

Article

Enhancing Target Speech Based on Nonlinear Soft Masking Using a Single Acoustic Vector Sensor

Yuexian Zou ^{1,*}, Zhaoyi Liu ¹ and Christian H. Ritz ²

¹ ADSPLAB, School of Electronic Computer Engineering, Peking University Shenzhen Graduate School, Shenzhen 518055, China; 1701213615@sz.pku.edu.cn

² School of Electrical, Computer, and Telecommunications Engineering, University of Wollongong, Wollongong, NSW 2500, Australia; critz@uow.edu.au

* Correspondence: zouyx@pkusz.edu.cn; Tel.: +86-75526032016

Received: 15 June 2018; Accepted: 25 July 2018; Published: 23 August 2018



Abstract: Enhancing speech captured by distant microphones is a challenging task. In this study, we investigate the multichannel signal properties of the single acoustic vector sensor (AVS) to obtain the inter-sensor data ratio (ISDR) model in the time-frequency (TF) domain. Then, the monotone functions describing the relationship between the ISDRs and the direction of arrival (DOA) of the target speaker are derived. For the target speech enhancement (SE) task, the DOA of the target speaker is given, and the ISDRs are calculated. Hence, the TF components dominated by the target speech are extracted with high probability using the established monotone functions, and then, a nonlinear soft mask of the target speech is generated. As a result, a masking-based speech enhancement method is developed, which is termed the AVS-SMASK method. Extensive experiments with simulated data and recorded data have been carried out to validate the effectiveness of our proposed AVS-SMASK method in terms of suppressing spatial speech interferences and reducing the adverse impact of the additive background noise while maintaining less speech distortion. Moreover, our AVS-SMASK method is computationally inexpensive, and the AVS is of a small physical size. These merits are favorable to many applications, such as robot auditory systems.

Keywords: Direction of Arrival (DOA); time-frequency (TF) mask; speech sparsity; speech enhancement (SE); acoustic vector sensor (AVS); intelligent service robot

1. Introduction

With the development of information technology, intelligent service robots will play an important role in smart home systems. Auditory perception is one of the key technologies of intelligent service robots [1]. Research has shown that special attention is currently being given to human–robot interaction [2], and especially speech interaction in particular [3,4]. It is clear that service robots are always working in noisy environments, and there are possible directional spatial interferences such as the competing speakers located in different locations, air conditioners, and so on. As a result, additive background noise and spatial interferences significantly deteriorate the quality and intelligibility of the target speech, and speech enhancement (SE) is considered the most important preprocessing technique for speech applications such as automatic speech recognition [5].

Single-channel SE and two-channel SE techniques have been studied for a long time, while practical applications have a number of constraints, such as limited physical space for installing large-sized microphones. The well-known single channel SE methods, including spectral subtraction, Wiener filtering, and their variations, are successful for suppressing additive background noise, but they are not able to suppress spatial interferences effectively [6]. Besides, mask-based SE methods have predominantly been applied in many SE and speech separation applications [7]. The key idea

behind mask-based SE methods is to estimate a spectrographic binary or soft mask to suppress the unwanted spectrogram components [7–11]. For binary mask-based SE methods, the spectrographic masks are “hard binary masks” where a spectral component is either set to 1 for the target speech component or set to 0 for the non-target speech component. Experimental results have shown that the performance of binary mask SE methods degrades with the decrease of the signal-to-noise ratio (SNR) and the masked spectral may cause the loss of speech components due to the harsh black or white binary conditions [7,8]. To overcome this disadvantage, the soft mask-based SE methods have been developed [8]. In soft mask-based SE methods, each time-frequency component is assigned a probability linked to the target speech. Compared to the binary mask SE methods, the soft-mask SE methods have shown better capability to suppress the noise with the aid of some priori information. However, the priori information may vary with time, and obtaining the priori information is not an easy task.

By further analyzing the mask-based SE algorithms, we have the following observations. (1) It is a challenging task to estimate a good binary spectrographic mask. When noise and competing speakers (speech interferences) exist, the speech enhanced by the estimated mask often suffers from the phenomenon of “musical noise”. (2) The direction of arrival (DOA) of the target speech is considered as a known parameter for the target SE task. (3) A binaural microphone and an acoustic vector sensor (AVS) are considered as the most attractive front ends for speech applications due to their small physical size. For the AVS, its physical size is about 1–2 cm³ and AVS also has the merits such as signal time alignment and a trigonometric relationship of signal amplitudes [12–16]. A high-resolution DOA estimation algorithm with a single AVS has been proposed by our team [12–16]. Some effort has also been made for the target SE task with one or two AVS sensors [17–21]. For example, with the minimum variance distortionless response (MVDR) criterion, Lockwood et al. developed a beamforming method using the AVS [17]. Their experimental results showed that their proposed algorithm achieves good performance for suppressing noise, but brings certain distortion of the target speech.

As discussed above, in this study, we focus on developing the target speech enhancement algorithm with a single AVS from a new technical perspective in which both the ambient noise and non-target spatial speech interferences can be suppressed effectively and simultaneously. The problem formulation is presented in Section 2. Section 3 shows the derivation of the proposed SE algorithm. The experimental results are given in Section 4, and conclusions are drawn in Section 5.

2. Problem Formulation

In this section, the sparsity of speech in the time-frequency (TF) domain is discussed first. Then, the AVS data model and the corresponding inter-sensor data ratio (ISDR) models are presented for completeness, which was developed by our team in a previous work [13]. After that, the derivation of monotone functions between ISDRs and the DOA is given. Finally, the nonlinear soft TF mask estimation algorithm is derived specifically.

2.1. Time-Frequency Sparsity of Speech

In the research of speech signal processing, the TF sparsity of speech is a widely accepted assumption. More specifically, when there is more than one speaker in the same spatial space, the speech TF sparsity implies the following [5]. (1) It is likely that only one speaker is active during certain time slots. (2) For the same time slot, if more than one speaker is active, it is probable that the different TF points are dominated by different speakers. Hence, the TF sparsity of speech can be modeled as:

$$S_m(\tau, \omega)S_n(\tau, \omega) = 0, m \neq n \quad (1)$$

where $S_m(\tau, \omega)$ and $S_n(\tau, \omega)$ are the speech spectral at (τ, ω) for the m th speaker and n th speaker, respectively. (3) In practice, at a specific TF point (τ, ω) , it is most probably true that only one

speech source with the highest energy dominates, and the contributions from the other sources can be negligible.

2.2. AVS Data Model

An AVS unit generally consists of J co-located constituent sensors, including one omnidirectional sensor (denoted as o -sensor) and $J-1$ orthogonally oriented directional sensors. Figure 1 shows the data capture setup with a single AVS. It is noted that the left bottom plot in Figure 1 shows a 3D-AVS unit implemented by our team, which consists of one o -sensor with three orthogonally oriented directional sensors depicted as the u -sensor, v -sensor, and w -sensor, respectively. In theory, the directional response of the oriented directional sensor has dipole characteristics, as shown in Figure 2a, while the omnidirectional sensor has the same response in all of the directions, as shown in Figure 2b. In this study, one target speaker is considered. As shown in Figure 1, the target speech $S(t)$ is impinging from (θ_s, ϕ_s) meanwhile, interference $S_i(t)$ are impinging from (θ_i, ϕ_i) , where $\phi_s, \phi_i \in (0^\circ, 360^\circ)$ are the azimuth angles, and $\theta_s, \theta_i \in (0^\circ, 180^\circ)$ are the elevation angles.

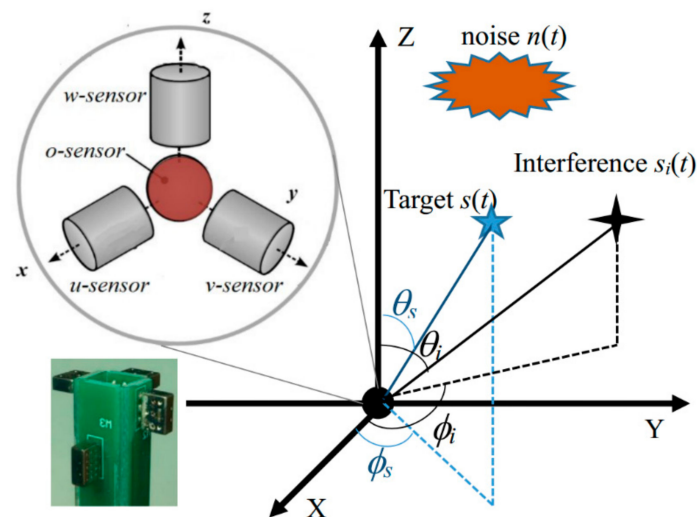


Figure 1. Illustration of a single acoustic vector sensor (AVS) for data capturing.

