# Investigating Imaginary Mask Estimation in Complex Masking for Speech Enhancement

Georgiana-Elena Sfeclis
*School of Computing Sciences*
*University of East Anglia*
Norwich, UK
G.Sfeclis@uea.ac.uk

Ben Milner
*School of Computing Sciences*
*University of East Anglia*
Norwich, UK
B.Milner@uea.ac.uk

Danny Websdale
*School of Computing Sciences*
*University of East Anglia*
Norwich, UK
D.Websdale@uea.ac.uk

***Abstract*—Speech enhancement methods have become effective at estimating a clean magnitude spectrum from a noisy speech signal. However, they are much less effective at recovering the noise-free phase. At higher signal-to-noise ratios (SNRs) this is unimportant, but at lower SNRs the noisy phase introduces a perceptible distortion to the enhanced speech that reduces quality and intelligibility. Complex masking methods have addressed this problem to some extent but they report underestimation of the imaginary mask which in turn limits the possible phase correction. This work first analyses the problem of imaginary mask estimation and examines further its effect on both phase and magnitude masking. Second, a CNN-DNN architecture is proposed for complex mask estimation that uses a new loss function aimed at giving errors in the imaginary mask component a greater contribution in model training. Experimental results are presented that consider variations to the loss function and demonstrate that improved speech quality and intelligibility can be achieved.**

***Index Terms*—Speech enhancement, complex masking, loss functions**

## I. Introduction

In most real-world speech processing applications, background noise is present and causes a reduction of speech quality and intelligibility, and, when used as input into a speech recogniser, a lower accuracy. Methods to enhance the noisy speech can be applied to a broad range of applications including telephony, hearing aids and robust speech recognition [1], [2]. Traditional methods of speech enhancement include spectral subtraction, Wiener filtering, minimum mean-square error estimation (MMSE) and masking which derive filters that are used to suppress noise [1], [3], [4]. More recent methods use deep learning architectures to learn a relationship typically between the input noisy speech spectrum and a set of target features. These target features may be clean spectral features that allow an enhanced speech signal to be recovered directly [5]–[7], or they may be a mask that is subsequently applied to the noisy speech to produce an enhanced signal [4], [8].

Masking has been shown to be highly effective and capable of outperforming systems that map the noisy speech onto a clean speech estimate [4], [9]. The most common approaches use binary masking or ratio masking to enhance the magnitude spectrum of the noisy speech, which is then combined with the noisy phase and returned to the time-domain for reconstruction. Current approaches to masking use various deep learning models to estimate the mask from noisy speech and include feedforward, recurrent and convolutional neural networks [10], [11]. At higher signal-to-noise ratios (SNRs) magnitude masking methods are effective, but at lower SNRs (typically below 7dB) the uncompensated phase introduces perceptible distortions that reduce quality and intelligibility [12]. Complex masks have been proposed to enhance both the magnitude and phase of the noisy speech signal [13]. However, a common issue with their estimation is that, while estimates of the real mask are relatively good, the imaginary mask is often underestimated. This is attributed to less structure being present in the ideal imaginary mask compared to the real mask. As this forms the target for model training, the lack of structure makes learning difficult [14]–[17].

The aim of this work is to improve estimation of the imaginary component of complex masks and thereby reduce phase distortion in the enhanced audio. An analysis of complex masks is made that examines how underestimation of the imaginary mask affects the magnitude and phase of the enhanced signal. A modification to the mean squared error (MSE) loss function is proposed that enhances the contribution that the imaginary mask has within a CNN-DNN architecture for complex mask estimation. This is motivated by several studies that have shown how modified loss functions have given improved estimation across a range of audio processing applications [18], [19]. The remainder of this paper begins in Section II with the analysis into the effect of the imaginary mask within complex masking. Section III introduces the proposed loss function and architecture for estimating the complex mask. Experimental results are presented in Section IV and a conclusion made in Section V.

