*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# Single-channel speech enhancement by subspace affinity minimization

*Dung N. Tran, Kazuhito Koishida*

Microsoft Corporation
{dung.tran, kazukoi}@microsoft.com

## Abstract

In data-driven speech enhancement frameworks, learning informative representations is crucial to obtain a high-quality estimate of the target speech. State-of-the-art speech enhancement methods based on deep neural networks (DNN) commonly learn a single embedding from the noisy input to predict clean speech. This compressed representation inevitably contains both noise and speech information leading to speech distortion and poor noise reduction performance. To alleviate this issue, we proposed to learn from the noisy input separate embeddings for speech and noise and introduced a subspace affinity loss function to prevent information leaking between the two representations. We rigorously proved that minimizing this loss function yields maximally uncorrelated speech and noise representations, which can block information leaking. We empirically showed that our proposed framework outperforms traditional and state-of-the-art speech enhancement methods in various unseen nonstationary noise environments. Our results suggest that learning uncorrelated speech and noise embeddings can improve noise reduction and reduces speech distortion in speech enhancement applications.

**Index Terms**: speech enhancement, noise reduction, deep neural network, convolutional neural network, regression, subspace affinity

## 1. Introduction

Speech enhancement aims to predict the target speech from its noisy counterpart without the knowledge of noise information. It plays a particularly important role in speech applications and has been extensively investigated for several decades. Numerous reliable speech enhancement techniques have been proposed in the literature. Classical speech enhancement methods typically employ a simple signal processing algorithm or heuristic to estimate a gain function, which is then applied to the noisy input to obtain the enhanced speech [1, 2, 3, 4]. Recent advances in deep learning have motivated several speech enhancement methods based on deep neural networks (DNN) [5, 6, 7, 8, 9, 10, 11], which outperform traditional signal processing based approaches.

Despite outperforming classical methods, DNN based speech enhancement frameworks deliver unsatisfactory noise reduction performance under challenging situations in which nonstationary noise severely degrades the target speech. The main reason for this limitation is that speech enhancement methods based on deep learning commonly utilize an encoder-decoder structure aiming to learn a single embedding via the encoder [5, 6, 7, 8, 11]. The decoder then maps this representation to the enhanced magnitude spectrum or raw speech. A significant drawback in this approach is that this compressed representation inevitably contains both noise and speech information resulting in speech distortion and poor noise reduction performance.

To resolve this issue, one can construct a network that learns the speech and noise representations separately in a supervised manner using a single encoder. Two separate decoders are then utilized to predict both the target speech and noise signals. However, another problem arises: without proper regularization, noise information unavoidably leaks into the speech embedding and vice versa. This information leaking problem can severely affect the performance of speech enhancement systems adopting this approach.

In this work, we proposed a framework that learns separate speech and noise embeddings with an inherent mechanism to prevent information leaking between the two representations. In particular, based on the hypothesis that speech and noise signals are uncorrelated to some extent in a high dimensional space, we designed a subspace affinity loss function for learning such embeddings. More specifically, this loss function encourages the speech and noise embeddings to reside in maximally uncorrelated subspaces. A consistency loss combined with the subspace affinity loss then guarantee that the speech information correctly propagates to the speech decoder and the noise information to the noise decoder. We theoretically proved that minimizing this loss function produces maximally uncorrelated speech and noise representations, which prevent leaking of information. We evaluated our proposed framework on a severely noisy speech dataset and showed that our approach outperforms state-of-the-art DNN based speech enhancement methods when handling unseen nonstationay noise.

The paper is organized as follows. Section 2 presents our network diagram and training loss. In Section 3, we introduce the subspace affinity loss. We rigorously prove that minimizing this loss yields uncorrelated representations, which prevent information leaking. In Section 4, we evaluate and benchmark our proposed framework against traditional and state-of-the-art speech enhancement approaches on a public speech enhancement dataset with various nonstationary noise settings. Section 5 summarizes our contributions and discusses future works.

## 2. Method

### 2.1. Problem statement

We consider the single-channel speech enhancement problem

$$\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{x}$, $\boldsymbol{s}$, and $\boldsymbol{n}$ are high dimensional vectors representing the observed noisy speech, the unknown target speech, and the unknown noise signals, respectively. We seek a high-quality estimate of the clean speech $\boldsymbol{s}$ from the measured noisy signal $\boldsymbol{x}$ without any noise information.

