IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 October 2023, accepted 28 October 2023, date of publication 6 November 2023, date of current version 10 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3330645

RESEARCH ARTICLE

Leveraging Sparse Approximation for Monaural Overlapped Speech Separation From Auditory Perspective

HIROSHI SEKIGUCHI, YOSHIAKI NARUSUE, (Member, IEEE), AND HIROYUKI MORIKAWA, (Member, IEEE)
Graduate School of Engineering, The University of Tokyo, Bunkyo, Tokyo 113-8654, Japan
Corresponding author: Hiroshi Sekiguchi (sekigh@mlab.t.u-tokyo.ac.jp)

ABSTRACT Neuroscience suggests that the sparse behavior of a neural population underlies the mechanisms of the auditory system for monaural overlapped speech separation. This study investigates leveraging sparse approximation to improve speech separation in a conventional deep learning algorithm. We develop a combined model that embeds a sparse approximation algorithm, a multilayered iterative soft thresholding algorithm (ML-ISTA), into a conventional time-domain-based speech separation algorithm, Conv-TasNet. Adopting ML-ISTA is a crucial enabler for the embedding process and helps avoid solving a bi-level optimization problem comprising sparse approximation and speech separation. ML-ISTA performs sparse approximation through forward calculations, thereby eliminating the optimization of sparse approximation. The combined model is trained with WSJ0-2mix, the Wall Street Journal English corpus for two-speaker mixed speech without noisy or reverberant interference, to clarify the proposed method's performance. The model demonstrates that sparse approximation improves separation performance regardless of the approximation setting. The peak performance of the model exceeds that of Conv-TasNet by 1.1% to 4.7% in four speech quality criteria. Moreover, sparse approximation accelerates the combined model performance gain at the early stages of learning relative to Conv-TasNet. The primary novelty of the study is embedding the sparse approximation algorithm, ML-ISTA, into a deep-learning-based speech separation framework and the experimental proof of improved separation performance in the proposed algorithm.

INDEX TERMS Deep learning, sparse approximation, sparsity, speech separation.

I. INTRODUCTION

Single-channel speech separation is crucial for restoring the quality and intelligibility of individual sources from overlapping speech in monaural recordings. This technology is the front end for speech recognition and sound scene analysis applications. Deep-learning-based speech separation has recently achieved remarkable performance among various engineering approaches owing to considerable advancements in image categorization. Conventional deep-learning-based speech separation technologies are reviewed in [1] and [2]. Reference [3] describes the comprehensive outlook of the state-of-the-art technologies for speech separation and presents a detailed performance comparison. However, these approaches have not yet reached the levels of separation capability of auditory systems. The challenge of the need for continuous improvement in the separation performance is pointed out in [4].

The auditory systems of mammals, particularly primates, can easily separate mixed sounds. This capability is demonstrated by the ''cocktail party problem,'' wherein researchers in several fields, including neuroscience, have collaborated to clarify the separation mechanism of the human auditory system from an auditory perspective. The mechanism for solving the ''cocktail party problem'' is analyzed from an auditory neuroscience point of view in [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

Neuroscience suggests that the mechanism of the human auditory system for distinguishing between temporally overlapping sounds is based on the neural population's sparse behavior. Sparse behavior and its merits in the auditory system are identified and presented in [6]. Sparse behavior is generated by the selectivity of the spectral-temporal receptive fields of auditory neurons to specific external stimuli. [7] reveals the spectral-temporal receptive fields of auditory neurons being adapted to natural sound statistics.

However, current state-of-the-art deep-learning-based speech separation algorithms do not focus on sparse behavior or its effect on separation performance. The algorithms concentrate solely on mapping overlapped speech to multiple separated speeches using neural networks and minimizing the difference between the target and separated speeches.

In contrast, speech enhancement, which is an application field relevant to speech separation, utilizes sparse behavior and demonstrates improvement in denoising performance. Sparse behavior produced by matching pursuit (MP) enforces a meaningful structure on the binary time-frequency mask, thus allowing to decrease estimation errors and maintain intelligibility [8]. The approach employing orthogonal matching pursuit (OMP) for producing sparse behavior outperforms the conventional methods in objective and subjective measures [9].

This study investigates sparse approximation's contribution to a deep-learning-based speech separation algorithm's separation performance. Sparse approximation is an optimization problem for representing sparse components in a latent space using external observations, and various methodologies for solving the sparse approximation problems are proposed [10], [11]. Sparse approximation is also employed to simulate sparse behaviors in the auditory systems. One example is the application of the sparse approximation to acquire auditory features in multilayered convolutional networks mimicking auditory systems [12].

We develop a combined model that embeds a sparse approximation algorithm into a deep-learning speech separation algorithm. We employ a multilayered iterative soft thresholding algorithm (ML-ISTA) [13] as a sparse approximation algorithm in conjunction with Conv-TasNet [14] as a conventional time-domain-based deep learning speech separation algorithm. In training and evaluation, we use the speech corpus WSJ0-2mix, which is the Wall Street Journal English version for the two-speaker mixed speech without additional noise or reverberant interference, to clarify the proposed method's performance. In experiments, the combined model's separation performance for various sparse approximation settings consistently surpasses that of Conv-TasNet. The novelty of this work is to propose embedding the sparse approximation algorithm, ML-ISTA, into a deep-learning-based speech separation framework and to experimentally demonstrate the improved performance of the proposed algorithm.

This study's prominent findings are summarized as follows:

1) We pioneer a model that embeds a sparse approximation algorithm into a deep-learning speech separation algorithm for better separation. The vital point in the embedding process is the adoption of ML-ISTA as a sparse approximation algorithm. This avoids solving a bi-level optimization problem comprising sparse approximation and speech separation. ML-ISTA performs sparse approximation through forward calculations, eliminating the optimization of sparse approximation. Furthermore, the calculation mechanism of ML-ISTA, with a strong affinity for Conv-TasNet's convolution-based encoder structure, facilitates end-to-end model training in a computationally efficient manner.
2) We reveal that the combined model achieves performance improvement of 1.1% to 4.7%, on average, over Conv-TasNet at higher sparseness in four quality criteria on the speech mixture dataset without noise or reverberate interference. These gains in performance are achieved by the increase in space and time complexity for ML-ISTA: an increase in space complexity, i.e., the model size, by 769 parameters (0.009% of Conv-TasNet's 8.64 million parameters) and in time complexity, i.e., the computational operation, by 0.36 giga floating point operations for a batch (0.7% per given iteration count of the Conv-TasNet's 55.34 giga floating point operations for a batch).
3) We uncover that ML-ISTA accelerates the combined model performance gain at the early stages of learning relative to Conv-TasNet.

