A Phoneme-Scale Assessment of Multichannel Speech Enhancement Algorithms

Nasser-Eddine Eddine Monir, Paul Magron, Romain Serizel

Trends in Hearing
Volume 28: 1–22

DOI: 10.1177/23312165241292205
journals.sagepub.com/home/tia
Sage

A Phoneme-Scale Assessment of Multichannel Speech Enhancement Algorithms

Nasser-Eddine Monir, Paul Magron and Romain Serizel

Abstract

In the intricate acoustic landscapes where speech intelligibility is challenged by noise and reverberation, multichannel speech enhancement emerges as a promising solution for individuals with hearing loss. Such algorithms are commonly evaluated at the utterance scale. However, this approach overlooks the granular acoustic nuances revealed by phoneme-specific analysis, potentially obscuring key insights into their performance. This paper presents an in-depth phoneme-scale evaluation of three state-of-the-art multichannel speech enhancement algorithms. These algorithms—filter-and-sum network, minimum variance distortionless response, and Tango—are here extensively evaluated across different noise conditions and spatial setups, employing realistic acoustic simulations with measured room impulse responses, and leveraging diversity offered by multiple microphones in a binaural hearing setup. The study emphasizes the fine-grained phoneme-scale analysis, revealing that while some phonemes like plosives are heavily impacted by environmental acoustics and challenging to deal with by the algorithms, others like nasals and sibilants see substantial improvements after enhancement. These investigations demonstrate important improvements in phoneme clarity in noisy conditions, with insights that could drive the development of more personalized and phoneme-aware hearing aid technologies. Additionally, while this study provides extensive data on the physical metrics of processed speech, these physical metrics do not necessarily imitate human perceptions of speech, and the impact of the findings presented would have to be investigated through listening tests.

Keywords

multichannel speech enhancement, phoneme-scale evaluation, binaural speech enhancement, hearing aids

Received 8 January 2024; Revised received 4 August 2024; accepted 27 September 2024

Introduction

Speech within a noisy environment is a complicated scenario that can substantially diminish the clarity of spoken words. Background noise can obscure important acoustic cues, challenging listeners in differentiating individual speech sounds and words. Speech enhancement is a solution to enhance speech intelligibility in noisy environments (Loizou, 2007). This technique estimates the speech signal from the noisy mixture, by relying on acoustic cues and temporal patterns inherent to the speech. Speech enhancement algorithms are broadly categorized into two types: single channel and multichannel, depending on the number of available microphones to record the sound.

In scenarios where audio is captured through a single microphone, speech enhancement algorithms concentrate on temporal, frequency, and spectro-temporal characteristics to filter out the noise (Loizou, 2007). Such single-channel speech enhancement is limited to the information captured by one reference point and often focuses on aspects like noise variance over time or spectral consistency.

Conversely, multichannel speech enhancement algorithms harness the power of spatial diversity by exploiting the various captures of speech across microphones (Benesty et al., 2008). This multifold capture allows for considering the spatial characteristics and the directionality of sound. By comparing the different signal channels obtained at the microphones, these algorithms offer a more robust reconstruction of the original speech, effectively mitigating the masking effects of background noise.

As a subset of multichannel speech enhancement algorithms, beamformers manipulate spatial sound attributes using microphone arrays (Benesty et al., 2008). Unlike broader multichannel algorithms that filter or cancel noise,

Université de Lorraine, CNRS, Inria, Loria, Nancy, France

Corresponding author:
Nasser-Eddine Monir, Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France.
Email: nasser-eddine.monir@inria.fr

beamformers enhance speech intelligibility by precisely manipulating the spatial attributes of the acoustic signal, such as sound's directionality. They amplify speech from a specific direction while reducing noise and reverberation from others. This targeted approach is particularly effective in noisy environments, where it isolates the speaker's voice from disruptive background sounds, enhancing speech clarity and intelligibility.

Current state-of-the-art multichannel speech enhancement systems are characterized by advanced beamforming algorithms and the integration of neural networks to improve the intelligibility of speech in noise. The minimum variance distortionless response (MVDR) beamformer optimizes noise reduction while preserving the desired speech directionality (Capon, 1969; Heymann et al., 2016). In hybrid algorithms, neural networks provide parameters for signal processing filters. These include the distributed multichannel Wiener filter (MWF) (Bertrand & Moonen, 2010; Furnon et al., 2021) and its adaptations like the generalized eigenvalue decomposition MWF (GEVD-MWF) (Serizel et al., 2014). Alternatively, some algorithms relying entirely on neural networks have been proposed, such as the filter-and-sum network (FaSNet) beamformer (Luo et al., 2019). This model uses neural networks to directly predict signals rather than the parameters of a spatial filter, which allows for enhanced flexibility in optimization. These developments reflect a shift towards sophisticated, binaural processing setups where hearing aids on both sides collaborate, leveraging spatial information to differentiate speech from noise effectively (Kollmeier & Koch, 1994; Van den Bogaert et al., 2009).

Usually, speech enhancement algorithms are evaluated at the utterance scale using objective signal-to-noise ratio (SNR)-like metrics, which offers a convenient way to quantify their performance at a coarse level and compare algorithms. However, this evaluation process does not capture the nuanced ways different phonemes interact with noise, nor the way algorithms process these phonemes, which potentially simplify their true effectiveness. Studies contrasting English phoneme recognition in noise for native and nonnative speakers reveal this complexity (Adachi et al., 2006). For instance, Miller and Nicely (1955) indicate that consonants vary in noise tolerance, suggesting that some phonemes are more susceptible to noise masking than others. This variance may significantly affect the perceived effectiveness of speech enhancement models. Furthermore, phoneme confusion observed in both human and automatic speech recognition systems suggests that consonants and vowels experience a different impact from information loss due to noise (Meyer et al., 2010; Zaar & Dau, 2017). Studies have shown that a degraded classification of voicing can lead to more confusion between voiced and unvoiced phonemes, such as /p/ and /b/. In contrast, phonemes differing in the place of articulation, like /p/ and /d/, remain distinguishable (Dubno & Levitt, 1981; Gelfand et al., 1985). Additionally, different amplification strategies affect phoneme perception in hearing-impaired listeners (Scheidiger Christoph, 2017).

Research on phoneme recognition, such as the study by Meyer et al. (2010), shows that intrinsic speech variations (e.g., speaking rate, effort, style, and dialect) significantly affect phoneme recognition in noisy environments. For instance, Li et al. (2010) showed that the robustness of stop consonants to noise relies on dominant acoustic features like bursts and F2 transitions. Woods et al. (2010) investigated consonant identification in consonant–vowel–consonant syllables presented in speech-spectrum noise, revealing that baseline SNRs required for consonant identification vary by more than 40 dB across different consonants. Furthermore, Phatak and Allen (2007) demonstrated that consonants can be grouped into three distinct sets based on their susceptibility to noise masking: low-scoring consonants such as /f/, /θ/, /v/, /ð/, /b/, and /m/; high-scoring consonants such as /t/, /s/, /z/, /b/, and /ʧ/; and an intermediate set including consonants such as /n/, /p/, /g/, /k/, and /d/. These groups highlight how different consonants exhibit a varying resilience to noise, with significant implications for the improvement and evaluation of speech enhancement algorithms.

In this paper, we propose evaluating three state-of-the-art speech enhancement algorithms at the phoneme scale for a nuanced analysis that aligns with the distinct acoustic properties of phonetic elements. Such detailed scrutiny can reveal the specific strengths and weaknesses of algorithms in preserving the fidelity of speech sounds. This approach also offers valuable insights for the design of future speech enhancement algorithms, ensuring they are tuned to enhance phonemic clarity by accounting for the unique acoustic characteristics of specific phonemes.

The rest of this paper is structured as follows. First, we provide an overview of multichannel speech enhancement by setting the problem and detailing the algorithms we use in our study. Then, the Methodology section delves into the process of data collection and generation, and notably highlights the phoneme classification. The next section describes our extensive experiments and discusses its results, with a particular emphasis on the phoneme-scale evaluation. Finally, the last section draws some concluding remarks.

Overview of Multichannel Speech Enhancement

Problem Statement and Notations

Consider an acoustic scenario with two punctuate sources and several distant microphones. One source is the target speech, while the other is some interfering noise. In the case of hearing aids, we have M microphones on each

hearing aid and two hearing aids: one on the left (L) and one on the right (R). This scenario is illustrated in Figure 1.

We note S and $N \in \mathbb{R}^{\tilde{T}}$ the time-domain speech and noise signals, where $\tilde{T}$ denotes the length (in samples) of these signals. Assuming the signals are band limited, we express these in the time–frequency domain using the short-time Fourier transform. The target speech source and interfering noise are denoted S and N, respectively.[1] The contribution of the speech signal recorded at the mth microphone of the right (respectively left) hearing aids is denoted $S_{R,m}$ (respectively $S_{L,m}$). Similarly, $N_{R,m}$ and $N_{L,m}$ denote the interfering noise contribution at the mth microphone of the right and left hearing aids, respectively. These signals are referred to as *images* of speech (respectively noise) at the microphones.

The noisy mixture signal at the hearing aids is the combination of the speech and noise images:

$$X_{i,m}=S_{i,m}+N_{i,m}, \quad i \in \{L,\ R\}.$$

We denote $\mathbf{X}_R[X_{R,1},\ \ldots,\ X_{R,M}]$ the set of noisy signals at the right hearing aid, and similarly for $\mathbf{X}_L$, and $\mathbf{X}_{\text{Bin}}$ as the set of all mixture signals. The same notation applies to the target speech and noise signals.

Speech enhancement encompasses noise reduction (Loizou, 2007), dereverberation (Naylor & Gaubitch, 2010), or both. Noise reduction aims to estimate the speech signal at a microphone $S_{i,m}$ given the recorded mixture $\mathbf{X}_i\ i \in \{L,\ R,\ \text{Bin}\}$. Dereverberation aims to estimate the speech source S from the speech recorded by one or several microphone $S_{i,m}$. Estimating the speech source S from the recorded mixture $\mathbf{X}_i$ combines noise reduction and dereverberation. This paper focuses on noise reduction.

In hearing aid scenarios, there are two possible speech enhancement setups (Figure 2). In the bilateral setup, there is no communication between the hearing aids, and each side is processed independently. Therefore, the input of the left hearing aid filter is $\mathbf{X}_L$, and the input of the right hearing aid filter is $\mathbf{X}_R$. Each hearing aid acts as a compact microphone array (Benesty et al., 2008).

In the binaural speech enhancement setup (Kollmeier & Koch, 1994), the hearing aids communicate with each other, giving each filter access to more diverse information improving effectiveness in asymmetric scenarios (Van den Bogaert et al., 2009). This is also known as distributed microphone arrays (Bertrand & Moonen, 2010). Each hearing aid filter processes the entire set $\mathbf{X}_{\text{Bin}}$ as input, but with different ordering $\mathbf{X}_{\text{Bin},R} = [\mathbf{X}_R, \mathbf{X}_L]$ for the right filter ($\mathbf{X}_{\text{Bin},L} = [\mathbf{X}_L, \mathbf{X}_R]$ for the left one). Each filter uses a different reference signal, usually from the ear of interest (a microphone from the left ear is used as a reference for the filter that will produce the output signal sent to the left ear) (Bronkhorst & Plomp, 1988). In this paper, we will focus on the binaural setup and assume perfect signal transmission without any latency or packet loss, the impact of which has been studied by Cornelis (2014).

Algorithms

Multichannel filters exploit the spatial information about the acoustic scene obtained through multiple microphones. Because of their ability to focus on one direction in space, these algorithms are commonly referred to as beamformers. Recently, the use of neural networks in multichannel speech enhancement algorithms has significantly improved their performance and applicability in realistic scenarios.

These algorithms can be divided into two categories. On the one hand, hybrid algorithms combine traditional signal processing spatial filters (obtained as a solution to an optimization problem—see below) with neural networks that estimate these filters' parameters (Carbajal et al., 2020; Hendriks & Gerkmann, 2011; Heymann et al., 2016; Nugraha et al., 2016). On the other hand, end-to-end algorithms use neural networks to directly estimate signals or multichannel filters, optimizing parameters on training sets (Dowerah et al., 2023; Luo et al., 2019; Tolooshams et al., 2020).

In this paper, we study the behavior of three different algorithms. The motivation for choosing these is threefold. Firstly, their source code and trained parameters are available publicly. Secondly, these algorithms can be applied in a binaural enhancement setup. Finally, they cover a wide variety of methodologies among the neural-based multichannel speech enhancement algorithms. The first two algorithms integrate neural networks within signal processing-based filtering; thus, they belong to the category of hybrid algorithms. One algorithm is designed principally for compact microphone arrays and relies on a single-channel neural network, while the other is designed for distributed arrays and relies on a multichannel neural network. The last algorithm is fully based on neural networks; thus, it is representative of the category of end-to-end algorithms.