## II. Analysis of complex masking framework

A noisy speech signal, $y(n)$, can be considered to be the sum of clean speech, $x(n)$, and additive noise, $d(n)$, where $n$ is the discrete-time sample. Transforming this into the spectral domain, the noisy speech signal is

$$Y(f,t) = X(f,t) + D(f,t) \qquad (1)$$

where $Y(f,t)$, $X(f,t)$ and $D(f,t)$ are the noisy speech, clean speech and noise complex spectra at frequency bin $f$ and time frame $t$. Conventional masking methods, e.g. [4], estimate a magnitude mask, $M^{MAG}(f,t)$, and multiply this by the noisy speech magnitude to estimate the clean speech magnitude spectrum, $|\widehat{X}(f,t)|$,

$$|\widehat{X}(f,t)| = |Y(f,t)|M^{MAG}(f,t) \tag{2}$$

No attempt is made to correct the phase, with the noisy phase used in the inverse Fourier transform to return the signal to the time-domain. At higher SNRs, this phase error is not significant but at lower SNRs, using the noisy phase introduces perceptible distortion into the enhanced speech [12].

### A. *Complex masking*

To avoid phase distortions, complex masking has been proposed as an alternative solution that enhances both the magnitude and phase of noisy speech [13],

$$\widehat{X}(f,t) = Y(f,t)M(f,t) \tag{3}$$

Two approaches to estimate the complex mask can be considered, estimating real and imaginary masks, $M_r(f,t) = \Re\{M(f,t)\}$ and $M_i(f,t) = \Im\{M(f,t)\}$, or estimating magnitude and phase masks, $|M(f,t)|$ and $\theta_M(f,t)$. Signals from the two approaches are illustrated in Figure 1 which first shows a clean speech spectrogram, taken from male speaker *s6* in the GRID database [20], then contaminated with white noise at an SNR of 0dB. Figures 1c and 1d show the corresponding ideal magnitude and phase masks. These illustrate a clear structure in the magnitude mask, while the phase mask has an almost random structure, except where the speech energy is high and consequently the ideal phase mask is close to zero as little phase compensation is necessary. Figures 1e and 1f show the ideal real and imaginary masks. The structure in the real mask is clear and follows closely that of the magnitude mask. Conversely, the structure is less clear and of lower amplitude in the imaginary mask. In terms of complex mask estimation, the unstructured phase mask has led most methods to favour real and imaginary mask estimation as,

$$M(f,t) = M_r(f,t) + jM_i(f,t) \tag{4}$$

where the $M_r(f,t)$ and $M_i(f,t)$ are the real and imaginary masks and $j = \sqrt{-1}$. However, studies have reported difficulties in estimating the imaginary mask due to its lack of structure and low amplitude in comparison to the real mask [14]–[16].

### B. *Impact of underestimating the imaginary mask*

The contribution of the real and imaginary masks in enhancement can be considered by representing the masking process of (3) in complex exponential form,

$$\begin{aligned}|\hat{X}(f,t)|e^{j\theta_{\hat{X}}(f,t)} &= |Y(f,t)|e^{j\theta_Y(f,t)}|M(f,t)|e^{j\theta_M(f,t)} \\ &= |Y(f,t)||M(f,t)|e^{j(\theta_Y(f,t)+\theta_M(f,t))}\end{aligned} \tag{5}$$

