INSIGHTS INTO MAGNITUDE AND PHASE ESTIMATION BY MASKING AND MAPPING IN DNN-BASED MULTICHANNEL SPEAKER SEPARATION

Alexander Bohlender[1]*, Ann Spriet*[2]*, Wouter Tirry*[2]*, Nilesh Madhu*[1]

[1] IDLab, Department of Electronics and Information Systems, Ghent University - imec, Belgium
[2] Goodix Technology (Belgium) B.V., Leuven, Belgium

ABSTRACT

Speakers are often separated by time-frequency masking in the short-time Fourier domain to take advantage of the high degree of sparsity of the individual speech spectrograms. Magnitude and phase can be jointly enhanced with complex masks, but prior work suggests that directly mapping the input to the complex spectrogram of the clean signal is a better alternative. For a setup with a compact microphone array, experiments conducted in this paper compare these paradigms with focus on magnitude and phase estimation. Whereas phase is enhanced effectively in general, differences between masking and mapping are minor in this regard. Spectral mapping causes the least target distortion. Complex masking better suppresses interference, but speech quality suffers due to artifacts. Combining magnitude masking with phase mapping presents a compromise, which amounts to the best performance regarding instrumental metrics.

Index Terms— multichannel speaker separation, phase-aware, time-frequency masking, complex spectral mapping, neural network

1. INTRODUCTION

Speech enhancement is used to extract clean signals from mixtures captured in noisy environments. When interfering talkers are present, a distinction between desired and undesired speech must be made. Given a microphone array, algorithms can take advantage of spatial information for this, e. g., a beamformer can be steered in the direction of interest [1]. Similarly, information on a target direction [2] or region [3] can be integrated into a deep neural network (DNN).

In the short-time Fourier transform (STFT) domain, the signal is represented in polar notation by its magnitude and phase. Traditionally, the focus was on magnitude estimation [4]. Based on the assumption that there is at most one dominant speaker in each bin [5], time-frequency (TF) masking is considered to be particularly effective for speaker separation. The phase spectrogram, in contrast, is devoid of a clear structure [6], and thus more difficult to estimate, especially when dealing with single-channel signals. In a multichannel setup, the interchannel phase differences depend on the sound source locations even when the phase of each individual channel appears random. This allows for a more substantial phase enhancement.

Various training targets that exhibit a clearer spectral structure were proposed to avoid estimating phase with a DNN directly [6–8]. To jointly enhance magnitude and phase, masking was extended to the complex domain in [6], where real and imaginary parts of a complex ideal ratio mask (cIRM) are estimated. However, its unboundedness makes it a less suitable training target than, e. g., the IRM. Also, upon comparing the two masks, one might argue that the structure of the cIRM is essentially that of the IRM, whereas phase enhancement remains difficult. Other work focuses on phase specifically, but real-valued masks are still used for magnitude enhancement. For example, the observed structure of the time and frequency derivatives of phase is exploited in [7, 8]. However, [9] shows that with an appropriate approach, even when the input is single-channel, estimating (cosine and sine of) the phase itself is possible, too.

The real and imaginary parts of the desired signal can also be estimated directly by a complex spectral mapping (CSM). According to [10], CSM enhances phase quite effectively, and outperforms the complex masking. When applied to microphone arrays, complex spectrograms can be mapped from multiple input channels to one output channel [11, 12]. Nevertheless, using a mask to selectively attenuate bins dominated by noise and interference may still benefit the separation, which motivates us to consider a hybrid approach that combines magnitude masking with phase mapping.

In this paper, we empirically compare masking, mapping, and their combination. Magnitude and phase are evaluated separately to gain a better understanding of the results. To the best of our knowledge, previously only [10] directly compared complex masking and CSM in such a manner for a monaural setup. More accurate phase estimation is possible with multiple microphones, whereby its benefit becomes clearer. Although our experiments verify that the phase enhancement contributes to an improved output signal quality, we find that mapping and masking actually differ *mainly* in terms of the magnitude estimates: (complex) masking gives rise to artifacts, but improves the separation. Instrumental metrics and an audio example indicate that especially the hybrid approach enables a good trade-off between interference suppression and target distortion.