Minimum Variance Distortionless Beamformer

The MVDR beamformer is a particular spatial filtering technique, which in general can be expressed as applying a filter $W \in \mathbb{C}^M$ to the vector of noisy mixtures X to yield an enhanced signal via $W^H X$, where H denotes the Hermitian transpose. More specifically, in MVDR, the goal is to design a filter that minimizes the noise contribution in the noisy mixture while the signal coming from the target direction (here, the target speech) is left unaltered (Capon, 1969). This can be formulated as solving the following constrained optimization process:

$$\min_{\boldsymbol{W}} ||\boldsymbol{W}^H \boldsymbol{N}_i||^2 \text{ subject to } \boldsymbol{W}^H \boldsymbol{d} = 1,$$

where $\mathbf{d} \in \mathbb{C}^M$ is the steering vector. This optimization problem is usually reformulated as follows:

$$\min_{\mathbf{W}} \mathbf{W}^H \mathbf{R}_N \mathbf{W},$$

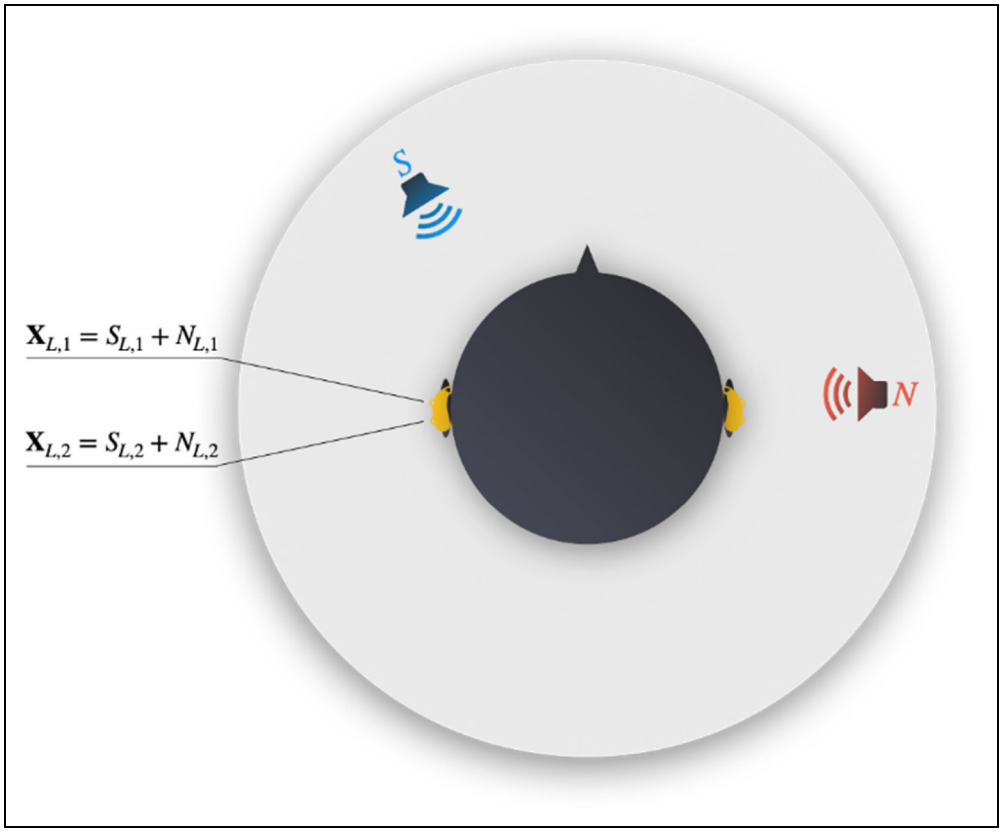

Figure 1. Spatialized acoustic scenarios with two sources: a speech source and a noise source. The acoustic sources are point sources. The signal $X_{i,m}$ recorded by each of the microphone is the sum of the reverberated images of these sources $S_{i,m}$ and $N_{i,m}$ at the hearing aid microphones.

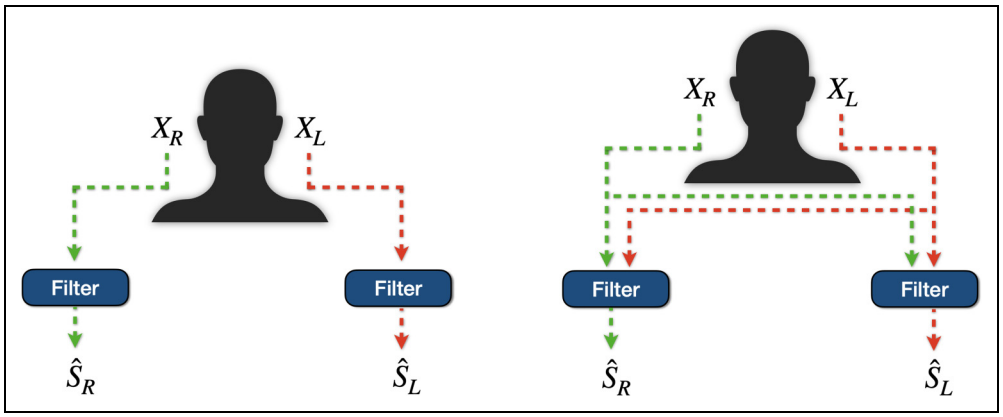

Figure 2. Speech enhancement setups: bilateral (left) and binaural (right).

where R_N is the correlation matrix of the noise component. Solving this optimization problem leads to the so-called MVDR filter:

$$\mathbf{W}_{\text{MVDR}} = \frac{\mathbf{R}_N^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{R}_N^{-1}\mathbf{d}}.$$

Instead of estimating the steering vector d and the noise correlation matrix R_N, to compute this filter, an alternative is to estimate the steering vector as the principal component of the correlation matrix of the speech component $\mathbf{R}_S$. Thus, computing the MVDR filter relies solely on estimating the speech and noise correlation matrices.

Heymann et al. (2016) proposed to estimate these matrices using time–frequency masks computed with a recurrent neural network (Heymann et al., 2016). The noisy mixture is input to the neural networks, which returns a speech

mask coefficient M_S indicating the speech presence in the mixture in each time–frequency point (Figure 1). The speech correlation matrix is then obtained as follows:

$$\mathbf{R}_S(f) = \frac{1}{T} \sum_{t=1}^{T} M_S(t,f)\mathbf{X}(t,f)^H\mathbf{X}(t,f). \quad (1)$$

A similar process is conducted to obtain the noise correlation component. As a result, the MVDR filter is calculated in each frequency channel, but it is time independent (Figure 3).

Distributed MWF

The goal of the MWF is to estimate the speech component at an arbitrary reference microphone S_{ref} (Doclo & Moonen, 2002). This can be formulated as minimizing the mean squared error:

$$\min_{\mathbf{W}} \|\mathbf{W}^H\mathbf{X} - S_{\text{ref}}^2\|,$$

where $S_{\text{ref}} = d_{\text{ref}}^H X$, and $\mathbf{d}_{\text{ref}}$ is a vector whose entries are equal to 0, except for the one corresponding to the reference channel where it is equal to 1. Solving this optimization problem leads to the MWF formula:

$$\mathbf{W}_{\text{MWF}} = \mathbf{R}_X^{-1}\mathbf{R}_S\mathbf{d}_{\text{ref}}.$$

The computation of the MWF relies on the correlation matrices $\mathbf{R}_X$ and $\mathbf{R}_S$, which can be estimated with neural networks as for the MVDR.

Variants of the MWF include the speech distortion weighted MWF for balancing the noise reduction and speech distortion (Spriet et al., 2004), and the GEVD-MWF for a more robust filtering in noisy conditions (Serizel et al., 2014). This latter variant is used in this paper.

Bertrand and Moonen (2010) proposed an algorithm adapted to microphone arrays. Initially, this algorithm assumed perfect voice activity detection to estimate the correlation matrices. It was later adapted to use a two-step neural network-based mask estimation algorithm called Tango (Furnon et al., 2021).

In the first stage, a mask is obtained for one local channel, similarly to the MVDR. This mask helps in isolating the primary speech signal by attenuating the background noise. Specifically, the mask is estimated by focusing on the time–frequency representation of the signal, where the neural network identifies and suppresses the noise components while preserving the speech components. In the second stage, the neural network uses signals from both ears to jointly estimate the masks, incorporating binaural cues to refine and enhance the initial mask. This process involves the neural network analyzing the spatial and spectral characteristics of the signals from both ears. By leveraging the interaural time and level differences, the network can better differentiate between the target speech and background noise. The binaural integration allows the model to exploit the spatial separation between the speech and noise sources, leading to a more precise and robust mask estimation (Figure 4).

FaSNet Beamformer

The last beamformer that we study in this paper is a beamformer that relies mainly on neural networks, the so-called FaSNet beamformer (Luo et al., 2019). More precisely, the beamformer itself is computed with neural networks and directly applied to the signals recorded by the microphones.

This algorithm also operates in two stages. In the first step, the algorithm computes a filtered speech signal at an arbitrary reference channel using the multichannel input mixture (here: four-channel). In the second step, this filtered reference signal is used to compute pairwise beamformers for all other channels. Unlike the previous hybrid algorithms, this one is directly trained in an end-to-end fashion by minimizing a loss between the clean and enhanced speech signals. In the paper, the FaSNet beamformer is trained to optimize a scale-invariant signal-to-distortion ratio (SI-SDR) (Le Roux et al., 2019).

Methodology

First, we present the pipeline used for simulating multichannel speech data in realistic noisy conditions. Then, we introduce the evaluation metrics. Finally, we describe the phoneme categories and the classification method that allows for a fine-grain assessment.

Data Generation

We simulate mixtures that replicate diverse acoustic scenarios with the target speech contaminated by interfering noise. To that end, we consider real-life speech excerpts and measured room impulse response (RIR), while the mixtures themselves are simulated. This approach allows for the creation of a large quantity of signals with various configurations, which would be complex and costly to obtain through real-world recordings.

Speech Data. The speech data used to simulate mixtures comprises 1,000 speech signals extracted from the test set of LibriSpeech (Panayotov et al., 2015). This dataset is a comprehensive corpus encompassing approximately 1000 h of English speech. The data is obtained from audiobooks within the LibriVox project, where audio recordings have been aligned with their corresponding texts and partitioned into short segments. The signals are sampled at 16 kHz.

Noise Types. We consider both synthetic and recorded noise types (see Figure 5). This approach allows us to observe and understand the different behavior of the speech enhancement algorithms described above.

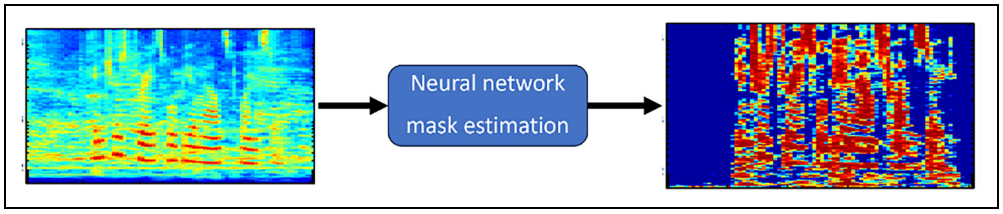

Figure 3. Neural network-based mask estimation. The neural network is fed noisy signal at the input and provides a mask that indicates the amount of speech in each time–frequency bin.

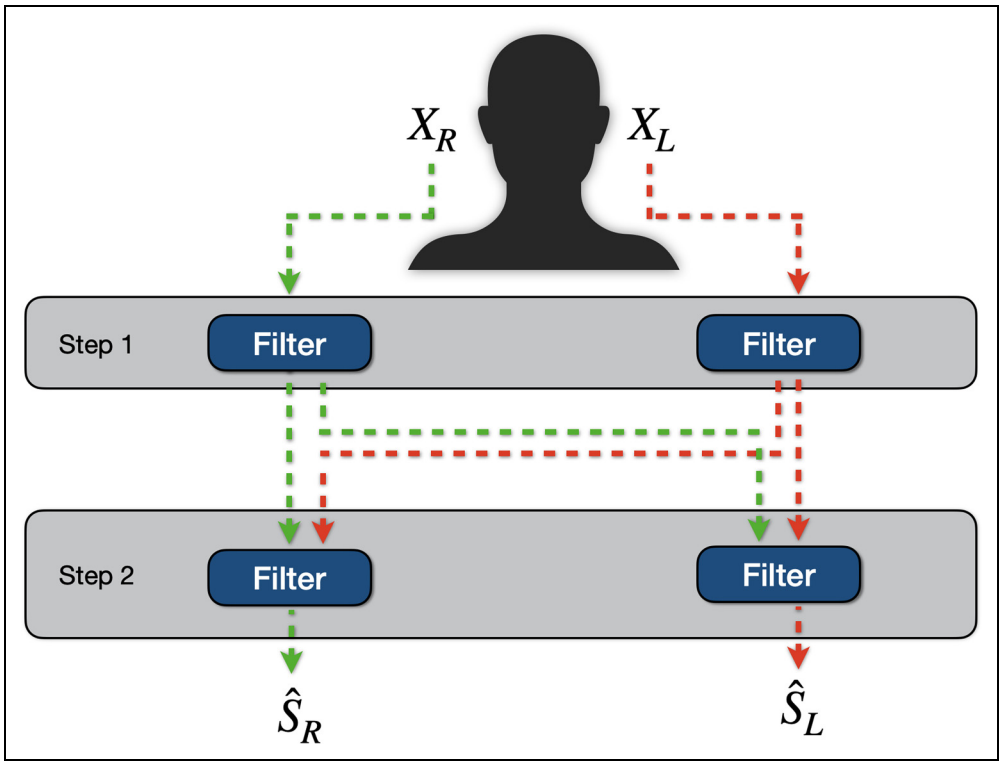

Figure 4. Distributed MWF in the binaural case.

First, we consider white noise, which is characterized by its uniform frequency distribution (Keith & Talis, 1972). Using white noise in our experiments provides a consistent baseline for evaluating speech enhancement algorithms. Indeed, testing our algorithms against white noise ensures that they can handle even simple noise types effectively, serving as a foundational benchmark. This approach aligns with many standard clinical tests that still employ white noise (Reynard et al., 2021).

We also consider speech-shaped noise,[2] a type of noise signal designed to mimic the average spectral characteristics of natural human speech. Unlike white noise, which has a uniform frequency distribution, speech-shaped noise is crafted to simulate the energy distribution across frequencies of speech sounds. To generate speech-shaped noise, we use five speech signals (from three females and two males) from LibriSpeech that are not part of our mixture subset. By transforming the signal (i.e., preserving its magnitude and randomizing its phase), we create a noise signal with the same spectral properties as the original speech.