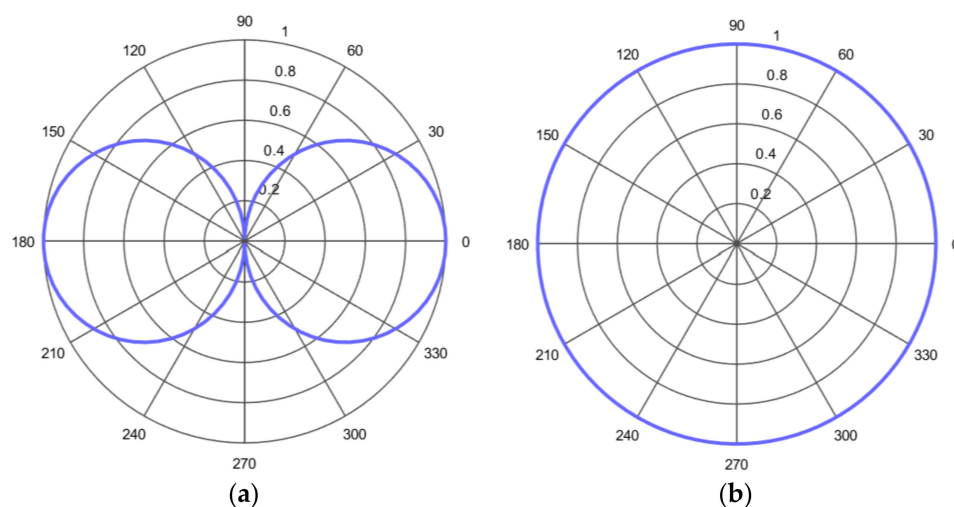


Figure 2. (a) The directional response of oriented directional sensor; (b) The directional response of omnidirectional sensor.

For simplifying the derivation, without considering room reverberation, the received data of the AVS can be modeled as [13]:

$$\mathbf{x}_{avs}(t) = \mathbf{a}(\theta_s, \phi_s)s(t) + \sum_{i=1}^{M_i} \mathbf{a}(\theta_i, \phi_i)s_i(t) + \mathbf{n}_{avs}(t) \quad (2)$$

where $\mathbf{x}_{avs}(t)$, $\mathbf{n}_{avs}(t)$ and $\mathbf{a}(\theta_s, \phi_s)$ are defined respectively as:

$$\mathbf{x}_{avs}(t) = [x_u(t), x_v(t), x_w(t), x_o(t)] \quad (3)$$

$$\mathbf{n}_{avs}(t) = [n_u(t), n_v(t), n_w(t), n_o(t)] \quad (4)$$

$$\mathbf{a}(\theta_s, \phi_s) = [u_s, v_s, w_s, 1]^T = [\sin \theta_s \cos \phi_s, \sin \theta_s \sin \phi_s, \cos \theta_s, 1]^T \quad (5)$$

$$\mathbf{a}(\theta_i, \phi_i) = [u_i, v_i, w_i, 1]^T = [\sin \theta_i \cos \phi_i, \sin \theta_i \sin \phi_i, \cos \theta_i, 1]^T \quad (6)$$

In Equation (3), $x_u(t)$, $x_v(t)$, $x_w(t)$, $x_o(t)$ are the received data of the u -sensor, v -sensor, w -sensor, and o -sensor, respectively; $n_u(t)$, $n_v(t)$, $n_w(t)$, $n_o(t)$ are assumed as the additive zero-mean white Gaussian noise captured at the u -sensor, v -sensor, w -sensor, and o -sensor, respectively; $s(t)$ is the target speech; $s_i(t)$ are the i th interfering speech; the number of interferences is M_i ; $\mathbf{a}(\theta_s, \phi_s)$ and $\mathbf{a}(\theta_j, \phi_j)$ are the steering vectors of $s(t)$ and $s_i(t)$, respectively. $[\cdot]^T$ denotes the vector/matrix transposition.

From the AVS data model given in Equation (2), taking the short-time Fourier transform (STFT), for a specific TF point (τ, ω) , we have:

$$\mathbf{X}_{avs}(\tau, \omega) = \mathbf{a}(\theta_s, \phi_s)S(\tau, \omega) + \sum_{i=1}^{M_i} \mathbf{a}(\theta_i, \phi_i)S_i(\tau, \omega) + \mathbf{N}_{avs}(\tau, \omega) \quad (7)$$

where $\mathbf{X}_{avs}(\tau, \omega) = [X_u(\tau, \omega), X_v(\tau, \omega), X_w(\tau, \omega), X_o(\tau, \omega)]^T$; $X_u(\tau, \omega)$, $X_v(\tau, \omega)$, $X_w(\tau, \omega)$, and $X_o(\tau, \omega)$ are the STFT of $x_u(t)$, $x_v(t)$, $x_w(t)$, and $x_o(t)$, respectively. Meanwhile, $\mathbf{N}_{avs}(\tau, \omega) = [N_u(\tau, \omega), N_v(\tau, \omega), N_w(\tau, \omega), N_o(\tau, \omega)]^T$; $N_u(\tau, \omega)$, $N_v(\tau, \omega)$, $N_w(\tau, \omega)$, and $N_o(\tau, \omega)$ are the STFT of $n_u(t)$, $n_v(t)$, $n_w(t)$, and $n_o(t)$, respectively. Since the target speech spectral is $S(\tau, \omega)$, let us define a quantity as follows:

$$\mathbf{N}_{total}(\tau, \omega) = \sum_{i=1}^{M_i} \mathbf{a}(\theta_i, \phi_i)S_i(\tau, \omega) + \mathbf{N}_{avs}(\tau, \omega) \quad (8)$$

where we define $\mathbf{N}_{total}(\tau, \omega) = [N_{tu}(\tau, \omega), N_{tv}(\tau, \omega), N_{tw}(\tau, \omega), N_{to}(\tau, \omega)]^T$ to represent the mixture of the interferences and additive noise. Therefore, from Equations (7) and (8), we have the following expressions:

$$X_u(\tau, \omega) = u_s S(\tau, \omega) + N_{tu}(\tau, \omega) \quad (9)$$

$$X_v(\tau, \omega) = v_s S(\tau, \omega) + N_{tv}(\tau, \omega) \quad (10)$$

$$X_w(\tau, \omega) = w_s S(\tau, \omega) + N_{tw}(\tau, \omega) \quad (11)$$

$$X_o(\tau, \omega) = S(\tau, \omega) + N_{to}(\tau, \omega) \quad (12)$$

In this study, we make the following assumptions. (1) $s(t)$ and $s_i(t)$ are uncorrelated and are considered as far-field speech sources; (2) $n_u(t)$, $n_v(t)$, $n_w(t)$ and $n_o(t)$ are uncorrelated. (3) The DOA of the target speaker is given as (θ_s, ϕ_s) ; the task of target speech enhancement is essentially to estimate $S(\tau, \omega)$ from $\mathbf{X}_{avs}(\tau, \omega)$.

2.3. Monotone Functions between ISDRs and the DOA

Definition and some discussions on the inter-sensor data ratio (ISDR) of the AVS are presented in our previous work [13]. In this subsection, we briefly introduce the definition of ISDR first, and then present the derivation of the monotone functions between the ISDRs and the DOA of the target speaker.

The ISDRs between each channel of the AVS are defined as:

$$I_{ij}(\tau, \omega) = X_i(\tau, \omega) / X_j(\tau, \omega) \text{ where } (i \neq j) \quad (13)$$

where i and j are the channel index, which refers to u , v , w , and o , respectively. Obviously, there are 12 different computable ISDRs, which are shown in Table 1. In the following context, we carefully evaluate I_{ij} , and it is clear that only three ISDRs (I_{uv} , I_{vu} and I_{wo}) hold the approximate monotone function between ISDR and the DOA of the target speaker.

Table 1. Twelve computable inter-sensor data ratios (ISDRs).

Sensor	u	v	w	o
u	NULL	I_{vu}	I_{wu}	I_{ou}
v	I_{uv}	NULL	I_{wv}	I_{ov}
w	I_{uw}	I_{vw}	NULL	I_{ow}
o	I_{uo}	I_{vo}	I_{wo}	NULL

According to the definition of ISDRs given in Equation (13), we look at I_{uv} , I_{vu} and I_{wo} first. Specifically, we have:

$$I_{uv}(\tau, \omega) = X_u(\tau, \omega) / X_v(\tau, \omega) \quad (14)$$

$$I_{vu}(\tau, \omega) = X_v(\tau, \omega) / X_u(\tau, \omega) \quad (15)$$

$$I_{wo}(\tau, \omega) = X_w(\tau, \omega) / X_o(\tau, \omega) \quad (16)$$

Substituting Equations (9) and (10) into Equation (14) gives:

$$I_{uv}(\tau, \omega) = \frac{u_s S(\tau, \omega) + N_{tu}(\tau, \omega)}{v_s S(\tau, \omega) + N_{tv}(\tau, \omega)} = \frac{u_s + N_{tu}(\tau, \omega) / S(\tau, \omega)}{v_s + N_{tv}(\tau, \omega) / S(\tau, \omega)} = \frac{u_s + \varepsilon_{tus}(\tau, \omega)}{v_s + \varepsilon_{tvs}(\tau, \omega)} \quad (17)$$

where $\varepsilon_{tus}(\tau, \omega) = N_{tu}(\tau, \omega) / S(\tau, \omega)$, and $\varepsilon_{tvs}(\tau, \omega) = N_{tv}(\tau, \omega) / S(\tau, \omega)$.