The remainder of this paper is organized as follows. Section II describes previous work on sparse behavior in supervised deep-learning-based speech separation and neuroscience. Section III explains the construction of the combined model of ML-ISTA and Conv-TasNet. Section IV presents the experimental setting and evaluates the performance differences between the combined model and Conv-TasNet. The rational factors that support the obtained results are presented in Section V. In Section VI, we conclude that integrating ML-ISTA as a sparse approximation explicitly influences the separation performance in Conv-TasNet.

II. RELATED WORK

We first review state-of-the-art supervised deep-learning-based speech separation algorithms, including Conv-TasNet [14], which was state-of-the-art in the past, and describe their treatment of sparse behavior. Then, we describe neuroscience-based findings regarding sparse behavior and its role in auditory speech separation. Finally, we discuss sparse approximation.

Various supervised deep-learning-based speech separation approaches have been proposed, yielding significant performance improvements [4]. Reference [15] proposed a separation-mask method for speaker-dependent scenarios, where inference was conducted on mixed speech from only the speakers seen during training. Frame-wise or utterance-wise permutation invariant training (PIT) (μPIT) [16] based on a separation mask has been proposed to address speaker-independent scenarios. Such an approach can separate speech uttered by individuals who do not participate in training, thereby expanding the applicability of speech separation in practice. One method not based on separation masks is deep clustering [17] to identify segmentation in the bin space by speaker clusters. Reference [18] proposed the algorithm jointly solving inferring a representation for each source by deep clustering technique and estimating each source signal given the inferred representations. Conv-TasNet is an algorithm that receives the mixture waveform as input and uses the separation-mask method with speaker-independent scenarios supported by μPIT. Like the methods above, it achieved state-of-the-art separation performance in the past [4]. Dual-path RNN also uses the mixture waveform as input and adopts a double-cross RNN structure in the separation block [19]. Reference [20] explored two-stage processing, one simultaneous grouping and the other sequential grouping, which mimics the speech separation in the auditory system and achieves state-of-the-art performance. Reference [21] utilized a transformer in speech separation, showing that the attention mechanism is effective in dealing with longer temporal dependency in speech separation. Other studies deal with the circumstances under noise and reverberation [22], [23], [24]. Reference [22] deals with noisy and reverberant speech separation by estimating a room impulse response. Reference [23] used robust principal component analysis and sparse nonnegative matrix factorization for reverberant speech separation. Reference [24] applies a diffusion-based generative technology to separate a mixture of reverberant speech.

However, these algorithms, including Conv-TasNet, do not consider the effect of the sparse behavior of hidden layer components on separation performance.

Neuroscience [6], [25] considers sparse behavior in sensory mechanisms. Sparse behavior is a state of sparse representation of neural codes [26], [27]. From a neurophysiological perspective, neural codes dictate the sensory neurons' transformation of an external stimulus from sensory receptors to the central brain for action. Sparse behavior indicates that few neurons simultaneously activate in response to an external stimulus at any given time. Sparse behavior is generated by the selectivity of neurons in response to specific external stimuli, causing sparse excitation in their physical organization.

Sparse behavior also occurs in the auditory system. The primary function of the auditory system is to distinguish between temporally overlapping sounds [5]. A practical implementation of auditory sparse excitation at the neural level is generated using the spectral-temporal receptive field characteristics of auditory neurons along the auditory pathway.

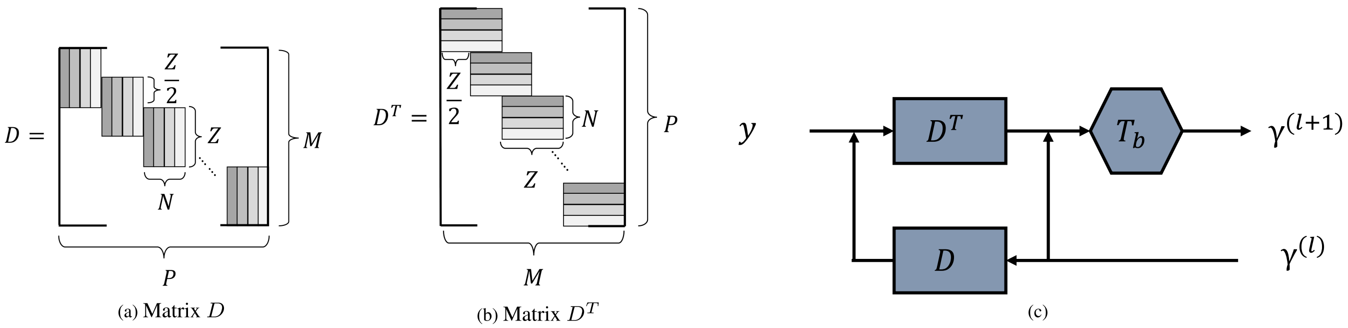
Sparse approximation, or sparse coding, is a promising method for simulating sparse behaviors at different stages of the auditory pathway [12], [28], [29], [30]. It seeks a small number of nonzero latent variables to represent a given observation when the observation space is mapped to a latent variable space with a given linear relationship [11]. Reference [12] used sparse coding to obtain auditory features in multilayered convolutional networks simulating an auditory system. Reference [28] employed binary sparse coding to estimate auditory features in constructing an auditory model between speech embedded in natural sound and the functional magnetic resonance imaging (fMRI) signals from the auditory cortex. Sparse coding has also been used to learn sparse features in midlevel auditory representations [29] and to simulate sparse features on the outputs of cochlear filter banks [30]. The latent variable space is often the time-feature domain, wherein a few nonzero features represent sparse behavior at any time.

III. COMBINING CONV-TASNET AND ML-ISTA

We first present an overview of Conv-TasNet and ML-ISTA. We review the rationale for selecting ML-ISTA as a solution to the sparse approximation in the combined model. Thereafter, we explain how ML-ISTA is embedded into Conv-TasNet to construct a combined model. Finally, we describe the optimization problem and how to acquire the combined model by solving the problem.

First, we describe the configuration of Conv-TasNet. This monaural supervised deep-learning speech separation algorithm consists of the encoder, separator, and decoder with a loss function of μ PIT-based scale-invariant signal-to-noise ratio (SI-SNR) between the estimated and target speech signals. It achieved state-of-the-art performance in the past. The input signal to the model is a mixture of time-series speech signals, and the model outputs are the estimated separated speech signals. The encoder consists of multiple one-dimensional (1-D) filters in a single layer and transforms its mixture input into a latent space. The separator consists of multiple blocks of dilated convolutions and estimates a mask for each speaker, which is applied multiplicatively to the latent representations at the encoder outputs to produce the estimated separated speech in the latent space. The decoder is a linear filter with overlap-and-add operation for the concatenation of speech signals and reverts the estimated latent signals to the time domain.