Our network is trained to produce from the noisy vector separate representations for speech and noise, which are then used to predict the target speech and the noise signals. In particular, we first use an encoder $f$ to map the noisy input into an



http://dx.doi.org/10.21437/Interspeech.2020-2982

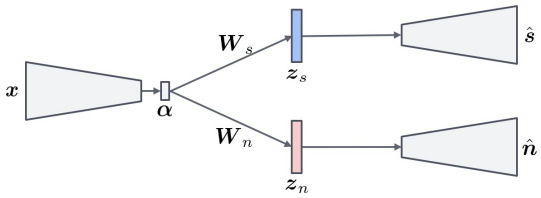


Figure 1: *High-level diagram of our network.*

encoding $\boldsymbol{\alpha} \in \mathbb{R}^d$:

$$\boldsymbol{\alpha} = f(\boldsymbol{x}). \tag{2}$$

The speech representation $\boldsymbol{z}_s \in \mathbb{R}^D$ and the noise representation $\boldsymbol{z}_n \in \mathbb{R}^D$ are then obtained from this common encoding using two linear mappings $\boldsymbol{W}_s \in \mathbb{R}^{D\times d}$ and $\boldsymbol{W}_n \in \mathbb{R}^{D\times d}$, where $D \geq d$:

$$\boldsymbol{z}_s = \boldsymbol{W}_s\boldsymbol{\alpha} \quad \text{and} \quad \boldsymbol{z}_n = \boldsymbol{W}_n\boldsymbol{\alpha}. \tag{3}$$

Finally, we use a speech decoder $g_{\boldsymbol{s}}$ to predict the target speech and a noise decoder $g_{\boldsymbol{n}}$ to predict the noise signal:

$$\hat{\boldsymbol{s}} = g_{\boldsymbol{s}}(\boldsymbol{z}_s) \quad \text{and} \quad \hat{\boldsymbol{n}} = g_{\boldsymbol{n}}(\boldsymbol{z}_n). \tag{4}$$

Here, $\hat{\boldsymbol{s}}$ is the target speech prediction and $\hat{\boldsymbol{n}}$ is the noise prediction. Fig 1 show a high-level diagram of our network structure.

We train our network using triplets $\{(\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{n})\}$ of noisy speech, target speech ground-truths, and noise ground-truths, respectively. The network is trained to match the predicted speech to the actual target speech, the predicted noise to the noise ground-truth. During inference time, we obtain the final speech prediction from the speech decoder $g_{\boldsymbol{s}}$ and discard the noise decoder $g_{\boldsymbol{n}}$.

The linear mappings $\boldsymbol{W}_s$ and $\boldsymbol{W}_n$ play a crucial role in our framework. Properly regularizing these transformations can prevent information leaking between the speech and noise embeddings, hence improving speech enhancement performance. In particular, $\boldsymbol{W}_s$ and $\boldsymbol{W}_n$ are constrained to extract *maximally uncorrelated information* from the common embedding $\boldsymbol{\alpha}$. We achieve this using a carefully-designed training loss function.

### 2.2. Training loss

Our training loss consists of a consistency loss and an affinity loss:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{consistency}}\left(f, g_{\boldsymbol{s}}, g_{\boldsymbol{n}}, \boldsymbol{W}_s, \boldsymbol{W}_n\right) + \lambda \mathcal{L}_{\text{affinity}}\left(\boldsymbol{W}_s, \boldsymbol{W}_n\right). \tag{5}$$

In Eq. (5), $\lambda > 0$ is a regularization parameter that balances the loss components. The consistency loss $\mathcal{L}_{\text{consistency}}$ encourages the speech and noise predictions to be consistent with the corresponding ground-truth counterparts and depends on all the parameters of the network. The subspace affinity loss $\mathcal{L}_{\text{affinity}}$, on the other hand, is imposed on $\boldsymbol{W}_s$ and $\boldsymbol{W}_n$ only. It forces these transformations to extract maximally uncorrelated information from the common embedding $\boldsymbol{\alpha}$. A combination of $\mathcal{L}_{\text{consistency}}$ and $\mathcal{L}_{\text{affinity}}$ balanced by $\lambda$ thus guarantees that speech information properly propagates to the speech decoder and noise information to the noise decoder.