where the magnitude and phase of the mask are calculated as,

$$|M(f,t)| = \sqrt{\Re\{M(f,t)\}^2 + \Im\{M(f,t)\}^2} \tag{6}$$

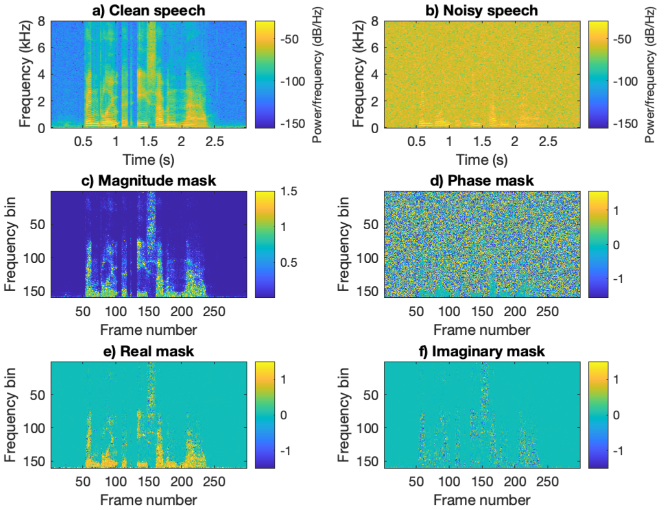

Fig. 1. *Clean and noisy speech spectrograms (panels a) and b)), the resulting ideal magnitude and phase masks (panels c) and d)), and real and imaginary masks (panels e) and f)).*

and

$$\theta_M(f,t) = \tan^{-1}\left(\frac{\Im\{M(f,t)\}}{\Re\{M(f,t)\}}\right) \tag{7}$$

In the situations where the estimated imaginary mask is underestimated and tends toward zero, this causes the magnitude and phase masks to also be underestimated,

$$\lim_{\Im\{M(f,t)\}\to 0} |M(f,t)| = |\Re\{M(f,t)\}| \tag{8}$$

and

$$\lim_{\Im\{M(f,t)\}\to 0} \theta_M(f,t) = 0 \tag{9}$$

When these are considered within (5), the magnitude mask is underestimated and consequently over-attenuates the noisy signal. The phase mask has minor effect that results in the estimate of the clean speech spectrum having a phase approaching that of the noisy speech phase,

$$|\hat{X}(f,t)|e^{j\theta_{\hat{X}}(f,t)} = |Y(f,t)||\Re\{M(f,t)\}|e^{j\theta_Y(f,t)} \tag{10}$$

This underestimation of the imaginary mask is the primary reason for complex masking methods to perform only marginally better, or even worse, than conventional magnitude masking.

### C. *Analysis of the imaginary mask*

To investigate why the imaginary mask is underestimated, the ideal real and imaginary masks in (4) can be expressed in terms of the real and imaginary clean speech and noisy spectra, $X_r$, $X_i$, $Y_r$ and $Y_i$, where, to simplify notation, the $f$ and $t$ subscripts are not shown,

$$M = \frac{Y_rX_r + Y_iX_i}{Y_r^2 + Y_i^2} + j\frac{Y_rX_i + Y_iX_r}{Y_r^2 + Y_i^2} \tag{11}$$

Combining (1) and (11), the real and imaginary masks can be expressed in terms of clean speech and noise spectra,

$$M_r = \frac{X_r^2 + X_i^2 + D_rX_r + D_iX_i}{X_r^2 + D_r^2 + 2X_rD_r + X_i^2 + D_i^2 + 2X_iD_i} \tag{12}$$

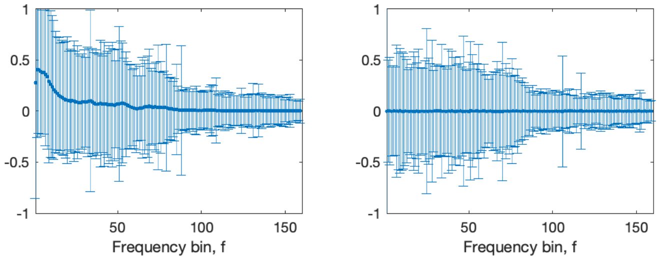

Fig. 2. *Mean and range of values within one standard deviation of the mean for a) real and b) imaginary masks, shown across each frequency bin.*

$$M_i = \frac{X_i D_r - X_r D_i}{X_r^2 + D_r^2 + 2X_r D_r + X_i^2 + D_i^2 + 2X_i D_i} \tag{13}$$

where $X_r$, $X_i$, $D_r$ and $D_i$ are all zero mean signals.

