49, 78]. With the advance of artificial intelligence, neural network based models for audio enhancement have been presented, being particularly efficient in source separation [18, 20, 35, 36, 68] and denoising [5, 27, 32, 34, 47, 60, 71] tasks—the performance of classic algorithms is often overtaken by artificial neural networks [7].

Despite the ongoing efforts and the already achieved outcomes [13, 14, 39], the enhancement of in-the-wild audio sources is still an open research topic for which more research is required. One of the still open challenges in audio enhancement is the need for more robust methods to be developed. Although an audio enhancement model relies on its noise generalisation—which is limited to the data size and diversity of the training noises—in real scenarios, audio may simultaneously be corrupted by multiple kinds of noises, including unseen noises [30]. Furthermore, the non-stationary nature of real-life noises yields a level of uncertainty in realistic applications unapproachable, for instance, by the existing speech enhancement methods, which typically process a single noise recording at once [5, 47, 49].

Another challenge still open in the audio enhancement domain is the development of methods able to efficiently handle interference components while preserving, at the same time, the essential features of the signal. The currently available enhancing technology, often characterised by the application of aggressive methods for the estimation of noise and other interference components [48, 63, 76] is indeed not yet able to preserve a signal's essential properties, such as, the speech's naturalness. An efficient and precise distinction of target and background becomes particularly challenging when the interfering components present similar acoustic properties to the target signal to be retained. For instance, within the speech domain, separating an undesired user in the background speaking the same language as the target one might be particularly challenging, especially in noisy audio samples [17, 36]. Finally, a third challenge to be faced is the need for further improvements in the development of intelligent audio enhancement technology equipped with an autonomous decision making system. This becomes crucial in specific circumstances where preserving environmental interferences, such as alarms, might be essential for security reasons; thus, making the target noise to be removed very specific. In such a scenario, a selective noise suppression system should be capable to identify and preserve the allowed noises, i. e., the "positive noises", while suppressing the undesired ones, i. e., the "negative noises". Although this mechanism is crucial in real-life scenarios, where the noisy audio samples contain often important signals, such as alarms or other acoustic warnings aimed to prevent, e. g., traffic accidents, existing technology for selective noise suppression is still unable to process problems with this level of complexity. The main limitation of prior approaches is that they mostly exploit spatial information where multi-channel recordings are available, e. g., binaural hearing aids using end-to-end (e2e) wireless technology [3, 72]. However, these approaches rely on assumptions regarding the spatial properties of the target signal and the different noise sources, which limits considerably their adaptability to unseen environments; thus, becoming almost inapplicable in real-world scenarios.

1.1 Contributions of the presented work

All in all, the future development of in-the-wild audio enhancement technology should mainly focus on three challenges: (i) robustness, developed by increasing a system's ability to handle unseen noises; (ii) efficiency, promoted by refining a system's capacity to preserve signal's essential properties; (iii) decision making, encouraged by improving a system's capability to autonomously identify the important signals. In order to contribute to the alleviation of the described challenges, we introduce the Neuro-Holistic Audio-eNhancement System[1] (`N-HANS`), a neural network-based toolkit for in-the-wild audio enhancement developed with Tensorflow in Python. The objectives of `N-HANS` are therefore three-fold: (i) successfully process unseen noises through a robust technology especially tailored for audio denoising; (ii) efficiently preserve signals' essential properties through their accurate separation from similar interfering sources; (iii) properly identify and retain important signals through a intelligent selective noise suppression system capable to autonomously discriminate between positive and negative noises. Hence, the main contributions of this work can be summarised as follows:

- We present, to the best of our knowledge, the first audio enhancement toolkit with the functionality of selective noise suppression.
- We propose a neural network architecture named ±Auxiliary Network using a novel fusion method to project information from auxiliary input references.
- We present `N-HANS`, an open-source audio enhancement toolkit specially tailored for in-the-wild applications trough a three-fold functionality: denoising, selective noise suppression, and source separation. Along with the toolkit, we also provide an user-friendly command line interface.

The rest of this manuscript is organised as follows. In Section 2, the related work is outlined. In Section 3, we give an overview of the `N-HANS` framework and introduce the system's input processing. Section 4 presents the proposed ±auxiliary network, which works as back end for the system. Section 5 discusses the performed experiments and their evaluation. Section 6 illustrates the system's performance by visualising a selection of audio examples processed by `N-HANS`. Finally, concluding remarks and future research directions are drawn in Section 7.

2 Related work

Although a variety of methods for speech enhancement have been presented (for an overview, cf. Table 1), the open-source toolkits currently available focus only on one specific task, i. e., either audio denoising or source separation, while methods presenting those functionalities in the same tool have not yet been developed. Furthermore, the performance

[1] https://github.com/N-HANS/N-HANS

Table 1 List of the most popular open-source toolkits for speech enhancement

	Methods			Functionalities			Adaptation	
Toolkit	SP	ML	DL	DE	SS	SNS	Spk	Env
VoiceBox[2]	✓			✓				
CtuCopy[3]	✓			✓				
SETK[4]	✓		✓	✓	✓			
SE Toolkit [25]			✓	✓				
SEDNN [75]			✓	✓				
SEGAN [47]			✓	✓				
openBlissart [70]		✓			✓			
FASST [51]		✓			✓			
GCC-NMF[5]		✓		✓	✓			
Asteroid [46]			✓		✓			
UNTWIST [50]		✓	✓		✓			
N-HANS[1]			✓	✓	✓	✓	✓	✓

Processing Methods: classic signal processing (SP), machine learning (ML) except deep learning, deep learning (DL); Functionalities: denoising (DE), speech separation (SS), selective noise suppression (SNS); and Adaptation ability to speaker (Spk) and speech surrounding environments (Env), are indicated

of many of them is limited by specific acoustic conditions, for instance, presenting a predisposition in handling only stationary noises. This is the case of VoiceBox[2], a speech processing toolkit that provides classic signal processing algorithms for a wide range of audio tasks, including denoising. Similarly, CtuCopy[3], based on the combination of Wiener filtering and spectra subtraction methods, was developed for audio feature extraction and speech denoising. These two tools, since using classical signal processing methods for denoising, expect accurate noise power estimation, which can only be assured under stationary noise conditions. Indeed, for the processing of non-stationary noises, these classic approaches are characterised by a decline in their performance.

With the always increasing use of artificial neural networks, promising denoising toolkits based on neural networks, such as SETK[4], SE Toolkit [25], SEDNN [75], SEGAN [47], and U-Net [5, 60] have been presented in the literature. However, these methods were specifically designed for audio denoising, thus, presenting difficulties to be used for source separation. Similarly, Untwist [50] and Asteroid [46], are two neural network-based toolkits for source separation: the former includes the most basic neural network architecture, i. e., Multi-Layer Perception (MLP); the latter—recently proposed and considered 'superior' in literature—integrates a variety of neural networks, such as ConvTasnet [36], Deep clustering [18], and Chimera++ [69]. Nevertheless, none of both present functionalities for denoising applications. Finally, source separation methods based on non-negative matrix factorisation (NMF) such as OpenBlissart [70], have been also presented. Similarly, Flexible Audio Source Separation Toolbox (FASST) [51], considers Gaussian mixture model (GMM) and hidden Markov model (HMM) for the NMF training. The NMF-based method

[2] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[3] http://noel.feld.cvut.cz/speechlab/share/download/ctucopy/ctucopy3.html

[4] https://github.com/funcwj/setk

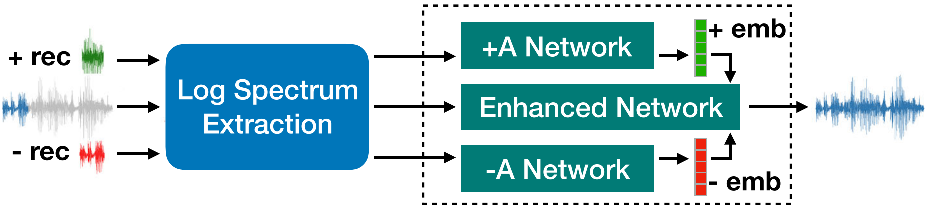


Fig. 1 System framework. From left to right, the input (noisy audio) and the recordings: positive (+rec), negative (−rec); the Log Spectrum extraction block; the Networks: ±Auxiliary (A), Enhanced; and the system's output (denoised audio). The +A Network processes the +rec to produce a positive embedding vector (+emb), i. e., the components to be preserved. The −A Network processes the −rec to obtain the negative embedding vector (−emb) that hints at the components to be suppressed. The enhanced network processes the noisy audio, as well as the positive and negative embeddings, in order to generate the desired output

for source separation has also been extended to a denoising task in GCC-NMF[4], which applies the generalised cross correlation (GCC) spatial localisation method; yet, GCC-NMF can be only considered for deonising, but not for source separation applications.

To the best of our knowledge, N-HANS is the first publicly available neural network based tool presenting both: audio denoising and source separation functionalities in one toolkit framework. In addition, N-HANS provides the solution to selective noise suppression, i. e., suppressing only unwanted noises while preserving others—pertaining a natural audio surrounding can be particularly important when relevant signals are involved, e. g., alarms or other acoustic warnings. Furthermore, the performance of the currently available machine learning based speech enhancement tools, such as those indicated in Table 1, is limited to the diversity of speakers and noise types in training set, which impairs their application in real-life scenarios, where unseen speakers can appear and multiple noise types exist simultaneously. Differently, N-HANS, by leveraging auxiliary networks that learn to identify and generalise the characteristics of unseen speakers and speech surroundings, presents a more adaptive performance, w. r. t. the existing methods, in real-life scenarios.