In our framework, we use the Mean Squared Error (MSE) function as the consistency loss:

$$\mathcal{L}_{\text{consistency}} = \frac{1}{|\mathcal{D}|}\sum_{\mathcal{D}}\left(\|\hat{\boldsymbol{s}} - \boldsymbol{s}\|_2^2 + \eta\,\|\hat{\boldsymbol{n}} - \boldsymbol{n}\|_2^2\right), \tag{6}$$

where $|\mathcal{D}|$ denotes the number of samples in the training set $\mathcal{D}$.

The main innovation in our work is the introduction of the subspace affinity loss $\mathcal{L}_{\text{affinity}}$ to prevent information leaking between the speech and noise representations. We devote the next section to the main idea and analysis of this loss function.

## 3. Subspace affinity loss

Our main assumption is that speech and noise can be represented by some uncorrelated embeddings in some high dimensional space. We aim to learn such unknown representations by regularizing the linear transforms $\boldsymbol{W}_s$ and $\boldsymbol{W}_n$.

Denote $\mathcal{W}_s$ and $\mathcal{W}_n$ as the subspaces spanned by the columns of $\boldsymbol{W}_s$ and $\boldsymbol{W}_n$, respectively. Eq. (3) implies that, in the ambient space $\mathbb{R}^D$, the speech encoding $\boldsymbol{z}_s$ lie in $\mathcal{W}_s$, the noise encoding $\boldsymbol{z}_n$ in $\mathcal{W}_n$. Furthermore, as $\boldsymbol{z}_s$ and $\boldsymbol{z}_n$ share the same encoding $\boldsymbol{\alpha}$ in these subspaces, if $\mathcal{W}_s$ and $\mathcal{W}_n$ are dissimilar, $\boldsymbol{z}_s$ and $\boldsymbol{z}_n$ will be dissimilar. Our subspace affinity is designed to encourage the dissimilarity between $\mathcal{W}_s$ and $\mathcal{W}_n$.

In this section, we will discuss the subspace affinity concept which characterizes the dissimilarity between two arbitrary subspaces. Then, we introduce our subspace affinity loss function built upon this concept and show that minimizing this loss function results in maximally uncorrelated representations.

### 3.1. Subspace affinity

Subspace affinity is built upon the concept of principal angles which naturally capture the notion of similarity/affinity between subspaces.

**Definition 3.1** (Principal angles [12])**.** *The principal angles $\{\theta_i\}_{i=1}^d$ between two subspaces $\mathcal{U}$ and $\mathcal{V}$ of dimensions $d_u$ and $d_v$, where $d = \min\{d_u, d_v\}$, are recursively defined by*

$$\cos\theta_i \coloneqq \max_{\boldsymbol{u}_i\in\mathcal{U}}\max_{\boldsymbol{v}_i\in\mathcal{V}} \frac{\boldsymbol{u}_i^T\boldsymbol{v}_i}{\|\boldsymbol{u}_i\|_2\,\|\boldsymbol{v}_i\|_2}, \tag{7}$$

*with the orthogonality constraints $\boldsymbol{u}_i^T\boldsymbol{u}_j = 0, \boldsymbol{v}_i^T\boldsymbol{v}_j = 0, j = 1, ..., i-1$.*

**Definition 3.2** (Subspace affinity [12])**.** *The affinity between two subspaces $\mathcal{U}$ and $\mathcal{V}$ of dimensions $d_u$ and $d_v$, respectively, is defined as*

$$\text{aff}\,(\mathcal{U}, \mathcal{V}) \coloneqq \left(\sum_{i=1}^{d}\cos^2\theta_i\right)^{\frac{1}{2}}, \tag{8}$$

*where $\{\theta_i\}_{i=1}^d$ are the principal angles between $\mathcal{U}$ and $\mathcal{V}$ and $d = \min\{d_u, d_v\}$.*

Intuitively, the subspace affinity compactly captures the notion of correlation between two subspaces. That is, the subspaces are dissimilar or uncorrelated when the principal angles are right angles, i.e., the affinity is small. On the other hand, large affinity due to small principal angles implies the subspaces are correlated.

The following lemmas provides an easy way to compute principal angles and subspace affinity.