Finally, we consider a babble noise signal taken from Freesound (Font et al., 2013). The selected clip was recorded in a restaurant during a lunch break. This audio signal consists of a soundscape where multiple people are conversing simultaneously. Unlike synthetic noise signals (e.g., white noise or speech-shaped noise), babble noise comprises overlapping speech from various speakers in the same acoustic scene, which introduces a higher amount of acoustic complexity. This complexity mirrors real-world scenarios to investigate how individuals with hearing loss navigate challenging auditory scenes.

RIRs. A RIR is a filter that describes the impact of sound propagation within a room from the position where the sound source is emitted to the microphone where it is recorded.

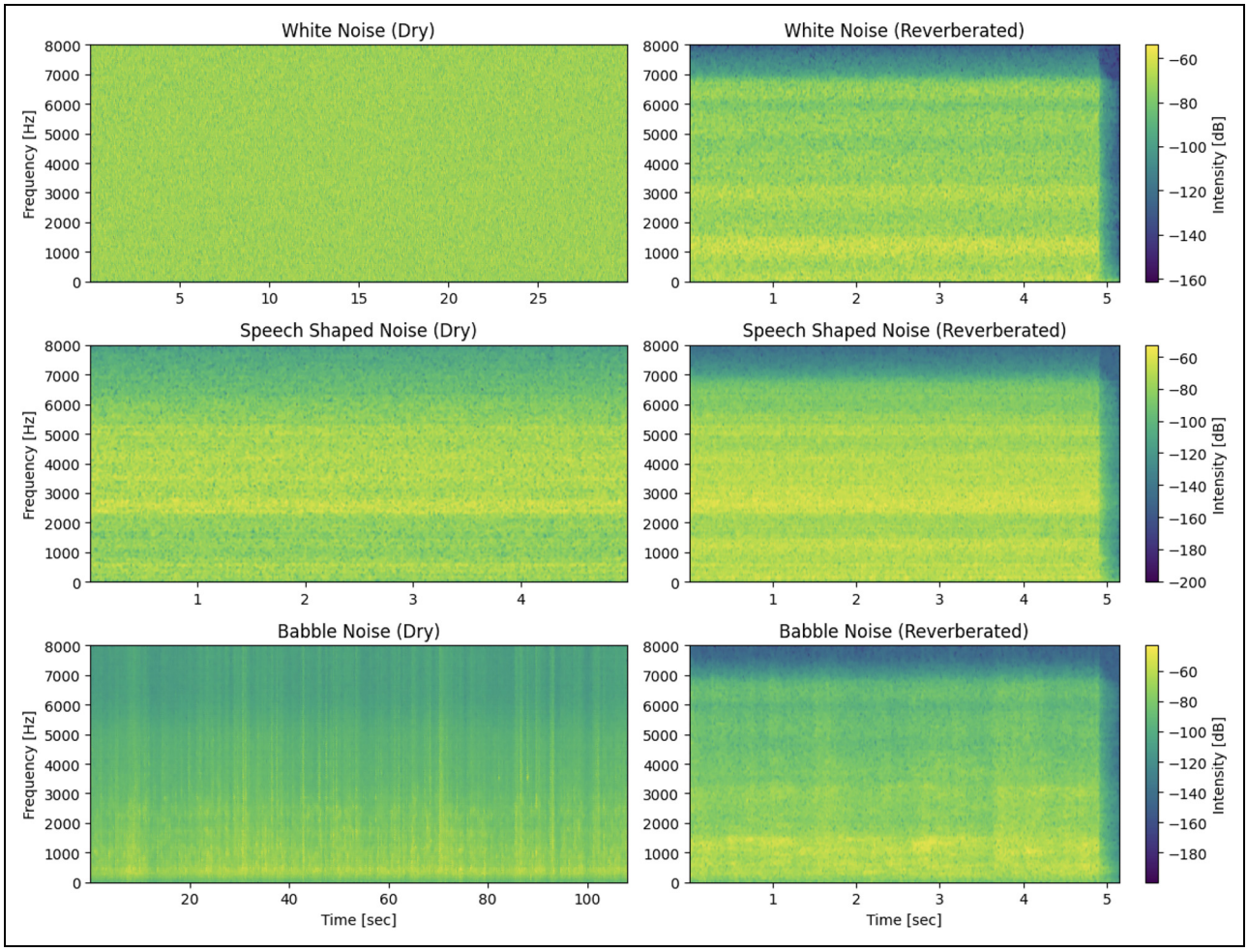

Figure 5. Spectrograms of white noise, speech-shaped noise, and babble noise.

Even though RIRs can be simulated using an acoustic model, in this work, we use measured RIRs since they allow for more realism. More specifically, we use the RIRs collected from Delebecque and Serizel (2023), which correspond to a typical hearing aid use-case scenario. In a nutshell, these RIRs are measured by playing a sweep signal at the source location and recording the reverberated signal at the listener's position (Novak et al., 2015). We use the RIRs obtained for a source signal angled at 0° (in front, relative to the forward-facing listener), 45°, and 90°. The reverberated signals were recorded by Delebecque and Serizel (2023) using a Portable Hearing Laboratory (PHL) placed on a KEMAR head and torso model simulator. The PHL device comprises two behind-the-ear hearing aid shells, each of which being equipped with two omnidirectional microphones; thus, it yields a realistic four-channel signal as typically processed by hearing aid devices. The room is rectangular with dimensions of 6.62 m (length) × 2.57 m (width) × 2.60 m (height), and the reverberation time (RT60) is 0.20 s. The KEMAR head is positioned 2.27 m from the wall along the length axis and centrally in the room along the width. Both the loudspeakers and the KEMAR head's ears are placed 1.48 m above the floor. The loudspeakers are placed 1 m away from the KEMAR. We refer the interested reader to the original publication (Delebecque & Serizel, 2023) for more details on the data acquisition setup.

Mixtures. We use RIRs to simulate mixtures, creating scenarios where the speech is at a 0° angle and the noise is angled at either 45° or 90° to the right (see Figure 6 for noise angled at 45°). To build these noisy mixtures, we first need to adjust the relative speech and noise amounts. To that end, we apply an amplification factor to the noise source, and we control the amount of noise via the gain, defined as:

$$\text{gain} = 10.\log 10\left(\frac{||s||^2}{||n||^2}\right),$$

and expressed in decibels (dB), where s and n denote the anechoic speech and noise signals, respectively. In practice, the gain is computed by only considering segments where

the speech signal is active. Note that the gain is adjusted by considering source signals (before applying the RIR). Subsequently, the clean speech and the scaled noise are convolved with the RIRs as detailed in the previous section, resulting in a four-channel-audio mixture signal by summing the two convolved signals. Noisy mixtures are built using a gain of −5, 0, or 5 dB. This process aims to replicate the acoustic characteristics of a real-world environment as closely as possible.

Evaluation Metrics

In our evaluation, we use objective metrics to compare the target speech with the speech estimated with the different speech enhancement algorithms at utterance and phoneme scales. Traditionally, speech enhancement algorithms are tested using metrics that measure the quantity of interference, artifacts, and distortions that remain in the estimated speech. To that end, we use the BSS eval metrics from Vincent et al. (2006), originally tailored for source separation applications, but widely used for speech enhancement. Note that other metrics are designed for evaluating speech signals in terms of intelligibility (Taal et al., 2010) or perceived quality (Rix, 2001). However, these operate on audio segments (about 300 ms to 1 s) that are significantly longer than the duration of individual phonemes. As such, they are not suitable for our experiments since assessing their relevance at the phoneme scale would require a dedicated study.

Let us consider the following decomposition of the error between the target speech signal s_{target} and its estimate $\hat{s}$ at a reference microphone:

$$\hat{s} = s_{\text{target}} + e_{\text{interf}} + e_{\text{artif}},$$

where e_{interf} and e_{artif} denote the interference and artifacts errors, that is, the contributions of the nontarget source(s). Here, the interference[3] is the contribution of the noise source, and the artifacts represent other types of distortions (e.g., burbling noise) in the speech estimate. From this decomposition, we define the following signal-to-distortion ratio (SDR), signal-to-artifact ratio (SAR), and SNR as follows:

$$\text{SDR} = 10.\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{interf}} + e_{\text{artif}}||^2},$$

$$\text{SAR} = 10.\log_{10}\frac{||s_{\text{target}} + e_{\text{interf}}||^2}{||e_{\text{artif}}||^2},$$

$$\text{SNR} = 10.\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{interf}}||^2}.$$

These ratios are expressed in decibels, and for all metrics, the higher the better. Since they are computed using estimated speech signals, that is, the outputs of the speech enhancement algorithms, we will refer to them as SDR_{out}, SAR_{out}, and SNR_{out}.

Note that to quantify the actual noise reduction achieved by the speech enhancement algorithm, it is necessary to compare a given output metric to a reference initial value. To that end, we calculate the metrics by replacing the estimated speech with the noisy mixture: since the resulting metrics are computed at the input of the algorithms (before any processing), we refer to them as SDR_{in}, SAR_{in}, and SNR_{in}.

In particular, SNR_{in} measures the ratio of desired speech to background noise as it is received at the ear. In our setup, since we do not consider any additional measurement noise (e.g., induced by the recording device), this is equivalent to a SNR, except it is calculated using the images instead of the source signals. As such, this metric is critical for understanding the impact of a room's acoustics on the listener's ear and serves as a reference point for the mixture signal quality (i.e., before enhancement). By comparing SNR_{out} and SNR_{in}, we can quantify the actual noise reduction achieved by the speech enhancement model.

Note that in theory, there are no artifacts in the input signals, so SAR_{in} is infinite and the SDR_{in} is equal to SNR_{in}. Therefore, we will not consider these metrics when presenting our results.

WSNR is a metric that evaluates the quality of speech by assigning a weight to each frequency segment of the speech signal (Greenberg et al., 1993). This approach ensures that the metric reflects the relative importance of different frequency components of speech, and it has demonstrated a strong correlation with the speech reception threshold. Similar to the previous metric, $WSNR_{\text{out}}$ refers to the evaluation using the estimated speech.

When evaluating speech enhancement algorithms, these metrics are typically computed at the utterance scale and aggregated over several sentences to obtain a consolidated metric. This process overlooks the potential performance

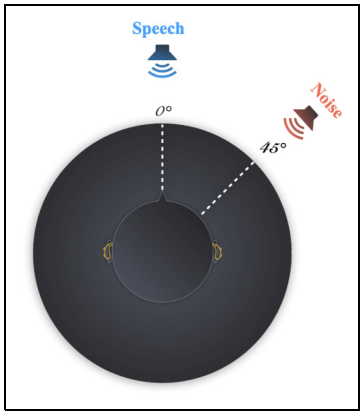

Figure 6. Spatial configuration of the speech and noise sources.

variability of algorithms depending on the phonetic content of the speech signals.

Phoneme Classes

In our study, we investigate the evaluation of speech enhancement algorithms at the phoneme scale. In this regard, we perform phoneme segmentation of clean speech signals. This process involves using a phoneme recognizer to estimate the boundaries of each phoneme within a speech signal.

We used the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) to align spoken audio recordings with their corresponding phonetic transcriptions. In our analysis, we utilized the "English MFA dictionary v2.2.1" version, which has been trained on a dataset comprising 95,278 words. This ensured the alignment model's proficiency in handling a diverse range of phonetic patterns and nuances present in the clean speech data.

As shown in Figure 7, the speech dataset comprises 55 phonemes, each with varying frequencies. Most frequent phonemes include /ə/ with 2,373 occurrences, /ɪ/ with 2,022, and /n/ with 1,972. The least frequent phonemes occur less than 10 times in the whole dataset. Hence, a study at the phoneme scale would hardly lead to any statistically significant outcome on these phonemes.

To simplify the analysis, we group phonemes into categories. This classification relies on a slightly modified version of the MFA IPA chart, as we included an additional vowel class for the near-close near-front unrounded phoneme /ɪ/ and the near-close near-back rounded phoneme /ʊ/. This decision was driven by the important presence of the unrounded phoneme within the dataset, prompting us to investigate the near-close performance and behavior in the evaluations. Moreover, we include both /e/ and /ej/ in the close-mid category as the number of /o/ and /ow/ is very low.

As illustrated in Figure 8, the most prominent phoneme categories in our dataset are plosives, open-mid, and nasals, occurring 6,214, 4,287, and 2,489 times, respectively. The close-mid, affricate, and tap phoneme categories add to the phonetic diversity, albeit as the least frequent, enriching the overall representation of speech.

Experimental Setup

As outlined in the section Overview of Multichannel Speech Enhancement, we selected three speech enhancement algorithms whose pretrained weights are available online. All the models are used in their default setup. For each ear, the front microphone of the speech enhancement algorithm is selected as the reference microphone. The MVDR and FaSNet models are implemented in the ESPnet toolbox (Li et al., 2021), and the corresponding weights can be readily downloaded from the toolbox. Both models have been trained on the CHiME-4[4] dataset, which includes 8,738 noisy utterances (1,600 recorded and 7,138 simulated) with speaker distribution of 4 speakers for recorded noisy speech mixtures and 83 speakers for simulated noisy speech mixtures (Vincent et al., 2017). The environments where noise signals were recorded include buses, cafes, pedestrian areas, and street junctions. Impulse responses are provided in the simulated training set to simulate reverberation effects.