3 System overview: methodology

N-HANS, embedded with two trained models sharing an identical architecture, faces the challenge of handling unseen noises by considering individual configurations, i. e., each model is conditioned on additional environmental backgrounds in order to adapt it to unseen noises from the real life. In addition, through its audio source separation and selective noise suppression system, based on an ±Auxiliary (A) Network (cf. Figure 1), N-HANS recovers a target audio while removing the interfering sources. To the best of our knowledge, the presented fusion method, used to inject the context information into the conditional residual network, has not been proposed in previous research (for further details cf. Section 4). First of all, the log magnitude spectrum is extracted from the input contaminated audio and from the positive and negative recordings by taking the logarithmic absolute value of the Short-Time Fourier Transformation (STFT)—extracted using a 25 ms Hanning window shifted by 10 ms—which are fed into the Enhanced and the ±Auxiliary Networks separately. Then, the +A Network processes the extracted positive spectrum to produce a positive embedding

[5] https://github.com/seanwood/gcc-nmf

Table 2 N-HANS overview. The input, i. e., the raw input and the positive (+) and negative (−) recordings; as well as the output, are indicated for the three considered tasks: denoising, source separation, and selective noise suppression

Task	Denoising	Source Separation	Selective Noise Suppression
raw input	noisy audio	overlapping sources	noisy audio
+recording	-	target source	noise to preserve
−recording	noise to suppress	interference source	noise to suppress
output	denoised audio	separated source	denoised audio

vector, while the −A Network processes the negative spectrum to emit a negative one. The positive and negative embedding vectors can be seen as the representations of the characteristics of the unseen audio contents and are then injected into the enhancement network to emit the denoised or separated audio. Through positive and negative context awareness, N-HANS encourages a system's adaptability and applicability to different unseen noisy environments and audio sources, e. g., different speakers.

3.1 Input audio processing

N-HANS processes contaminated audio conditioned on additional positive and negative recordings, which indicate the audio content to be preserved and suppressed, respectively.[1] The audio files used in our experiments (cf. Section 5) are: Librispeech [43] and Audioset [12] for denoising and selective noise suppression; and VoxCeleb Corpus [6, 41] for source separation. The sampling frequency of all audio files is 16 kHz, i. e., each frame consists of 400 samples with a resulting feature vector of 201 frequencies.[2] In Table 2, an overview of N-HANS's input and output information for the considered tasks is given.

The three inputs, i. e., the raw input (original audio file to be enhanced), the positive recording (containing interferences to be preserved), and the negative recording (containing interferences to be suppressed), are processed by the Enhanced and ±Auxiliary Networks (cf. Section 4). The *contaminated* segment M, consisting of N successive frames ($N = 35$ in the experiments) from the log magnitude spectrum of the contaminated audio, leads to $M \in \mathbb{R}^{N \times F}$. The *positive context* C_+ is a B frames segment of the log magnitude spectrum extracted from the positive recording. The *negative context* C_- is retrieved from the log magnitude spectrum of the negative recording using the same process, leading to $C_+, C_- \in \mathbb{R}^{L \times F}$ ($L = 200$ in the experiments). The positive and the negative contexts, containing the information to be preserved and suppressed, respectively, from the raw input, are used to create the positive and negative embeddings involved in the enhancement process (cf. Section 4). In order to aggregate the acoustic characteristics of the audio content to be preserved and suppressed, sufficient acoustic information should be considered—the larger the size of the positive and negative contexts, the more information would be supplied to the system. The *target* segment $T \in \mathbb{R}^{N \times F}$, with the same size of the contaminated segment,

[1] Note that the conditioning content, i. e., that from the positive and negative recordings, does not necessarily need to appear in the contaminated audio.

[2] The N-HANS model can process audio files that operate at other sampling rates, leading to the corresponding length variation of the resulting feature vector. We denote the number of frequencies of a frame as F for further derivation.

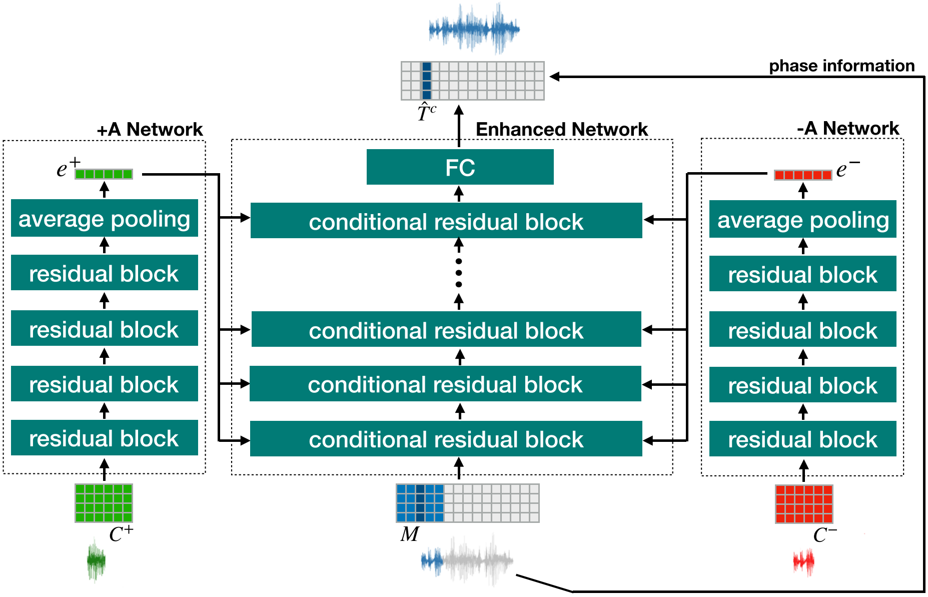


Fig. 2 Architecture of the ±Auxiliary (A) Networks. The +A and −A Networks process the positive and negative contexts (C^+ and C^-) via a sequence of 4 residual blocks to produce positive and negative embeddings (e^+ and e^-). To estimate the *contamination frame* (CF), the Enhanced Network processes the contaminated segment M (noisy or overlapping segment) through a sequence of 8 residual blocks, each additionally conditioned by the e^+ and e^-

represents the ideal output segment: denoised audio for the denoising and selective noise suppression tasks[3]; and a separated source for the source separation task. The centre frames are indicated as M^c, $T^c \in \mathbb{R}^{1\times F}$, for the *contaminated* and *target* segments, respectively.

4 Approach: ±auxiliary networks

The architecture of the proposed ±Auxiliary (A) Networks is based on stacks of residual blocks [17] as depicted in Fig. 2. Residual networks (Resnets), which introduce skip-connections to the conventional neural networks framework—resulting in smoother loss landscape and enabling a substantially deeper architecture [31]—have shown to be successful in both the computer vision and audio domains [17, 23, 67]. A basic residual block contains two convolutional layers, where batch normalisation [21, 52] followed by a rectified linear unit (ReLU) [42] are applied between the convolutional layers. The residual block's input is then added to the output of the second convolutional layer after channel conversion via a 1 × 1 convolution. Again, batch normalisation and ReLU activation are applied to produce the block's output.

The N-HANS architecture consists of three subnetworks, each containing a sequence of residual blocks. An embedding network processes the positive context, i. e., the +A Network, in order to emit the *positive embedding*. Similarly, another embedding network with

[3] Note that the target segment for selective noise suppression contains the speech component and the positive noise, both to be preserved from the contaminated segment.

Table 3 Specifications of the auxiliary embedding network

Block	Kernel	Stride	#Channels
1	(8, 4)	(3, 2)	64
2	(8, 4)	(3, 2)	128
3	(4, 4)	(1, 1)	256
4	(4, 4)	(1, 2)	512

For each residual block the kernel, stride, and number (#) of channels are indicated

the same architecture (−A Network) processes the negative context to emit the *negative embedding*. Then, the enhanced network processes the two embeddings and the contaminated segment, by this emitting a *contamination frame* (CF), which estimates the audio components that need to be eliminated in the centre frame of the contaminated segment. Finally, the *estimated target frame* (F),[4] i. e., the difference in the centre frame between the contaminated segment and the estimated contamination frame, is computed. To minimise the mean squared error between the estimated target frame and the true target frame, i. e., the centre frame of the target segment, the model is trained considering stochastic gradient descent (cf. Section 4.2).

4.1 Auxiliary embedding network

In order to enable an individual management of the positive and negative contexts, these are separately processed in two embedding networks which share an identical structure but present different training parameters. Each of the two embedding networks, made up of a sequence of four residual blocks, takes an audio context as input and emits an embedding vector that may contain valuable acoustic information obtained from the context segment. The specifications of each embedding network are given in Table 3.

The output feature map of the last residual block in each embedding network is averaged across all locations (time steps and frequency bins), leading to a positive and negative embedding vector (cf. (1) and (2), respectively). The positive embedding vector is defined as

$$e^{+} = \mathbf{avg}(f^{+A}(C^{+})), \tag{1}$$

while the negative embedding vector is defined as

$$e^{-} = \mathbf{avg}(f^{-A}(C^{-})), \tag{2}$$

both with a fixed length of 512. f^{+A} and f^{-A} denote the operation of the residual blocks sequence in the positive and negative networks (+A and −A), which through their own learning parameters separately process the positive and negative contexts (C^{+} and C^{-}) to produce the positive and negative embeddings: $e^{+}, e^{-} \in \mathbb{R}^{512}$. The two embeddings are subsequently injected into the enhanced network to assist the audio denoising, source separation, and selective noise suppression tasks.