Sec. 2 describes the speaker separation problem, for which we use the DNN-based approach of Sec. 3. Signal reconstruction by masking and mapping, as outlined in Sec. 4, is evaluated in Sec. 5 regarding magnitude and phase estimation. Sec. 6 concludes the paper.

2. SELECTIVE SPEAKER SEPARATION

We use $Y_n(f,t)$ to denote the STFT domain signal captured by the nth microphone of a compact array, where $n \in \{1,...,N\}$, $f \in \{0,...,F-1\}$ is the discrete frequency and t the frame index. This signal is assumed to be a mixture of J concurrent talkers and background noise $V_n(f,t)$, where the speech signals $S^{\mathrm{dr}}_{j,n}(f,t)$ are composed of direct-path and reverberation components, such that $S^{\mathrm{dr}}_{j,n}(f,t) = S^{\mathrm{d}}_{j,n}(f,t) + S^{\mathrm{r}}_{j,n}(f,t)$. This gives us the signal model

$$Y_n(f,t) = \sum_{1\leq j\leq J} S^{\mathrm{dr}}_{j,n}(f,t) \; + V_n(f,t). \tag{1}$$

The microphone index will be omitted when referring to the reference channel. Here, instead of selecting one arbitrary microphone as the reference, for $X \in \{Y, S^{\mathrm{d}}, S^{\mathrm{dr}}\}$ we choose

$$X(f,t) = \tfrac{1}{\sqrt{N}} \left\|\mathbf{X}(f,t)\right\| \mathrm{e}^{\jmath \angle X_1(f,t)}, \tag{2}$$

where $\mathbf{X}(f,t) = [X_1(f,t),...,X_N(f,t)]^T$ and $\|\cdot\|$ is the ℓ_2-norm.

We define the target signal for a desired speaker $j^* \in \{1,...,J\}$:

$$S(f,t) = \gamma_{j^*} S^{\mathrm{d}}_{j^*}(f,t). \tag{3}$$

This work is partially supported by the Research Foundation - Flanders (FWO) under grant numbers 11G0721N and G081420N.

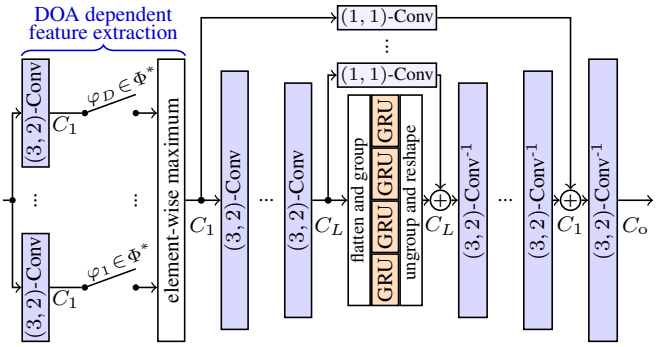

Fig. 1: U-Net based on [14]. Here, we choose $C_1 = 64$, $C_2 = 128$, and $C_\ell = 256$ for $\ell \in \{3, 4, 5\}$. The target speaker is determined by a DOA dependent feature extraction [3] in the first encoder layer.

The factor γ_j is chosen so that $\gamma_j S_j^{\mathrm{d}}(f,t)$ and $S_j^{\mathrm{dr}}(f,t)$ are matched in energy [13]. The purpose of this scaling is to prevent a strong attenuation of the signal at low direct-to-reverberant ratios (DRRs).

3. NEURAL NETWORK ARCHITECTURE

We make use of the network depicted in Fig. 1. Its basis is the convolutional recurrent U-Net for speech enhancement (CRUSE) of [14], which consists of an encoder with L convolution layers and a decoder with L transposed convolution layers. Leaky ReLU is used as activation function. The kernel size is 3 for the frequency and 2 for the time dimension. Convolutions are causal to allow for real-time processing. With a stride of 2, a compression is gradually performed only along the frequency axis while the number of channels grows. At the network bottleneck, features from all subbands are divided into 4 groups that are processed by parallel gated recurrent units (GRUs). Additive skip connections with a learnable scaling and bias are inserted between encoder and decoder. These are implemented as fully grouped (channel-wise) convolution layers.