**Lemma 3.1.** *Let the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ be orthonormal bases for subspaces $\mathcal{U}$ and $\mathcal{V}$ of dimensions $d_u$ and $d_v$, respectively. Let $\{\sigma_i\}_{i=1}^d$ be the singular values of $\boldsymbol{U}^T\boldsymbol{V}$, where $d = \min\{d_u, d_v\}$, then*

$$\cos\theta_i = \sigma_i, \quad i = 1, ..., d. \tag{9}$$

**Lemma 3.2.** *The affinity between two subspaces $\mathcal{U}$ and $\mathcal{V}$ can be calculated by*

$$\text{aff}\left(\mathcal{U},\mathcal{V}\right)=\left\|\boldsymbol{U}^{T}\boldsymbol{V}\right\|_{F}. \tag{10}$$

Lemma 3.2 allows us to construct our subspace affinity loss function.

### 3.2. Subspace affinity loss

Based on Lemma 3.2, we propose to minimize the following subspace affinity loss to encourage the dissimilarity between the subspaces $\mathcal{W}_s$ and $\mathcal{W}_n$:

$$\begin{aligned}\mathcal{L}_{\text{affinity}}=&\left\|\boldsymbol{W}_{s}^{T}\boldsymbol{W}_{n}\right\|_{F}^{2} \\ &+\mu\left(\left\|\boldsymbol{W}_{s}^{T}\boldsymbol{W}_{s}-\boldsymbol{I}\right\|_{F}^{2}+\left\|\boldsymbol{W}_{n}^{T}\boldsymbol{W}_{n}-\boldsymbol{I}\right\|_{F}^{2}\right).\end{aligned} \tag{11}$$

In Equation (11), the last two terms force the columns of $\boldsymbol{W}_s$ and $\boldsymbol{W}_n$ to be orthonormal bases of the subspaces $\mathcal{W}_s$ and $\mathcal{W}_n$, respectively. The first term of the loss penalizes the squared affinity of the two subspaces. Minimizing the subspace affinity loss therefore promotes the dissimilarity between $\mathcal{W}_s$ and $\mathcal{W}_n$.

In our framework, minimizing the subspace affinity loss function leads to an attractive properties of the speech and noise representations which prevents information leaking between them.

**Theorem 3.1.** *Assume $\boldsymbol{W}_s$ and $\boldsymbol{W}_n$ are orthonormal bases of $\mathcal{W}_s$ and $\mathcal{W}_n$, respectively. The correlation between the nonzero speech and noise embeddings is bounded by the affinity between $\mathcal{W}_s$ and $\mathcal{W}_n$:*

$$\left|\cos(\boldsymbol{z}_s,\boldsymbol{z}_n)\right|\leq\text{aff}\left(\mathcal{W}_s,\mathcal{W}_n\right). \tag{12}$$

*Proof.* Let $\boldsymbol{W}=\boldsymbol{W}_s^T\boldsymbol{W}_n$, we have

$$\left|\cos(\boldsymbol{z}_s,\boldsymbol{z}_n)\right|=\frac{\left|\boldsymbol{z}_s^T\boldsymbol{z}_n\right|}{\left\|\boldsymbol{z}_s\right\|_2\left\|\boldsymbol{z}_n\right\|_2}=\frac{\left|\boldsymbol{\alpha}^T\boldsymbol{W}\boldsymbol{\alpha}\right|}{\left\|\boldsymbol{\alpha}\right\|_2\left\|\boldsymbol{\alpha}\right\|_2}=\left|\frac{\boldsymbol{\alpha}^T\boldsymbol{W}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\boldsymbol{\alpha}}\right|, \tag{13}$$

where the second equality is due to Eq. (3) and the assumption that $\boldsymbol{W}_s$ and $\boldsymbol{W}_n$ are orthonormal bases of $\mathcal{W}_s$ and $\mathcal{W}_n$, respectively. Let $\lambda_1\leq\lambda_2\leq...\leq\lambda_d$ be the eigenvalues of the symmetric matrix $\boldsymbol{W}$, the min-max theorem implies that the Reyleigh quotient $\frac{\boldsymbol{\alpha}^T\boldsymbol{W}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\boldsymbol{\alpha}}$ is bounded [13]:

$$\lambda_1\leq\frac{\boldsymbol{\alpha}^T\boldsymbol{W}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\boldsymbol{\alpha}}\leq\lambda_d. \tag{14}$$