Similarly, we get I_{uv} and I_{wo} :

$$I_{vu}(\tau, \omega) = \frac{v_s S(\tau, \omega) + N_{tv}(\tau, \omega)}{u_s S(\tau, \omega) + N_{tu}(\tau, \omega)} = \frac{v_s + N_{tv}(\tau, \omega) / S(\tau, \omega)}{u_s + N_{tu}(\tau, \omega) / S(\tau, \omega)} = \frac{v_s + \varepsilon_{tvs}(\tau, \omega)}{u_s + \varepsilon_{tus}(\tau, \omega)} \quad (18)$$

$$I_{wo}(\tau, \omega) = \frac{w_s S(\tau, \omega) + N_{tw}(\tau, \omega)}{S(\tau, \omega) + N_{to}(\tau, \omega)} = \frac{w_s + N_{tw}(\tau, \omega) / S(\tau, \omega)}{1 + N_{to}(\tau, \omega) / S(\tau, \omega)} = \frac{w_s + \varepsilon_{twS}(\tau, \omega)}{1 + \varepsilon_{tos}(\tau, \omega)} \quad (19)$$

In Equation (19), $\varepsilon_{twS}(\tau, \omega) = N_{tw}(\tau, \omega) / S(\tau, \omega)$ and $\varepsilon_{tos}(\tau, \omega) = N_{to}(\tau, \omega) / S(\tau, \omega)$.

Based on the assumption of TF sparsity of speech shown in Section 2.1, we can see that if the TF points (τ, ω) are dominated by the target speech from (θ_s, ϕ_s) , the energy of the target speech is high, and the value of $\varepsilon_{tus}(\tau, \omega)$, $\varepsilon_{tvs}(\tau, \omega)$, $\varepsilon_{twS}(\tau, \omega)$ and $\varepsilon_{tos}(\tau, \omega)$ tends to be small. Then, Equations (17)–(19) can be accordingly approximated as:

$$I_{uv}(\tau, \omega) \approx u_s / v_s + \varepsilon_1(\tau, \omega) \quad (20)$$

$$I_{vu}(\tau, \omega) \approx v_s / u_s + \varepsilon_2(\tau, \omega) \quad (21)$$

$$I_{wo}(\tau, \omega) \approx w_s + \varepsilon_3(\tau, \omega) \quad (22)$$

where ε_1 , ε_2 , and ε_3 can be viewed as the ISDR modeling error with zero-mean introduced by interferences and background noise. Moreover, $\varepsilon_i(\tau, \omega)$ ($i = 1, 2, 3$) is inversely proportion to the local SNR at (τ, ω) .

Furthermore, from Equation (5), we have $u_s = \sin\theta_s \cdot \cos\phi_s$, $v_s = \sin\theta_s \cdot \sin\phi_s$ and $w_s = \cos\theta_s$. Then, substituting Equation (5) into Equations (20)–(22), we obtain the following equations:

$$I_{uv}(\tau, \omega) \approx \frac{\sin\theta_s \cos\phi_s}{\sin\theta_s \sin\phi_s} + \varepsilon_1(\tau, \omega) = \cot\phi_s + \varepsilon_1(\tau, \omega) \quad (23)$$

$$I_{vu}(\tau, \omega) \approx \frac{\sin\theta_s \sin\phi_s}{\sin\theta_s \cos\phi_s} + \varepsilon_2(\tau, \omega) = \tan(\phi_s) + \varepsilon_2(\tau, \omega) \quad (24)$$

$$I_{wo}(\tau, \omega) \approx w_s + \varepsilon_3(\tau, \omega) = \cos(\theta_s) + \varepsilon_3(\tau, \omega) \quad (25)$$

From Equations (23)–(25), it is desired to see that the approximate monotone functions between I_{uv} , I_{vu} , and I_{wo} and the DOA (θ_s or ϕ_s) of the target speaker have been obtained since arccot , arctan , and arccos functions are all monotone functions.

However, except for I_{uv} , I_{vu} , and I_{wo} , other ISDRs do not hold such a property. Let's take I_{uw} as an example. From the definition in Equation (13), we can get:

$$I_{uw}(\tau, \omega) = \frac{u_s S(\tau, \omega) + N_{Iu}(\tau, \omega)}{w_s S(\tau, \omega) + N_{Iw}(\tau, \omega)} = \frac{u_s + N_{Iu}(\tau, \omega)/S(\tau, \omega)}{w_s + N_{Iw}(\tau, \omega)/S(\tau, \omega)} = \frac{u_s + \varepsilon_{Ius}(\tau, \omega)}{w_s + \varepsilon_{Iws}(\tau, \omega)} = \frac{u_s}{w_s} + \varepsilon_4(\tau, \omega) \quad (26)$$

where ε_4 can be viewed as the ISDR modeling error with zero-mean introduced by unwanted noise. Obviously, Equation (26) is valid when w_s is not equal to zero. Substituting Equation (5) into Equation (26) yields:

$$I_{uw}(\tau, \omega) \approx \frac{\sin\theta_s \cos\phi_s}{\cos\theta_s} + \varepsilon_4(\tau, \omega) = \tan\theta_s \cos\phi_s + \varepsilon_4(\tau, \omega) \quad (27)$$

From Equation (27), we can see that I_{uw} is a function of both θ_s and ϕ_s .

In summary, after analyzing all of the ISDRs, we find that the desired monotone functions between ISDRs and θ_s or ϕ_s , which are given in Equations (23)–(25), respectively. It is noted that Equations (23)–(25) are derived conditioned by assuming v_s , u_s , and w_s are not equal to zero. Therefore, we need to find out where v_s , u_s , and w_s are equal to zero. For presentation clarity, let's define an ISDR vector $I_{\text{isdr}} = [I_{uv}, I_{vu}, I_{wo}]$.

From Equation (5), it is clear that when the target speaker is at angles of 0° , 90° , 180° , and 270° , one of v_s , u_s , and w_s becomes zero, and it means that I_{isdr} is not fully available. Specifically, we need to consider the following cases:

Case 1: the elevation angle θ_s is about 0° or 180° . In this case, $u_s = \sin\theta_s \cdot \cos\phi_s$ and $v_s = \sin\theta_s \cdot \sin\phi_s$ are close to zero. Then, the denominator in Equations (20) and (21) is equal to zero, and we cannot obtain I_{uv} and I_{vu} , but we can get I_{wo} .

Case 2: θ_s is away from 0° or 180° . In this condition, we need to look at ϕ_s carefully.

- (1) ϕ_s is about 0° or 180° . Then, $v_s = \sin\theta_s \sin\phi_s$ is close to zero, and the denominator in Equation (20) is equal to zero, which leads to I_{uv} being invalid. In this case, we can compute I_{vu} and I_{wo} properly.
- (2) ϕ_s is about 90° or 270° . Then, $u_s = \sin\theta_s \cdot \cos\phi_s$ is close to zero, and the denominator in Equation (21) is equal to zero, which leads to I_{vu} being invalid. In this case, we can obtain I_{uv} and I_{wo} properly.
- (3) ϕ_s is away from 0° , 90° , 180° , and 270° , we can obtain all of the I_{uv} , I_{vu} and I_{wo} values properly.

To visualize the discussions above, a decision tree of handling the special angles in computing I_{isdr} is plotted in Figure 3.

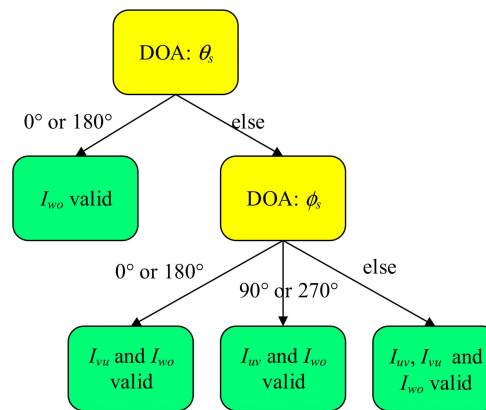


Figure 3. The decision tree of handling the special angles in computing I_{isdr} .