To examine the distribution of mask amplitudes, Figure 2 shows the mean and range of values within one standard deviation for the real and imaginary masks across each of the $F$=160 frequency bins. These are calculated from the 1000 utterances spoken by speaker *s6* in the GRID database, each contaminated with white noise at an SNR of 0dB, and including both speech and non-speech regions. The mean of the real mask is all positive and has higher amplitude at lower frequency bins. This corresponds to generally higher levels of speech energy being present and consequently higher local SNR that requires less mask attenuation (i.e. higher mask amplitudes). Conversely, the mean of the imaginary mask is very close to zero across all frequency bins, irrespective of the speech energy. In terms of standard deviation, the real and imaginary masks have similar values which are higher at lower frequencies that have greater speech energy. These different distributions are seen clearly in the example shown in Figure 1e and 1f and show more clear structure to be present in the real mask than in the imaginary mask.

## III. Mask Estimation

This section presents the combined CNN-DNN architecture to estimate a complex mask from input noisy spectral features and in particular introduces the proposed loss function to improve imaginary mask estimation.

### A. Feature extraction

Masking methods have traditionally relied on hand-crafted input features, such as cochleagrams, mel-filterbanks, MFCCs and complementary sets of these ( [14], [21], [22]). Recent research has demonstrated that deep-learning based features outperform these classic features, as models have the power to learn more discriminative representations of the signals, hence this method has been adopted [14], [23]. First, log-power spectral features are extracted from the input signal using a window length of 20ms and a 50% overlap. Initial experiments found that more accurate masks were estimated from log power spectral features as opposed to using real and imaginary spectra, which we attribute to the log compression that cannot be applied to the negative values of the complex components. To capture temporal information, a set of spectral features are stacked into mini-spectrograms. From preliminary tests, a stack of 47 frames was found to give highest PESQ scores, and corresponds to a mini-spectrogram 480ms wide. This is input into a 5-layer CNN, shown in Figure 3, which learns relevant features from the input mini-spectrograms and comprises five two-dimensional convolutional layers with the first two both followed by max-pooling layers. The first layer comprises 16 2 × 2 filters, which is increased to 16 3 × 3 filters in the second layer, with the remaining layers each having 64 2 × 2 filters.

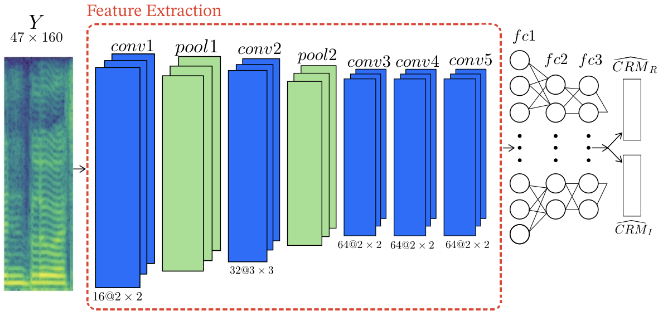

Fig. 3. Proposed complex mask architecture that takes as input a set of stacked log-power spectral vectors, applies CNN feature extraction followed by four feedforward layers to output estimates of the real and imaginary masks.