Deviating from the original CRUSE, we use our location dependent feature extraction (LDE) of [3] to select one particular target speaker based on their (known or estimated) direction of arrival (DOA). Different features are generated from the same input for each target DOA $\varphi_d \in \Phi^*$. These target DOAs are selected out of the predefined discrete grid $\Phi = \{\varphi_1, ..., \varphi_D\}$, such that $\Phi^* \subseteq \Phi$. The angular region defined by Φ^* controls the trade-off between spatial resolution and robustness to uncertainty in the speaker locations. From the feature tensors obtained for all $\varphi_d \in \Phi^*$, the maximum is taken independently for each frequency, frame, and channel. Our results of [3] indicate that an effective DOA conditioning is achieved even when LDE is only applied in the first encoder layer (see Fig. 1), so that the rest of CRUSE can remain as in [14].

The network output consists of C_{o} elements, which are used to obtain an estimate $\widehat{S}(f,t)$. As the different strategies to accomplish this are the focus of this work, they are discussed in more detail in Sec. 4. During training, we process batches comprising 5 signals of 2 s. The error is determined by comparing $S(f,t)$ with $\mathcal{P}\{\widehat{S}(f,t)\}$, where $\mathcal{P}\{\cdot\} = \mathrm{STFT}\{\mathrm{ISTFT}\{\cdot\}\}$ is applied for STFT consistency [15]. The loss function

$$\begin{aligned}\mathcal{L} = &\tfrac{1}{2}\,\mathrm{MSE}\left[|S(f,t)|^c, |\mathcal{P}\{\widehat{S}(f,t)\}|^c\right] + \\ &\tfrac{1}{2}\,\mathrm{MSE}\left[|S(f,t)|^c \mathrm{e}^{\jmath \angle S(f,t)}, |\mathcal{P}\{\widehat{S}(f,t)\}|^c \mathrm{e}^{\jmath \angle \mathcal{P}\{\widehat{S}(f,t)\}}\right]\end{aligned} \tag{4}$$

is the same as that of [14], but with equally weighted magnitude and phase-aware terms. In the above, magnitude is compressed with $c = 0.3$, and $\mathrm{MSE}[\cdot]$ returns the mean square error between the two arguments. The learning rate of the AdamW optimizer is set to $8 \cdot 10^{-5}$ and the weight decay to 0.1.

4. MAGNITUDE AND PHASE ESTIMATION

Real-valued masks $\mathcal{M}(f,t) \in [0,1]$ are often used to attenuate bins dominated by noise, interference, and reverberation while preserving bins dominated by the target signal. The resulting output is given by

$$\widehat{S}(f,t) = \mathcal{M}(f,t) \cdot Y(f,t). \tag{5}$$

This only has an effect on the magnitude. Phase can be enhanced as well with a complex mask, or by direct estimation of the clean signal.

4.1. Complex mask estimation (CME)

As the cIRM is an unbounded mask, [6] proposed to compress its real and imaginary parts by a hyperbolic tangent. With a signal approximation loss like (4), however, no explicit definition of a target mask is necessary. Instead, we decompress the $C_{\mathrm{o}} = 2$ network outputs $\mathcal{O}_\Re(f,t)$ and $\mathcal{O}_\Im(f,t)$ with the corresponding inverse function:

$$\mathcal{M}(f,t) = \frac{1}{C}\left[\log\frac{K + \mathcal{O}_\Re(f,t)}{K - \mathcal{O}_\Re(f,t)} + \jmath \log\frac{K + \mathcal{O}_\Im(f,t)}{K - \mathcal{O}_\Im(f,t)}\right]. \tag{6}$$

The hyperparameters are set to $K = 10$ and $C = 0.1$ [6].