Second, we review ML-ISTA. ML-ISTA has recently been proposed as an algorithm for solving the least absolute shrinkage and selection operator (LASSO) problem, an ℓ_1 regularization problem for sparse approximation. Algorithms for LASSO constitute a well-studied, computationally efficient category of solutions for sparse approximation, compared to greedy algorithms such as MP and OMP, which solve an ℓ_0 regularization problem [11]. ISTA, which is the basis for ML-ISTA, has mainly been employed in various image applications such as image denoising and image reconstruction of magnetic resonance (MR) modality in [31]



(a) Matrix D (b) Matrix D^T (c)

FIGURE 1. **(a) Matrix D is a union of vertically banded matrices to which N local one-dimensional (1-D) convolutional filters with filter length Z and filter shift $Z/2$ are reformulated. Segments with the same shade degree represent 1-D convolutional filters with the same coefficients. (b) Matrix D^T is the transpose of D. (c) The ℓth recursive forwarding calculation on the single-layer recurrent convolutional neural network (RCNN) is equivalent to (2) with $\gamma^{(0)} = 0$. D and D^T undergo matrix multiplications, and T_b is a rectified linear unit (ReLU) activation function with bias b.**

and [32]. The former uses ISTA for the restoration of MR images. The latter uses ISTA to shorten data acquisition time to ease patient fatigue and mitigate target movement during MR scans. Our work is the first time that ML-ISTA has been employed for speech separation.

ML-ISTA is a variant of multilayered convolutional sparse coding [33], which is a sparse approximation method that maps the input to the output as a series of matrix multiplications with formatted convolution matrices. For simplicity, the authors of [13] derived ML-ISTA using a two-layer convolutional sparse coding case as an example. In contrast, we proceed with our discussion by concentrating on a single-layer case because a single-layer network is the encoder of Conv-TasNet, into which ML-ISTA is embedded.

In the single-layer case, the true sparse solution $\boldsymbol{\gamma}^*$ is obtained in the minimization problem as follows:

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}} \frac{1}{2}\|\boldsymbol{y}-D\boldsymbol{\gamma}\|_2^2 + \lambda\|\boldsymbol{\gamma}\|_1, \tag{1}$$

where the given observation $\boldsymbol{y} \in \mathbb{R}^M$ and its sparse latent representations $\boldsymbol{\gamma} \in \mathbb{R}^P$ form a linear combination with a matrix $D \in \mathbb{R}^{M\times P}$. One example of D has a union of vertically banded matrices to which N local 1-D convolutional filters with length Z and shift $Z/2$ are reformulated. The matrix D is shown in Fig. 1(a). $\| * \|_2^2$ and $\| * \|_1$ are the squared ℓ_2-norm and ℓ_1-norm, respectively, of the term $*$, and λ is a Lagrange multiplier. The first term on the right-hand side of (1) is the reconstruction error against $\boldsymbol{y}$ and the second term is the ℓ_1-norm of $\boldsymbol{\gamma}$. The norm serves as a sparse regularization for the solution of $\boldsymbol{\gamma}$. Therefore, the obtained solution $\boldsymbol{\gamma}^*$ is the optimal sparse representation of the input signal $\boldsymbol{y}$.

ML-ISTA determines $\boldsymbol{\gamma}^{(L)}$ for a given total iteration count of L as the solution to the aforementioned minimization problem, which approximates the true $\boldsymbol{\gamma}^*$ as follows:

$$\boldsymbol{\gamma}^{(l+1)} = \mathrm{ReLU}(\boldsymbol{\gamma}^{(l)} - D^T(D\boldsymbol{\gamma}^{(l)} - \boldsymbol{y}) + \boldsymbol{b}), \tag{2}$$

$$\boldsymbol{\gamma}^{(0)} = \boldsymbol{0}, \tag{3}$$

where $\boldsymbol{\gamma}^{(l+1)}, l = 0, \cdots, L-1$ is the latent representation at the $(l+1)$th iteration. The rectified linear unit (ReLU) with bias $\boldsymbol{b}$, corresponding to λ, is used to add a nonnegative constraint to the sparse representation $\boldsymbol{\gamma}$. $\boldsymbol{0}$ is a vector with all components zero. Note that $\boldsymbol{\gamma}^{(l+1)}$ can be obtained in (2) once $\boldsymbol{\gamma}^{(l)}$ is known. Thus, L iterations of the equation yield $\boldsymbol{\gamma}^{(L)}$ for a given $\boldsymbol{y}$. Furthermore, (2) is equivalent to a forwarding calculation consisting of a pair of a forward pass with D^T, the transpose of D in Fig. 1(b), and a backward pass with D on the recurrent convolutional neural network (RCNN), as depicted in Fig. 1(c). Therefore, $\boldsymbol{\gamma}^{(L)}$ can be obtained using the L-time recursive forwarding calculation of the RCNN.

ML-ISTA is selected because it exhibits the following properties: a capability of approximating solutions to an optimization problem via forwarding calculations, a strong affinity for the convolution-based encoder structure of Conv-TasNet, technically guaranteed convergence, guaranteed solution uniqueness, and a high convergence speed. These properties are favorable for solving high-dimensional sparse approximation problems, including speech separation.

To embed ML-ISTA into Conv-TasNet and obtain a combined model, the RCNN of ML-ISTA in the one-layer case replaces the original encoder of Conv-TasNet, which consists of multiple 1-D convolutional filters in a single layer, whereas the two networks of Conv-TasNet, that is, the separator and decoder, remain unchanged. This replacement is efficient because ML-ISTA has the same convolutional neural network (CNN) basis as the encoder of Conv-TasNet.

In a precise sense, a matrix $\gamma^{(L)}_{\text{matrix}}$, which is reformatted from the RCNN output vector $\boldsymbol{\gamma}^{(L)}$ as

$$\gamma^{(L)}_{\text{matrix}} = \mathcal{P}_{\mathcal{N}}(\boldsymbol{\gamma}^{(L)}) \in \mathbb{R}^{N\times K}, \tag{4}$$

replaces the output of the Conv-TasNet encoder. Here, $\boldsymbol{\gamma}^{(L)} \in \mathbb{R}^{NK}$ is a vector obtained using (2) L times. $\alpha\beta$ denotes the multiplication of scalars α and β, and $\mathbb{R}^{\alpha\times\beta}$ denotes a real matrix with α dimensions on one axis and β dimensions on the other axis. $\mathcal{P}_{\mathcal{N}}(x)$ is a reformatting operator over vector $\boldsymbol{x}$ to a matrix such that $\{\mathcal{P}_{\mathcal{N}}(x)\}_{i,j} = x_{i+N(j-1)}, i = 1, \cdots, N, j = 1, \cdots, K$. N represents the number of 1-D convolutional filters in the original Conv-TasNet encoder, and K represents the number of overlapping segments of the mixed signal $\boldsymbol{y}$ with length Z and shift $Z/2$. Z is the length of each 1-D convolutional filter in the original Conv-TasNet encoder. The mixed signal $\boldsymbol{y}(t) \in \mathbb{R}^M, t = 1, \cdots, M$ can be denoted as the summation of the speech signals of C speakers $\boldsymbol{s}_c(t), t = 1, \cdots, M, c = 1, \cdots, C$ such that $\boldsymbol{y}(t) = \sum_{c=1}^{C} \boldsymbol{s}_c(t)$.