The Tango model (Furnon et al., 2021) is trained on the same dataset as in the original paper, including several babble noise recordings as well as speech-shaped noise, and its code and pretrained weights are also available online. This training data includes noise sources amplified by a random gain between −6 and 0 dB (after convolution, most of the gains range from −10 to 10 dB). The simulated environments include a meeting room and a living room. In the meeting room setup, two sources (target and interference) are placed around a circular table (0.5–1 m radius), and four nodes[5] with microphones are positioned at 90° angles around the table, between 5 and 20 cm from the edge. In the living room scenario, the nodes are placed within 50 cm of the walls to mimic shelf placement, while the sources are randomly positioned at least 50 cm away from the nodes and walls.

In our study,[6] we compute the BSS eval metrics using the mir_eval library (Raffel et al., 2014).

Experimental Results

First, we detail the results at the utterance scale, as speech enhancement evaluations are typically conducted. Then, we delve into a finer-grain evaluation at the phoneme scale.

Evaluation at the Utterance Scale

Comparison between the Left and Right Microphones. Let us first recall that each algorithm produces an enhanced signal at a reference channel that can be on either the left or right device (we arbitrarily chose it to be the front microphone of the device). This first experiment compares the results obtained in these two cases. We operate in an asymmetrical scenario since the noise source is placed on the right side of the head in our setup (see Figure 6). This investigation aims to clarify how the noise's spatial orientation affects the binaural algorithms' performance on each microphone.

Figure 9 displays the results of the algorithms' effectiveness to enhance the speech at each ear (note that these results are averaged across noise types and positions, gain factors, and models). In line with our expectations, the influence of noise is more pronounced on the right ear. The right ear shows lower input SNR, which can be attributed to the proximity of the right reference microphone to the noise source. On the contrary, the head shadow effect impacts the sound propagation to the left ear and the input SNR is larger.

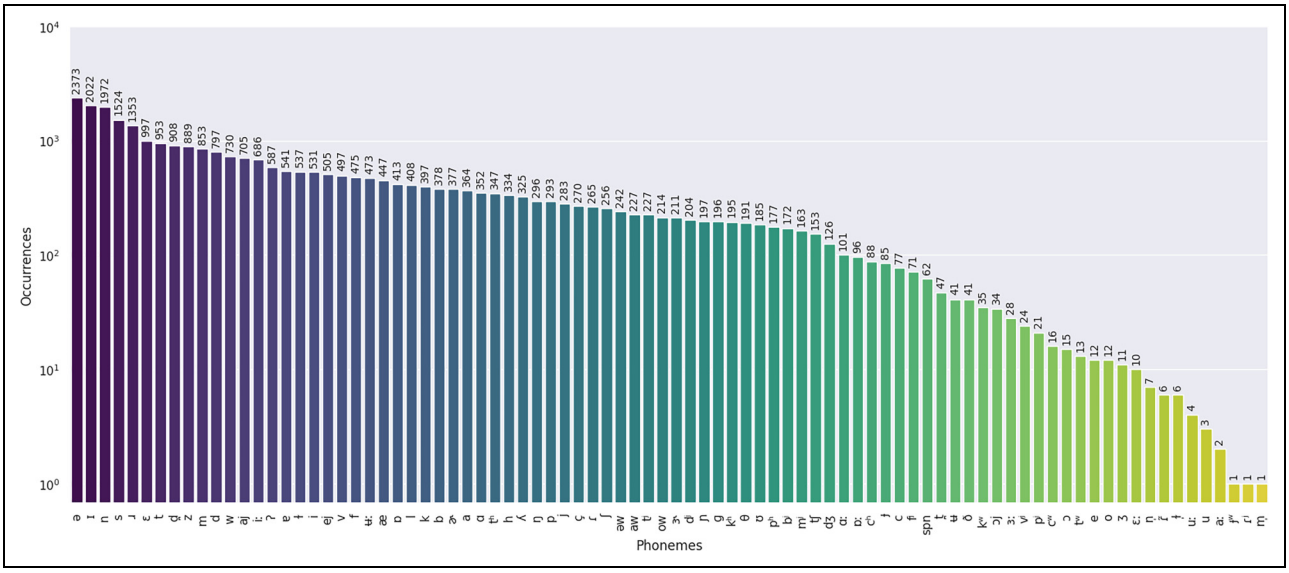

Figure 7. Phoneme distribution in our speech dataset.

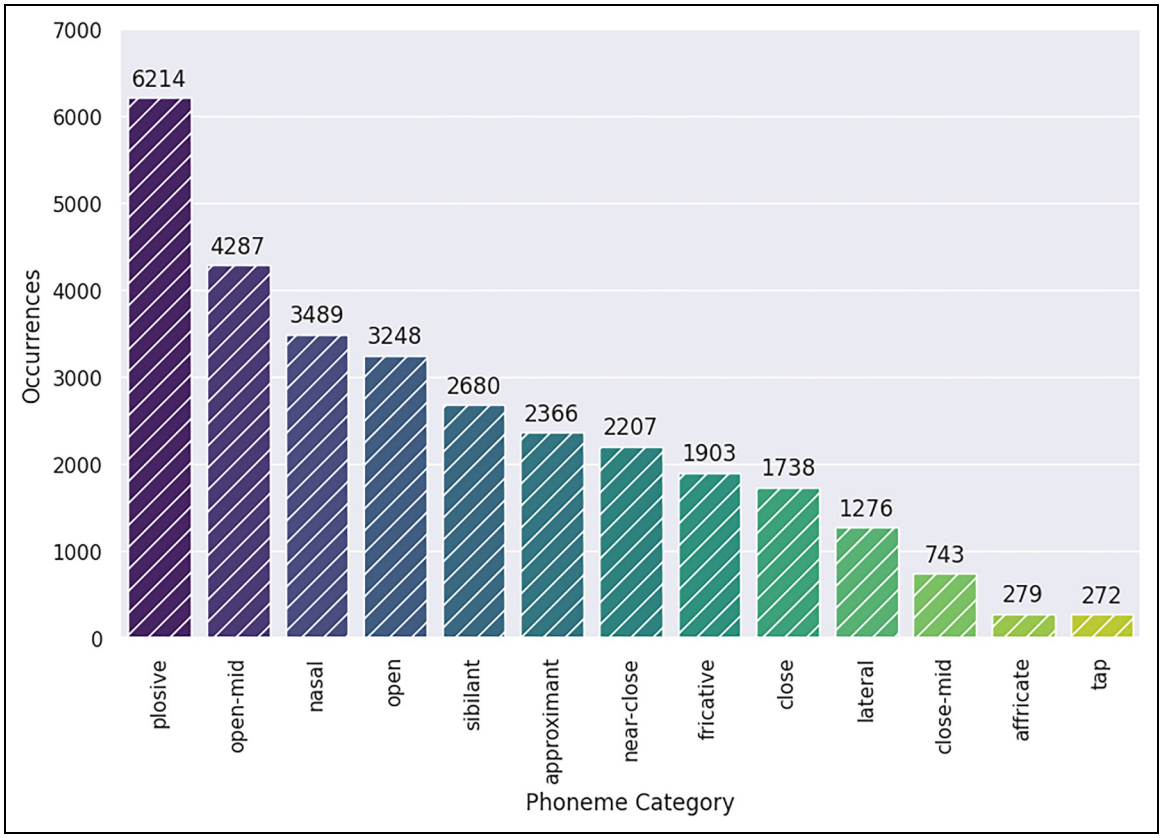

Figure 8. Distribution of phoneme per categories in our speech dataset.

Even though assessments of the estimated speech (output) indicate a better absolute performance on the left microphone, it is noteworthy that the relative improvement is more important for the right microphone. Specifically, we observe that the SNR improvement is more substantial on the right side (10.60 dB) than on the left side (2.86 dB). This outlines that the right-side microphone benefits more from the binaural property of the speech enhancement

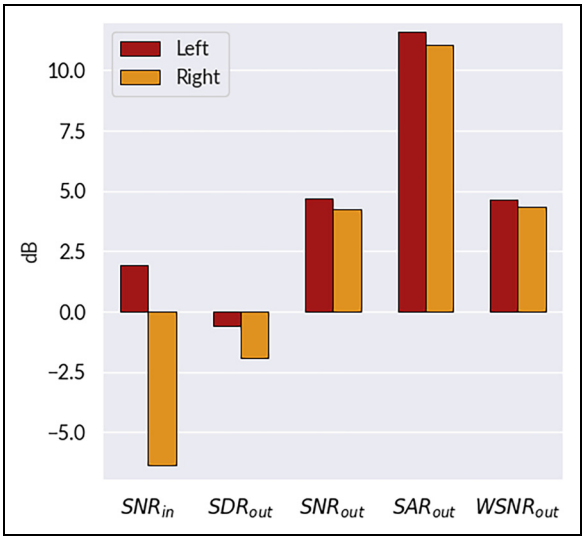

Figure 9. Comparison between the left and right microphones.

Figure 11. Comparison between noise angles at 45° and at 90°.

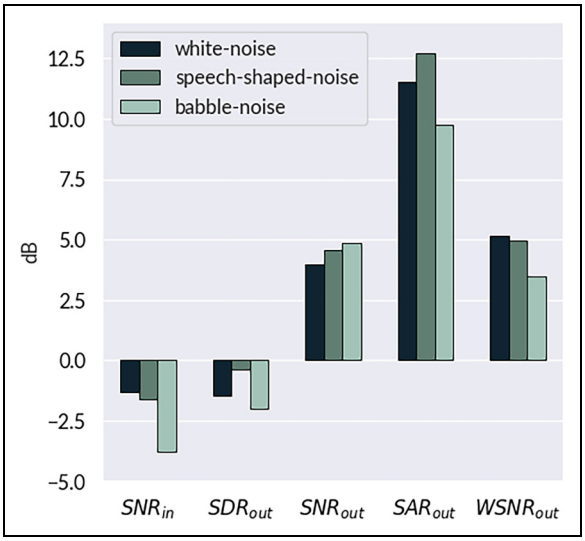

Figure 10. Comparison between the white noise, speech-shaped noise, and babble noise.

algorithm, albeit with a relatively greater amount of distortion compared to the left ear. The $WSNR_{\text{out}}$ results show a similar trend to the output SNR, with the right ear benefiting more from the algorithm's enhancements. Nonetheless, prior studies in audiology have outlined the importance of reducing the amount of noise at the best ear (here, the left) (Bronkhorst & Plomp, 1988). Thus, in the phoneme-scale evaluation that follows, we will focus our analysis on the results for the left reference microphone. We leave to future work a fully binaural evaluation of speech enhancement that would account for both noise and artifact reduction at both ears.

Comparison between Noise Types. In this experiment, we investigate the influence of the noise type on performance. We consider three noise types: white noise, speech-shaped noise, and babble noise. The results are averaged across gain factors and models and displayed in Figure 10.

First, we observe an overall consistent performance for the white noise and the speech-shaped noise. The babble noise is more challenging for models to deal with than the other noise types, as indicated by the corresponding low values of input SDR, SNR, and SAR. As the focus of the paper is not to analyze the performance of speech enhancement algorithms under challenging scenarios, but rather to understand their behavior at a fine-grained scale, we will focus on white noise and speech-shaped noise. As demonstrated by Stone et al. (2012), the presence of temporal modulations in noise undermines the intelligibility of speech. While white noise and speech-shaped noise exhibit less complex temporal modulations compared to babble noise, speech-shaped noise can serve as a proxy for babble noise due to its similar spectral characteristics.

Regarding $WSNR_{\text{out}}$, we observe an opposite trend compared to the output SNR, with white noise showing the highest values and babble noise the lowest. This outcome is expected, as $WSNR_{\text{out}}$ is designed to be more sensitive to the frequency content of the residual noise, which is more prominent in audible regions for babble noise.

Impact of the Noise Location. This experiment analyzes the impact of the noise location on the performance of the speech enhancement algorithms. The noise source can be positioned at either 45° or 90° relative to the forward-facing listener (see Figure 6). The results are averaged across noise types, gain factors, and models, and presented in Figure 11.

First, we remark that the input SNR is slightly higher when the noise is oriented at 90° relative to the listener. This was expected since these results correspond to the left-

side microphone, which is less contaminated with noise when the source is placed on the opposite side of the head rather than at 45°. Likewise, we also observe that speech enhancement algorithms exhibit a higher performance when the noise is placed at 90° compared to 45°, as indicated by the output SDR and SNR. Nevertheless, the differences between the two scenarios are not very important in terms of both input and output metrics. Therefore, selecting either one of the two scenarios would not impact the overall analysis greatly. In the rest of our study at the phoneme scale, we will focus on the 45° scenario that is potentially more challenging for the speech enhancement algorithms. In terms of $WSNR_{\text{out}}$, we observe that the trend is consistent with the output SNR, showing better performance at 90° compared to 45°.

Impact of the Amount of Noise. Figure 12 presents the impact of varying the mixture noise gain factors on both the input and the output evaluation metrics.

First, comparing the gain factor (computed using the anechoic sources) and the input SNR (computed using the image sources) highlights the influence of room acoustics on the mixtures at the ears' position. On average, this phenomenon results in a drop of approximately 2 dB in terms of amount of noise for the three scenarios. This is justified by the propagation in the room and the relative position of the sources with respect to the walls (Delebecque & Serizel, 2023). This underscores the importance of computing the SNR at the microphone as a reference value and not relying on the gain that is set on the dry sources.

Overall, we observe that the algorithms' performance improves as the mixture gain increases. We can also note an important drop in all the metrics for the scenarios where the gain is −5 dB. This indicates that it will probably be useful to control the performance of the algorithms at different gain. Yet, for the simplicity of the analysis, in our phoneme-scale experiments, we will focus mainly on the scenario with a gain of 0 dB in the phoneme-scale evaluations. Indeed, this setting is sufficiently challenging for the task at hand and allows us to examine the relative performance of speech enhancement algorithms without inducing an excessive degradation of the signal. Still, we will examine the impact of the mixture gain on specific phonemes, where it might deserve some finer-grain analysis.