Therefore,

$$\left|\cos(\boldsymbol{z}_s,\boldsymbol{z}_n)\right|\leq\max_i\left|\lambda_i\right|. \tag{15}$$

As $\boldsymbol{W}$ is symmetric, $\max_i|\lambda_i|=\sigma_{\max}$, where $\sigma_{\max}$ is the largest singular value of $\boldsymbol{W}$. We conclude that

$$\left|\cos(\boldsymbol{z}_s,\boldsymbol{z}_n)\right|\leq\sigma_{\max}\leq\sqrt{\sum_{i=1}^{d}\sigma_i^2}=\left\|\boldsymbol{W}\right\|_F=\text{aff}\left(\mathcal{W}_s,\mathcal{W}_n\right), \tag{16}$$

where $\sigma_i$'s are the singular values of $\boldsymbol{W}$. □

**Corollary 3.1.1.** *The speech embedding $\boldsymbol{z}_s$ and the noise embedding $\boldsymbol{z}_n$ are uncorrelated, i.e., $\left|\cos(\boldsymbol{z}_s,\boldsymbol{z}_n)\right|=0$, when the affinity between $\mathcal{W}_s$ and $\mathcal{W}_n$ vanishes.*

We can achieve zero affinity when the total dimension of $\mathcal{W}_s$ and $\mathcal{W}_n$ is less than or equal the ambient dimension $D$, e.g., two perpendicular lines and $\mathbb{R}^3$. A special case is $\mathcal{W}_s$ and $\mathcal{W}_n$ are orthogonal subspaces of $\mathbb{R}^D$, e.g., a line perpendicular to a 2D plane in $\mathbb{R}^3$. In these situations, the embedding correlation is zero, which implies the representations are uncorrelated.

It is important to note that although it is possible to manually design $\mathcal{W}_s$ and $\mathcal{W}_n$ to be orthogonal subspaces of $\mathbb{R}^D$ so that $\boldsymbol{z}_s$ and $\boldsymbol{z}_n$ are uncorrelated, e.g. by sampling the elements of $\boldsymbol{W}_s$ and $\boldsymbol{W}_n$ from a Normal distribution, the subspace affinity loss provides a flexible way to balance the prediction consistency and the correlation of the encodings, especially when the uncorrelated assumption between speech and noise is violated.

**Remark.** One can directly enforce a correlation constraint on the embeddings, by minimizing the dot-product of the speech and noise embeddings, to obtain uncorrelated representations. However, this only guarantees the uncorrelation between the speech and noise embeddings asymptotically, instead of on individual samples as such cost functions aim to minimize the average of the embedding dot-products of all data samples.

## 4. Experimental results

In this section, we empirically evaluate and benchmark our affinity minimization framework against state-of-the-art speech enhancement methods on a public speech enhancement dataset.

### 4.1. Dataset

We use the popular VCTK dataset by Valentini et al [14] which is publicly available at [15]. The dataset includes clean and noisy speech data sampled at 48 kHz. In our experiments, we downsample the data to 16 kHz.

For training, we use the clean speech audio data of 28 speakers selected from the Voice Bank corpus [16]. The noisy training data are created by adding to the clean speech ten different types of noise at four signal-to-noise ratios (SNR), yielding 40 noise conditions. The ten noise types include eight real noise samples selected from the Demand data [17] and two artificially ones. The four training SNRs are 0 dB, 5 dB, 10 dB, and 15 dB.

The test data is different from the training data. The noisy test set is created by adding five different types of noises from the Demand database to the clean speech of two speakers from the Voice Bank corpus. The noise types and speakers in the test set are different from the ones used in training. The four test SNR values are 2.5 dB, 7.5 dB, 12.5 dB, and 17.5 dB. Consequensely, there are 20 different noise conditions.