[4] For denoising and selective noise suppression, F refers to the *estimated denoised frame*; for source separation, to the *estimated separated frame*.

4.2 Enhanced network

The enhanced network, aimed to process the contaminated segment and the positive and negative embeddings, comprises a sequence of 8 conditional residual blocks, each of them presenting different kernel size, stride, and number of channels (cf. Table 4). Each conditional residual block, made up of two convolutional layers, processes the block input $M_{in} \in \mathbb{R}^{T \times F \times C_{in}}$ (cf. Figure 3). In the first convolutional layer, the learnt positive and negative embeddings are projected to a vector with a length equals to the number of output feature maps in the layer by applying a trainable fully-connected layer. The projected embeddings are then added to every locations of the feature maps, leading to

$$M_1 = \mathbf{conv}(M_{in}) + e_1^+ + e_1^-, \tag{3}$$

which has the shape of $T \times F \times C_1$, and

$$e_1^+ = e^+ W_1^+ + b_1^+, \tag{4}$$

$$e_1^- = e^- W_1^- + b_1^-, \tag{5}$$

denote the linearly projected embedding vectors with the length of C_1. W_1^+, b_1^+, and W_1^-, b_1^- are trainable parameters. The projected embedding vectors are extended to the size of the convolution output by using array broadcasting.

Further, for the second convolutional layer, M_1 is processed similarly, resulting in

$$M_2 = \mathbf{conv}(M_1) + e_2^+ + e_2^-, \tag{6}$$

with the shape of $T \times F \times C_2$, where

$$e_2^+ = e^+ W_2^+ + b_2^+, \tag{7}$$

$$e_2^- = e^- W_2^- + b_2^- \tag{8}$$

are projected embedding vectors with the length of C_2. By doing this, all convolutional layers in the enhanced network are conditioned on the information from the positive and negative contexts, allowing the model to better estimate the components that need to be preserved and suppressed in the contaminated segment. Besides, in the skip connection path, the input of the conditional residual block is converted to have the same channels as the output of the second convolutional layer through 1 × 1 convolution, leading to

$$M_{sc} = \mathbf{conv}_{1\times1}(M_{in}), \tag{9}$$

Table 4 Specifications of the enhanced network

Block	Kernel	Stride	#Channels
1	(4, 4)	(1, 1)	64
2	(4, 4)	(1, 1)	64
3	(4, 4)	(2, 2)	128
4	(4, 4)	(1, 1)	128
5	(3, 3)	(2, 2)	256
6	(3, 3)	(1, 1)	256
7	(3, 3)	(2, 2)	512
8	(3, 3)	(1, 1)	512

For each residual block the kernel, stride, and number (#) of channels are indicated

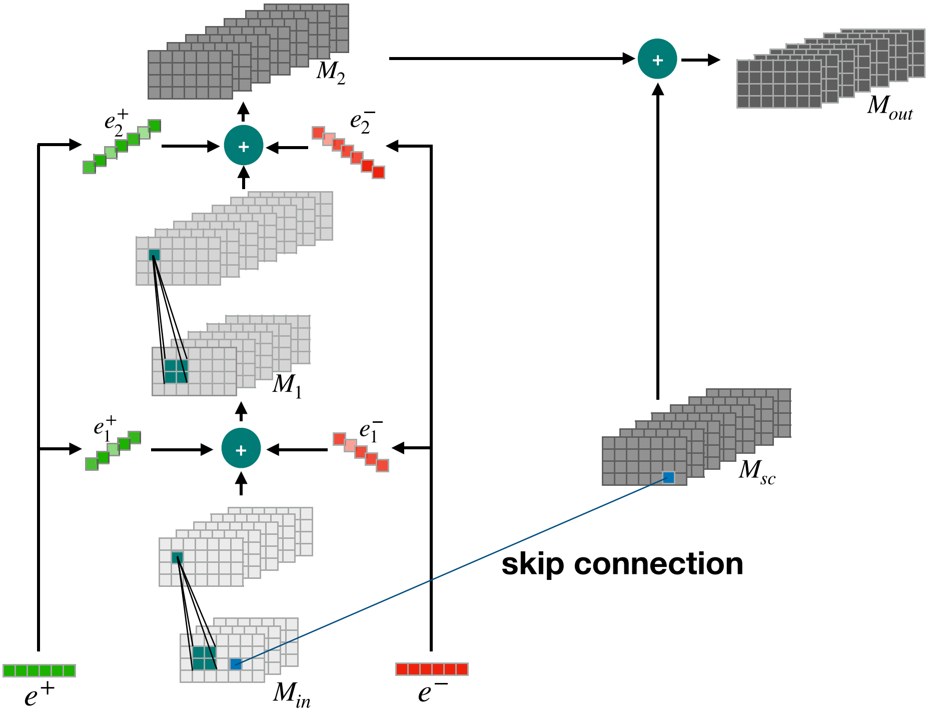


Fig. 3 Conditional residual block: the learnt positive and negative embeddings (e^+ and e^-) are injected in the two convolutional layers of the enhanced network. Block's input (M_{in}), output of skip connection path (M_{sc}), first convolutional layer output (M_1), second convolutional layer output (M_2), and block's output (M_{out}) are also indicated

which has the shape of $T \times F \times C_2$ and is added to the main path to achieve the block output

$$M_{out} = M_2 + M_{sc}. \tag{10}$$

Again, batch normalisation is applied for each convolutional layer of the conditional residual block, followed by ReLU activation functions (cf. Figure 3).

The output of the last conditional residual block is additionally convolved along the time axis, and then flattened to a vector. The flattened vector is projected to the length of F ($F = 201$ in experiments), through a fully-connected layer, representing the estimated contamination frame:

$$\hat{S}^c = \mathbf{conv_T}(f^{enh}(M, e^+, e^-))W_o + b_o, \tag{11}$$

where f^{enh} denotes the operation of the conditional residual blocks in the enhanced network, and $\mathbf{conv_T}$ indicates the convolution only along time direction. W_o, b_o are the fully connected layer's learnable parameters. We subtract the estimated contamination frame from the central frame of the contaminated spectrum to obtain the estimated target frame:

$$\hat{T}^c = M^c - \hat{S}^c. \tag{12}$$

During the training phase, we optimise the network parameters using stochastic gradient descent (SGD) with a learning rate of 0.1, to minimise the weighted mean squared error (MSE) between the estimated target frame and the true centre frame of the target spectrum:

$$L = ||(\hat{T}^c(f) - T^c(f)) \times w(f)||^2, \tag{13}$$

where $f \in [1, F]$ stands for each frequency bin in the target frame. Further, $w(f)$ is defined as

$$w(f) = 2 - \frac{f}{F}, \tag{14}$$

and hence, the low frequencies are given more weight to better follow speech characteristics. At evaluation time, the same positive and negative contexts are used to process all contaminated segments that belong to a given audio sample from the test set. Thus, each estimated target centre frame takes into account the entire information of positive and negative contexts. Concatenating the target centre frames leads to the estimated target spectrum, and inverse short-time Fourier transform (iSTFT) is then used to reconstruct the target audio using the phase of the contaminated audio.

5 Usability: experimental results & evaluation

N-HANS, initially developed using Python 3 and TensorFlow 1.14, has been also made compatible with TensorFlow 2 according to the code migration official guidance provided by the platform. Its source code is freely available for developers in a GitHub public repository, and for users who want to directly apply N-HANS, trained models are also accessible via command line interface.[5] Although N-HANS was implemented to process 16 kHz Waveform Audio File Format (WAV) for input and output, i. e., one of the most standard and broadly used audio formats, input files in other formats or sample rates are also handled through an embedded format conversion based on pysox [4], which internally transfer them into WAV format.[6] Both, single or multiple audio files organised into a directory, can be provided as an input for N-HANS, which achieves an optimal performance with GPU-acceleration[7] but is also capable of running with CPU only. To test N-HANS functionalities, a series of experiments, presented in the following, were conducted.

5.1 Speech denoising & selective noise suppression

5.1.1 Dataset & evaluation metrics

To evaluate the performance of N-HANS in the denoising and selective noise suppression tasks, the LibriSpeech [43] and AudioSet [12] databases were considered. The LibriSpeech database provides large-scale clean speech utterances, which consist of approximately 1 000 hours of read speech derived from over 8 000 public domain audiobooks, containing its own train, development, and test splits. The AudioSet corpus contains more than two million human-labelled 10-second environmental sound clips drawn from YoutTube videos.

[5] Source code and functionalities are documented at: https://github.com/N-HANS/N-HANS

[6] Note that for audio files with a quality lower than WAV, such as mp3 in most configurations, the performance of N-HANS may decrease to some degree.

[7] To carry out any of the considered tasks with a single NVIDIA Titan X Pascal GPU, N-HANS takes about 0.272 seconds to operate one second of audio, i. e., roughly resembling a Real-Time-Factor (RTF) of 0.3.

According to AudioSet's ontology, excluding the noise recordings labelled as 'Human sounds', we considered 16 198 samples for training, 636 for development, and 714 for test.[8] A variety of evaluation metrics, including log spectral distortion (LSD), signal-to-distortion ratio (SDR), perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), Mel cepstral distortion (MCD), and segmental SNR (SSNR), which are widely used in prior work [22], were taken into account to assess the performance of N-HANS in several Signal-to-Noise Ratio (SNR) conditions.