As in [3], we use input features $\mathbf{Z}_{\mathrm{CME}}(f,t) \in \mathbb{R}^{2N+1}$ with

$$\mathbf{Z}_{\mathrm{CME}}(f,t) = \begin{bmatrix} \Re\{\mathbf{Y}(f,t)\}/\|\mathbf{Y}(f,t)\| \\ \Im\{\mathbf{Y}(f,t)\}/\|\mathbf{Y}(f,t)\| \\ \log|Y(f,t)| - E\{\log|Y(f,t)|\} \end{bmatrix}. \tag{7}$$

The first $2N$ entries capture the noisy phases and normalized amplitudes (spatial information). The last entry represents the spectral information in the form of the log-magnitude. The expected value $E\{\cdot\}$ is subtracted for insensitivity to scaling of the input signals. It is approximated by averaging over all bins within the last 0.3 s.

4.2. Complex spectral mapping (CSM)

The idea of CSM is to map the complex input spectrograms directly to the clean spectrogram of the target signal. Experiments of [10, 16] indicate that CSM outperforms CME, no matter whether the cIRM is estimated or a signal approximation loss is used.

Whereas a mask-based approach can simply identify the dominant signal component to assign a weight to each bin, CSM requires knowledge of the complex spectrograms. To obtain a valid input, the normalization by $\|\mathbf{Y}(f,t)\|$ in (7) is therefore omitted. Instead, to maintain scale invariance, the signals are rescaled to unit root mean square with $\alpha(t) = \sqrt{E\{|Y(f,t)|^2\}}$. This yields as CSM input

$$\mathbf{Z}_{\mathrm{CSM}}(f,t) = \begin{bmatrix} \Re\{\mathbf{Y}(f,t)\}/\alpha(t) \\ \Im\{\mathbf{Y}(f,t)\}/\alpha(t) \\ \log|Y(f,t)| - E\{\log|Y(f,t)|\} \end{bmatrix}. \tag{8}$$

The last (log-magnitude) entry is kept for a fair comparison between CME and CSM. The correct scaling is restored in the end, such that

$$\widehat{S}(f,t) = \alpha(t)\left(\mathcal{O}_\Re(f,t) + \jmath\mathcal{O}_\Im(f,t)\right). \tag{9}$$

Again, $\mathcal{O}_\Re(f,t)$ and $\mathcal{O}_\Im(f,t)$ are the $C_{\mathrm{o}} = 2$ network outputs.

4.3. Hybrid phase mapping and magnitude masking

Estimating phase either indirectly [7, 8] or directly [9] leaves the option to use a real-valued mask for magnitude enhancement. Here, such a hybrid phase mapping and magnitude masking is therefore considered as a third alternative. We can motivate this approach with the efficacy of masks in speech separation due to the forced assignment of a weight to each TF bin, even when CSM achieves better joint magnitude and phase estimation than CME [10].

Concretely, we now use the DNN to obtain $C_{\mathrm{o}} = 3$ elements: $\mathcal{O}_\mathcal{M}(f,t)$, $\mathcal{O}_\Re(f,t)$ and $\mathcal{O}_\Im(f,t)$. For the real-valued masks, instead of (6), we choose an output nonlinearity

$$\mathcal{M}(f,t) = 10^{\mathcal{O}_\mathcal{M}(f,t)}, \tag{10}$$

whereby $\mathcal{O}_\mathcal{M}(f,t)$ is interpreted as a log-mask [3] so as to better capture values $\mathcal{M}(f,t) \ll 1$ for a strong interference suppression.

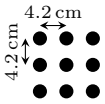

Fig. 2: Experiments are performed with a URA ($N = 9$).

Additionally, we confine the masks to $\mathcal{M}(f,t) \in [0.01, 4]$. Values above 1 may be needed especially at low DRRs due to the use of the amplified direct-path signal as the target according to (3).

From the two remaining outputs, we construct the phase estimate

$$\widehat{\phi}(f,t) = \arctan2\left[\mathcal{O}_{\Im}(f,t), \mathcal{O}_{\Re}(f,t)\right]. \tag{11}$$

Thus, $\mathcal{O}_{\Re}(f,t)$ and $\mathcal{O}_{\Im}(f,t)$ can be seen as estimates of $\beta\cos(\angle S(f,t))$ and $\beta\sin(\angle S(f,t))$, respectively, where β is arbitrary. Finally, the output signal is computed with

$$\widehat{S}(f,t) = \mathcal{M}(f,t) \cdot |Y(f,t)|\, \mathrm{e}^{\jmath\widehat{\phi}(f,t)}. \tag{12}$$

As input for the hybrid method, we select $\mathbf{Z}_{\mathrm{hyb}}(f,t) = \mathbf{Z}_{\mathrm{CME}}(f,t)$.