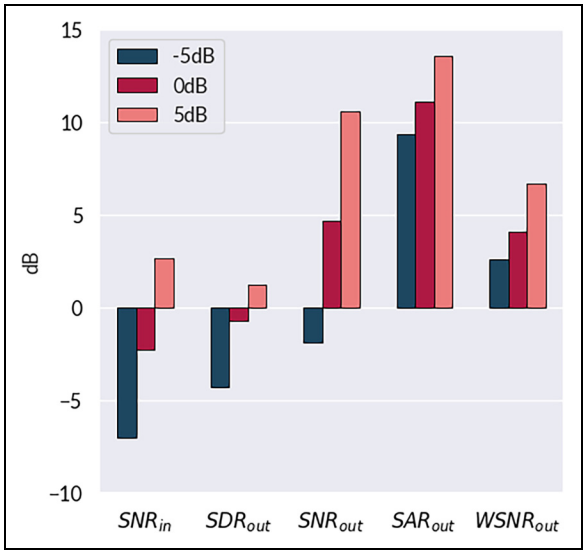

Figure 12. Comparison between gain factors (−5, 0, and 5 dB).

The output *WSNR* results reveal that higher gain factors consistently lead to better speech enhancement performance. Additionally, $WSNR_{\text{out}}$ follows a similar performance pattern as the output SNR, indicating consistent improvements in overall speech quality across different amounts of noise.

Evaluation at the Phoneme Scale

We now delve into our evaluation at the phoneme scale, which we illustrate by the following introductive example. Figure 13 displays the spectrogram of a clean speech signal, as well as the spectrograms of this same signal contaminated with noise, and where some specific phonemes are highlighted.

The spectrograms display for instance the phonemes "g" and "s" across different noise conditions. In the clean speech, the phoneme "g" shows distinct horizontal striations representing its voiced nature with a rich harmonic structure. The phoneme "s" is characterized by a high-frequency, almost texture-like pattern, indicating its sibilant, unvoiced nature.

When white noise is added to the clean speech, we observe that the low-frequency harmonics of the phoneme "g" remains relatively intact. On the other hand, the sibilant "s" is strongly affected by this noise, since its energy is more concentrated in the higher frequencies where the white noise also contains energy.

With speech-shaped noise, the impact on the phoneme "g" is less uniform. Similar to white noise, speech-shaped noise fills in the temporal gaps of stopped consonants, making it difficult to discern the phoneme's harmonic patterns. The phoneme "s" remains relatively discernible, but its crisp edges are somewhat softened, and the definition between silence and sibilance is less clear.

The presence of babble noise introduces a more complex interference. The phoneme "g" is disrupted by the varying intensities and frequencies of overlapping speech, obfuscating its harmonic structure. Conversely, while still visible due to its high-frequency content, the phoneme "s" competes with similar sounds from the babble, which can make it challenging to isolate from the background chatter. Note that in addition to this so-called *energetic* masking, babble noise is known to introduce an *informational* masking (Brungart, 2001) due to its cognitive interference with the speech

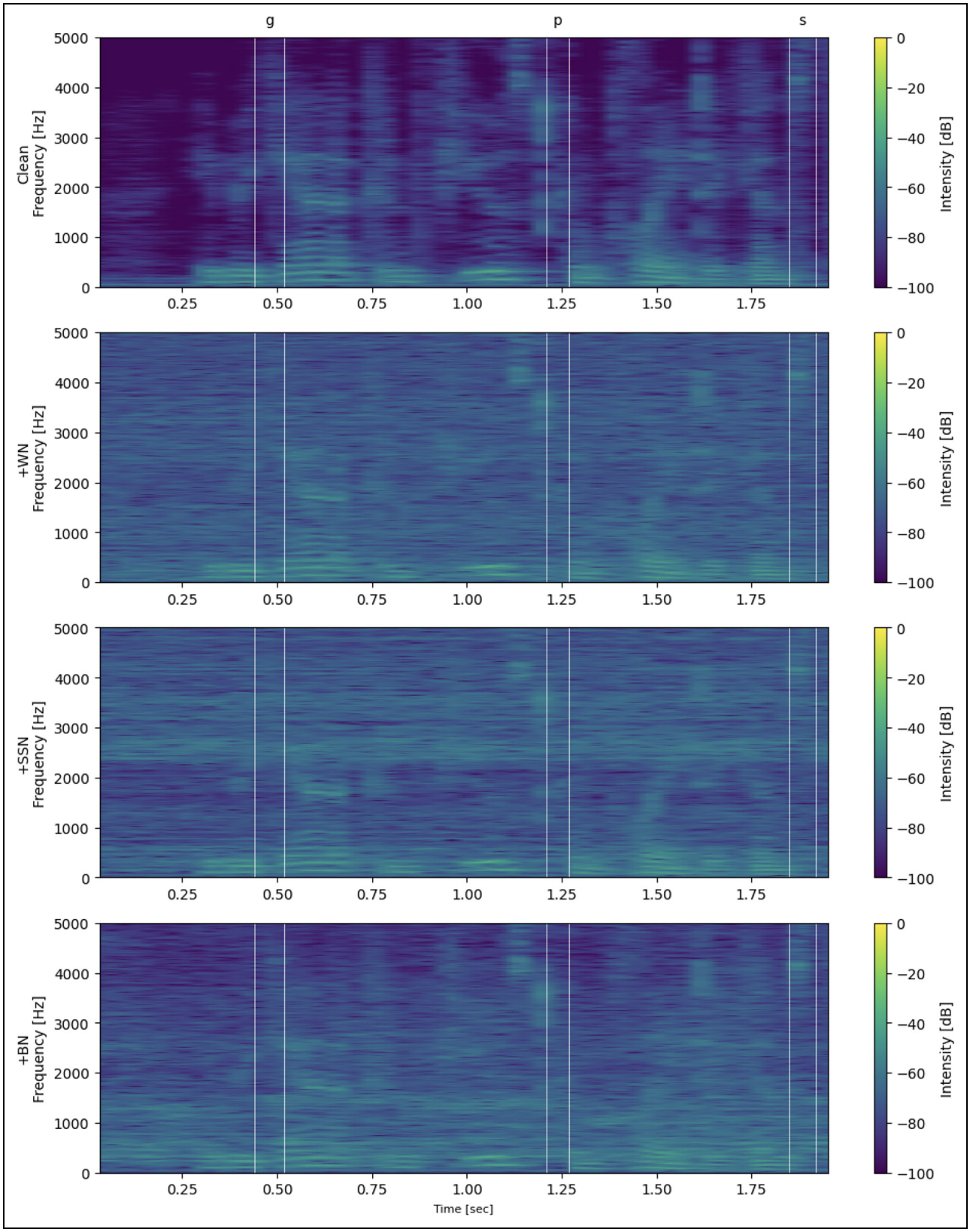

Figure 13. Spectrogram of clean speech and mixtures with diverse types of noise on the utterance "He began a confused complaint against the wizard who had vanished behind the curtain on the left" segmented into phonemes. Note: The white noise has been filtered with an RIR, altering its characteristics, and the low-frequency region of the speech-shaped noise appears less intense due to intensity scaling.

signal. While accounting for such a masking is necessary in listening tests, it is outside the scope of this paper.

This analysis underscores the critical importance of evaluating speech enhancement algorithms at the phoneme scale. The differences in how various phonemes are affected by several types of noise highlight the nuanced challenges faced by the speech enhancement systems. Voiced phonemes, with their rich harmonic structures, and unvoiced phonemes, with their high-frequency energy, require different enhancement strategies to overcome the masking effects of noise. Understanding these varied impacts is essential for improving speech enhancement algorithms that can effectively disentangle and clarify the essential elements of speech, ensuring that each phoneme, regardless of its unique acoustic properties, is accurately reproduced and easily discernible, even is adverse listening conditions.

An Overview of the Results at the Phoneme Scale. We first present an overview of the results at the phoneme scale. In all the following experiments (except for the last one), the mixtures are at 0 dB. The results are displayed in Figure 14. A notable observation is that plosives, fricatives, and taps are the most impacted by the noise. These are also the phoneme categories on which the speech enhancement algorithms perform the worst.

The experiment reveals that, on average, the speech enhancement models yield a substantial improvement in all the other phoneme categories. They particularly reduce the amount of noise while introducing only a controlled amount of distortion and artifacts. However, this positive trend in phoneme categorization contrasts with the findings at the utterance scale. Indeed, the evaluation of artifacts at the utterance scale tends to overestimate the performance across all phoneme categories. This suggests that while the models perform well at refining speech at the phoneme scale, their effectiveness may be overstated when considering the broader context of complete utterances.

Some phoneme categories show different performance in $WSNR_{\text{out}}$ versus output SNR. For instance, in the case of sibilants, the output SNR indicates that speech enhancement algorithms perform better at enhancing sibilants compared to the overall utterance scale. In contrast, the $WSNR_{\text{out}}$ for sibilants suggests that the frequency-weighted SNR of sibilants is lower than the overall utterance evaluation.

Overall, we observe different trends per phoneme categories, which motivates us to analyze these results in more depth. We conduct such an analysis in the following experiments, for which we select specific categories of phonemes such that the comparison is made clearer.

Impact of the Noise Type on Plosive, Approximant, and Open Phonemes. In this experiment, we analyze the results (displayed in Figure 15) for opens, approximant, and plosive phonemes with respect to the noise type (white noise or speech-shaped noise). In examining the outcomes across various metrics, it is apparent that the general trend persists regardless of the noise type. Nonetheless, there is a slightly higher SNR improvement when speech-shaped noise is present as compared to white noise. This can be explained by the fact that speech-shaped noise is commonly used for training speech enhancement models. Additionally, the spectral density differences between the two noise types result in differential effects on phonemes, such as fricatives and vowels. It is interesting however to see that the performance in terms of SAR remains consistent across phoneme categories, regardless of the noise type. This indicates that the SNR improvement does not occur at the costs of a lower SAR, while such a trade-off is usually observed in speech enhancement algorithms.

$WSNR_{\text{out}}$ exhibits the same trend for both white noise and speech-shaped noise. However, while $WSNR_{\text{out}}$ aligns with output SNR trends, plosives show better quality in both noise types than the output SNR suggests. This difference indicates that plosives contain a higher amount of noise relative to the utterance scale, despite appearing to have better quality.

Comparison of the Algorithms on Nasals, Affricates, and Sibilants. The input SNR indicates that the nasals are the least degraded by the noise (with the white noise having slightly less impact than the speech-shaped noise).

As for the residual interference and distortions in the estimated speech, the performance varies with the noise type. Tango outperforms other models in mitigating interference and distortions with the presence of white noise. In contrast, when using speech-shaped noise, MVDR appears to be the best at reducing interference and artifacts, whereas Tango is superior for reducing distortions. FaSNet appears to deteriorate the speech signal when white noise is involved, since the output SNR is larger than its input value. This could potentially be due to a mismatch between the training conditions of the model and the testing setup considered here. Indeed, end-to-end algorithms have been shown to exhibit less robustness to these conditions (types of noise, amounts of noise, acoustic environments, etc.) than hybrid algorithms (Ditter & Gerkmann, 2020). Nevertheless, FaSNet notably improves the SNR in the presence of speech-shaped noise, especially for nasal phonemes, indicating its effectiveness in enhancing certain aspects of speech.

Across different noise conditions, Tango exhibits robustness, consistently improving the SNR, more so for sibilants than for nasals and affricates. This suggests that Tango presents a balanced performance across various acoustic noise scenarios on the three phonemes categories. MVDR performs well on nasals regardless of the noise type, but its performance on affricates and sibilants is always lower than for nasals. Besides, $WSNR_{\text{out}}$ reveals that Tango consistently improves the frequency-weighted SNR for sibilants across both white noise and speech-shaped noise, while MVDR better improves nasals and affricates (Figure 16).