As selective noise suppression has not been explored in the literature yet, it is not possible to compare N-HANS performance on selective noise suppression with previous work. To this end, in order to assure a fair comparison, we performed a baseline considering the same model as the one proposed in N-HANS, but being conditioned only on the negative noise contexts. We leverage only the negative embedding subnetwork to learn a negative noise embedding, this is subsequently fed into the denoising subnetwork to assist in the identification of the noise to be suppressed.

5.1.2 Data processing

For selective noise suppression, to create a large and diverse dataset of in-the-wild speech corrupted by two different types of daily-life noise, we mixed each clean spoken utterance from LibriSpeech with two randomly selected environmental recordings from AudioSet. The two environmental recordings were considered as positive and negative noise, respectively.[9] The positive noise, negative noise, and spoken utterance, were truncated—by removing the exceeding signal tails—in order to set them to the same length. Subsequently, to create contaminated audio for training, the positive and negative noises were mixed with each utterance with two randomly selected SNRs in the range of $-3, 0, 1, 3, 5, 8$dB: one selected for the positive noise, i. e., SNR(+); the other for the negative one, i. e., SNR(−). Afterwards, the contaminated segments were randomly selected from the log magnitude spectrum of the contaminated audio, and the positive and negative contexts were created from the parts that did not appear in the contaminated segment.

For test, each pair of positive and negative noises was mixed with each utterance by considering all the possible permutations of SNR pairs in the range of $0, 3, 5, 8$dB. To encourage the model's robustness, a larger variety of SNRs were considered in the training process. Positive and negative contexts were chosen from the beginning of the positive and negative noises, respectively. The test and validation sets were created once, and were consistent across all experiments.

5.1.3 Results on selective noise suppression

The experimental results show that the baseline model, trained on exactly the same data as our proposed architecture, is outperformed by N-HANS on the selective noise suppression task. In Table 5, baseline results (given in parentheses) are indicated for all the evaluation metrics and all the different SNR combinations. The performance gains over the baseline are attributed to the introduction of the complementary auxiliary network that learns the

[8] The partitioning considered for the experiments with AudioSet can be found in the N-HANS's Github repository.

[9] Note that when the positive audio is set to mute, the selective noise suppression system turns into a denoising system capable of adapting to different unseen environments [53, 80].

Table 5 Test results for the selective noise suppression task. SNR(+) and SNR(−) represent the Signal-to-Noise ratio (SNR) applied to the positive and negative noises, respectively

SNR							
(+)	(−)	LSD	SDR	PESQ	STOI	MCD	SSNR
0dB	0dB	0.76 (0.91)	7.72 (6.38)	2.86 (2.70)	0.79 (0.76)	5.36 (5.45)	7.13 (5.51)
	3dB	0.69 (0.83)	9.49 (8.07)	3.09 (2.89)	0.84 (0.81)	4.96 (5.10)	8.55 (6.76)
	5dB	0.65 (0.76)	10.64 (9.35)	3.23 (3.09)	0.87 (0.85)	4.67 (4.65)	9.46 (7.75)
	8dB	0.59 (0.69)	12.12 (11.00)	3.40 (3.28)	0.90 (0.88)	4.29 (4.23)	10.83 (8.88)
3dB	0dB	0.78 (0.92)	7.16 (5.96)	2.78 (2.61)	0.78 (0.75)	5.58 (5.69)	6.56 (5.06)
	3dB	0.73 (0.82)	8.93 (7.81)	2.98 (2.84)	0.83 (0.81)	5.29 (5.25)	7.79 (6.37)
	5dB	0.68 (0.79)	10.06 (8.82)	3.12 (2.98)	0.85 (0.83)	4.97 (5.03)	8.90 (7.19)
	8dB	0.64 (0.72)	11.46 (10.56)	3.29 (3.19)	0.88 (0.87)	4.68 (4.64)	9.93 (8.55)
5dB	0dB	0.81 (0.93)	7.19 (5.80)	2.74 (2.57)	0.78 (0.75)	5.71 (5.87)	6.30 (4.73)
	3dB	0.75 (0.87)	8.68 (7.61)	2.93 (2.76)	0.82 (0.79)	5.44 (5.57)	7.44 (5.91)
	5dB	0.72 (0.82)	9.76 (8.67)	3.06 (2.90)	0.84 (0.82)	5.28 (5.34)	8.27 (6.69)
	8dB	0.65 (0.73)	11.33 (10.36)	3.26 (3.11)	0.88 (0.86)	4.83 (4.86)	9.85 (8.20)
8dB	0dB	0.86 (0.98)	7.01 (5.63)	2.67 (2.48)	0.77 (0.74)	5.99 (6.06)	5.83 (4.04)
	3dB	0.79 (0.89)	8.63 (7.35)	2.86 (2.68)	0.81 (0.79)	5.71 (5.70)	7.07 (5.40)
	5dB	0.74 (0.92)	9.62 (6.38)	2.99 (2.68)	0.84 (0.76)	5.44 (5.54)	7.92 (5.46)
	8dB	0.68 (0.78)	11.30 (10.47)	3.18 (3.02)	0.87 (0.86)	5.09 (5.14)	9.36 (7.77)

For each condition, i. e., a pair of SNR(+) and SNR(−), the following evaluation metrics are given: log spectral distortion (LSD), signal-to-distortion ratio (SDR), perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), Mel cepstral distortion (MCD), and segmental SNR (SSNR). For comparability, the performance of the baseline model is given in parentheses

positive noise embedding. Differently, the baseline model, which does not present the positive embedding, is agnostic to the noisy sources that should be preserved; thus, it ends up by removing some of them. This comparison makes evident the importance of introducing a positive embedding to guarantee an efficient selective noise suppression. The best N-HANS performance on this task is achieved in conditions in which the speech surroundings contain more energy for the positive noise than for the negative, i. e., lower SNR for the positive noise, or higher SNR for the negative one. Indeed, considering the lowest level of SNR on the positive noise (i. e., 0dB), a higher SNR on the negative one yielded generally to a better performance in all the evaluation metrics; cf. the results for SNR(+) = 0dB and SNR(−) = 8dB in Table 5. This is probably due to the fact that the intensity difference between positive and negative noises provides an additional cue for discriminating them; thus, less negative noise is more easily suppressed by the system while consistently preserving the positive one.

5.1.4 Results on speech denoising

By supplying as positive recording a silent audio segment, the N-HANS selective noise suppression system works as an environment-aware speech denoising system. The system processes an in-the-wild speech audio samples and attempts to remove the noise to its greatest extent, based on the identification of the speech surroundings, which is indicated by the negative recording. Considering the same test set evaluation metrics, the experimental

Table 6 Test results for the speech denoising task with N-HANS trained on the LibriSpeech and AudioSet corpora considering the Evaluation Metrics: LSD, SDR, PESQ, STOI, MCD, and SSNR (cf. the caption of Table 5)

SNR	LSD	SDR	PESQ	STOI	MCD	SSNR
0dB	1.17	7.02	2.49	0.81	6.79	4.06
3dB	1.10	8.72	2.70	0.84	6.51	5.10
5dB	1.05	9.60	2.84	0.86	6.40	5.90
10dB	0.93	11.86	3.12	0.90	5.98	7.80
15dB	0.84	13.35	3.34	0.92	5.49	9.58

results achieved for the LibriSpeech and AudioSet corpora indicate that N-HANS produces an audio output of reliable quality—in comparison to other systems [22, 74]—in terms of speech distortion, as indicated by the levels of LSD, SDR, and MCD (cf. Table 6). Furthermore, our system yielded to high STOI results for all the evaluated conditions, even for the lower SNR (STOI = 0.81 for 0dB; cf. Table 6), which indicates the high speech intelligibility of the output.

In order to gain understanding of the proposed method w. r. t. existing approaches, we compare the denoising performance of N-HANS with several state-of-the-art methods recently presented, including: SEGAN [47], Wavenet [49], MMSE-GAN [57], and DCUnet-20 [5]. The comparison is based on two publicly available database: The Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [59], and the Voice Bank corpus [64]; the training and test splits are according to Choi et al. [5]. The comparisons of results are given in terms of PESQ, SSNR, and three additional evaluation metrics [19]: CSIG, i. e., the Mean Opinion Score (MOS) predictor of signal distortion; CBAK, i. e., the MOS predictor of background-noise intrusiveness; and COVL, i. e., the MOS predictor of the overall signal quality. In Table 7, the comparisons of results are given.