5. EVALUATION

As the main contribution of this work, we experimentally compare complex masking, mapping, and their combination with regard to magnitude and phase estimation in a multichannel setup.

Therefore, similar to the comparison of complex masking and mapping for the use case of monaural speech enhancement in [10], scores are computed for different pairs of noisy (input), DNN-enhanced, and clean magnitude $|X'(f,t)|$ and phase $\angle X''(f,t)$, where $X', X'' \in \{Y, \mathcal{P}\{\widehat{S}\}, S\}$. For example, $\mathrm{ISTFT}\{|\mathcal{P}\{\widehat{S}(f,t)\}|\, \mathrm{e}^{\jmath\angle S(f,t)}\}$ is considered to evaluate the *estimated* magnitude paired with the *clean* phase. Note that enhancing either the magnitude *or* the phase of $\widehat{S}$ has an effect on magnitude *and* phase of $\mathcal{P}\{\widehat{S}\}$. Thus, by considering $\mathcal{P}\{\widehat{S}\}$ instead of $\widehat{S}$, we evaluate the *actually realized* magnitude and phase enhancement. A study based on $\widehat{S}$ may lead to slightly different conclusions.

The time-invariant mask-based minimum variance distortionless response (MVDR) beamformer [17] provides another reference for enhanced magnitude and phase. The speech power spectral density (PSD) matrix is estimated in a noncausal manner with

$$\mathbf{\Psi}_{SS}(f) = \sum_{t} |\mathcal{M}(f,t)|\, \mathbf{Y}(f,t)\mathbf{Y}^{H}(f,t), \tag{13}$$

where $\mathcal{P}\{\widehat{S}(f,t)\}/Y(f,t)$ is used as the mask $\mathcal{M}(f,t)$ in the case of CSM. Mask values are clipped at a maximum of 1. For the noise-plus-interference PSD, $\mathcal{M}(f,t)$ is replaced by $1 - \mathcal{M}(f,t)$ in (13).

The considered uniform rectangular array (URA) geometry with $N = 9$ microphones is depicted in Fig. 2. To define the grid of discrete DOAs for LDE, we consider only the azimuth angle with a resolution of 5°, which yields $\Phi = \{0°, 5°, ..., 355°\}$ ($D = 72$).

The sampling rate is set to $f_s = 16\,\mathrm{kHz}$. For the STFT, we use a 512-point discrete Fourier transform (DFT). Frames are 512 samples (32 ms) long with a hop size of 160 samples (10 ms). Square-root von Hann is used as the analysis and synthesis window.

5.1. Training setup

Our training setup is as described in [3]. Signals are generated online according to the mixture model of (1). Clean speech from the TIMIT [18] and PTDB-TUG [19] databases (7 430 training and 2 280 validation utterances) is convolved with room impulse responses (RIRs) simulated using [20] for various room dimensions, reverberation times, and array positions. The DOA (from the grid Φ defined above) and the distance from the array are selected randomly for each of the $J \in \{1, 2\}$ sources. Spectrally white, diffuse noise (spherically isotropic noise field) generated as described in [21] is mixed with the speech signals at a signal-to-noise ratio (SNR) drawn from a uniform distribution $\mathcal{U}(0\,\mathrm{dB}, 30\,\mathrm{dB})$.