Finally, to acquire the combined model, the following optimization problem is defined and solved with respect to D, $\boldsymbol{b}$, Sep, and Dec:

$$\max_{D,\boldsymbol{b},\text{Sep, Dec}} \frac{1}{C} \sum_{c=1}^{C} \text{SI-SNR}^{(c)}(\hat{\boldsymbol{s}}_c(\gamma_{\text{matrix}}^{(L)}), \boldsymbol{s}_c), \tag{5}$$

where $\text{SI-SNR}^{(c)}(\hat{\boldsymbol{s}}_c(\gamma_{\text{matrix}}^{(L)}), \boldsymbol{s}_c)$ denotes speaker c's part of the scale-invariant signal-to-noise ratio (SI-SNR) loss function, and Sep and Dec denote the network parameters of the separator and decoder, respectively.

The role of $\text{SI-SNR}^{(c)}(\hat{\boldsymbol{s}}_c(\gamma_{\text{matrix}}^{(L)}), \boldsymbol{s}_c)$ as a function of $\gamma_{\text{matrix}}^{(L)}$ stems from the fact that all of the variables Q, Q_c, $\hat{S}_c$, $\hat{\boldsymbol{s}}_c$, and $\text{SI-SNR}^{(c)}$ can be derived from $\gamma_{\text{matrix}}^{(L)}$. We explain the derivation below. $\text{SI-SNR}^{(c)}(\hat{\boldsymbol{s}}_c(\gamma_{\text{matrix}}^{(L)}), \boldsymbol{s}_c)$ can be calculated as

$$\text{SI-SNR}^{(c)}(\hat{\boldsymbol{s}}_c(\gamma_{\text{matrix}}^{(L)}), \boldsymbol{s}_c) \triangleq 10 \log_{10} \frac{\|\boldsymbol{s}_{\text{target}}^{(c)}\|^2}{\|\boldsymbol{e}_{\text{noise}}^{(c)}\|^2}, \tag{6}$$

$$\boldsymbol{s}_{\text{target}}^{(c)} \triangleq \frac{< \hat{\boldsymbol{s}}_c(\gamma_{\text{matrix}}^{(L)}), \boldsymbol{s}_c > \boldsymbol{s}_c}{\|\boldsymbol{s}_c\|^2}, \tag{7}$$

$$\boldsymbol{e}_{\text{noise}}^{(c)} \triangleq \hat{\boldsymbol{s}}_c(\gamma_{\text{matrix}}^{(L)}) - \boldsymbol{s}_{\text{target}}^{(c)}, \tag{8}$$

where $< \alpha, \beta >$ is the inner product of the two vectors α and β. Here, $\hat{\boldsymbol{s}}_c(\gamma_{\text{matrix}}^{(L)})$ is the separated individual representation for speaker c in the time domain as estimated by the combined model. It results from an overlap-and-add operation on K segments of a Z-long-sequence with a shift $Z/2$ of $V\hat{S}_c(\gamma_{\text{matrix}}^{(L)}) \in \mathbb{R}^{Z \times K}$, where $V \in \mathbb{R}^{Z \times N}$ is the multiplication matrix of the decoder. In turn, $\hat{S}_c(\gamma_{\text{matrix}}^{(L)})$ denotes the separated individual representations for speaker c in the latent space, which is calculated as $\hat{S}_c = \gamma_{\text{matrix}}^{(L)} \odot Q_c \in \mathbb{R}^{N \times K}$. Here, Q_c represents the mask pertinent to speaker c and is defined as $\{Q_c\}_{n',k'} = \{Q\}_{c'=c,n',k'} \in \mathbb{R}^{N \times K}$. $Q = \text{Sep}(\gamma_{\text{matrix}}^{(L)})$ is the output of the separator, which is the full separation mask and a function of $\gamma_{\text{matrix}}^{(L)}$. $\odot$ denotes element-wise multiplication. Therefore, $\hat{S}_c(\gamma_{\text{matrix}}^{(L)})$ is a function of $\gamma_{\text{matrix}}^{(L)}$. Consequently, all of these variables Q, Q_c, $\hat{S}_c$, $\hat{\boldsymbol{s}}_c$, and $\text{SI-SNR}^{(c)}$ can be derived from $\gamma_{\text{matrix}}^{(L)}$.

Equation (5) is regarded as optimizing a single speech separation problem, where all model parameters, including D, $\boldsymbol{b}$, Sep, and Dec, can be updated and learned via back-propagation in an end-to-end manner. The single problem is constructed by embedding the ML-ISTA forwarding calculation (2) into (5) through (4), thereby inserting the updated values of D and $\boldsymbol{b}$ into the back-propagated loss function. The original problem to be solved is a bi-level optimization problem that consists of (5) and (1). The adoption of ML-ISTA avoids solving the bi-level problem and solves the single problem of (5) by approximating the solution of the second minimization problem of (1) by L iterations of (2), which is equivalent to the L-time recursive forwarding calculation in the RCNN.

The number of updates for D in the training of the combined model is the same as that of Conv-TasNet because ML-ISTA purely involves the network forwarding calculation with a fixed D. D is updated only when the loss function is updated. Therefore, the performance comparison between the combined model and Conv-TasNet is fair with regard to the number of updates for D.

IV. EXPERIMENTS

A. DATASET

The WSJ0-2mix English corpus [17] is used for training, validation, and evaluation. In this study, the WSJ0 corpus, containing speech signals recorded under different conditions on microphone properties and settings (WV1 and WV2), is used for enjoying the variety of speech quality out of WSJ0. 30-hour training and 10-hour validation datasets are created from the WSJ0 data under the si_tr_s directory. Two utterances that are randomly selected from different speakers are mixed with a random SNR that varies from −5 to +5 dB to create mixed speech. A five-hour evaluation dataset is similarly compiled using utterances from 16 unseen speakers under the si_dt_05 and si_et_05 directories. Thereafter, all speech signals are converted to 8 kHz and are fed to the model.

B. EXPERIMENTAL PROCEDURE

We train Conv-TasNet as a baseline model and train the combined models for comparison to investigate the quantitative effect of sparse approximation on separation performance. We evaluate and compare the separation performance and sparsity of the models.