Since N-HANS is initially trained for selective noise suppression, the functionality of 'conventional' denoising, i. e., removing all the surrounding noises while extracting only clean speech, is a side-product of supplying a silent segment as positive recording. Besides,

Table 7 Test results for the speech denoising task with SEGAN, Wavenet, MMSE-GAN, DCUnet-20, and N-HANS considering the Evaluation Metrics: CSIG, CBAK, COVL, PESQ, and SSNR (cf. the caption of Table 5)

	CSIG	CBAK	COVL	PESQ	SSNR
SEGAN	3.48	2.94	2.80	2.16	7.73
Wavenet	3.62	3.23	2.98	–	–
MMSE-GAN	3.80	3.12	3.14	2.53	–
DCUnet-20	4.24	4.00	3.69	3.13	15.95
N-HANS (Train 1)	3.60	2.84	2.83	2.05	6.42
N-HANS (Train 2)	4.00	3.18	3.23	2.44	8.24

For N-HANS, results are given considering the Librispeech and AudioSet corpora (Train 1), as well as Voice Bank and DEMAND (Train 2), for training. Note that the other evalauted methods are trained with Voice Bank and DEMAND, thus, results for Train 2 enable a fairer comparison our denoising model is originally trained on Librispeech and AudioSet, hence, its test performance on the Voice Bank and DEMAND corpora is not optimised. Nevertheless, even in such non-optimised conditions, `N-HANS` can still achieve comparable results w. r. t. state-of-the art methods, as shown in Table 7. The proposed denoising model performs slightly better than SEGAN in terms of CSIG and COVL, indicating that our method achieves better signal quality and less distortion. Differently, concerning the other three evaluation metrics, `N-HANS` seems to under-perform in removing noises of the DEMAND corpus. In order to carry out a fair comparison, we trained `N-HANS` with the training set of DEMAND and the Voice Bank corpus, which improved its performance on the test set for all the considered evaluation metrics. Despite a general improvement, our denoising method cannot reach the performance of DCUnet-20, which is due to the fact that DCUnet is trained using the loss function wSDR [5]—specially designed for enhancing hearing experience. However, to benefit from this kind of loss, the model needs as input an audio sample sufficiently long, resulting, in practice, in a lower RTF of the model for inference. Although applying wSDR and employing longer input has the potential to further improve the evaluation results from `N-HANS`, we consider that using wMSE (cf. (13)) ensures real-time processing, which enables the use of the proposed model for many to most realistic applications while keeping reliable denoising performance.

5.2 Source separation

5.2.1 Dataset & evaluation metrics

In order to evaluate the performance of `N-HANS` in the source separation task, the outcomes from our system were compared to two of the recently proposed state-of-the-art baselines for speech source separation [18, 35], which are not conditioned on any additional recordings. The experiments were conducted on the large and diverse VoxCeleb data set [6, 41], which provides more than 2 000 hours of single-channel recordings, encompassing more than one million utterances ($4 - 12$ seconds length each) extracted from Youtube interviews, including more than 7 000 speakers from different nationalities. Since the dataset contains two versions, i. e., VoxCeleb1 and VoxCeleb2—each of them with its own training and test partitioning (consisting of distinct speakers)—we considered as training and test sets the union of the two corresponding sets from both versions. To assess the system performance, according to previous work [65], the three objective evaluation metrics signal-to-distortion ratio (SDR), signal-to-artefacts ratio (SAR), and signal-to-interference ratio (SIR), were considered.

5.2.2 Data processing

In order to improve the separation quality, we enlarged the size of the training set by randomly creating the model inputs. At each iteration, we randomly selected two speakers, i. e., the *target speaker* and the *interference speaker*, and an utterance from each of them, i. e., the *target utterance* and the *interference utterance*. To create a *mixture utterance*, the two utterances were truncated to have the same length and subsequently mixed using a random SNR selected from a wide range (-5 to 25dB), i. e., either $-5, 0, 5, 10, 15, 20$ or 25dB. The positive and negative contexts, which we also refer to as target and interference contexts, were created from the parts that did not appear in the mixture utterance. For creating the test set, target and interference utterances were mixed using an more constricted range of SNR (from -5 to 5dB), i. e., either $-5, -3, -1, 0, 1, 3$ or 5dB—these SNR values have

been selected in order to ensure a fair comparison between our algorithm and previous work [18]. In order to encourage the model's capability of handling real-life scenarios, a wider range of SNRs was used for training. The positive (target) and negative (interference) contexts were chosen from the beginning of the target and interference utterances, respectively. Following the procedure considered for selective noise suppression and speech denoising, the test and validation sets for source separation were created once, and were consistently used across all experiments.
used in previous works [18, 33, 36], since it contains much larger and more diverse daily-life environments; thus, promoting a more realistic understanding of the performance of the model in real-life audio applications. Experiments were performed for female and male speakers (both as target and interference) separately and together, i. e., considering two female speakers (f+f), two male speakers (m+m), and speakers of different genders (f+m); overall result including all speakers (all) are also reported in Table 8.

5.2.3 Results

The performance of N-HANS as a speech separation system was compared with the outcomes of two baseline models re-implemented on the VoxCeleb dataset [18]: one based on Deep Clustering (DC) [10, 11]; the other based on Conv-Tasnet [36]. In separating the speech signals, concerning signal-to-distortion ratio, signal-to-artifacts ratio, and signal-to-interference ratio— evaluation metrics computed using the BSSEval toolbox [58]—N-HANS considerably outperforms the DC baseline by a large margin: two tailed t-test yielded $p < .0004$ for the three evaluated metrics across the evaluated groups (cf. SDR, SAR, and SIR, for DC and N-HANS in Table 8). Concerning the Conv-Tasnet baseline, N-HANS presents also a significant improvement for SDR and SIR: $p < .008$; differently, despite N-HANS outperforms Conv-Tasnet in all the evaluated conditions concerning SAR, the improvement for this evaluation metric is not statistically significant: $p = .558$ (cf. the results for all the conditions of SAR for Conv-Tasnet and N-HANS in Table 8). The presented results show that although DC and Conv-Tasnet had achieved good separation outcomes on the WSJ0 and the TIMIT corpus [10, 11], its performance decreased noticeable when processing the VoxCeleb corpus, which presents a higher complexity. Indeed, the superior performance of N-HANS on this challenging dataset indicates the robustness of the method herein presented.

Concerning speaker's gender, both models, i. e., the baseline (DC) and N-HANS, perform better on speakers of different genders w. r. t. speakers of the same gender (cf. f+m in Table 8). This phenomenon can be attributed to the fact that speech signals from speakers with the same gender share similar acoustic properties, which makes the mixed spectrum more challenging to separate. Comparing the outcomes of the experiments performed on speakers with the same gender w. r. t. those on speakers with different genders, we observe a larger average performance gap between our proposed model and the baseline methods for all three evaluation metrics. We conclude from this observation that especially in the challenging same gender condition, conditioning our model on the additional context recordings resulted in valuable information assisting the separation process. In addition, our source separation system overtakes the label permutation problem [29]—a question that has been only recently afforded [10, 29, 77]. The enhanced network, by learning the additional target and interference recordings, receives indications of the speaker labels, i. e., 'target' and 'interference', which encourages the separation of a mixture utterance and prevents, at the same time, the label permutation problem.

Table 8 Test results for the speech separation task considering the evaluation metrics signal-to-distortion ratio (SDR), signal-to-artifacts ratio (SAR), and signal-to-interference ratio (SIR), for the baseline with Deep Clustering (DC), Conv-Tasnet, and N-HANS methods

	SDR				SAR				SIR			
Method	f+f	m+m	f+m	all	f+f	m+m	f+m	all	f+f	m+m	f+m	all
DC	3.05	2.52	4.33	3.48	5.73	5.32	6.59	6.00	9.46	9.05	11.31	10.21
Conv-Tasnet	4.18	6.16	6.76	6.38	8.29	9.10	10.46	9.94	8.12	9.75	10.83	10.19
N-HANS	**7.62**	**8.92**	**10.58**	**9.47**	**8.61**	**9.71**	**11.04**	**10.15**	**17.13**	18.93	**22.48**	**20.21**

Results for combining female (f) and male (m) speakers are given, followed by their overall average (all) Results for combining female (f) and male (m) speakers are given, followed by their overall average (all)

6 Performance visualisation

To illustrate the performance of N-HANS for its three functionalities, i. e., selective noise suppression, speech denoising, and speech source separation, we depict its processing procedure for some noisy speech samples in Figs. 4, 5, and 6, respectively.