### 4.2. Objective metrics

To evaluate the enhanced speech, we use four popular objective measures for speech enhancement. Each metric is computed by comparing the enhanced speech with the corresponding clean reference of each of the test samples. The first three metrics predict the Mean Opinion Score (MOS) that would result from human perceptual trials. They include (1) CSIG which predicts the signal distortion MOS, (2) CBAK which is a MOS predictor of background-noise intrusiveness, and (3) COVL which computes the MOS value of the overall signal quality. CSIG, CBAK, and COVL all produce MOS values from 1 to 5. Finally, PESQ, which stands for Perceptual Evaluation of Speech Quality, is a broadly used objective measure for speech quality. It provides a score in the range between -0.5 to 4.5. For all of the metrics, the higher value corresponds to the better quality of the enhanced speech.

Table 1: *Encoder architecture.*

| Type | Kernel | Stride | Output | Activation |
|---|---|---|---|---|
| Input | | | $16 \times 256 \times 1$ | |
| Conv2d | $5 \times 3$ | $1 \times 1$ | $16 \times 256 \times 64$ | LReLU |
| Conv2d | $3 \times 3$ | $1 \times 2$ | $16 \times 128 \times 128$ | LReLU |
| Conv2d | $3 \times 3$ | $1 \times 2$ | $16 \times 64 \times 128$ | LReLU |
| Conv2d | $3 \times 3$ | $1 \times 2$ | $16 \times 32 \times 128$ | LReLU |
| Conv2d | $3 \times 3$ | $1 \times 2$ | $16 \times 16 \times 128$ | LReLU |
| Conv2d | $3 \times 3$ | $1 \times 2$ | $16 \times 8 \times 128$ | LReLU |
| Conv2d | $3 \times 3$ | $1 \times 2$ | $16 \times 4 \times 128$ | LReLU |
| Conv2d | $3 \times 3$ | $1 \times 2$ | $16 \times 2 \times 128$ | LReLU |
| Conv2d | $3 \times 3$ | $1 \times 2$ | $16 \times 1 \times 128$ | LReLU |
| Conv2d | $1 \times 3$ | $2 \times 1$ | $8 \times 1 \times 256$ | LReLU |
| Conv2d | $1 \times 3$ | $2 \times 1$ | $4 \times 1 \times 256$ | LReLU |
| Conv2d | $1 \times 3$ | $2 \times 1$ | $2 \times 1 \times 256$ | LReLU |
| Conv2d | $1 \times 1$ | $2 \times 1$ | $1 \times 1 \times 256$ | |

### 4.3. Experimental setup

**Data format.** Our network predicts the log power spectrum of clean speech from that of the corresponding noisy signal. Then, the predicted power spectrum is combined with the noisy phase extracted from the noisy signal to produce the enhanced speech. To obtain the signal power spectrum, we apply the Short-Time-Fourier-Transform (STFT) to the raw audio using 512 FFT points with a hop size of 256 and Hann window. This produces $16 \times 257$ overlapping time-frequency frames, where 16 frames are equivalent to 256 ms. Finally, we remove the last frequency bin yielding $16 \times 256$ time-frequency frames.

**Network architecture.** Our general network structure consists of an encoder, a speech decoder and a noise decoder, and two bias-free fully-connected layers to split the bottleneck at the end of the encoder into a speech encoding and a noise encoding. For this speech enhancement experiment, we propose a time-frequency separable architecture for the encoder and the decoders. In particular, our encoder consists of a series of 13 convolutional layers which downsample the input time-frequency frame along the frequency and the time axes separately. The first 9 layers downsample the frequency axis, and the last 4 layers perform downsampling along the time axis. The detailed configuration of the encoder is shown in Table 1. Similarly, the decoders reverse the downsampling process in the encoder by upsampling the data along the frequency and the time axes separately. The decoders use pixel shufflers [18] for upsampling data. Each layer in the encoder and the decoders uses the Leaky ReLU activation function and batch normalization. We use skip connections between the encoder layers and the corresponding layers in both the speech and noise decoders. As the last encoder layer in Table 1 produces an encoding of length $d = 256$, we set $D = 2d = 512$ so that the affinity between $\boldsymbol{W}_s$ and $\boldsymbol{W}_n$ vanishes when the training converges.

**Training setup.** We train our network using the ADAM optimizer with a learning rate of 0.0001, decay rates $\beta_1 = 0.5$ and $\beta_2 = 0.9$, and a batch size of 64. The leaky ReLU constant is 0.2. We set $\eta = 1$, $\lambda = 0.1$ and $\mu = 10$. These values are chosen using grid search. To prevent overfitting, we apply $\ell_2$ regularization to the convolutional weights with a value of 0.1. For each configuration, the network is trained for 200 epochs.