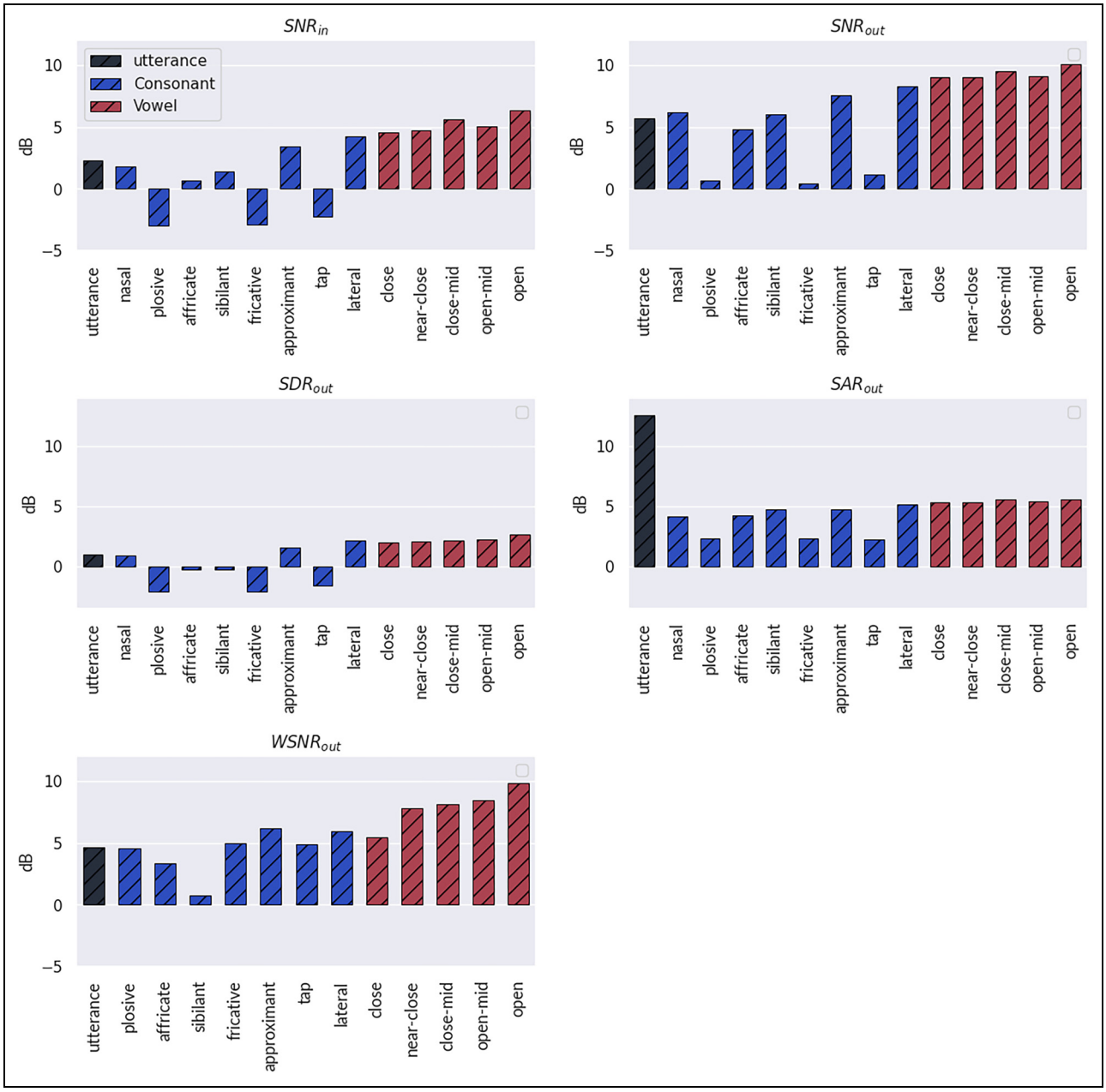

Figure 14. Evaluation results across phoneme categories (the results at the utterance scale are also reported). Results are averaged over algorithms, with a gain of 0 dB and the noise source oriented at 45° relative to the listener.

Comparison of the Algorithms on Vowels. In terms of SNR_{in}, the initial conditions for all vowel categories are favorable under both noise types, indicating that the vowels are less affected by noise than the consonant at the input stage. Besides, the input SNR remains consistent across all vowel categories.

The analysis of the speech enhancement algorithms reveals that Tango is the most effective model for minimizing distortions, performing with both white noise and speech-shaped noise. This suggests Tango's processing techniques are well suited for maintaining the integrity of the speech signal even in the presence of various noise types. Tango also stands out for reducing the interference of vowels in white noise environments. This is particularly true for open phonemes, which are more vulnerable to interference due to their wider spectral spread. MVDR, on the other hand, outperforms other systems in reducing the interference on vowels in scenarios with speech-shaped noise. FaSNet is overall outperformed by MVDR and Tango, except under speech-shaped noise where it outperforms Tango. Tango's performance appears to depend on the initial amount of interference. The SNR improvement is almost constant for all

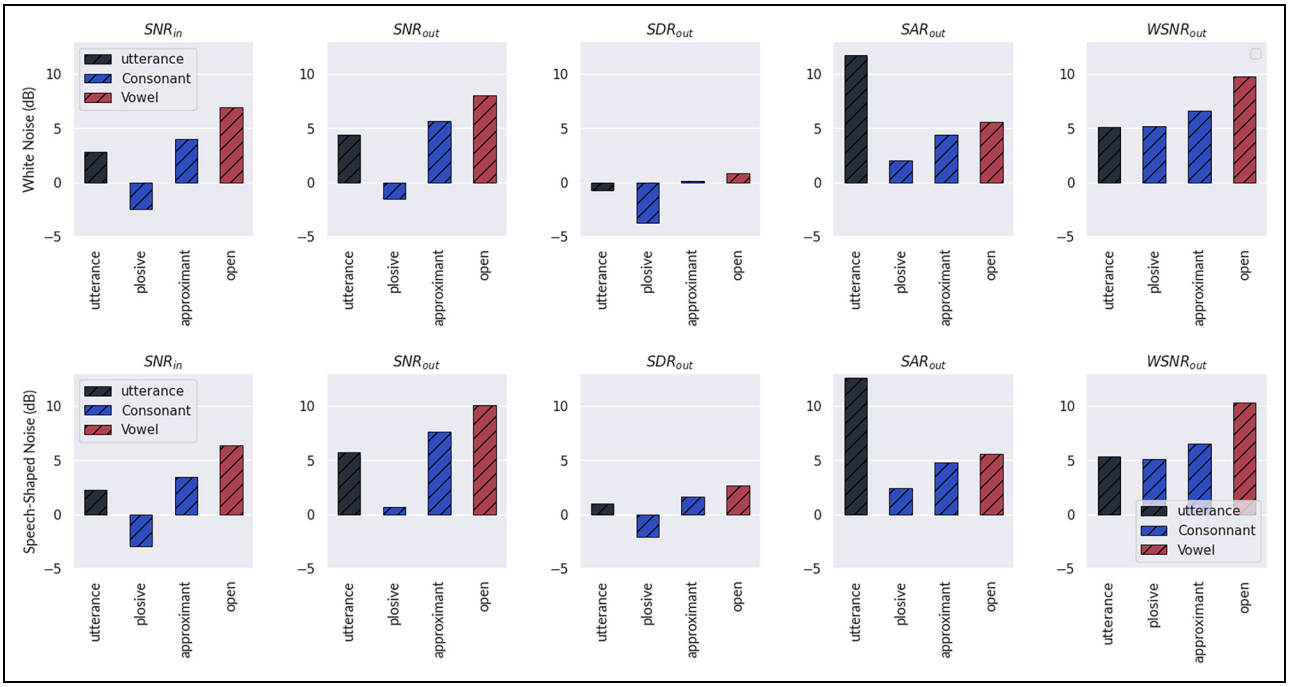

Figure 15. Evaluation results per noise types on plosive, approximant, and open phonemes. Results are averaged over algorithms, with a gain of 0 dB and the noise source oriented at 45° relative to the listener.

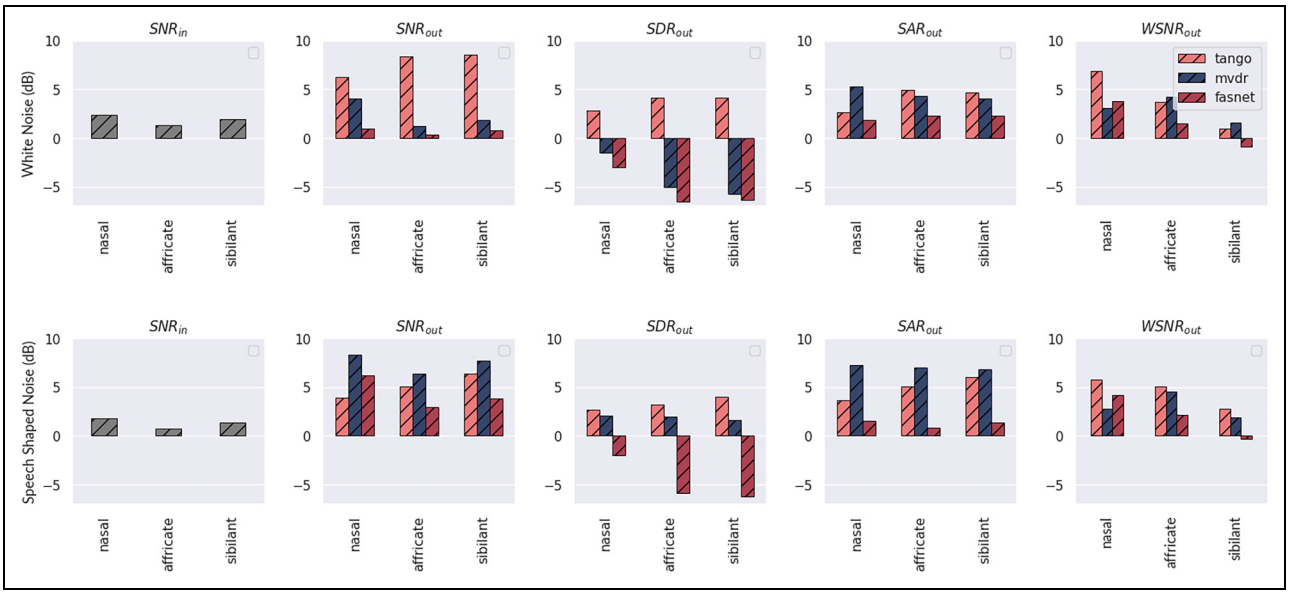

Figure 16. Performance of the algorithms on nasals and sibilants on mixtures, with a gain of 0 dB and the noise source oriented at 45° relative to the listener.

categories and the output SNR then depends on the phoneme category. MVDR, in contrast, improves the SNR more uniformly, potentially offering a more predictable enhancement outcome, especially when the initial amounts of noise in vowels are varying.

MVDR also prevails at controlling the presence of artifacts in the estimated speech, which is crucial for the overall perceived quality and intelligibility of the enhanced speech, since a low amount of artifacts implies that the speech signal retains more of its natural characteristics

postestimation. The output SAR obtained with MVDR is once again almost constant regardless of the phoneme category, which could lead to more predictable behavior. The output SAR obtained by Tango depends on the input SNR which makes it potentially less predictable than the MVDR. FaSNet is outperformed by the other systems by a large margin in terms of SAR_{out}. However, the $WSNR_{out}$ shows that while an algorithm might reduce noise effectively on specific type of noise, this does not always enhance vowel quality, and vice versa. In particular, if Tango is better at reducing white noise on vowels, and MVDR outperforms with speech-shaped noise, the vowels appear to have similar quality after enhancement (Figure 17).

Comparison of the Algorithms on Plosives, Fricatives, and Taps. As in prior experiments, the analysis is conducted in environments with two types of noise conditions: white noise and speech-shaped noise. The results are presented in Figure 18.

The results reflect the distinct acoustic challenges that plosive, fricative, and tap consonants face in the room environment. These consonant types are intrinsically affected by interference due to their articulatory characteristics, with plosives being particularly vulnerable due to the transient nature of sound production, which can be easily masked by environmental noise.

Tango's performance with white noise stands out in its ability to reduce interference, especially in the case of plosives. Despite the challenging initial conditions, Tango manages to enhance plosives' clarity while keeping distortion reasonably low, highlighting its efficacy in dealing with the abrupt and high-intensity nature of plosive sounds. Under speech-shaped noise, Tango does not exhibit any remarkable improvement, yet it maintains its proficiency in interference reduction for plosives. The model's average performance in artifact reduction suggests that while Tango can mitigate some noise elements, there's a trade-off in terms of introducing new artifacts into the signal.

The MVDR exhibits more nuanced results, as we observe a persistence of high amounts of distortion and low interference improvements when dealing with white noise; however, it consistently reduces interference across all phoneme categories and maintains a high SAR under speech-shaped noise. While Tango performed the best on plosives, the MVDR provides the highest amount of noise reduction on tap phonemes.

Similarly, as on vowels, FaSNet fails to reduce interference and introduces an important amount of distortion in the presence of white noise. With speech-shaped noise, FaSNet exhibits improvement in the SNR. Like for the MVDR, these improvements are larger on tap phonemes than on plosives and fricatives. Finally, like vowels, the $WSNR_{out}$ demonstrates that even though an algorithm may efficiently reduce noise, this does not necessarily improve the frequency-weighted SNR of the three consonants.

Comparison of the Algorithms on Approximants and Laterals. In this experiment, we focus on the performance of the speech enhancement algorithms on approximant and lateral phonemes. The results provide insights into each algorithm's ability to preserve the integrity of these phoneme categories amidst the interference, distortion, and artifacts. The results are presented in Figure 19.

In the presence of white noise, Tango demonstrates proficiency in reducing interference while concurrently

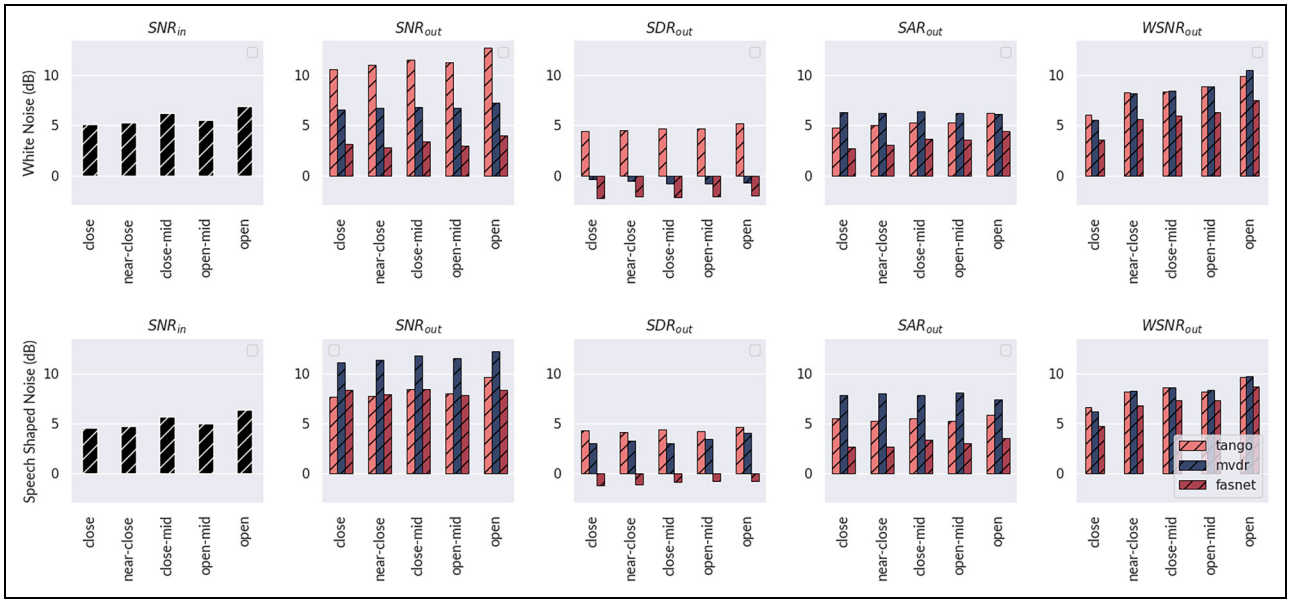

Figure 17. Performance of the algorithms on close and open phonemes, with a gain of 0 dB and the noise source oriented at 45° relative to the listener.