When $I_{\text{isdr}} = [I_{uv}, I_{vu}, I_{w0}]$ has been computed properly, with simple manipulation from Equations (23)–(25), we get:

$$\phi_s(\tau, \omega) = \text{arccot}(I_{uv}(\tau, \omega) - \varepsilon_1(\tau, \omega)) \quad (28)$$

$$\phi_s(\tau, \omega) = \text{arctan}(I_{vu}(\tau, \omega) - \varepsilon_2(\tau, \omega)) \quad (29)$$

$$\theta_s(\tau, \omega) = \text{arccos}(I_{w0}(\tau, \omega) - \varepsilon_3(\tau, \omega)) \quad (30)$$

From Equations (28)–(30), we can see that arccot , arctan , and arccos functions are all monotone functions, which are what we expected. Besides, we also note that (θ_s, ϕ_s) is given, and I_{uv} , I_{vu} and I_{w0} can be computed by Equations (14)–(16). However, ε_1 , ε_2 , and ε_3 are unknown, which reflect the impact of noise and interferences. According to the assumptions made in Section 2.1, if we are able to select the TF components (θ_s, ϕ_s) dominated by the target speech, and the local SNR at this (τ, ω) is high, then ε_1 , ε_2 , and ε_3 can be ignored, since they will have values approaching zero at these (τ, ω) points. In such conditions, we obtain the desired formulas to compute (θ_s, ϕ_s) :

$$\phi_s(\tau, \omega) \approx \text{arccot}(I_{uv}(\tau, \omega)), \phi_s(\tau, \omega) \approx \text{arctan}(I_{vu}(\tau, \omega)) \text{ and } \theta_s(\tau, \omega) \approx \text{arccos}(I_{w0}(\tau, \omega)) \quad (31)$$

2.4. Nonlinear Soft Time-Frequency (TF) Mask Estimation

As discussed above, Equation (31) is valid when the (τ, ω) points are dominated by target speech with high local SNR. Besides, we have three equations to solve two variables, θ_s and ϕ_s . In this study, from Equation (31), we estimate θ_s and ϕ_s in the following way:

$$\hat{\phi}_{s1}(\tau, \omega) = \text{arccot} I_{uv}(\tau, \omega) + \Delta\eta_1 \quad (32)$$

$$\hat{\phi}_{s2}(\tau, \omega) = \text{arctan} I_{vu}(\tau, \omega) + \Delta\eta_2 \quad (33)$$

$$\hat{\phi}_s(\tau, \omega) = \text{mean}(\hat{\phi}_{s1}, \hat{\phi}_{s2}) \quad (34)$$

$$\hat{\theta}_s(\tau, \omega) = \text{arccos} I_{w0}(\tau, \omega) + \Delta\eta_3 \quad (35)$$

where $\Delta\eta_1$ and $\Delta\eta_2$ are estimation errors. Comparing Equation (31) and Equations (32)–(35), we can see that if the estimated DOA values $(\hat{\phi}_s(\tau, \omega), \hat{\theta}_s(\tau, \omega))$ approximate to the real DOA values (θ_s, ϕ_s) , then $\Delta\eta_1$ and $\Delta\eta_2$ should be small. Therefore, for the TF points (τ, ω) dominated by the target speech, we can derive the following inequality:

$$|\hat{\phi}_s(\tau, \omega) - \phi_s| < \delta_1 \quad (36)$$

$$|\hat{\theta}_s(\tau, \omega) - \theta_s| < \delta_2 \quad (37)$$

where $\hat{\phi}_s(\tau, \omega)$ and $\hat{\theta}_s(\tau, \omega)$ are the target speaker's DOA estimated by Equations (34) and (35), respectively. θ_s and ϕ_s are given the DOA of the target speech for the SE task. The parameters δ_1 and δ_2 can be set as the predefined permissible parameters (referring to an angle value). Following the derivation up to now, if Equations (36) and (37) are met at (τ, ω) points, we can infer that these (τ, ω) points are dominated by the target speech with high probability. Therefore, using Equations (36) and (37), the TF points (τ, ω) can be extracted, and a mask associated with these (τ, ω) points dominated by the target speech can be designed accordingly. In addition, we need to take the following facts into account. (1) The value of ϕ_s belongs to $(0, 2\pi]$. (2) The principal value interval of the arccot function is $(0, \pi)$, and the arctan function is $(-\pi/2, \pi/2)$. (3) The value range of θ_s is $(0, 2\pi]$. (4) The principal value interval of the arccos function is $[0, \pi]$. (5) To make the principal value of the anti-trigonometric function match the value of θ_s and ϕ_s , we need to add $L\pi$ to avoid ambiguity. As a result, a binary TF mask for preserving the target speech is designed as follows:

$$mask(\tau, \omega) = \begin{cases} 1, & \text{if } \begin{cases} \Delta\phi(\tau, \omega) = |\hat{\phi}_s(\tau, \omega) - \phi_s + L\pi| < \delta_1 \\ \Delta\theta(\tau, \omega) = |\hat{\theta}_s(\tau, \omega) - \theta_s + L\pi| < \delta_2 \end{cases} \\ 0, & \text{else} \end{cases} \quad (38)$$

where $L = 0, \pm 1$. $(\Delta\phi(\tau, \omega), \Delta\theta(\tau, \omega))$ is the estimation difference between the estimated DOA and the real DOA of the target speaker at TF point (τ, ω) . Obviously, the smaller the value of $(\Delta\phi(\tau, \omega), \Delta\theta(\tau, \omega))$, the more probable it is that the TF point (τ, ω) is dominated by the target speech. To further improve the estimation accuracy and suppress the impact of the outliers, we propose a nonlinear soft TF mask as:

$$mask(\tau, \omega) = \begin{cases} \frac{1}{1 + e^{-\xi(1 - (\Delta\phi(\tau, \omega)/\delta_1 + \Delta\theta(\tau, \omega)/\delta_2)/2)}} & \Delta\phi < \delta_1 \ \& \ \Delta\theta < \delta_2 \\ \rho & \text{else} \end{cases} \quad (39)$$

where ξ is a positive parameter and ρ ($0 \leq \rho < 1$) is a small positive parameter tending to be zero, which reflects the noise suppression effect. The parameters Δ_1 and Δ_2 control the degree of the estimation difference $(\Delta\phi(\tau, \omega), \Delta\theta(\tau, \omega))$. When parameters Δ_1 , Δ_2 , and ρ become larger, the capability of suppressing noise and interferences degrades, and the possibility of the (τ, ω) being dominated by the target speech also degrades. Hence, selecting the values of ρ , Δ_1 , and Δ_2 is important. In our study, these parameters are determined through experiments. Future work could focus on selecting these parameters based on models of human auditory perception. In the end, we need to emphasize that the mask designed in Equation (39) has the ability to suppress the adverse effects of the interferences and background noise, and preserve the target speech simultaneously.

3. Proposed Target Speech Enhancement Method

The diagram of the proposed speech enhancement method (termed as AVS-SMASK) is shown in Figure 4, which is processed in the time-frequency domain. The details of each block in Figure 4 will be addressed in the following context.

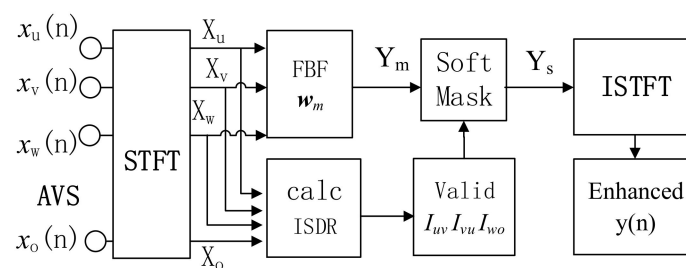


Figure 4. Block diagram of our proposed AVS-SMASK algorithm (STFT: Short-Time Fourier Transform; FBF: a fixed beamformer; ISTFT: inverse STFT; $y(n)$: enhanced target speech).

3.1. The FBF Spatial Filter

As shown in Figure 4, the input signals to the FBF spatial filter are the data captured by the u , v , and w -sensor of the AVS. With the given DOA (θ_s, ϕ_s) , the spatial matched filter (SMF) is employed as the FBF spatial filter, and its output can be described as:

$$Y_m(\tau, \omega) = \mathbf{w}_m^H \mathbf{X}_{avs}(\tau, \omega) \quad (40)$$

where $\mathbf{w}_m^H = \mathbf{a}^H(\theta_s, \phi_s) / \|\mathbf{a}(\theta_s, \phi_s)\|^2$ is the weight vector of the SMF, and $\mathbf{a}(\theta_s, \phi_s)$ is given in Equation (5). $[\cdot]^H$ denotes the vector/matrix conjugate transposition. Substituting the expressions in Equations (5), (3), and (9)–(11) in Equation (40) yields:

$$\begin{aligned} Y_m(\tau, \omega) &= u_s X_u(\tau, \omega) + v_s X_v(\tau, \omega) + w_s X_w(\tau, \omega) \\ &= u_s^2 S(\tau, \omega) + u_s N_{tu}(\tau, \omega) + v_s^2 S(\tau, \omega) + v_s N_{tv}(\tau, \omega) + w_s^2 S(\tau, \omega) + w_s N_{tw}(\tau, \omega) \\ &= (u_s^2 + v_s^2 + w_s^2) S(\tau, \omega) + N_{tuvw}(\tau, \omega) \\ &= S(\tau, \omega) + N_{tuvw}(\tau, \omega) \end{aligned} \quad (41)$$

where $N_{tuvw}(\tau, \omega)$ is the total noise component given as:

$$\begin{aligned} N_{tuvw}(\tau, \omega) &= u_s N_{tu}(\tau, \omega) + v_s N_{tv}(\tau, \omega) + w_s N_{tw}(\tau, \omega) \\ &= u_s (u_i S_i(\tau, \omega) + N_u(\tau, \omega)) + v_s (v_i S_i(\tau, \omega) + N_v(\tau, \omega)) \\ &\quad + w_s (w_i S_i(\tau, \omega) + N_w(\tau, \omega)) \\ &= (u_s u_i + v_s v_i + w_s w_i) S_i(\tau, \omega) + u_s N_u(\tau, \omega) + v_s N_v(\tau, \omega) + w_s N_w(\tau, \omega) \end{aligned} \quad (42)$$

It can be seen that $N_{tuvw}(\tau, \omega)$ in Equation (42) consists of the interferences and background noise captured by directional sensors, while $Y_m(\tau, \omega)$ in Equation (41) is the mix of the desired speech source $S(\tau, \omega)$ and unwanted component $N_{tuvw}(\tau, \omega)$.