1) TRAINING PROCEDURE

We confirm that the hyperparameter settings of the baseline model [1] are optimal for speech performance. In particular, we confirm the best hyperparameter settings for the number of convolutional filters and the size of the convolutional kernel in the encoder by grid search. The actual parameters for training are as follows: the number of convolutional filters in the encoder $N = 256$, the convolutional filter kernel size $Z = 20$ with shift 10, the number of channels to the separator $B = 256$, the inner channels in the dilated-convolution $H = 512$, dilated-convolution filter size $P_{\text{dilated}} = 3$, the number of blocks of dilated-convolution X=8, and the number of repetitions of X-dilated-blocks $R = 4$. The optimizer is Adams with an initial learning rate l_r=0.001 and the learning scheduler with a learning rate decreasing by half once three consecutive epochs show no loss reduction. The batch is 3. The number of workers in dataloader is 4.

[1] The code of the baseline model used is downloaded from https://github.com/kaituoxu/Conv-TasNet

The combined model with the given iteration count L for ML-ISTA is trained to maximize the SI-SNR of the target speech for all network parameters, including those of ML-ISTA. The iteration count L is initially regarded as a sparse approximation setting related to the level of convergence. It also specifies the number of recursive forwarding calculations in the RCNN in practical network implementations. During training, all hyperparameters except the iteration count for the combined model are the same as those for the baseline model. We train the combined model over 100 epochs on the training dataset and validate the model improvement in the objective function every epoch on the validation dataset. Moreover, we generate intermediate model snapshots after every ten epochs. This training yields the best model with optimal weights for all networks, including the bias of the ReLU in ML-ISTA for maximizing the objective function after every ten epochs for a given iteration count. We train the combined models with different sparsity levels by varying the iteration counts among 3, 6, 9, and 12 to examine the difference in separation performance.

2) EVALUATION PROCEDURE

We employ SI-SNR improvement (SI-SNRi) as a primary quality criterion for separated speech to indicate the separation performance because it is related to the loss function to optimize training. In addition, we use three complementary criteria: signal-to-distortion ratio improvement (SDRi), perceptual evaluation of speech quality (PESQ), and short-time objective intelligibility (STOI). SI-SNRi and SDRi represent improved differences in SI-SNR and SDR from mixed speech. Each criterion is measured by the average over-per-utterance metrics produced by 2616 utterances composing the evaluation dataset.

We use population sparseness S_{pop} as the sparsity measure. This metric measures the sparsity of brain neuron activity in neuroscience [34], [35] and is defined as $S_{pop} = E_j[S_{pop,j}]$, where $E_j[*]$ represents the expectation for stimuli and $S_{pop,j}$ is formulated as

$$S_{pop,j} = 1 - \frac{(E_i[\|r_{i,j}\|])^2}{E_i[r_{i,j}^2]}. \tag{9}$$

Here, $S_{pop,j}$ represents the population sparseness to stimulus j ranging from zero to one. $r_{i,j}$ ranges between $(-\infty, \infty)$ and denotes the response of neuron i in the population to stimulus j. $E_i[*]$ represents the expectation for the neurons. In our experiment, $r_{i,j}$ represents the encoder output $\{\gamma_{\text{matrix}}^{(L)}\}_{i,j}$ within the range $[0, \infty)$. $S_{pop,j}$ describes the inequality in the distribution of $\{\gamma_{\text{matrix}}^{(L)}\}_{i,j}$ along neuron i for a given stimulus j. A higher inequality leads to $S_{pop,j}$ increasing to one, whereas a lower inequality leads to $S_{pop,j}$ decreasing to zero. $S_{pop,j}$ also corresponds to the degree of nonzero N-dimensional encoder output as a population activated for a given stimulus j. S_{pop} indicates the average of $S_{pop,j}$ for all utterances in the evaluation dataset.

The relationship between sparseness and quality criteria is investigated from pairs obtained for each model using the evaluation dataset. We use Conv-TasNet as the baseline model for comparison with the combined models. The baseline model exhibits sparseness owing to the implicit sparsity incorporated by ReLU. Therefore, a pair of sparseness and quality criteria for the baseline model can also be calculated for comparison. We examine whether the combined models achieve higher performance and sparseness than the baseline model.

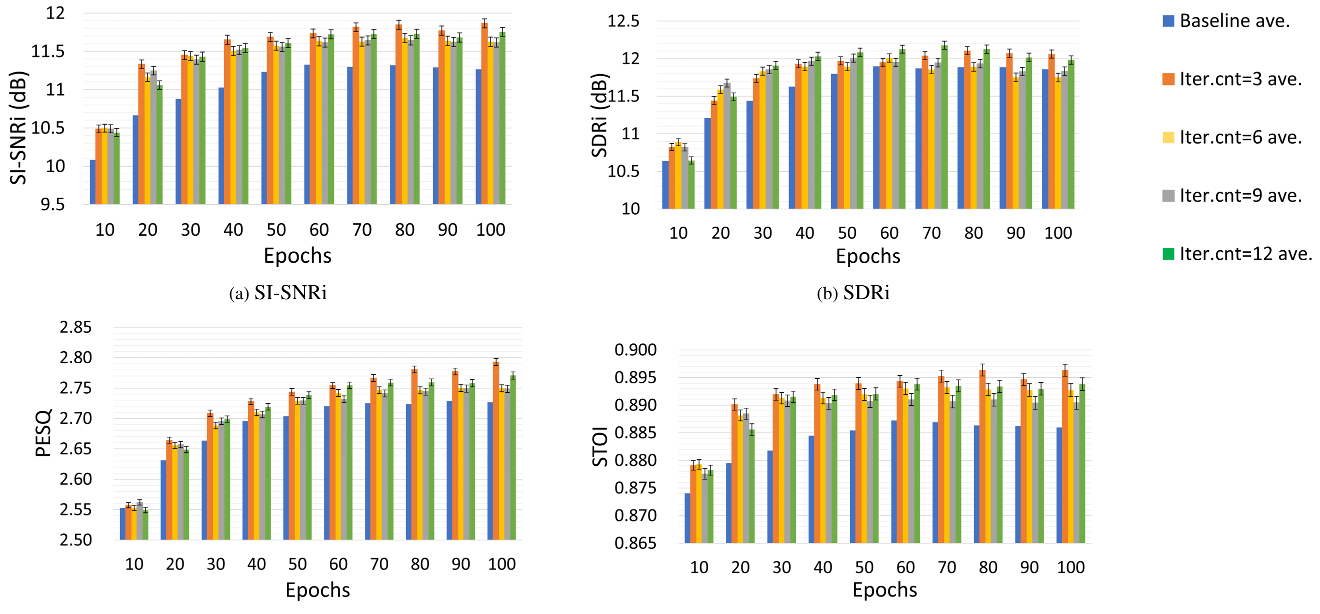
C. RESULTS

1) SEPARATION PERFORMANCE VS. NUMBER OF EPOCHS

We first evaluate the speech separation performance trend for the first 100 epochs in the baseline model, and four combined models with iteration counts of 3, 6, 9, and 12 for sparse approximation settings. We train and validate the five models with five distinct training runs with random weight initializations to generate deviation among the runs. Thus, we train 25 networks over 100 epochs, with network snapshots obtained every 10 epochs. We evaluate the separation performance of the 25 networks on the evaluation dataset using the four quality criteria every 10 epochs.