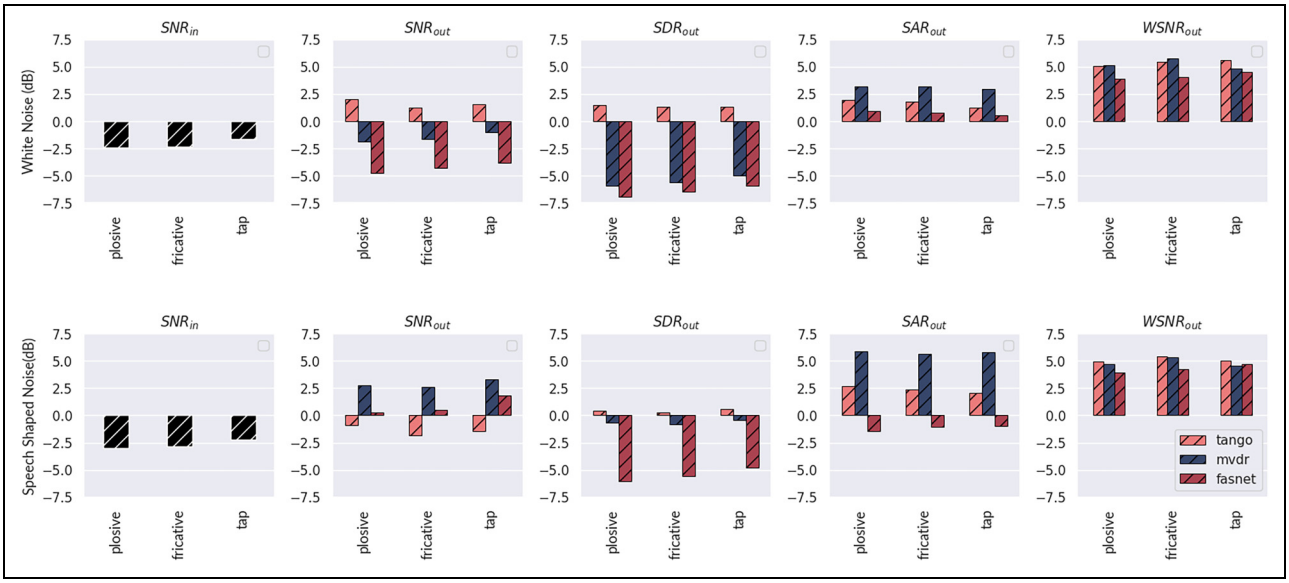

Figure 18. Performance of the algorithms on plosives, fricatives, and taps, with a gain of 0 dB and the noise source oriented at 45° relative to the listener.

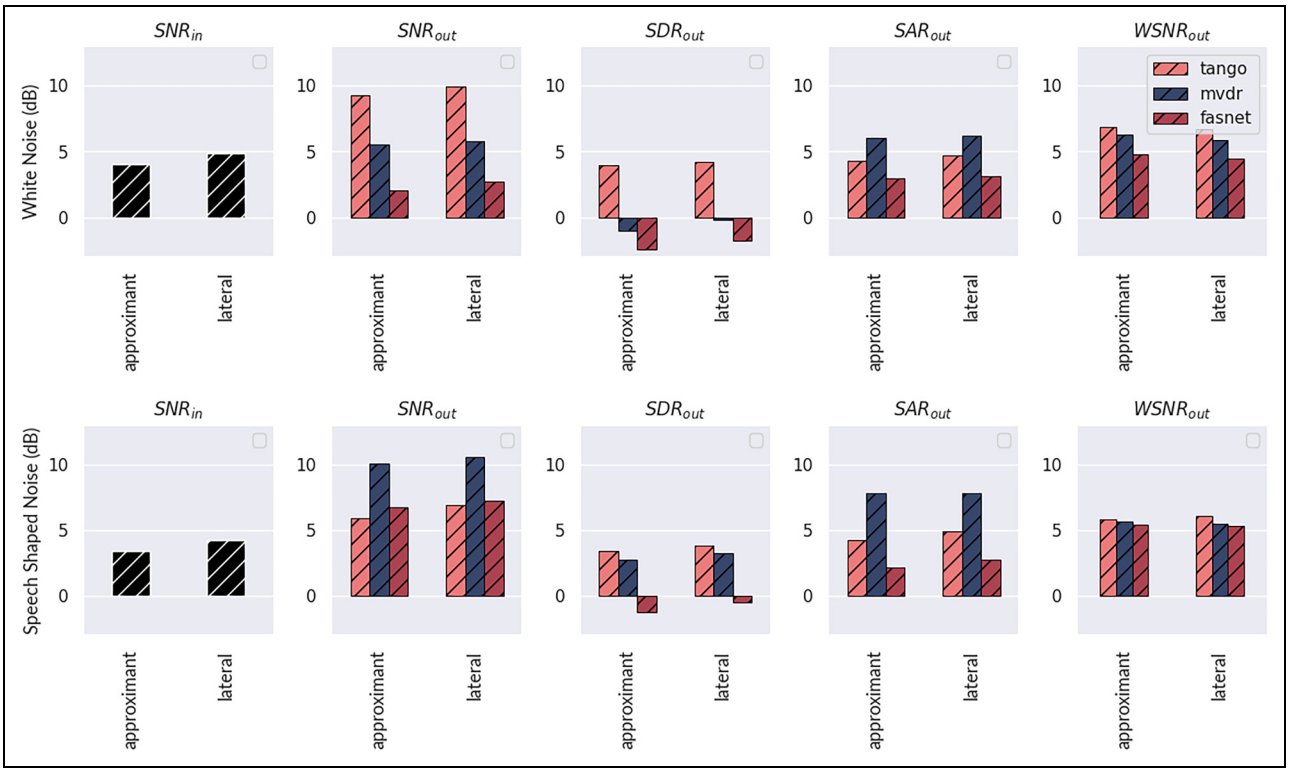

Figure 19. Performance of the algorithms on approximant and lateral phonemes, with a gain of 0 dB and the noise source oriented at 45° relative to the listener.

maintaining a low amount of distortion, outperforming the other models for both phoneme types. The trend is inverted in the presence of speech-shaped noise where MVDR exhibits the best SNR improvement for both approximants and laterals. Once again, FaSNet exhibits a deficient performance on white noise and performs on par with Tango in

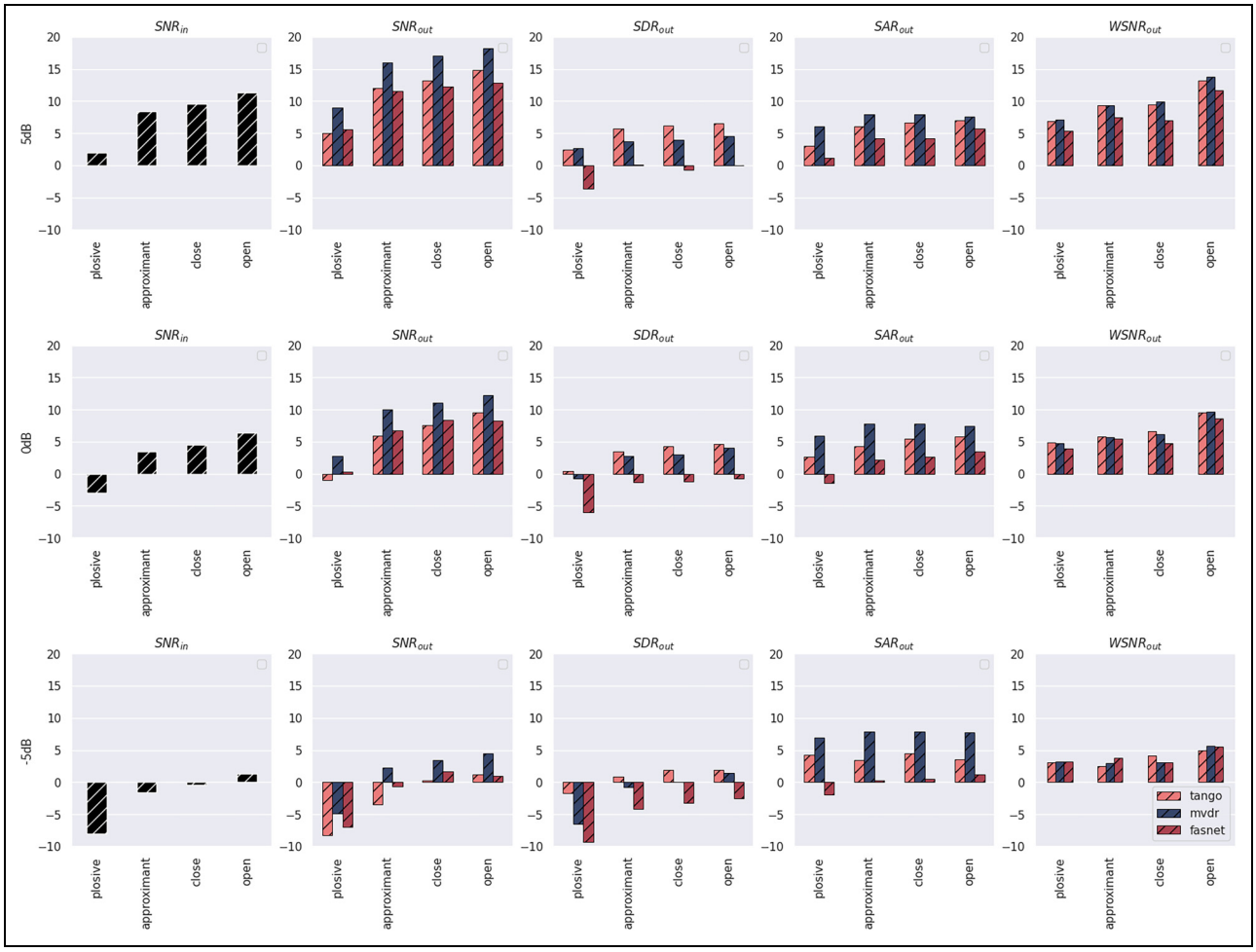

Figure 20. Performance of the algorithms on plosives, fricatives, closes, and opens with respect to the amount of input noise. The noise source is oriented at 45° relative to the listener.

terms of SNR when speech-shaped noise is present. MVDR stands out for its superior control over distortion across both phoneme categories, suggesting its advanced capability to preserve speech quality. Tango and FaSNet, while demonstrating lower SARs, successfully maintain a reduced quantity of artifacts for both types of noise, confirming the algorithms' capabilities in artifact control. It is interesting to note that each model performs consistently across the phoneme class considered on each separate noise condition.

In summary, these experiments show that each algorithm exhibits a different behavior that depends not only on the noise type but also on the phoneme class.

Impact of the Amount of Noise on Algorithms' Performance. Finally, this experiment aims to understand how each model responds to different amounts of background noise across a range of speech sounds, from plosives to open vowels. Figure 20 presents a comparative analysis of the algorithms, evaluating their performance on mixtures with speech-shaped noise.

The plosives, which are characterized by a complete obstruction of the vocal tract, are particularly susceptible to the acoustic interferences, as seen by the consistently lower input SNR compared to the gain of the mixtures. This suggests that the high modulation frequencies at the onset of plosives make them more prone to interference. All models grapple with reducing interference for plosives across scenarios. Even the best performing model (MVDR) only yields a slight SNR improvement. Tango exhibits the worst SNR improvement at −5 dB. This could indicate a limit to how much speech enhancement models can counteract the acoustic masking for these rapidly changing phonemes, especially in environments with strong background noise.

In contrast, approximants, close, and open vowels, which involve less abrupt articulatory gestures and more continuous airflow, seem to be easier to enhance even for a gain below 0 dB. This could possibly be due to their more sustained and resonant acoustic signatures that are less easily masked by noise. MVDR yields a noticeable improvement in reducing interference for these phonemes, particularly at 0 and

5 dB, highlighting its strength in enhancing such sounds. Even though this is not as drastic as for the plosive, all the models struggle to improve the SNR when the gain is −5 dB. MVDR once again is the algorithm that performs the best in this case and with speech-shaped noise. Tango and FaSNet on the other hand show no improvement for open vowels at −5 dB.

MVDR and Tango produce low artifacts in the estimated speech across gains from 5 to 0 dB for all phonemes. As in previous scenarios, FaSNet introduces the largest amount of artifacts regardless of the gain, as observed for plosives. For some phonemes, the gap between output SNR and $WSNR_{out}$ widens as the gain factor increases. For instance, although the output SNR at −5 dB indicates a high amount of noise remaining in the signal, the frequency-weighted SNR of plosives is as good as that of approximations and close phonemes.