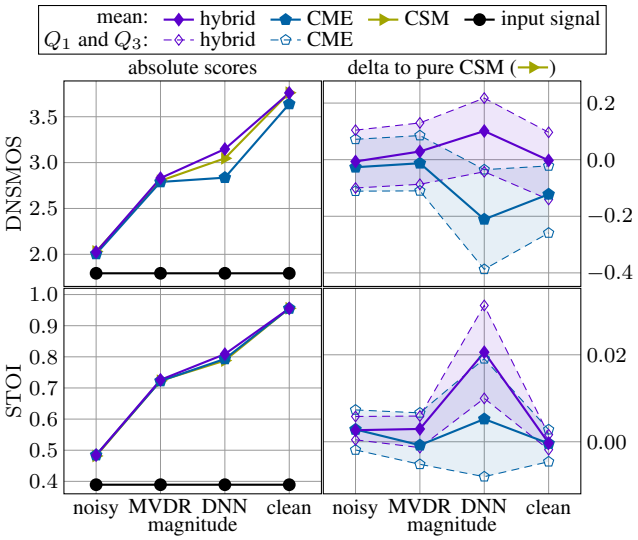

Fig. 3: The DNN-estimated *phase* $\angle\mathcal{P}\{\widehat{S}\}$ is paired with different *magnitudes* (as indicated on the x-axis). Based on the results with a fixed (noisy or clean) magnitude, we find that CSM and hybrid yield phase estimates of comparable quality, but CME also comes close.

During training, the choice of the target DOAs Φ^* is randomized under the following constraints: (i) Φ^* describes a contiguous angular range comprising *at least* one speaker, (ii) the extent of this region can vary, but small widths are more common. When more than one source DOA is inside Φ^*, the target signal is a mixture of multiple speakers. This general case is discussed in more detail in [3].

5.2. Experimental setup

To create realistic test conditions, we recorded RIRs in a meeting room (approximately $7.50 \times 5.00 \times 2.65\,\mathrm{m}^3$), for which we measured a reverberation time of 0.66 s. A loudspeaker was placed at azimuth angles $\varphi \in \{0°, 20°, ..., 180°\}$ in one of two different distances (1 m, 2 m) from the array. As additive noise, we re-recorded the pub noise of [22] under diffuse conditions in a reverberant room.

The reverberant signals of $J = 2$ concurrent talkers and the noise are additively mixed. For the source signals, we concatenate 5 utterances of one random speaker from the TSP speech database [23]. The time-invariant DOAs are selected randomly (but uniquely for each speaker) out of the available recordings. As the speech signals are not rescaled, the source-to-interference ratio is around 0 dB, but may vary depending on source signals and RIRs.

Our findings of [3] indicate that a suitable choice for LDE is $\Phi^* = \{\varphi \in \Phi : |\varphi_{\mathrm{c}} - \varphi| \leq 10°\}$, where the center direction $\varphi_{\mathrm{c}} \in \Phi$ is the DOA of the target speaker (here assumed known). Speakers are extracted sequentially, each time choosing a different φ_{c}.

We consider the DNSMOS P.835 [24] overall score and extended STOI [25] as speech quality and intelligibility measures, respectively. By selecting the *personalized* version, DNSMOS accounts for interfering speakers although the metric is nonintrusive. 25 experiments are conducted for 2 SNRs (5 dB, 30 dB) and 2 source-array distances (1 m, 2 m), i. e., 100 experiments in total. The SNR is defined as the ratio of all reverberant speech to additive noise.

5.3. Magnitude and phase estimation performance

For the results in Fig. 3, we paired the phase estimated by the DNN ($X'' = \mathcal{P}\{\widehat{S}\}$) with either of the four different options for the magnitude. In Fig. 4, in contrast, the magnitude estimated by the DNN is used in all cases ($X' = \mathcal{P}\{\widehat{S}\}$), but the phase varies. Estimated magnitude and phase are always obtained with one configuration:

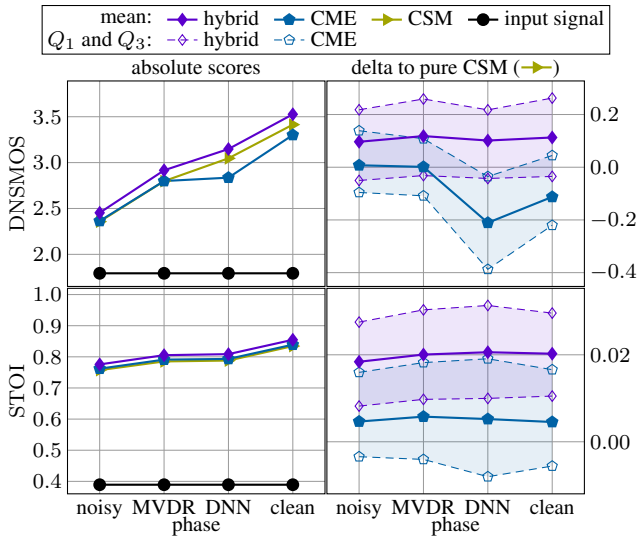

Fig. 4: The DNN-estimated *magnitude* $|\mathcal{P}\{\widehat{S}\}|$ is paired with different *phases*. For estimating magnitude, hybrid yields the best scores.