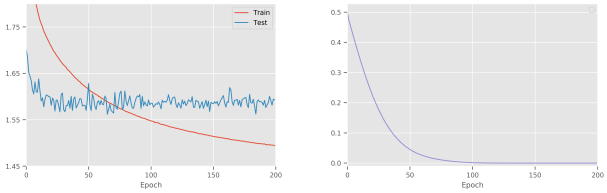


Figure 2: *Losses curves. Left:* $\mathcal{L}_{\text{consistency}}$. *Right:* $\mathcal{L}_{\text{affinity}}$.

### 4.4. Results

We benchmark our proposed framework against classical and state-of-the-art speech enhancement methods which include: (1) Wiener filtering [2] based on a priori SNR estimation; (2) SEGAN [5]; (3) Wavenet [7]; (4) MMSE-GAN [6]; (5) D+M [19]; and (6) UNet [11]. Table 2 shows the numerical results of the aforementioned objective metrics on the test dataset for all benchmarked frameworks. For reference, we also report the objective measures computed for the noisy test signals. The results indicate that our affinity minimization framework outperforms state-of-the-art speech enhancement methods in all the metrics.

Table 2: *Speech enhancement benchmark*

| | PESQ | CSIG | CBAK | COVL |
|---|---|---|---|---|
| Noisy | 1.97 | 3.35 | 2.44 | 2.63 |
| Wiener [2] | 2.22 | 3.23 | 2.68 | 2.67 |
| SEGAN [5] | 2.16 | 3.48 | 2.94 | 2.80 |
| WaveNet [7] | N/A | 3.62 | 3.23 | 2.98 |
| MMSE-GAN [6] | 2.53 | 3.80 | 3.12 | 3.14 |
| D+M [19] | 2.73 | 3.94 | 3.35 | 3.33 |
| UNet [11] | 2.90 | 4.22 | 3.32 | 3.58 |
| Ours | **3.04** | **4.30** | **3.42** | **3.69** |

Fig. 2 shows the consistency loss for the train and test sets and the affinity loss. Notice that the affinity loss starts vanishing when the training consistency loss starts crossing below the test consistency loss. This phenomenal suggests that our network achieves zero affinity at convergence. This leads to uncorrelated representations as proved in Theorem 3.1.

## 5. Conclusions

We presented a speech enhancement framework that learns separate representations for speech and noise and proposed the subspace affinity loss function to prevent information leaking between the two representations. We theoretically proved that minimizing this loss function produces maximally uncorrelated speech and noise representations, allowing the speech information to correctly propagate to the speech decoder and the noise information to the noise decoder. Experimental results indicated that our framework is significantly more robust than state-of-the-art DNN based speech enhancement approaches in unseen nonstationary noise settings.

Our results for unseen noise conditions suggest that uncorrelated representations learned by the subspace affinity minimization enable the network to generalize to unseen noise distributions. Precisely quantify this behaviour of our framework allows a better understanding of the generalization of DNN based speech enhancement. We reverse this task for future works.

# 6. References


[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1929.

[2] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[3] ——, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

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

[5] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *Interspeech*, 2017.

[6] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, 2017.

[7] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," *ICASSP*, 2018.

[8] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," *Arxiv*, 2018.

[9] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," *ICASSP*, 2018.

[10] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," *Interspeech*, 2018.

[11] A. E. Bulut and K. Koishida, "Low-latency single channel speech enhancement using u-net convolutional neural networks," *ICASSP*, 2020.

[12] M. Soltanolkotabi, E. Elhamifar, and E. J. Cand'es, "Robust subspace clustering," *The annals of Statistics*, vol. 42, no. 2, pp. 669–699, 2014.

[13] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge: Cambridge University Press, 2012.

[14] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," *Interspeech*, 2016.

[15] "https://datashare.is.ed.ac.uk/handle/10283/2791."

[16] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," *Proc. Int. Conf. Oriental COCOSDA*, 2013.

[17] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 3591–3591, 2013.

[18] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," *CVPR*, 2016.

[19] J. Yao and A. Al-Dahle, "Coarse-to-fine optimization for speech enhancement," *Interspeech*, 2019.