Comparisons of averages over five training runs for the baseline model and four combined models in all epochs for the four separation speech quality criteria are presented in Fig. 2(a), Fig. 2(b), Fig. 2(c), and Fig. 2(d). All combined models converge toward epoch 100. They are superior to the baseline model in all epochs and for all four criteria except SDRi, with model iteration counts 6 and 9 at epochs 90 and 100. The peak performance of all models is not necessarily achieved by epoch 100 because performance saturation or over-fitting occurs midway through training, with peak performance attained at that point.

Paired t-tests between two groups, the combined model and a baseline model, are conducted to show two statistical metrics, Cohen's d [36] and 95% confidence intervals of the averaged performance difference between them. Cohen's d, a standardized effect size metric, measures the distance between the two distributions. Farther distributions lead to the metric increasing to one and closer distributions lead to the metric decreasing to zero. For example, Fig. 3 shows Cohen's d metrics for SI-SNRi, indicating the distance between the distributions of the combined model with each iteration count of 3, 6, 9, and 12 and the baseline model at the same epoch. The combined model with all iteration counts shows Cohen's d values in the range of 0.06 to 0.1 at all epochs in SI-SNRi. The 95% confidence interval is illustrated as a vertical black line at the top of each combined model's performance bar in the subfigures of Fig. 2. The differences in performance between the combined and baseline models in SI-SNRi, PESQ, and STOI are distinct over all epochs except epoch 10 of PESQ, as there is no overlap between 95% confidence intervals and the average baseline performance regardless of the iteration counts. In comparison, differences



(a) SI-SNRi

(b) SDRi

(c) PESQ

(d) STOI

FIGURE 2. Four averaged levels of performance with 95% confidence intervals vs. epochs for the baseline model and four combined models with iteration counts (iter.cnts) of 3, 6, 9, and 12. The vertical black lines indicate 95% confidence intervals.

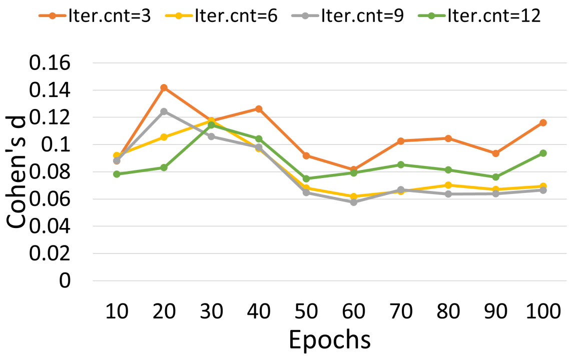


FIGURE 3. Cohen's d effect size for SI-SNRi, indicating how far two distributions of each iteration count of 3,6,9, and 12 and the baseline is (far:1, close:0).

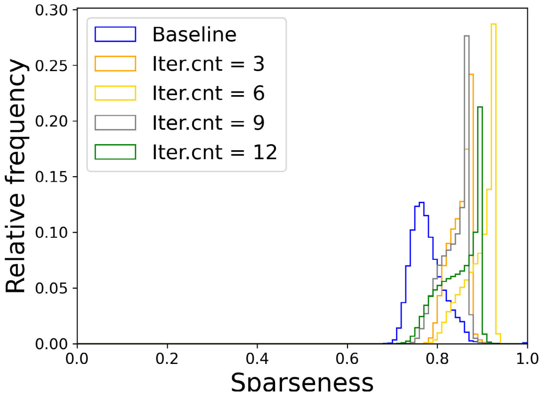


FIGURE 4. Typical examples of population sparseness histograms for five models with iteration counts (iter.cnts) of 3, 6, 9, and 12 at epochs exhibiting their highest performance.

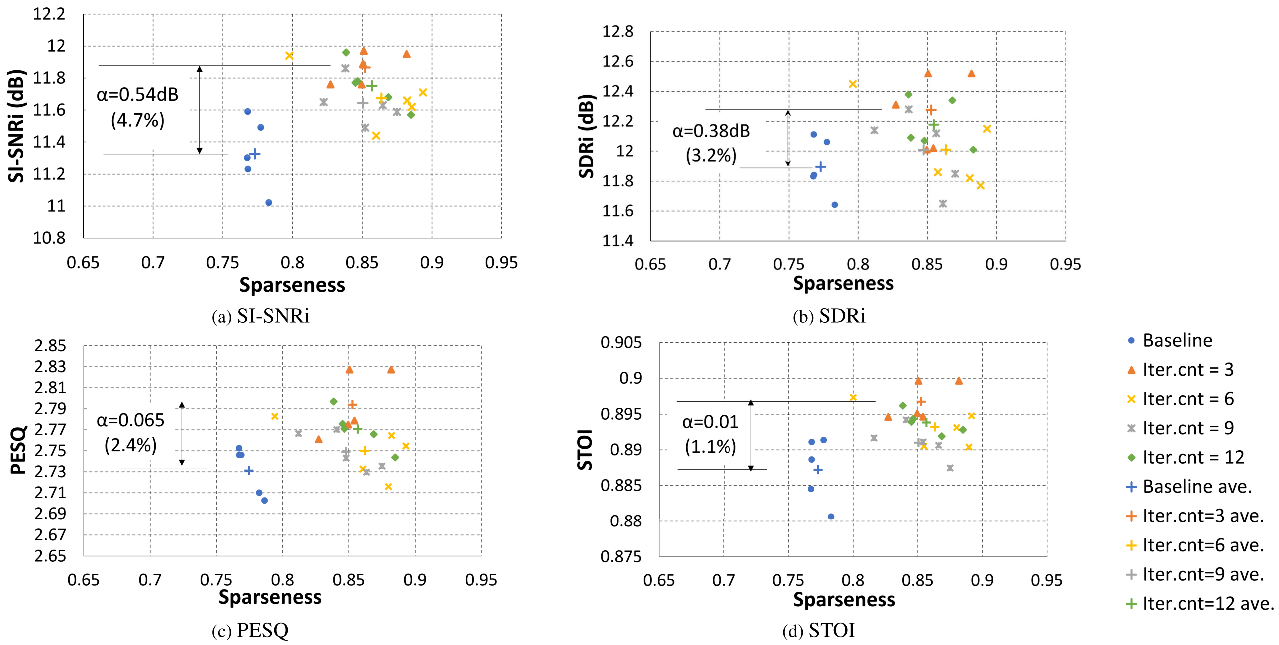
in SDRi are less prominent owing to several overlaps. The combined model with an iteration count of three exhibits the best performance for all four quality criteria in the later stages of training.