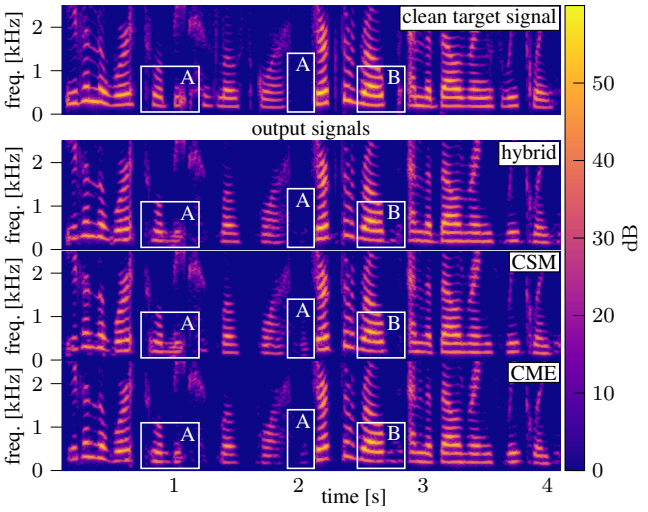

Fig. 5: Separating a mixture (not shown) of two talkers and noise (5 dB SNR). The (complex) masking reduces interference leakage (A), but causes more prominent speech-shaped artifacts (B).

we *never* pair, e. g., CME magnitude with CSM phase. As a reference, the scores obtained with noisy magnitude *and* phase are also included in both figures (—●—). Scores are aggregated over all 100 experiments. In addition to the mean absolute values (left), the difference compared to CSM is shown on the right, where values greater than 0 indicate an improvement. To measure how consistent the results are, first and third quartile (Q_1 and Q_3) of the 100 experiments are indicated in the difference plots as well. Below, we first discuss how well magnitude and phase are estimated in general, before looking at the differences between masking and mapping more closely.

Magnitude estimation efficacy: We consider Fig. 3 with regard to how the scores change along the x-axis (variable magnitude, fixed phase). Unsurprisingly, enhancing magnitude improves both quality and intelligibility significantly. Compared to the MVDR beamformer, the estimate of the DNN raises the scores by another 0.3 in terms of DNSMOS (except CME) and almost 0.1 in terms of STOI.

Phase estimation efficacy: We consider Fig. 4 with regard to how the scores change along the x-axis (fixed magnitude, variable phase). Clearly, the DNN-based phase estimation is effective: although we find that DNSMOS tends to overestimate the benefit of enhancing phase (e. g., hybrid (—◆—): gain from 2.5 to 3.2), even STOI increases by about 0.03 over the noisy phase score (hybrid: 0.78 to 0.81). The improvement over the beamformer's phase is more subtle: 0.2 in DNSMOS (except CME), less than 0.01 in STOI.

Comparing CME, CSM, and hybrid phase estimates: To evaluate phase independently of magnitude estimation, we consider only the *noisy* and *clean* magnitude results in Fig. 3. As CSM (—►—) and the hybrid approach both estimate phase by mapping, they perform comparably. CME (—⬟—) also achieves similar scores (except DNSMOS for the clean magnitude), which suggests that complex masking can, in fact, enhance phase (almost) as affectively as CSM.

Comparing CME, CSM, and hybrid magnitude estimates: To evaluate magnitude independently of phase estimation, we consider only the *noisy* and *clean* phase results in Fig. 4. CME performs *slightly* better than CSM in terms of STOI, but worse in terms of DNSMOS (based on the clean phase results). An explanation may be that the mask-based weighting of bins according to the dominance of desired and undesired signals is an effective tool for interference suppression, but can give rise to an increased target distortion. The hybrid approach achieves the best performance with an average improvement over CSM of 0.1 in DNSMOS and 0.02 in STOI.