### B. Target mask pre-processing

The training target for the proposed model is the complex ideal ratio mask (CIRM). While the range of values in the real and imaginary mask components extend to $[-\infty; \infty]$, the majority of amplitudes are concentrated near zero, as evidenced in Figure 2. To avoid optimisation problems caused by some of the large values present in the masks, the real and imaginary components are truncated to the $[-5; 5]$ range. After evaluating various methods (e.g. [13]), a sigmoid function was determined to be the most effective approach for compressing the mask values. After prediction, the real and imaginary amplitudes are expanded back to the truncated range using the inverse sigmoid function, such that,

$$\text{CIRM}_z = \frac{1}{1+e^{-M_z}}, \quad \widehat{M}_z = \log_e\left(\frac{\widehat{CIRM}_z}{1-\widehat{CIRM}_z}\right) \tag{14}$$

where $M_z$ and $\text{CIRM}_z$ are the original and compressed target masks, $\widehat{\text{CIRM}}_z$ and $\widehat{M}_z$ are the estimated and expanded mask amplitudes and $z \in \{r, i\}$ denotes the real or imaginary mask. The truncation applied before compression is non-invertible, and it restricts a small subset of mask amplitudes, though not using it limits the model's ability to learn mask patterns.

### C. Model architecture

The overall model architecture maps each 480ms mini-spectrogram into an estimate of the complex mask vector using the two-stage CNN-DNN model shown in Figure 3. The first stage uses a CNN for feature extraction as explained in Section III-A. The output of the CNN is flattened and input into the second stage which comprises three fully-connected layers, containing 1024, 512 and 256 nodes respectively, all preceded by a batch normalisation layer followed by a 0.2 rate dropout. A fourth layer outputs a vector that contains the real and imaginary mask estimates, $\widehat{CIRM}_r$ and $\widehat{CIRM}_i$, and these correspond in time to the central vector in the mini-spectrogram input. Rectified linear units (ReLU) are used throughout the model [24], with the exception of the output layer which uses a sigmoid activation (14).

### D. Loss function

The mean squared error (MSE) is the conventional metric used within loss functions for training neural networks and has been used within complex mask estimation methods. This applies an equal weighting across all components of the real and imaginary masks. To address the issue of underestimation of the imaginary mask, a weighted loss function is proposed that aims to direct model training towards the imaginary mask. This is achieved by separating the mask error into individual real and imaginary components, where a tunable weight, $\alpha^{\text{IMAG}}$, can be applied to the imaginary error to increase its contribution within the loss function. Further, as ultimately better phase correction by the mask is desired, a third term is introduced into the loss function that is the phase error of the mask and this is also given its own adjustable weight parameter, $\alpha^{\text{PH}}$. The proposed loss function, $\mathcal{L}_w$, is calculated as,

$$\begin{aligned}\mathcal{L}_w = \frac{1}{2N}\sum_{t=1}^{N}\Bigg[&\Big(\sum_{f=1}^{F}(\Re\{CIRM(f,t)\} - \Re\{\widehat{CIRM}(f,t)\})^2\Big)\\ &+\alpha^{\text{IMAG}}\Big(\sum_{f=1}^{F}(\Im\{CIRM(f,t)\} - \Im\{\widehat{CIRM}(f,t)\})^2\Big)\\ &+\alpha^{\text{PH}}\Big(\sum_{f=1}^{F}\Big|\theta_{CIRM}(f,t) - \theta_{\widehat{CIRM}}(f,t)\Big|\Big)\Bigg]\end{aligned} \tag{15}$$

where $CIRM(f,t)$ and $\widehat{CIRM}(f,t)$ are the ideal (target) and estimated masks that follow truncation and compression in (14). The summation is made over the $N$ mask vectors used in training, and $F = 160$ is the number of frequency bins. The factor of 2 in the denominator is used to normalize the MSE by the number of output dimensions. The error from the real mask is implicitly one, as this was found best for preserving magnitude, while the imaginary and phase component weights, $\alpha^{\text{IMAG}}$ and $\alpha^{\text{PH}}$, can be adjusted to increase or decrease their contribution to the overall error.

## IV. Experimental results

The aim of the experiments is to examine the effectiveness of the proposed loss function and the contribution made by the imaginary and phase components in terms of improving speech quality and intelligibility.