3.2. Enhancing Target Speech Using Estimated Mask

With the estimated mask in Equation (39) and the output of the FBF spatial filter $Y_m(\tau, \omega)$ in Equation (42), it is straightforward to compute the enhanced target speech as follows:

$$Y_s(\tau, \omega) = Y_m(\tau, \omega) \times \text{mask}(\tau, \omega) \quad (43)$$

where $Y_s(\tau, \omega)$ is then the spectra of the enhanced speech or an approximation of the target speech.

For presentation completeness, our proposed speech enhancement algorithm is termed as an AVS-SMASK algorithm, which is summarized in Table 2.

Table 2. The pseudo-code of our proposed AVS-SMASK algorithm.

-
- | | |
|------|--|
| (1) | Segment the output data captured by the u -sensor, v -sensor, w -sensor, and o -sensor of the AVS unit by the N -length Hamming window; |
| (2) | Calculate the STFT of the segments: $X_u(\tau, \omega)$, $X_v(\tau, \omega)$, $X_w(\tau, \omega)$ and $X_o(\tau, \omega)$; |
| (3) | Calculate the ISDR vector $\mathbf{I}_{\text{isdr}} = [I_{uv}, I_{vu}, I_{wo}]$ by Equations (14)–(16); |
| (4) | Obtain the valid $\mathbf{I}_{\text{isdr}}$ according to the known direction of arrival (DOA) (θ_s, ϕ_s) and the summary of Section 2.3; |
| (5) | Utilize the valid $\mathbf{I}_{\text{isdr}}$ to estimate the DOA $(\hat{\theta}_s, \hat{\phi}_s)$ of the target speech for each time-frequency (TF) point; |
| (6) | Determine TF points belong to the target speech by Equations (36) and (37); |
| (7) | Calculate the nonlinear soft TF mask: $\text{mask}(\tau, \omega)$ by Equation (39); |
| (8) | Calculate the output of the FBF $Y_m(\tau, \omega)$ by Equation (40); |
| (9) | Compute the enhanced speech spectrogram by Equation (43); |
| (10) | Get the enhanced speech signal $y(n)$ by ISTFT. |
-

4. Experiments and Results

The performance evaluation of our proposed AVS-SMASK algorithm has been carried out with simulated data and recorded data. Five commonly used performance measurement metrics—SNR, the signal-to-interference ratio (SIR), the signal-to-interference plus noise ratio (SINR), log spectral deviation (LSD), and the perceptual evaluation of speech quality (PESQ)—have been adopted. The definitions are given as follows for presentation completeness.

- (1) Signal-to-Noise Ratio (SNR):

$$SNR = 10 \log \left(\|s(t)\|^2 / \|n(t)\|^2 \right) \quad (44)$$

- (2) Signal-to-Interference Ratio (SIR)

$$SIR = 10 \log \left(\|s(t)\|^2 / \|s_i(t)\|^2 \right) \quad (45)$$

- (3) Signal-to-Interference plus Noise Ratio (SINR):

$$SINR = 10 \log \left(\|s(t)\|^2 / \|x(t) - s(t)\|^2 \right) \quad (46)$$

where $s(t)$ is the target speech, $n(t)$ is the additive noise, $s_i(t)$ is the i th interference, and $x(t) = s(t) + s_i(t) + n(t)$ is the received signal of the o -sensor. The metrics are calculated by averaging over frames to get more accurate measurement [22].

- (4) Log Spectral Deviation (LSD), which is used to measure the speech distortion [22]:

$$LSD = \|\ln(\psi_{ss}(f) / \psi_{yy}(f))\| \quad (47)$$

where $\psi_{ss}(f)$ is the power spectral density (PSD) of the target speech, and $\psi_{yy}(f)$ is the PSD of the enhanced speech. It is clear that smaller LSD values indicate less speech distortion.

- (5) Perceptual Evaluation of Speech Quality (PESQ). To evaluate the perceptual enhancement performance of the speech enhancement algorithms, the ITU-PESQ software [23] is utilized.

In this study, the performance comparison is carried out with the comparison algorithm AVS-FMV [17] under the same conditions. We do not take other SE methods into account since they use different transducers for signal acquisition. One set of waveform examples that is used in our experiments is shown in Figure 5, where $s(t)$ is the target speech, $s_i(t)$ is the i -th interference speech, $n(t)$ is the additive noise, and $y(t)$ is the enhanced speech.

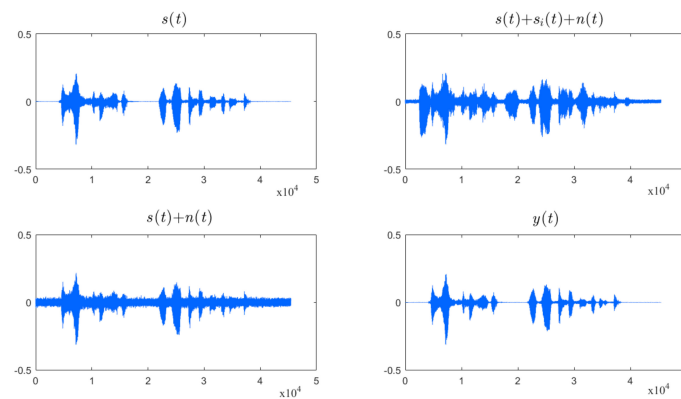


Figure 5. Waveform examples: $s(t)$ is the target speech, $s_i(t)$ is the interference speech, $n(t)$ is the additive noise, and $y(t)$ is the enhanced speech signal.

4.1. Experiments on Simulated Data

In this section, three experiments have been carried out. The simulated data of about five seconds duration is generated, where the target speech $s(t)$ is male speech, and two speech interferences $s_i(t)$ are male and female speech, respectively. Moreover, the AURORA2 database [24] was used, which includes subway, babble, car, exhibition noise, etc. Without loss of generality, all of the speech sources are placed one meter away from the AVS.

4.1.1. Experiment 1: The Output SINR Performance under Different Noise Conditions

In this experiment, we have carried out 12 trials (numbered as trial 1 to trial 12) to evaluate the performance of the algorithms under different spatial and additive noise conditions following the experimental protocols in Ref. [25]. The details are given below:

(1) The DOAs of target speech, the first speech interference (male speech) and the second speech interference (female speech) are at $(\theta_s, \phi_s) = (45^\circ, 45^\circ)$, $(\theta_1, \phi_1) = (90^\circ, 135^\circ)$, and $(\theta_2, \phi_2) = (45^\circ, 120^\circ)$, respectively. The background noise is chosen as babble noise $n(t)$;

(2) We evaluate the performance under three different conditions: (a) there exists only additive background noise: $n(t) \neq 0$ and $s_i(t) = 0$; (b) there exists only speech interferences: $n(t) = 0$ and $s_i(t) \neq 0$; (c) there exists both background noise and speech interferences: $n(t) \neq 0$ and $s_i(t) \neq 0$;

(3) The input SINR (denoted as SINR-input) is set as -5 dB, 0 dB, 5 dB, and 10 dB, respectively.

Following the setting above, 12 different datasets are generated for this experiment.

In addition, the parameters of algorithms are set as follows. (1) The sampling rate is 16 kHz, 1024 -point FFT (Fast Fourier Transform), and 1024 -point Hamming window with 50% overlapping are used. (2) For our proposed AVS-SMASK algorithm, we set $\delta_1 = \delta_2 = 25^\circ$, $\rho = 0.07$, and $\zeta = 3$. (3) For comparing algorithm AVS-FMV: $F = 32$, $M = 1.001$ followed Ref. [17]. The experimental results are given in Table 3.

Table 3. Output signal-to-interference plus noise ratio (SINR) under different noise conditions.

Noise Conditions	SINR-Input (dB)	AVS-FMV [17] (dB)		AVS-SMASK (dB)	
		SINR-Out	Average	SINR-Out	Average
Trial 1 ($n(t) = 0$ and $s_i(t) \neq 0$)	-5	4.96	4.88	7.32	8.14
Trial 2 ($n(t) = 0$ and $s_i(t) \neq 0$)	0	5.60		9.38	
Trial 3 ($n(t) = 0$ and $s_i(t) \neq 0$)	5	7.81		11.53	
Trial 4 ($n(t) = 0$ and $s_i(t) \neq 0$)	10	11.15		14.31	
Trial 5 ($n(t) \neq 0$ and $s_i(t) = 0$)	-5	4.77	4.97	6.70	9.11
Trial 6 ($n(t) \neq 0$ and $s_i(t) = 0$)	0	5.51		10.17	
Trial 7 ($n(t) \neq 0$ and $s_i(t) = 0$)	5	6.76		13.03	
Trial 8 ($n(t) \neq 0$ and $s_i(t) = 0$)	10	12.83		16.55	
Trial 9 ($n(t) \neq 0$ and $s_i(t) \neq 0$)	-5	3.66	4.42	4.70	6.66
Trial 10 ($n(t) \neq 0$ and $s_i(t) \neq 0$)	0	5.70		7.22	
Trial 11 ($n(t) \neq 0$ and $s_i(t) \neq 0$)	5	7.10		10.46	
Trial 12 ($n(t) \neq 0$ and $s_i(t) \neq 0$)	10	11.20		14.27	

As shown in Table 3, for all of the noise conditions (Trial 1 to Trial 12), our proposed AVS-SMASK algorithm outperforms AVS-FMV [17]. From Table 3, we can see that our proposed AVS-SMASK algorithm gives about 3.26 dB, 4.14 dB, and 2.25 dB improvement compared with that of AVS-FMV under three different experimental settings, respectively. We can conclude that our proposed AVS-SMASK is effective in suppressing the spatial interferences and background noise.