2) RELATIONSHIP BETWEEN SPARSITY AND SEPARATION PERFORMANCE

We analyze the contribution of sparse approximation to performance based on a peak-to-peak performance comparison. We determine the peak performance levels over all epochs for each of the five models and compare them using a scatterplot for each quality criterion. The aim is to independently identify the separation performance potential of the combined models for the four criteria.

We collect the encoder output associated with peak performance for each model for each quality criterion and calculate the population sparseness S_{pop} in the output using the method described in IV-B2. Examples of histograms for $S_{pop,j}$ as expressed in (9) over the evaluation dataset for the five models are shown in Fig. 4. The figure shows that the distributions of population sparseness for the combined models are situated at higher positions than that for the baseline model, regardless of sparse approximation settings, that is, iteration counts.

Four sparseness vs. performance scatterplots for all five models are presented in Fig. 5(a), Fig. 5(b), Fig. 5(c), and Fig. 5(d), where the plot for each model depicts the results for five training runs and their average as dots.



(a) SI-SNRi

(b) SDRi

(c) PESQ

(d) STOI

FIGURE 5. Sparseness vs. performance plots for five training runs of the baseline and four combined models, along with their averages over iteration counts at epochs with the best performance. α indicates the most significant average improvement over the baseline model among all combined models. α values indicate the differences between the baseline and combined models for all four quality criteria for an iteration count three.

We consider two evaluation perspectives: groups of dot distributions and one model's average performance. Furthermore, two sub-perspectives are considered under each view, one along the sparseness axis and the other along the performance axis.

For the dot distribution groups, we observe one for the baseline model and the other for the four combined models for the four quality criteria. The latter group is consistently distributed within a higher sparseness than the baseline for all four quality criteria along the sparseness axis. Most dots in the group for the four combined models are at higher performance levels than those in the baseline model along the performance axis for SI-SNRi and STOI. However, some dots for the four combined models overlap with those in the baseline model group for SDRi and PESQ. Thus, the combined models exhibit more distinct differences from the baseline model in SI-SNRi and STOI than in SDRi and PESQ. We suspect that the distinct differences in SI-SNRi stem from its role as an objective training function.

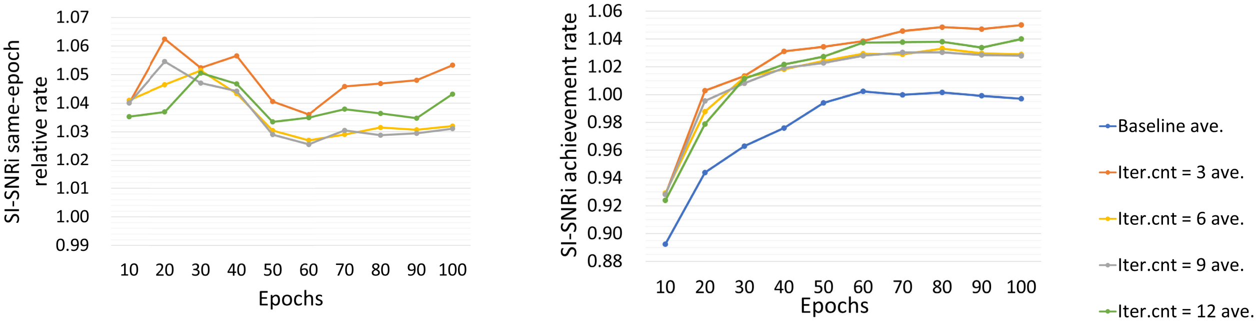
According to one model's average performance, all averages of the combined models exhibit higher peak performance at higher sparseness than the baseline model in all four quality criteria. The best peak performance improvement over the baseline along the performance axis is α for each criterion in the sub-figure of Fig. 5. All α values are achieved by the combined model for the iteration count of three, with 0.54 dB (4.7%) for SI-SNRi, 0.38 dB (3.2%) for SDRi, 0.065 (2.4%) for PESQ, and 0.01 (1.1%) for STOI. The average sparseness values of all four combined models consistently exceed those of the baseline model along the sparseness axis, irrespective of quality criterion. This implies that sparse approximation produces higher sparseness than the baseline model.

V. DISCUSSION

We discuss four issues in this section: the acceleration of the performance gain in the early stages of learning, the mechanism of sparse approximation that promotes improved separation, the penalty of space and time complexity in the combined models when compared with the baseline model, and the result of an ℓ_1-regularized model as an alternative sparse implementation.

For the performance gain acceleration, we demonstrate that sparse approximation facilitates the acceleration. We investigate the same-epoch relative rate for SI-SNRi as a function of epochs. The rate is defined as the ratio of the averaged performance of the combined model to that of the baseline model at the same epoch. A rate of more than one indicates a superior performance gain of the combined model over that of the baseline during the same training epochs. A higher rate indicates a faster acceleration in the performance gain.

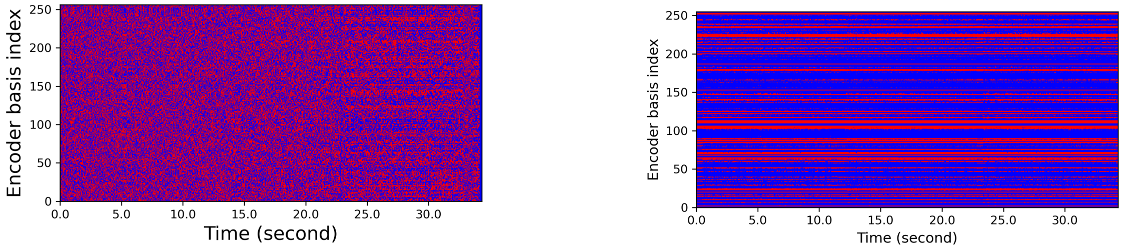
Fig. 6(a) presents the same-epoch relative rate for SI-SNRi, derived from the outcomes in Fig. 2 (a). The SI-SNRi rate exceeds one for all epochs up to 100, decreasing with increased epochs. For SI-SNRi, all combined models exhibit higher rates in the early epochs, from 20 to 50, than in the later epochs, from 60 to 100. This implies that the effect of sparse approximation emerges strongly as an accelerator in the performance gain during the early training stages and continues to influence the separation improvement in the later



(a) Same-epoch relative rate for SI-SNRi

(b) Achievement rate for SI-SNRi

FIGURE 6. Acceleration in performance gain during the early stages of learning for SI-SNRi: (a) same-epoch relative rate for SI-SNRi against baseline performance and (b) achievement rate for SI-SNRi against saturated performance of baseline model, averaged over epochs 60 to 100.