6.1 Selective noise suppression

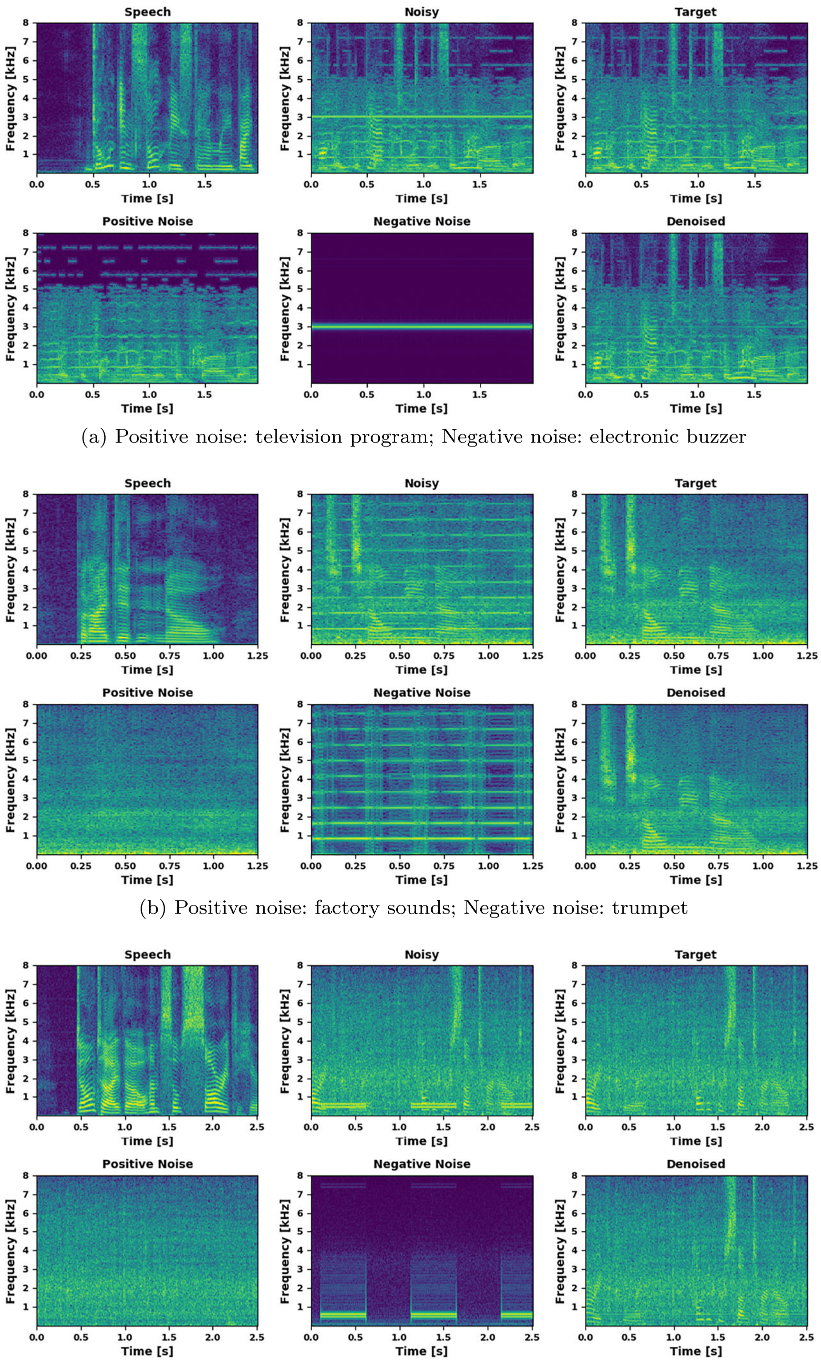
In each sub-figure of Fig. 4, an example of the spoken clean utterance, the noisy background (mixture of positive and negative noises), the target (mixture of the spoken utterance and the positive noise), the positive and negative noises, and the denoised sample (the system output), are presented. The N-HANS selective noise suppression system takes the noisy spectrum as input and removes only the negative noise; thus, the output is expected to be closest to the target spectrum, which includes only speech and positive noise. For a consistent negative noise that concentrates in some narrow frequency range, e. g., that shown in Fig. 4a, our system is able to sort out the noise and maximally retain the speech components and positive noise. Furthermore, processing noise that skips across a wide range of frequency axis, as shown in Fig. 4b, is usually a big challenge for most denoising systems; N-HANS shows also a good performance in such conditions. Finally, the system's capability to recover speech signals under strong noise conditions is displayed in Fig. 4c, where the speech components masked by the negative noise in the noisy spectrum reappear in the denoised output.

6.2 Denoising

In each sub-figure of Fig. 5, an example of the clean utterance, the noisy utterance, the background noise, and the denoised sample (the system output), are displayed to represent the processing procedure of the N-HANS system for speech denoising. In Fig. 5a, a spoken utterance severely covered with a strong industrial noise at an SNR of 0dB is shown. Despite the difficulty in visually recognising the speech content in the noisy spectrum, the system is able to recover the main voice components, enhancing therefore the speech quality of the noisy audio sample. When processing non-continuous noises, i. e., those characterised by specific and isolated impulses, such as those shown in Fig. 5b and in c, the denoising system is capable to remove the noise based on additional noisy recordings. Note that when the surrounding environment contains a noise type with acoustic properties similar to those from the voice (cf. Figure 5c), in order to suppress the noise as much as possible, the system might distort in some extent the estimated speech spectrum. Yet, such distortions have very limited influence to normal human hearing perception. In addition, although the examples given for selective noise suppression referred only to narrow band and non-stationary noises (cf. Fig. 4), the high performance of N-HANS in the suppression of wide-band stationary noise (cf. Figure 5), indicates that wide-band stationary noises, characterised by reach contexts, promote the system's ability to capture the noise to be suppressed— a principle that keeps valid for selective noise suppression too.

6.3 Speech separation

In each sub-figure of Fig. 6, an example of the mixture utterance, the target and interference utterances, and the resulting separated output, are displayed in order to represent N-HANS source separation performance. For each example, the mixture speech was composed from

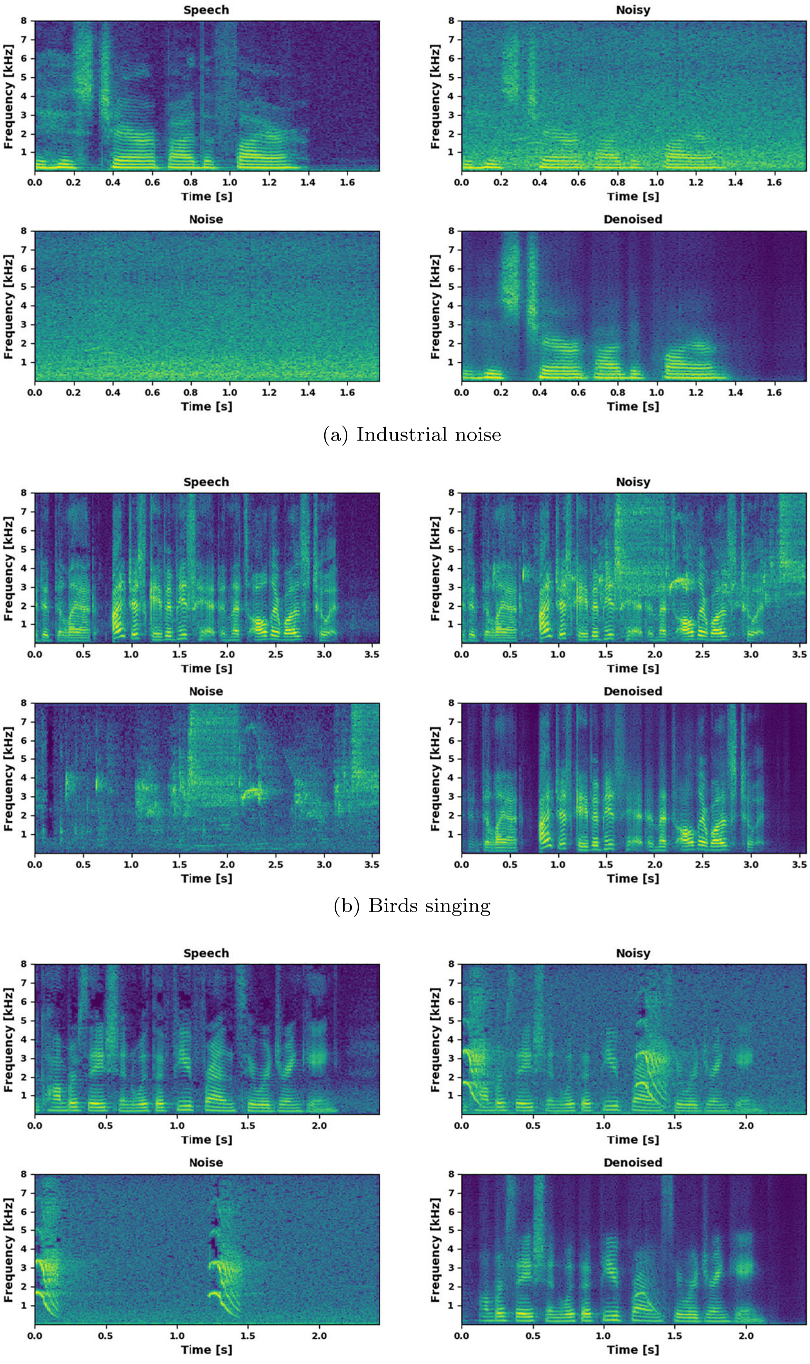


(a) Positive noise: television program; Negative noise: electronic buzzer

(b) Positive noise: factory sounds; Negative noise: trumpet

(c) Positive noise: train; Negative noise: telephone busy signal

Fig. 4 Spectrograms illustrating the audio components involved in the N-HANS selective noise suppression system, i. e., the clean spoken utterance (speech), the contaminated audio (noisy), the ideal result (target), the negative and positive noises, and the achieved outcome (denoised)