4.1.2. Experiment 2: The Performance versus Angle Difference

This experiment evaluates the performance of SE methods versus the angle difference between the target and interference speakers. Let's define the angle difference as $\Delta\phi = \phi_s - \phi_i$ and $\Delta\theta = \theta - \theta_i$

(here, the subscripts s and i refer to the target speaker and the interference speaker, respectively). Obviously, the closer the interference speaker is to the target speaker, the speech enhancement is more limited. The experimental settings are as follows. (1) PESQ and LSD are used as metrics. (2) The parameters of algorithms are set as the same as those used in *Experiment 1*. (3) Without loss of generality, the SIR-input is set 0 dB, while SNR-input is set 10 dB. (4) We consider two cases.

- Case 1: $\Delta\theta$ is fixed and $\Delta\phi$ is varied, $(\theta_1, \phi_1) = (45^\circ, 0^\circ)$, the DOA of the target speaker moves from $(\theta_s, \phi_s) = (45^\circ, 0^\circ)$ to $(\theta_s, \phi_s) = (45^\circ, 180^\circ)$ with 20° increments. Hence, the angle difference $\Delta\phi$ changes from 0° to 180° with 20° increments. Figure 6 shows the results of Case 1. From Figure 6, it is clear to see that when $\Delta\phi \rightarrow 0^\circ$ (the target speaker moves closer to the interference speaker), for both algorithms, the PESQ drops significantly, and the LSD values are also big. These results indicate that the speech enhancement is very much limited if $\Delta\phi \rightarrow 0^\circ$. However, when $\Delta\phi > 20^\circ$, the PESQ gradually increases, and LSD drops. It is quite encouraging to see that the performance of PESQ and LSD of our proposed AVS-SMASK algorithm is superior to that of the AVS-FMV algorithm for all of the angles. Moreover, our proposed AVS-SMASK algorithm has the absolute advantage when $\Delta\phi \geq 40^\circ$.
- Case 2: $\Delta\phi$ is fixed and $\Delta\theta$ is varied, $(\theta_1, \phi_1) = (10^\circ, 75^\circ)$, the DOA of the target speaker moves from $(\theta_s, \phi_s) = (10^\circ, 75^\circ)$ to $(\theta_s, \phi_s) = (170^\circ, 75^\circ)$ with 20° increments. Then, the angle difference $\Delta\theta$ changes from 0° to 160° with 20° increments. Figure 7 shows the results of Case 2. From Figure 7, we can see that when $\Delta\theta \rightarrow 0^\circ$ (the target speaker moves closer to the interference speaker), for both algorithms, the performance of PESQ and LSD are also poor. This means that the speech enhancement is very much limited if $\Delta\theta \rightarrow 0^\circ$. However, when $\Delta\theta > 20^\circ$, it is quite encouraging to see that the performance of PESQ and LSD of our proposed AVS-SMASK algorithm outperforms that of the AVS-FMV algorithm for all of the angles. In addition, it is noted that the performance of two algorithms drops again when the $\Delta\theta > 140^\circ$ (the target speaker moves closer to the interference speaker around a cone). However, from Figure 6, this phenomenon does not exit.

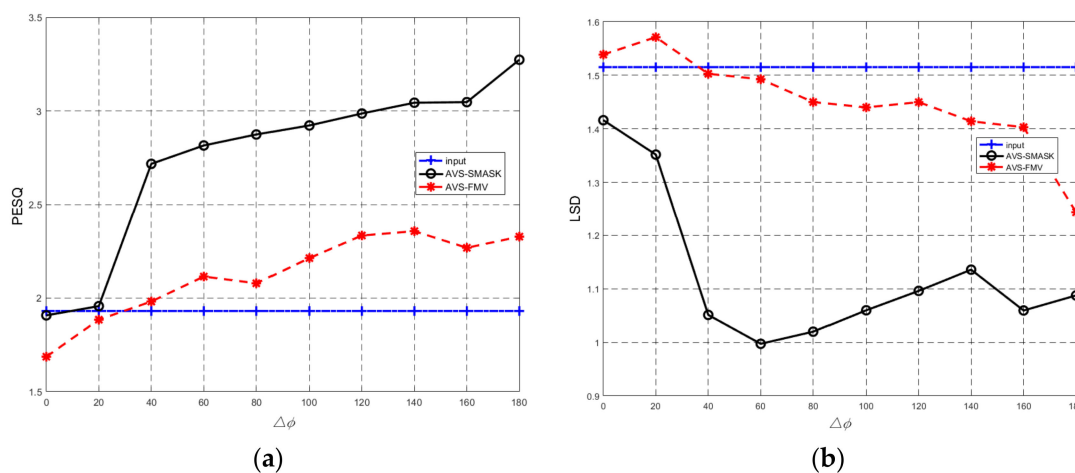


Figure 6. (Experiment 2) The performance versus $\Delta\phi$. (a) Perceptual evaluation of speech quality (PESQ) results and (b) Log spectral division (LSD) results (Case 1: ϕ_s of the target speaker changes from 0° to 180°) (Case 1).

In summary, from the experimental results, it is clear that our proposed AVS-SMASK algorithm is able to enhance the target speech and suppress the interferences when the angle difference between the target speaker and the interference is larger than 20° .

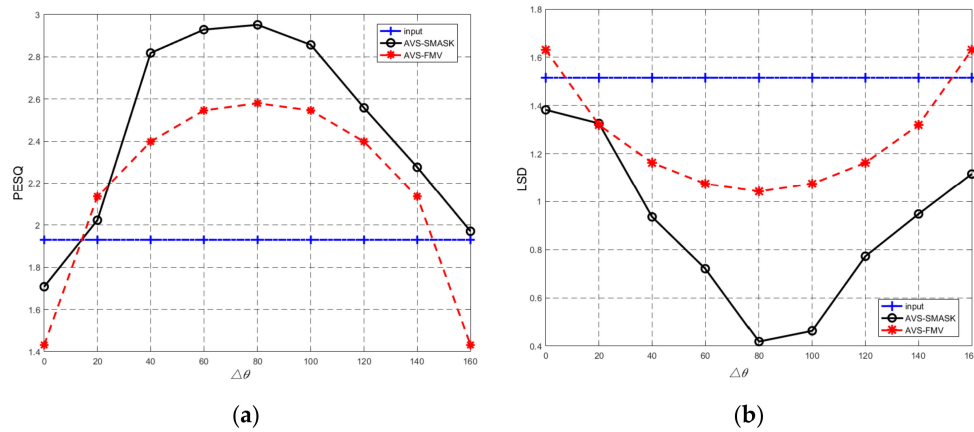


Figure 7. (Experiment 2) The performance versus $\Delta\theta$. (a) PESQ results and (b) LSD results (Case 2: θ_s of the target speaker changes from 0° to 160°).

4.1.3. Experiment 3: The Performance versus DOA Mismatch

In practice, the DOA estimation of the target speaker may be inaccurate or the target speaker may make a small movement that causes the DOA mismatch problem. Hence, this experiment evaluates the impact of the DOA mismatch on the performance of our proposed speech enhancement algorithm. The experimental settings are as follows. (1) The parameters of algorithms are set as same as the *Experiment 1*. (2) $(\theta_s, \phi_s) = (45^\circ, 45^\circ)$ and $(\theta_1, \phi_1) = (90^\circ, 135^\circ)$. (3) The SIR-input is set to 0 dB, while the SNR-input is set to 10 dB; the performance measurement metrics are chosen as SINR and LSD. (4) We consider two cases:

Case 1: Only ϕ_s is mismatched, and the mismatch ($\partial\phi_s$) ranges from 0° to 30° with 5° increments.

Case 2: Only θ_s is mismatched, and the mismatch ($\partial\theta_s$) ranges from 0° to 30° with 5° increments.

Experimental results are given in Figures 8 and 9 for Case 1 and Case 2, respectively. From these results, we can clearly see that when the DOA mismatch is less than 20° , our proposed AVS-SMASK algorithm is not sensitive to DOA mismatch. Besides, our AVS-SMASK algorithm outperforms the AVS-FMV algorithm under all of the conditions. However, when the DOA mismatch is larger than 20° , the performance of our proposed AVS-SMASK algorithm drops significantly. Fortunately, it is easy to achieve 20° DOA estimation accuracy.