5.4. Qualitative comparison

To understand how the outputs generated by masking and mapping differ, we now consider an example. A female speaker is located at azimuth $\varphi = 160°$ and a male speaker at $\varphi = 100°$, both in a distance of 2 m from the array. Additive noise is present in the mixture with an SNR of 5 dB. Spectrograms of clean speech and output are presented in Fig. 5 for the female speaker. For clarity, only the first seconds of the signal at frequencies up to 2.5 kHz are shown. Here, we focus on the unaltered results of (5), (9) and (12), but audio files for different magnitude and phase pairs are also available at `https://aspire.ugent.be/demos/HSCMA2024AB/`.

As expected based on the numerical results of Sec. 5.3, the CME, CSM, and hybrid results are *generally similar*, but some differences can be observed upon closer inspection. Residual interference (however distorted) is sometimes still noticeable especially in the case of CSM. Comparison with CME and hybrid indicates that the two talkers indeed tend to be more effectively separated by TF masking. In Fig. 5, this is seen in the amount of leakage from the male speaker in the different outputs at times 1.0 s and 2.0 s (boxes labeled "A").

However, we also observe that the mask-based approach is more prone to continue emphasizing certain frequencies during inactivity of the target speaker. Such hallucination of speech-like patterns gives rise to audible artifacts. This can be seen particularly well when comparing the clean signal with the CME output at a time of 2.8 s (box labeled "B"). The problem can also be observed in the output of the hybrid approach, but to a lesser extent.

6. CONCLUSIONS

This paper examined masking, spectral mapping, and their combination regarding magnitude and phase enhancement. Although the phase spectrogram lacks structure, in the case of localized sources the resulting relation between the phases of different microphones facilitates its estimation in a setup with a compact array. Comparison with the output of an MVDR beamformer confirms that the DNN enhances phase effectively, but masking and mapping yield comparable results. Masking reduces interference leakage, which can improve intelligibility, but artifacts related to the hallucination of speech during absence of the target signal deteriorate the speech quality. With a hybrid magnitude masking and phase mapping, the two can be combined to take advantage of the strengths of both paradigms.

7. REFERENCES

[1] P. Vary and R. Martin, *Digital Speech Transmission - Enhancement, Coding & Error Concealment*, John Wiley & Sons, Ltd., Chichester, West Sussex, U.K., Jan. 2006.

[2] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 558–565.

[3] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Spatially selective speaker separation using a DNN with a location dependent feature extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 930–945, 2024.

[4] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[5] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, vol. 1, pp. I–529–I–532.

[6] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[7] N. Zheng and X.-L. Zhang, "Phase-aware speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 63–76, 2019.

[8] M. Hasannezhad, H. Yu, W.-P. Zhu, and B. Champagne, "PACDNN: A phase-aware composite deep neural network for speech enhancement," *Speech Communication*, vol. 136, pp. 1–13, 2022.

[9] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network," in *Proc. AAAI Conference on Artificial Intelligence*, 2020, pp. 9458–9465.

[10] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2020.

[11] Z.-Q. Wang and D. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 486–490.

[12] K. Tan, Z.-Q. Wang, and D. Wang, "Neural spectrospatial filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 605–621, 2022.

[13] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu, "Neural networks using full-band and subband spatial features for mask based source separation," in *Proc. 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 346–350.

[14] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 656–660.

[15] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 900–904.

[16] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[17] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, D. N. L., and Z. V., "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1," *Linguistic Data Consortium*, 1993.

[19] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proc. 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011, pp. 1509–1512.

[20] E. A. P. Habets, "RIR generator," `https://github.com/ehabets/RIR-Generator`, Accessed: December 13, 2023.

[21] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

[22] European Telecommunications Standards Institute, "Speech processing, transmission and quality aspects (STQ); speech quality performance in the presence of background noise; part 1: Background noise simulation technique and background noise database," ETSI EG 202 396-1, 2005.

[23] P. Kabal, "TSP speech database," Tech. Rep., McGill University, Montreal, Quebec, Canada, 2002.

[24] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 886–890.

[25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.