Several studies, such as those by Phatak and Allen (2007) and Phatak et al. (2008), focus on how different consonants and vowels are confused in noisy conditions. By examining the variability in error rates across different consonants and within individual tokens, Toscano (2014) highlights the importance of accounting for these differences when assessing speech recognition. Our analysis underscores the need for speech enhancement models to be tailored to the acoustic properties of phonemes, considering not only the amount of background noise but also the phonetic and articulatory characteristics that define each phoneme's vulnerability to acoustic interference.

Conclusions and Perspectives

In this paper, we conducted a comprehensive evaluation of three state-of-the-art multichannel speech enhancement algorithms (FaSNet, MVDR, and Tango), with a particular emphasis on the phoneme-scale analysis. This study revealed that speech enhancement algorithms perform differently depending on the phonemes, underlining the limitations of traditional utterance scale evaluations. Specifically, it was found that specific phonemes like plosives are heavily impacted by environmental acoustics, whereas nasals and sibilants show more resistance to noise, especially when it is speech shaped.

This phoneme-scale evaluation framework reveals the need for these algorithms to consider the differential impact of noise on various phonemes and adapt accordingly. This research direction can focus on integrating phoneme-specific characteristics into the training of these models, potentially enhancing their effectiveness in real-world noisy environments. Particularly, this could lead to enhanced speech intelligibility in real-world scenarios, offering interesting observations for developing more effective, personalized hearing aid technologies.

An in-depth analysis would also be interesting to establish the correlation between the speech quality and the perceived quality of phonemes. While our current study focuses on evaluating algorithms at the phoneme level, it is important to note that human auditory perception may operate at an even finer granularity. The auditory system might detect subtle subphonemic cues, which our phoneme-based approach does not fully capture. This framework is also applicable to other tasks involving speech enhancement, such as vocal assistants, where phoneme-scale evaluation and/or processing could be beneficial. Furthermore, acknowledging that physical measurements do not always align with human perceptions of speech, the results should be examined through listening tests.

Acknowledgments

This work was made with the support of the French National Research Agency, in the framework of the project REFINED "REal-time artiFicial INtelligence for hEaring aiDs" (ANR-21-CE19-0043). Experiments presented in this paper were partially carried out using the Grid5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER, and several universities as well as other organizations (see https://www.grid5000).

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Agence Nationale de la Recherche (grant number ANR-21-CE19-0043).

ORCID iDs

Nasser-Eddine Monir https://orcid.org/0009-0006-7531-2051
Paul Magron https://orcid.org/0000-0002-8561-0961

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Notes

1. In practice, these time–frequency signals are represented as complex-valued matrices of dimensions $F \times T$, where F and T denote the number of frequency channels and time frames, respectively. Nonetheless, for brevity, the notation we employ in this paper does not account for the frequency channel and time frame indices, e.g., $S(f,t)$ will be simply denoted S.
2. Note that the correct terminology is "speech spectrum-shaped (random) noise," but in this paper, we use "speech-shaped noise" for brevity.
3. In the following, interference will be represented by the SNR.
4. Information about gain factors and angles was not available on the CHiME-4 documentation website.

5. In this context, a *node* refers to a device with several microphones. Each node can communicate signals with other nodes without delays or packets loss.
6. For reproducibility purposes, our code is available online at https://github.com/Nasseredd/SE-Ph-Eval/.

References

Adachi, T., Akahane-Yamada, R., & Ueda, K. (2006). Intelligibility of English phonemes in noise for native and non-native listeners. *Acoustical Science and Technology*, *27*(5), 285–289. https://doi.org/10.1250/ast.27.285

Benesty, J., Chen, J., & Huang, Y. (2008). *Microphone array signal processing*. Springer Science & Business Media. https://doi.org/10.1007/978-3-540-78612-2

Bertrand, A., & Moonen, M. (2010). Distributed adaptive node-specific signal estimation in fully connected sensor networks—Part I: Sequential node updating. *IEEE Transactions on Signal Processing*, *58*(10), 5277–5291. https://doi.org/10.1109/TSP.2010.2052612

Bronkhorst, A. W., & Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. (A. S. America, Ed.). *The Journal of the Acoustical Society of America*, *83*(4), 1508–1516. https://doi.org/10.1121/1.395906

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109. https://doi.org/10.1121/1.1345696.

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, *57*(8), 1408–1418. https://doi.org/10.1109/PROC.1969.7278

Carbajal, G., Serizel, R., Vincent, E., & Humbert, E. (2020). Joint NN-supported multichannel reduction of acoustic echo, reverberation and noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2158–2173. https://doi.org/10.1109/TASLP.2020.3008974

Cornelis, B. M. (2014). Reduced-bandwidth multi-channel Wiener filter based binaural noise reduction and localization cue preservation in binaural hearing aids. *IEEE Transactions on Audio, Speech, and Language Processing*, 1–16.

Delebecque, L., & Serizel, R. (2023). Binaurec: A dataset to test the influence of the use of room impulse responses on binaural speech enhancement. European Signal Processing Conference (pp. 126–130). Nancy, France: Eurasip. https://doi.org/10.23919/EUSIPCO58844.2023.10289772

Ditter, D., & Gerkmann, T. (2020). A multi-phase gammatone filterbank for speech separation via tasnet. ICASSP, 86–90.

Doclo, S., & Moonen, M. (2002). GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Transactions on Signal Processing*, *50*(9), 2230–2244. https://doi.org/10.1109/TSP.2002.801937

Dowerah, S., Kulkarni, A., Serizel, R., & Jouvet, D. (2023). Self-supervised learning with diffusion-based multichannel speech enhancement for speaker verification under noisy conditions. Interspeech (pp. 3849–3853). Dublin, Ireland: ISCA. https://doi.org/10.48550/arXiv.2307.02244

Dubno, J. R., & Levitt, H. (1981). Predicting consonant confusions from acoustic analysis. *The Journal of the Acoustical Society of America*, *69*(1), 249–261. https://doi.org/10.1121/1.385345

Font, F., Roma, G., & Serra, X. (2013). Freesound technical demo. International conference on Multimedia (pp. 411–412). Barcelona, Spain: ACM. https://doi.org/10.1145/2502081.2502245

Furnon, N., Serizel, R., Essid, S., & Illina, I. (2021). DNN-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 2310–2323. https://doi.org/10.1109/TASLP.2021.3092838

Gelfand, S. A., Piper, N., & Silman, S. (1985). Consonant recognition in quiet as a function of aging among normal hearing subjects. *The Journal of the Acoustical Society of America*, *78*(4), 1198–1206. https://doi.org/10.1121/1.392888

Greenberg, J. E., Peterson, P. M., & Zurek, P. M. (1993). Intelligibility-weighted measures of speech-to-interference ratio and speech system performance. *Journal of the Acoustical Society of America*, *94*(5), 3009–3010. https://doi.org/10.1121/1.407334.

Hendriks, R., & Gerkmann, T. (2011). Noise correlation matrix estimation for multi-microphone speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 223–233. https://doi.org/10.1109/TASL.2011.2159711

Heymann, J., Drude, L., & Haeb-Umbach, R. (2016). Neural network based spectral mask estimation for acoustic beamforming. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 196–200). Shanghai, China: IEEE. https://doi.org/10.1109/ICASSP.2016.7471664

Keith, R. W., & Talis, H. P. (1972). The effects of white noise on PB score of normal and hearing-impaired listeners. *International Journal of Audiology*, *11*(3–4), 177–186. https://doi.org/10.3109/00206097209089294

Kollmeier, B., & Koch, R. (1994). Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *The Journal of the Acoustical Society of America*, *95*(3), 1593–1602. https://doi.org/10.1121/1.408546

Le Roux, J., Wisdom, S., Erdogan, H., & Hershey, J. R. (2019). SDR—Half-baked or well done? International Conference on Acoustics, Speech and Signal Processing (pp. 626–630). Brighton, UK: IEEE. https://doi.org/10.48550/arXiv.1811.02508

Li, F., Menon, A., & Allen, J. B. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *The Journal of the Acoustical Society of America*, *127*(4), 2599–2610. https://doi.org/10.1121/1.3295689.

Li, C., Shi, J., Zhang, W., Subramanian, A. S., Chang, X., & Kamo, N. (2021). ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration. IEEE Spoken Language Technology Workshop (pp. 785–792). Shenzhen, China: IEEE. https://doi.org/10.1109/SLT48900.2021.9383615

Loizou, P. (2007). *Speech enhancement: Theory and practice*. CRC Press. https://doi.org/10.1201/b14529

Luo, Y., Han, C., Mesgarani, N., Ceolini, E., & Liu, S.-C. (2019). FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing. IEEE automatic speech recognition and understanding workshop (ASRU) (pp. 260–267). Sentosa, Singapore: IEEE. https://doi.org/10.1109/ASRU46091.2019.9003849

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech

alignment using Kaldi. Interspeech (pp. 498–502). Stockholm, Sweden: ISCA. https://doi.org/10.21437/Interspeech.2017-1386

Meyer, B. T., Jürgens, T., Wesker, T., Brand, T., Kollmeier, B., & Meyer, E. (2010). Human phoneme recognition depending on speech-intrinsic variability. *The Journal of the Acoustical Society of America*, *128*(5), 3126–3141. https://doi.org/10.1121/1.3493450

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338–352. https://doi.org/10.1121/1.1907526.

Naylor, P., & Gaubitch, N. (2010). *Speech dereverberation*. Springer. https://doi.org/10.1007/978-1-84996-056-4

Novak, A., Lotton, P., & Simon, L. (2015). Synchronized swept-sine: Theory, application, and implementation. *Journal of the Audio Engineering Society*, *63*(10), 786–798. https://doi.org/10.17743/jaes.2015.0071

Nugraha, A. A., Liutkus, A., & Vincent, E. (2016). Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(9), 1652–1664. https://doi.org/10.1109/TASLP.2016.2580946

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. IEEE international conference on acoustics, speech and signal processing (pp. 5206–5210). South Brisbane, QLD, Australia: IEEE. https://doi.org/10.1109/ICASSP.2015.7178964

Phatak, S. A., & Allen, J. B. (2007). Consonant and vowel confusions in speech-weighted noise. *The Acoustical Society of America*, 2312–2326. https://doi.org/10.1121/1.2642397

Phatak, S. A., Lovitt, A., & Allen, J. B. (2008). Consonant confusions in white noise. *The Journal of the Acoustical Society of America*, 1220–1233. https://doi.org/10.1121/1.2913251

Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., & Raffel, C. C. (2014). MIR_EVAL: A transparent implementation of common MIR metrics. ISMIR (pp. 367–372). Taipei, Taiwan: ISMIR.

Reynard, Pierre, Lagacé, Josée, Joly, Charles-Alexandre, Dodelé, Léon, Veuillet, Evelyne, & Thai-Van, Hung (2022). Speech-in-Noise Audiometry in Adults: A Review of the Available Tests for French Speakers. *Audiology and Neurotology*, *27*(3), 185–199. http://dx.doi.org/10.1159/000518968

Rix, A. B. (2001). Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 749–752.

Scheidiger, C., Allen, J. B., & Dau, Torsten (2017). Assessing the efficacy of hearing-aid amplification using a phoneme test. *Journal of Speech and Hearing Research*, 1739.

Serizel, R., Moonen, M., Van Dijk, B., & Wouters, J. (2014). Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(4), 785–799. https://doi.org/10.1109/TASLP.2014.2304240

Spriet, A., Moonen, M., & Wouters, J. (2004). Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction. *Signal Processing*, *84*(12), 2367–2387. https://doi.org/10.1016/j.sigpro.2004.07.028

Stone, M. A., Füllgrabe, C., & Moore, B. C. (2012). Notionally steady background noise acts primarily as a modulation masker of speech. *The Journal of the Acoustical Society of America*, 317–326. https://doi.org/10.1121/1.4725766

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 4214–4217.

Tolooshams, B., Giri, R., Song, A. H., Isik, U., & Krishnaswamy, A. (2020). Channel-attention dense u-net for multichannel speech enhancement. IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 836–840). Barcelona, Spain: IEEE. https://doi.org/10.1109/ICASSP40776.2020.9053989

Toscano, J. C. (2014). Across- and within-consonant errors for isolated syllables in noise. *Journal of Speech Language and Hearing Research*, 2293–2307. https://doi.org/10.1044/2014_JSLHR-H-13-0244

Van den Bogaert, T., Doclo, S., Wouters, J., & Moonen, M. (2009). Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids. *The Journal of the Acoustical Society of America*, *125*(1), 360–371. https://doi.org/10.1121/1.3023069

Vincent, E., Gribonval, R., & Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE transactions on Audio, Speech, and Language Processing*, *14*(4), 1462–1469. https://doi.org/10.1109/TSA.2005.858005

Vincent, E., Watanabe, S., Nugraha, A. A., Barker, J., & Marxer, R. (2017). An analysis of environment, microphone and data simulation mismatches in robust speech recognition. (Elsevier, ed.). *Computer Speech & Language*, *46*, 535–557. https://doi.org/10.1016/j.csl.2016.11.005

Woods, D. L., William Yund, E., Herron, T. J., & Ua Cruadhlaoich, M. A. I. (2010). Consonant identification in consonant-vowel-consonant syllables in speech-spectrum noise. *The Journal of the Acoustical Society of America*, *127*(3), 1609–1623. https://doi.org/10.1121/1.3293005.

Zaar, J., & Dau, T. (2017). Predicting consonant recognition and confusions in normal-hearing listeners. *The Journal of the Acoustical Society of America*, *141*(2), 1051–1064. https://doi.org/10.1121/1.4976054.