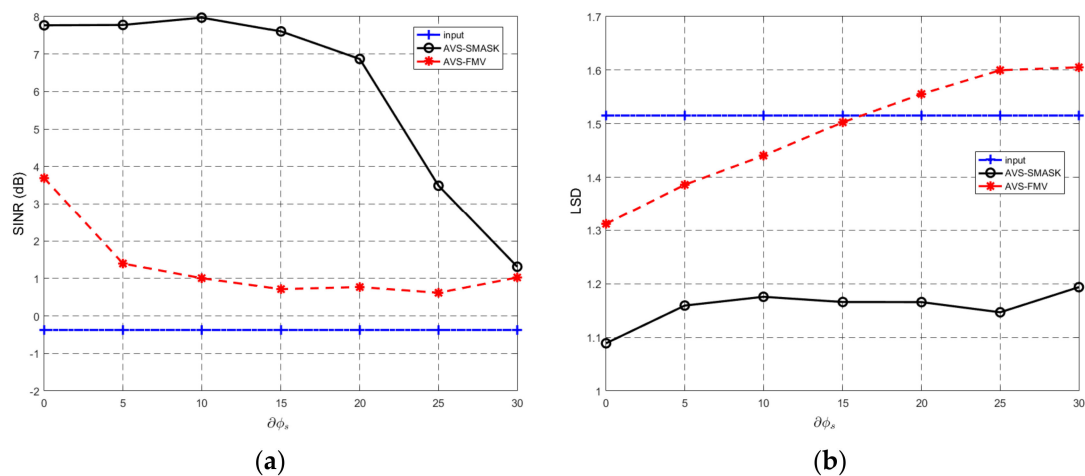


Figure 8. (Experiment 3) The performance versus the $\partial\phi_s$. (a) SINR results and (b) LSD results (Case 1).

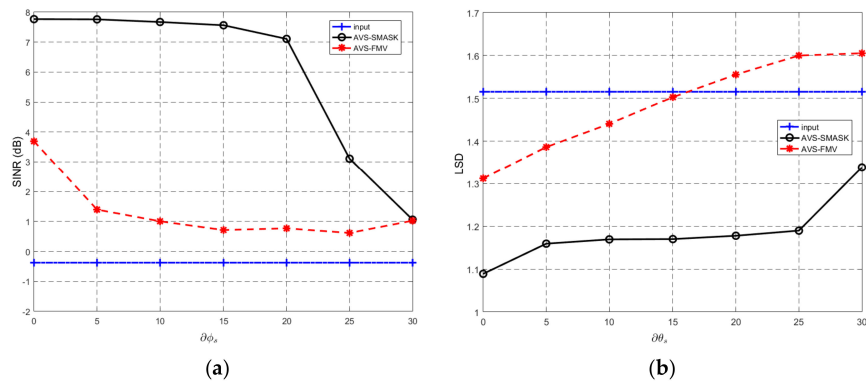


Figure 9. (Experiment 3, Case 2) The performance versus the $\Delta\theta_s$. (a) SINR results and (b) LSD results (Case 2).

4.2. Experiments on Recorded Data in an Anechoic Chamber

In this section, two experiments have been carried out with the recorded data captured by an AVS in an anechoic chamber [25]. Every set of recordings lasts about six seconds, which is made by the situation that the target speech source and the interference source are broadcasting at the same time along with the background noise, as shown in Figure 1. The speech sources taken from the Institute of Electrical and Electronic Engineers (IEEE) speech corpus [26] are placed in the front of the AVS at a distance of one meter, and the SIR-input is set to 0 dB, while the SNR-input is set to 10 dB, and the sampling rate was 48 kHz, and then down-sampled to 16 kHz for processing.

4.2.1. Experiment 4: The Performance versus Angle Difference with Recorded Data

In this experiment, the performance of our proposed method has been evaluated versus the angle difference between the target and interference speakers ($\Delta\phi = \phi_s - \phi_i$ and $\Delta\theta = \theta_s - \theta_i$). The experimental settings are as follows. (1) PESQ is taken as the performance measurement metric. (2) The parameters of algorithms are set as the same as that of *Experiment 1*. (3) Considering page limitation, here, we only consider the changing of azimuth angle ϕ_s while $\theta_s = 90^\circ$. The interfering speaker $s_1(t)$ is at $(\theta_1, \phi_1) = (90^\circ, 45^\circ)$. ϕ_s varies from 0° to 180° with 20° increments. Then, there are 13 recorded datasets. The experimental results are shown in Figure 10. It is noted that the x-axis represents the azimuth angle ϕ_s . It is clear to see that the overall performance of our proposed AVS-SMASK algorithm is superior to that of the comparing algorithm. Specifically, when ϕ_s approaches $\phi_1 = 45^\circ$, the PESQ degrades quickly for both algorithms. When the angle difference $\Delta\phi$ is larger than 30° (ϕ_s is smaller than 15° or larger than 75°), the PESQ of our proposed AVS-SMASK algorithm goes up quickly, and is not sensitive to the angle difference.

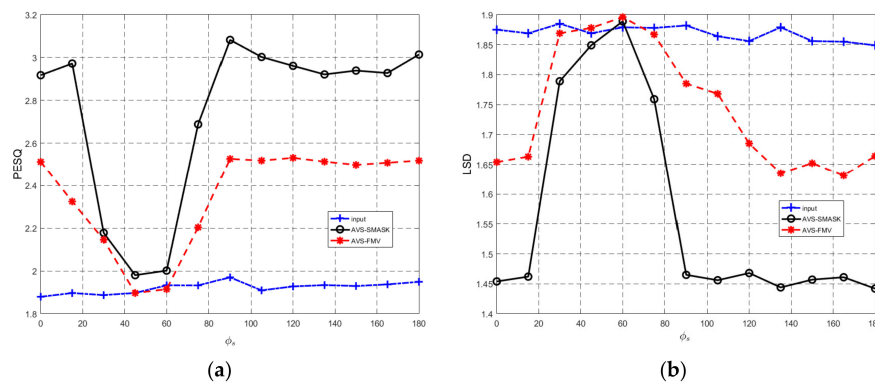


Figure 10. (Experiment 4) The performance versus ϕ_s . (a) PESQ results and (b) LSD results.

4.2.2. Experiment 5: Performance versus DOA Mismatch with Recorded Data

This experiment is carried out to evaluate the performance of speech enhancement algorithms when there are DOA mismatches. The experimental settings are as follows. (1) PESQ and LSD are taken as the performance measurement metric. (2) The parameters of algorithms are set the same as those of *Experiment 1*. (3) The target speaker is at $(\theta_s, \phi_s) = (45^\circ, 45^\circ)$, and the interference speaker is at $(\theta_1, \phi_1) = (90^\circ, 135^\circ)$. The azimuth angle ϕ_s is assumed to be mismatched. We consider the mismatch of ϕ_s (denoted as ϕ_s'') varying from 0° to 30° with 5° increments. The experimental results are shown in Figure 11, where the x-axis is the mismatch of the azimuth angle ϕ_s (ϕ_s''). It is noted that our proposed AVS-SMASK is superior to the compared algorithm under all conditions. It is clear to see that our proposed algorithm is not sensitive to DOA mismatch when the DOA mismatch is smaller than 23° .

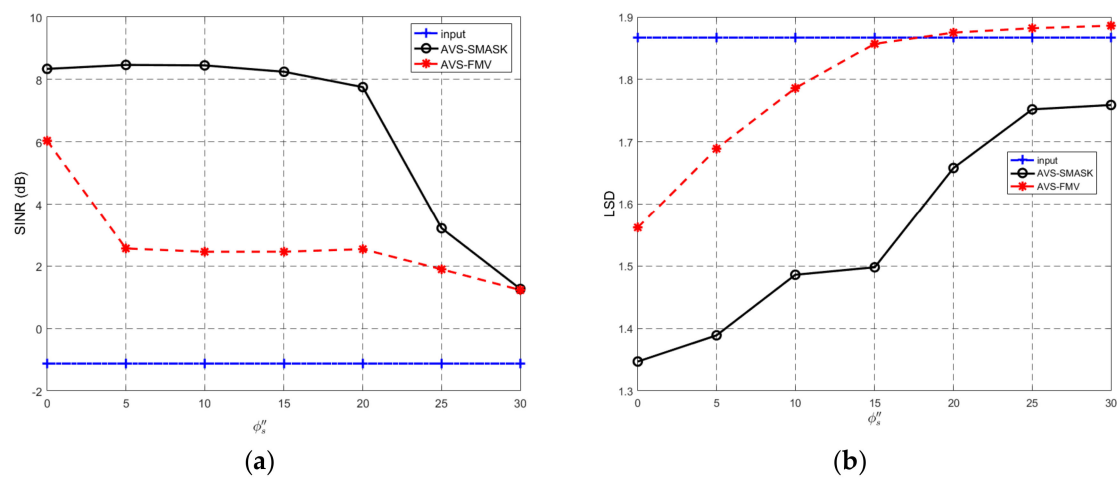


Figure 11. (Experiment 5) The performance versus the ϕ_s mismatch ϕ_s'' . (a) PESQ results and (b) LSD results.