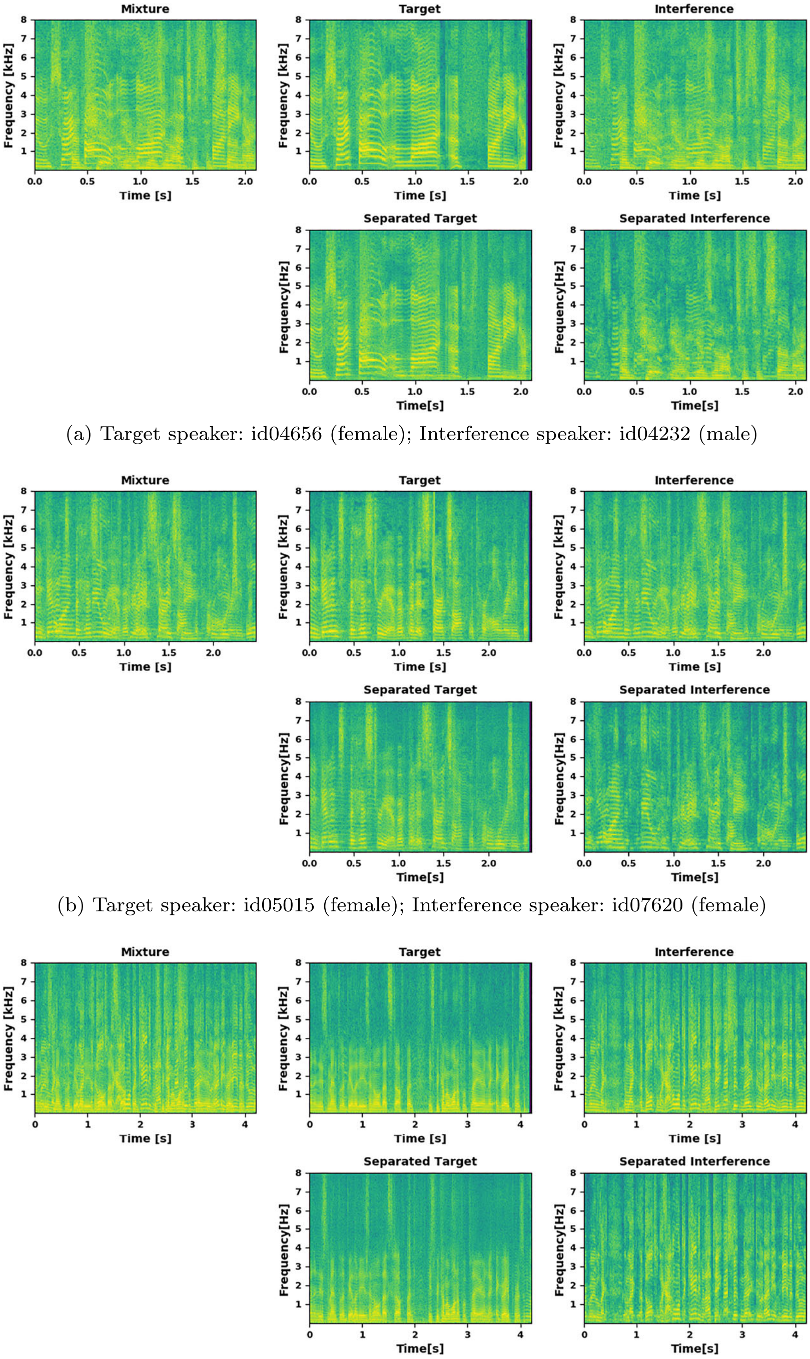


(a) Industrial noise

(b) Birds singing

(c) Bird tweet

Fig. 5 Spectrograms illustrating the audio components involved in the N-HANS denoising system, i. e., the clean spoken utterance (speech), the contaminated audio (noisy), the interfering noise, and the achieved outcome (denoised)



(a) Target speaker: id04656 (female); Interference speaker: id04232 (male)

(b) Target speaker: id05015 (female); Interference speaker: id07620 (female)

(c) Target speaker: id07494 (male); Interference speaker: id01989 (male)

Fig. 6 Spectrograms illustrating the audio components involved in the N-HANS source separation system, i. e., the mixture between the two speakers, and the target and interference speakers before (above) and after (below) to be separated by the system

two utterances, each produced by a different speaker from the VoxCeleb test set. The source separation system takes the mixture as input and produces the separated target and interference speech, which are depicted on the right column. Figure 6a shows the separation performance for two speakers of different genders, while Fig. 6b and c do so for two speakers of the same gender. For the three mixture conditions, the system successfully separates the target from the interference speaker, as shown by the comparison between the separated target and the separated interference w. r. t. their original spectrum, i. e., the target and the interference (cf. plots below w. r. t. the plots above, for each sub-figure). This is particularly clear in the Fig. 6a, i. e., in the separation of speakers from different genders. For instance, in the mixture utterance, although the target speech is particularly disturbed by the interference speech at 0.8 s and 1.5 s, the system is able to suppress the interference components to its maximum extent (cf. Fig. 6a). In addition, the target utterance presents high resolution in the sound wave at low-frequency (under 1 kHz), which is smeared by the interference speech in the mixture spectrum. The source separation system can jointly estimate the amounts of speech components in each time-frequency bin in order to recover the target speech with high clarity in low-frequency range.

7 Conclusions and outlook

We have shown `N-HANS`—an open source toolkit for audio denoising, source separation, and selective noise suppression, based on our proposed ±Auxiliary Network. As such, to the best of the authors' knowledge it is the first toolkit to provide selective noise suppression. Conditioned on reference recordings, `N-HANS` is capable of adapting to different unseen environments and audio sources, such as speakers. `N-HANS` can perform audio enhancement as front end to interface with other audio-related tools such as openXBOW [54] and auDeep [8], both of which have been broadly applied for audio features extraction. Future work for audio enhancement should focus on improving speech intelligibility in extreme low SNR cases, to overcome the distortions that occasionally introduced in audio. For audio separation, more work will be needed to extend the system to any number of audio sources, including, for instance, music source separation.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Atmaca E, Peker I, Altin A (2005) Industrial noise and its effects on humans. Polish J Environ Stud 14(6):721–726
2. Avila AR, Alam MJ, O'Shaughnessy DD, Falk TH (2018) Investigating speech enhancement and perceptual quality for speech emotion recognition. In: Proceedings of INTERSPEECH, Hyderabad, pp 3663–3667

3. Bharitkar S, Kyriakakis C (2003) Selective signal cancellation for multiple-listener audio applications using eigenfilters. IEEE Trans Multimed 5(3):329–338
4. Bittner EJHRM, Bello JP (2018) pysox: Leveraging the audio signal processing power of sox in python. In: Proceedings ISMIR, New York City, pp 3
5. Choi H.-S., Kim J.-H., Huh J, Kim A, Ha J.-W., Lee K (2019) Phase-aware speech enhancement with deep complex u-net. In: ProceedingsICLR, New Orleans, pp 20
6. Chung J, Nagrani A, Zisserman A (2018) VoxCeleb2: Deep Speaker recognition. In: Proceedings of INTERSPEECH, Hyderabad, pp 1086–1090
7. Delic V, Peric Z, Secujski M, Jakovljevic N, Nikolic J, Miskovic D, Simic N, Suzic S, Delic T (2019) Speech technology progress based on new machine learning paradigm. Comput Intell Neurosci 2019:1–19
8. Freitag M, Amiriparian S, Pugachevskiy S, Cummins N, Schuller B (2018) auDeep: Unsupervised Learning of representations from audio with deep recurrent neural networks. J Mach Learn Res 18(173):1–5
9. Fritschi L, Brown A, Kim R, Schwela D, Kephalopoulos S (2011) Burden of disease from environmental noise: Quantification of healthy life years lost in Europe. Bonn. World Health Organization, Germany
10. Garofolo JS, Graff D, Paul D, Pallett D (1993) CSR-i (WSJ0) other. In: Philadelphia: Linguistic data consortium
11. Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS (1993) DARPA TIMIT Acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon Techn Rep 93:27403
12. Gemmeke J, Ellis D, Freedman D, Jansen A, Lawrence W, Moore R, Plakal M, Ritter M (2017) Audio Set: An Ontology and human-labeled dataset for audio events. In: Proceedings of ICASSP, New Orleans, pp 776–780
13. Girin L, Gannot S, Li X (2018) Audio source separation into the wild. Comput Vis Pattern Recogn:53–78
14. Goehring T, Bolner F, Monaghan JJ, Van Dijk B, Zarowski A, Bleeck S (2017) Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users. Hearing Res 344:183–194
15. Goines L, Hagler L (2007) Noise pollution: a modem plague. South Med J 100(3):287–94
16. Gustafsson S, Jax P, Vary P (1998) A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics. In: Proceedings of ICASSP, Seattle, pp 397–400
17. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of CVPR, Las Vegas, pp 770–778
18. Hershey J, Chen Z, Roux J, Watanabe S (2016) Deep clustering: Discriminative embeddings for segmentation and separation. In: Proceedings of ICASSP, Shanghai, pp 31–35
19. Hu Y, Loizou PC (2008) Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio Speech Lang Process 16(1):229–238
20. Huang P, Kim M, Hasegawa-Johnson M, Smaragdis P (2014) Deep learning for monaural speech separation. In: Proceedings of ICASSP, Florence, pp 1562–1566
21. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings ICML, Lille, pp 448–456
22. Jeon K, Kim H (2017) Audio enhancement using local SNR-based sparse binary mask estimation and spectral imputation. Digit Signal Process (68):138–151
23. Jung H, Choi M-K, Jung J, Lee J-H, Kwon S, Young Jung W (2017) Resnet-based vehicle classification and localization in traffic surveillance systems. In: Proceedings of CVPR, Honolulu, pp 61–67
24. Keren G, Han J, Schuller B (2018) Scaling speech enhancement in unseen environments with noise embeddings. In: Proceedings of CHiME, Hyderabad, pp 25–29
25. Kim J, Hahn M (2019) Speech enhancement using a two-stage network for an efficient boosting strategy. IEEE Signal Process Lett 26(5):770–774
26. Kim M, Smaragdis P (2013) Collaborative audio enhancement using probabilistic latent component sharing. In: Proceedings of ICASSP, Vancouver, pp 896–900
27. Kolbæk M, Tan ZH, Jensen J (2016) Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. IEEE/ACM Trans Audio Speech Lang Process 25(1):153–167
28. Kolbæk M, Tan Z, Jensen J (2016) Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification. In: ProceedingsSLT, San Diego, pp 305–311
29. Kolbæk M, Yu D, Tan Z-H, Jensen J (2017) Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Trans Audio Speech Lang Process 25(10):1901–1913
30. Kumar A, Florêncio D (2016) Speech enhancement in multiple-noise conditions using deep neural networks. In: Proceedings of INTERSPEECH, San Francisco, pp 3738–3752