### A. Dataset and evaluation metrics

The dataset used comprises two female and two male speakers taken from the GRID corpus [20]. The dataset contains a total of 4000 utterances which has been partitioned using a 70:30 ratio for training/validation and testing, with all speakers used for training and testing, but with the respective sets extracted randomly with no overlap. The contaminating noise is taken from the DEMAND dataset [25] and uses the 'Public' and 'Street' noises with no samples shared across the training, validation and test sets. Evaluation of the quality of the enhanced speech is made objectively using PESQ while intelligibility is measured using ESTOI [26], [27]. Two SNRs are used for evaluation, 0dB and -10dB, chosen specifically as with these high noise conditions the effect of phase error is perceptible.

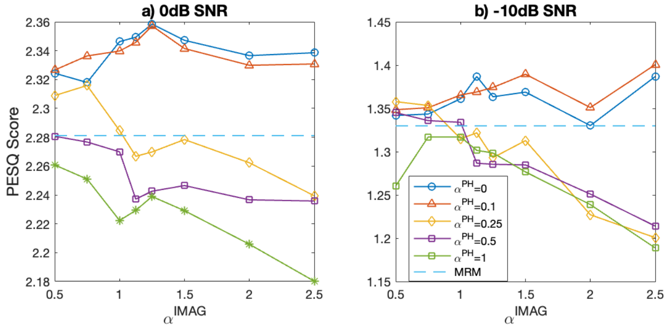

Fig. 4. Effect of varying $\alpha^{\text{IMAG}}$ and $\alpha^{\text{PH}}$ in terms of PESQ scores at SNRs of a) 0dB and b) -10dB.

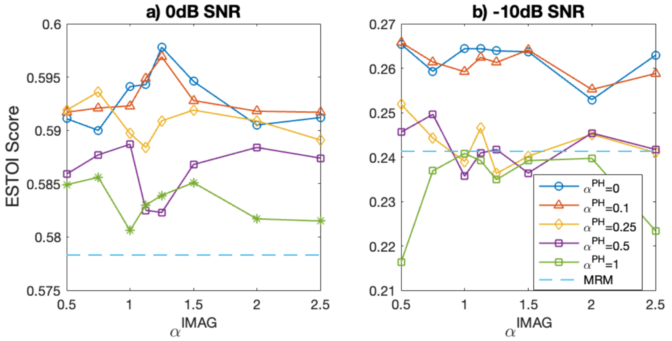

Fig. 5. Effect of varying $\alpha^{\text{IMAG}}$ and $\alpha^{\text{PH}}$ in terms of ESTOI scores at SNRs of a) 0dB and b) -10dB.

### B. Quality and intelligibility

To investigate the effect that the proposed loss function has on the enhanced speech signal, quality and intelligibility are measured for different combinations of the imaginary and phase weights in (15). Figures 4a and 4b show PESQ scores at SNRs of 0dB and -10dB for $0.5 \leq \alpha^{\text{IMAG}} \leq 2.5$ and $0 \leq \alpha^{\text{PH}} \leq 1$. With no enhancement applied, the PESQ scores for the noisy speech at 0dB are 1.74 and at -10 dB are 1.04. Applying enhancement using the proposed loss function, at 0dB the highest PESQ score of 2.36 is attained with no contribution from the phase mask error, i.e. $\alpha^{\text{PH}}$=0, although at -10dB including a small contribution ($\alpha^{\text{PH}}$=0.1) improves scores. The effect of increasing the contribution of the imaginary mask error is more beneficial, with increased scores that peak at $\alpha^{\text{IMAG}}$=1.25 at 0dB and $\alpha^{\text{IMAG}}$=1.5 at -10dB. These are both higher than the conventional magnitude ratio masking (MRM) scores that are shown in the figure.