We are encouraged to conclude that our proposed algorithm will offer a good speech enhancement performance in practical applications when the DOA may not be accurately estimated.

5. Conclusions

In this paper, aiming at the hearing technology of service robots, a novel target speech enhancement method has been proposed systematically with a single AVS to suppress spatial multiple interferences and additive background noise simultaneously. By exploiting the AVS signal model and its inter-sensor data ratio (ISDR) model, the desired monotone functions between ISDR and the DOA of the target speaker is derived. Accordingly, a nonlinear soft mask has been designed by making use of speech time-frequency (TF) sparsity with the known DOA of the target speaker. As a result, a single AVS-based speech enhancement method (named as AVS-SMASK) has been formulated and evaluated. Comparing with the existing AVS-FMV algorithm, extensive experimental results using simulated data and recorded data validate the effectiveness of our AVS-SMASK algorithm in suppressing spatial interferences and the additive background noise. It is encouraging to see that our AVS-SMASK algorithm is able to maintain less speech distortion. Due to page limitations, we did not show the derivation of the algorithm under reverberation. The signal model and ISDR model under reverberant conditions will be presented in our paper [27]. Our preliminary experimental results show that the PESQ of our proposed AVS-SMASK degrades gradually when the room reverberation becomes stronger ($RT60 > 400$ ms), but LSD is not sensitive to the room reverberation. Besides, there is an argument that learning-based SE methods achieve the state-of-art. In our opinion, in terms of SNR, PESQ, and LSD, this is true. However, learning-based SE methods ask for large amounts of training data, and require much larger memory size and a high computational cost. In contrast, the application scenarios of this

research are different to learning-based SE methods, and our solution is more suitable for low-cost embedded systems. A real demo system was established in our lab, and the conducted trials further confirmed the effectiveness of our method where room reverberation is moderate ($RT60 < 400$ ms). We are confident that with only four-channel sensors and without any additional training data collected, the subjective and objective performance of our proposed AVS-SMASK is impressive. Our future study will investigate the deep learning-based SE method with a single AVS to improve its generalization and capability to handle different noise and interference conditions.

Author Contributions: Original draft preparation and writing, Y.Z. and Z.L.; Review & Editing, C.H.R., Y.Z. and Z.L. carried out the studies of the DOA estimation and speech enhancement with Acoustic Vector Sensor (AVS), participated in algorithm development, carried out experiments as well as drafted the manuscript. C.H.R. contributed to the design of the experiments, analyzed the experimental results and helped to review and edit the manuscript. All authors read and approved the final manuscript.

Funding: This research was funded by National Natural Science Foundation of China (No: 61271309), Shenzhen Key Lab for Intelligent MM and VR (ZDSYS201703031405467) and the Shenzhen Science & Technology Fundamental Research Program (JCYJ20170817160058246).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, Y.; Song, H.; Liu, J. Service robot speech enhancement method using acoustic micro-sensor array. In Proceedings of the International Conference on Advanced Intelligence and Awareness Internet (IET), Beijing, China, 23–25 October 2010; pp. 412–415.
2. Gomez, R.; Ivanchuk, L.; Nakamura, K.; Mizumoto, T.; Nakadai, K. Utilizing visual cues in robot audition for sound source discrimination in speech-based human-robot communication. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Hamburg, Germany, 28 September–2 October 2015; pp. 4216–4222.
3. Atrash, A.; Kaplow, R.; Villemure, J.; West, R.; Yamani, H.; Pineau, J. Development and validation of a robust speech interface for improved human-robot interaction. *Int. J. Soc. Rob.* **2009**, *1*, 345–356. [[CrossRef](#)]
4. Chen, M.; Wang, L.; Xu, C.; Li, R. A novel approach of system design for dialect speech interaction with NAO robot. In Proceedings of the International Conference on Advanced Robotics (ICAR), Hong Kong, China, 10–12 July 2017; pp. 476–481.
5. Philipos, C.; Loizou, P.C. *Speech Enhancement: Theory and Practice*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2017.
6. Chen, J.; Benesty, J.; Huang, Y. New insights into the noise reduction Wiener filter. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1218–1234. [[CrossRef](#)]
7. Reddy, A.M.; Raj, B. Soft mask methods for single-channel speaker separation. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 1766–1776. [[CrossRef](#)]
8. Lightburn, L.; De Sena, E.; Moore, A.; Naylor, P.A.; Brookes, M. Improving the perceptual quality of ideal binary masked speech. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 661–665.
9. Wang, Z.; Wang, D. Mask Weighted Stft Ratios for Relative Transfer Function Estimation and Its Application to Robust ASR. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
10. Xiao, X.; Zhao, S.; Jones, D.L.; Chng, E.S.; Li, H. On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 3246–3250.
11. Heymann, J.; Drude, L.; Haeb-Umbach, R. Neural network based spectral mask estimation for acoustic beamforming. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 196–200.
12. Li, B.; Zou, Y.X. Improved DOA Estimation with Acoustic Vector Sensor Arrays Using Spatial Sparsity and Subarray Manifold. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 2557–2560.

13. Zou, Y.X.; Shi, W.; Li, B.; Ritz, C.H.; Shujau, M.; Xi, J. Multisource DOA Estimation Based On Time-Frequency Sparsity and Joint Inter-Sensor Data Ratio with Single Acoustic Vector Sensor. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 4011–4015.
14. Zou, Y.X.; Guo, Y.; Zheng, W.; Ritz, C.H.; Xi, J. An effective DOA estimation by exploring the spatial sparse representation of the inter-sensor data ratio model. In Proceedings of the IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP), Xi'an, China, 9–13 July 2014; pp. 42–46.
15. Zou, Y.X.; Guo, Y.; Wang, Y.Q. A robust high-resolution speaker DOA estimation under reverberant environment. In Proceedings of the International Symposium on Chinese Spoken Language Processing (ISCSLP), Singapore, 13–16 December 2014; p. 400.
16. Zou, Y.; Gu, R.; Wang, D.; Jiang, A.; Ritz, C.H. Learning a Robust DOA Estimation Model with Acoustic Vector Sensor Cues, In Proceedings of the Asia-Pacific Signal and Information Processing Association (APSIPA). Kuala Lumpur, Malaysia, 12–15 December 2017.
17. Lockwood, M.E.; Jones, D.L.; Bilger, R.C.; Lansing, C.R.; O'Brien, W.D.; Wheeler, B.C.; Feng, A.S. Performance of Time- and Frequency-domain Binaural Beamformers Based on Recorded Signals from Real Rooms. *J. Acoust. Soc. Am.* **2004**, *115*, 379–391. [[CrossRef](#)] [[PubMed](#)]
18. Lockwood, M.E.; Jones, D.L. Beamformer Performance with Acoustic Vector Sensors in Air. *J. Acoust. Soc. Am.* **2006**, *119*, 608–619. [[CrossRef](#)] [[PubMed](#)]
19. Shujau, M.; Ritz, C.H.; Burnett, I.S. Speech Enhancement via Separation of Sources from Co-located Microphone Recordings. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSAP), Dallas, TX, USA, 14–19 March 2010; pp. 137–140.
20. Wu, P.K.T.; Jin, C.; Kan, A. A Multi-Microphone SPE Algorithm Tested Using Acoustic Vector Sensors. In Proceedings of the International Workshop on Acoustic Echo and Noise Control, Tel-Aviv, Israel, 30 August–2 September 2010.
21. Zou, Y.X.; Wang, P.; Wang, Y.Q.; Ritz, C.H.; Xi, J. Speech enhancement with an acoustic vector sensor: An effective adaptive beamforming and post-filtering approach. *EURASIP J. Audio Speech Music Process* **2014**, *17*. [[CrossRef](#)]
22. Zou, Y.X.; Wang, P.; Wang, Y.Q.; Ritz, C.H.; Xi, J. An effective target speech enhancement with single acoustic vector sensor based on the speech time-frequency sparsity. In Proceedings of the 19th International Conference on Digital Signal Processing (DSP), Hong Kong, China, 20–23 August 2014; pp. 547–551.
23. Gray, R.; Buzo, A.; Gray, A.; Matsuyama, Y. Distortion measures for speech processing. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 367–376. [[CrossRef](#)]
24. ITU-T. 862-Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs; International Telecommunication Union-Telecommunication Standardization Sector (ITU-T): Geneva, Switzerland, 2001.
25. Hirsch, H.G.; Pearce, D. The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In Proceedings of the Automatic Speech Recognition: Challenges for the Next Millennium, Paris, France, 18–20 September 2000; pp. 29–32.
26. Shujau, M.; Ritz, C.H.; Burnett, I.S. Separation of speech sources using an Acoustic Vector Sensor. In Proceedings of the IEEE International Workshop on Multimedia Signal Processing, Hangzhou, China, 19–22 October 2011; pp. 1–6.
27. Rothaus, E.H. IEEE Recommended Practice for Speech Quality Measurements. *IEEE Trans. Audio Electroacoust.* **1969**, *17*, 225–246.