31. Li H, Xu Z, Taylor G, Studer C, Goldstein T (2018) Visualizing the loss landscape of neural nets. In: Proceedings NeurIPS, Montreal, pp 6389–6399
32. Liu D, Smaragdis P, Kim M (2014) Experiments on deep learning for speech denoising. In: Proceedings of INTERSPEECH, Singapore, pp 2685–2689
33. Liu Y, Wang D (2019) Divide and conquer: A deep casa approach to talker-independent monaural speaker separation. IEEE/ACM Trans Audio Speech Lang Process 27(12):2092–2102
34. Lu X, Tsao Y, Matsuda S, Hori C (2013) Speech enhancement based on deep denoising autoencoder. In: Proceedings INTERSPEECH, Lyon, pp 436–440
35. Luo Y, Mesgarani N. (2018) TaSNet: Time-domain audio separation network for real-time, single-channel speech separation. In: Proceedings of ICASSP, Calgary, pp 696–700
36. Luo Y, Mesgarani N (2019) Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Trans Audio Speech Lang Process 27(8):1256–1266
37. Michelsanti D, Tan Z-H (2017) Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. In: Proceedings of INTERSPEECH, Stockholm, pp 2008–2012
38. Miedema H, Oudshoorn C (2001) Annoyance from transportation noise: Relationships with exposure metrics DNL, and DENL and their confidence intervals. Environ Health Perspect 109(4):409–416
39. Ming J, Srinivasan R, Crookes D (2011) A corpus-based approach to speech enhancement from nonstationary noise. IEEE Trans Audio Speech Lang Process 19(4):822–836
40. Monaghan J, Goehring T, Yang X, Bolner F, Wang S, Wright GM, Bleeck S (2017) Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners. J Acoust Soc Amer 141(3):1985–1998
41. Nagrani A, Chung J, Zisserman A (2017) VoxCeleb: A Large-scale speaker identification dataset. In: Proceedings of INTERSPEECH, Stockholm, pp 2616–2620
42. Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. In: Proceedings ICML, Haifa, pp 807–814
43. Panayotov V, Chen G, Povey D, Khudanpur S (2015) Librispeech: an ASR corpus based on public domain audio books. In: Proceedings of ICASSP, Brisbane, pp 5206–5210
44. Parada-Cabaleiro E, Baird A, Batliner A, Cummins N, Hantke S, Schuller B (2017) The perception of emotions in noisified nonsense speech. In: Proceedings INTERSPEECH, Stockholm, pp 3246–3250
45. Parada-Cabaleiro E, Batliner A, Baird A, Schuller B (2020) The perception of emotional cues by children in artificial background noise. Int J Speech Technol 23:169–182
46. Pariente M, Cornell S, Cosentino J, Sivasankaran S, Tzinis E, Heitkaemper J, Olvera M, Stöter F-R, Hu M, Martín-Doñas JM, Ditter D, Frank A, Deleforge A, Vincent E (2020) Asteroid: the PyTorch,-based audio source separation toolkit for researchers. arXiv:2005.04132
47. Pascual S, Bonafonte A, Serrà J (2017) SEGAN: Speech enhancement generative adversarial network. In: Proceedings INTERSPEECH, Stockholm, pp 3642–3646
48. Pascual S, Serrà J, Bonafonte A (2019) Towards generalized speech enhancement with generative adversarial networks. In: Proceedings of INTERSPEECH, Graz, pp 1791–1795
49. Rethage D, Pons J, Serra X (2018) A wavenet for speech denoising. In: Proceedings ICASSP, Calgary, pp 5069–5073
50. Roma G, Grais E, Simpson A, Sobieraj I, Plumbley M (2016) Untwist: A new toolbox for audio source separation. In: ProceedingsISMIR, New York City, pp 4
51. Salaün Y, Vincent E, Bertin N, Souviraà-Labastie N, Jaureguiberry X, Tran D, Bimbot F (2014) The flexible audio source separation toolbox version 2.0. In: ProceedingsICASSP, Florence, pp 3
52. Santurkar S, Tsipras D, Ilyas A, Madry A (2018) How does batch normalization help optimization? In: ProceedingsneurIPS, Montreal, pp 2483–2493
53. Sari L, Hasegawa-Johnson M (2018) Speaker adaptation with an auxiliary network. In: ProceedingsML-SLP, Hyderabad, pp 3
54. Schmitt M, Schuller B (2017) OpenXBOW – introducing the Passau open-source crossmodal Bag-of-Words toolkit. J Mach Learn Res 18(96):1–5
55. Shenoy R, Patwardhan PP, Putraya GG (2017) Spatial audio enhancement apparatus. United States Patent 9769588
56. Shon S, Tang H, Glass JR (2019) VoiceID loss: Speech enhancement for speaker verification. In: Proceedings INTERSPEECH, Graz, pp 2888–2892
57. Soni MH, Shah N, Patil HA (2018) Time-frequency masking-based speech enhancement using generative adversarial network. In: ProceedingsICASSP, Calgary, pp 5039–5043
58. Stöter F-R, Liutkus A, Ito N (2018) The 2018 signal separation evaluation campaign. In: Proceedings LVA/ICA, Guildford, pp 293–305
59. Thiemann J, Ito N, Vincent E (2013) The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings. J Acoust Soc Amer 133(5):6

60. Tolooshams B, Giri R, Song AH, Isik U, Krishnaswamy A (2020) Channel-attention dense u-net for multichannel speech enhancement. In: Proceedings of ICASSP, Barcelona, pp 836–840
61. Triantafyllopoulos A, Keren G, Wagner J, Steiner I, Schuller B (2019) Towards robust speech emotion recognition using deep residual networks for speech enhancement. In: Proceedings of INTERSPEECH, Graz, pp 1691–1695
62. Tzivian L, Dlugaj M, Winkler A, Weinmayr G, Hennig F, Fuks KB, Vossoughi M, Schikowski T, Weimar C, Erbel R et al (2016) Long-term air pollution and traffic noise exposures and mild cognitive impairment in older adults: A cross-sectional analysis of the Heinz Nixdorf recall study. Environ Health Perspect 124(9):1361–1368
63. Valin J (2018) A hybrid DSP/deep learning approach to real-time full-band speech enhancement. In: ProceedingsMMSP, Vancouver, pp 1–5
64. Veaux C, Yamagishi J, King S (2013) The voice bank corpus: design, collection and data analysis of a large regional accent speech database. Proceedings O-COCOSDA/CASLRE, pp 1–4
65. Vincent E, Gribonval R, Févotte C (2006) Performance measurement in blind audio source separation. IEEE/ACM Trans Audio Speech Lang Process 14(4):1462–1469
66. Vincent E, Virtanen T, Gannot S (2018) Audio source separation and speech enhancement. Wiley, Hoboken
67. Vydana HK, Vuppala AK (2017) Residual neural networks for speech recognition. In: Proceedings of EUSIPCO, Kos island, pp 543–547
68. Wang D, Chen J (2018) Supervised speech separation based on deep learning: An overview. IEEE/ACM Trans Audio Speech Lang Process 26(10):1702–1726
69. Wang Z, Roux JL, Hershey JR (2018) Alternative objective functions for deep clustering. In: Proceedings of ICASSP, Calgary, pp 686–690
70. Weninger F, Lehmann A, Schuller B (2011) OpenBliSSART: Design and evaluation of a research toolkit for blind source separation in audio recognition tasks. In: Proceedings of ICASSP), Wuhan, pp 1625–1628
71. Westhausen NL, Meyer BT (2020) Dual-signal transformation LSTM network for real-time noise suppression. In: Proceedings of INTERSPEECH, Shanghai, pp 2477–2481
72. Wittkop T, Hohmann V (2003) Strategy-selective noise reduction for binaural digital hearing aids. Speech Commun (39):111–138
73. Wright B, Peters E, Ettinger U, Kuipers E, Kumari V (2014) Understanding noise stress-induced cognitive impairment in healthy adults and its implications for schizophrenia. Noise Health 16(70):166–176
74. Xu Y, Du J, Dai L, Lee C (2014) An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process Lett 21(1):65–68
75. Xu Y, Du J, Dai L, Lee C (2015) A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 23(1):7–19
76. Xu R, Wu R, Ishiwaka Y, Vondrick C, Zheng C (2020) Listening to sounds of silence for speech denoising. In: ProceedingsNeurIPS, Vancouver, pp 6
77. Yu D, Kolbæk M, Tan Z-H, Jensen J (2017) Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In: Proceedings ICASSP, pp 241–245
78. Yu G, Mallat S, Bacry E (2008) Audio denoising by time-frequency block thresholding. IEEE Trans Signal Process 56(5):1830–1839
79. Zannin PH, Calixto A, Diniz FB, Ferreira JA (2003) A survey of urban noise annoyance in a large Brazilian city: The importance of a subjective analysis in conjunction with an objective analysis. Environ Impact Assess Rev 23(2):245–255
80. Zhang J, Tian G, Mu Y, Fan W (2014) Supervised deep learning with auxiliary networks. In: Proceedings of KDD, New York, pp 353–361