ESTOI scores are shown in Figures 5a and 5b for SNRs of 0dB and -10dB for the same combinations of $\alpha^{\text{IMAG}}$ and $\alpha^{\text{PH}}$ values. With no enhancement applied, the ESTOI scores for the noisy speech are 0.378 and 0.110 at 0dB and -10dB respectively. Altering the imaginary and phase error weights shows a largely similar trend of scores as observed for PESQ, with best scores of 0.598 at 0dB and 0.264 at -10dB, although phase error weights of both $\alpha^{\text{PH}}$=0 and $\alpha^{\text{PH}}$=0.1 have almost identical performance. Compared to magnitude ratio masking, the proposed loss function gives increased ESTOI scores.

The results show that an increase in loss function weight for the imaginary mask error improves both PESQ and ESTOI, while emphasising the phase mask error has less effect and if set too large is detrimental. This shows that prioritising imaginary mask error in model training leads directly to complex masks better able to enhance the noisy speech. Conversely, giving more weight to the phase error reduces the relative contribution of the real and imaginary components, leading to degraded speech enhancement.

## V. Conclusion

In this study, an assessment of the theory behind complex masking has been made, showing why the imaginary mask is predisposed to underestimation during model training and how this affects the associated magnitude and phase of the complex mask. A parameterised loss function has been evaluated across a range of values to measure the effect of the imaginary and phase components during learning. Evaluation done on the GRID dataset indicates that assigning more weight to the imaginary component directs the model towards producing better masks. This is evident in the enhanced speech quality and intelligibility, which improved at both 0dB and -10dB.

## References

[1] P. C. Loizou, *Speech enhancement: theory and practice.* CRC press, 2013.

[2] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. V. Muñoz, "Clarity-2021 Challenges: Machine Learning Challenges for Advancing Hearing Aid Processing," in *Proc. Interspeech 2021*, 2021, pp. 686–690.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[4] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[5] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

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

[7] Y. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "A maximum likelihood approach to deep neural network based nonlinear spectral mapping for single-channel speech separation," in *Proc. Interspeech 2017*, 2017, pp. 1178–1182.

[8] C. Hummersone, T. Stokes, and T. Brookes, *On the Ideal Ratio Mask as the Goal of Computational Auditory Scene Analysis*, 05 2014, pp. 349–368.

[9] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

[10] M. Hasannezhad, Z. Ouyang, W.-P. Zhu, and B. Champagne, "An integrated CNN-GRU framework for complex ratio mask estimation in speech enhancement," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2020, pp. 764–768.

[11] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[12] R. Chappel, B. Schwerin, and K. Paliwal, "Phase distortion resulting in a just noticeable difference in the perceived quality of speech," *Speech Communication*, vol. 81, pp. 138–147, 2016, phase-Aware Signal Processing in Speech Communication.

[13] 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.

[14] M. Hasannezhad, Z. Ouyang, W.-P. Zhu, and B. Champagne, "Speech enhancement with phase sensitive mask estimation using a novel hybrid neural network," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 136–150, 2021.

[15] S. Xia, H. Li, and X. Zhang, "Using optimal ratio mask as training target for supervised speech separation," *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 163–166, 2017.

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

[17] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712.

[18] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 72–76, 2020.

[19] D. Websdale, S. Taylor, and B. Milner, "Speaker-independent speech animation using perceptual loss functions and synthetic data," *IEEE Transactions on Multimedia*, vol. 24, pp. 2539–2552, 2022.

[20] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 150, no. 5, pp. 2421–2424, Nov. 2006.

[21] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE Trans. Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1993–2002, Dec. 2014.

[22] ETSI, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," ETSI STQ-Aurora DSR Working Group, ES 202 050 version 1.1.1, Oct. 2002.

[23] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6989–6993.

[24] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[25] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multichannel recordings of acoustic noise in diverse environments," Jun. 2013, Supported by Inria under the Associate Team Program VERSAMUS.

[26] R. Rix, J. Beerands, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment telephone networks and codecs," in *ICASSP*, 2001, pp. 749–752.

[27] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE Trans. Audio, Speech and Language Processing*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.