(a) Baseline model

(b) Combined model with an iteration count of three (iter.cnt=3)

FIGURE 7. Comparison in intensity heatmaps on encoder outputs for a 34-second utterance. Blue and red dots represent zero and nonzero data: (a) Baseline model shows scattered intensity among encoder outputs. (b) The combined model with an iteration count of three shows encoder bases with differently weighted intensity.

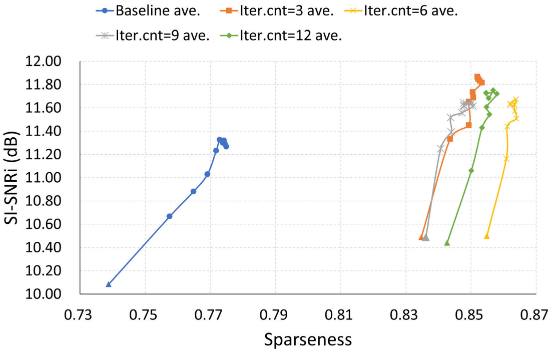


FIGURE 8. Sparseness vs. SI-SNRi trend as training updates every 10 epochs for the baseline and combined models with iteration counts (iter.cnts) of 3, 6, 9, and 12. Each line represents plots of the averaged sparseness and SI-SNRi over five runs per model. The triangle symbol indicates the first data at epoch 10 for each model. All combined models exhibit higher SI-SNRi at higher sparseness than the baseline.

training stages. The advantage of improved acceleration is observed at epoch 30 for SI-SNRi, as illustrated in Fig. 6(b). All combined models surpass the saturated performance of the baseline averaged over epochs from 60 to 100.

The mechanism of sparse approximation is the extraction of fewer nonzero encoder data, which facilitates speech separation compared with the baseline. For the models in Fig. 4, the number of nonzero encoder data at each time frame of the latent space in the combined model on the evaluation dataset is always smaller than those at the corresponding time frame in the baseline for the same speech input, regardless of iteration counts. Compared to the baseline, the average reduction ratios over all time frames are 36.3%, 45.4%, 29.6%, and 23.1% for iteration counts of 3, 6, 9, and 12, respectively. Thus, the combined models' nonzero encoder data are reduced by an average of 33.6% over all time frames and iteration counts. The combined models have 84.9 nonzero encoder data out of a total of 256 encoder data on average over all time frames and iteration counts, whereas the baseline has 128.0 nonzero encoder data. Nevertheless, all the combined models achieve a better separation performance than the baseline model. The observation validates that fewer nonzero encoder data extracted by sparse approximation facilitates speech separation.

Fig. 7 demonstrates the qualitative comparisons concerning the output of the encoder for a 34-second utterance, for the two models: one from the baseline, the other from the combined model with an iteration count of three. Blue and red colors are for zero and nonzero data, respectively. The baseline shows scattered activation, whereas the ML-ISTA model displays bases with differently weighted intensity. The plots show sparser encoder outputs in the ML-ISTA embedded model than in the baseline. Fig. 8 shows the trend of the learning in sparseness vs. SI-SNRi performance on the baseline and combined models with iteration counts of 3, 6, 9, and 12 as the epochs increase. Each line represents

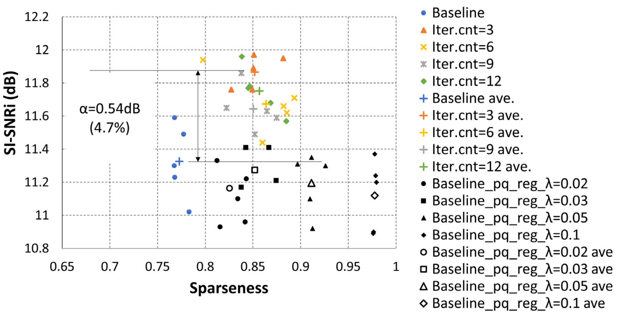


FIGURE 9. Sparseness vs. performance plots for five training runs of ℓ_1-regularized models, combined models, and the baseline. ℓ_1-regularized models gain sparseness but decrease the performance compared with the baseline as the λ increases.

plots of the averaged sparseness and SI-SNRi over five runs per model. The plots indicate that the combined models simultaneously learn sparseness and performance in SI-SNRi and achieve better performance than the baseline model at higher sparseness.

The requirement in space complexity for the ML-ISTA is 769 parameters, which is 0.009% of the 8.64 million parameters of the baseline model. Moreover, the requirement in time complexity for the ML-ISTA is 0.36 giga floating point operations per each iteration count for a batch, which is 0.7% of the 55.34 giga floating point operations for the baseline model. Thus, the penalties in the space and time complexity are small compared with the improvement in the performance by 1.1%–4.7%.

We conducted the imposition of an ℓ_1-regularization in the activation space on training loss, a straightforward sparse implementation before the proposed method. The result showed no improvement in separation performance compared with the baseline. This method is not effective in sparse implementation for facilitating separation. We share the experimental results below for interested readers.

The ℓ_1-regularized model is trained with a loss function containing the smoothed ℓ_p-over-ℓ_q(SPOQ) as an ℓ_1-regularization which is a good indicator for ℓ_1 sparsity with the existence of a derivative at input zero [37]. Here, all other hyperparameter settings of the model are the same as those of the baseline. ℓ_1-regularized model is acquired by minimization of the loss function as:

$$\min_{\text{Enc,Sep,Dec}} -\frac{1}{C}\sum_{c=1}^{C} \text{SI-SNR}^{(c)}(\hat{\boldsymbol{s}}_c(\boldsymbol{\gamma}), \boldsymbol{s}_c) + \lambda \cdot \Psi(\boldsymbol{\gamma}), \tag{10}$$

$$\Psi(\boldsymbol{x}) = \log\left(\frac{\left(\ell_{p,\alpha}^{p}(\boldsymbol{x}) + \beta^{p}\right)^{\frac{1}{p}}}{\ell_{q,\eta}(\boldsymbol{x})}\right), \tag{11}$$

$$\ell_{p,\alpha}(\boldsymbol{x}) = \left(\sum_{n=1}^{N}\left((x_n^2 + \alpha^2)^{\frac{p}{2}} - \alpha^p\right)\right)^{\frac{1}{p}}, \tag{12}$$

$$\ell_{q,\eta}(\boldsymbol{x}) = \left(\eta^q + \sum_{n=1}^{N}|x_n|^q\right)^{\frac{1}{q}}, \tag{13}$$

where Enc denotes the encoder's network parameters of Conv-TasNet, and $\boldsymbol{\gamma}$ is the encoder output. $\hat{\boldsymbol{s}}_c(\boldsymbol{\gamma})$ is the estimated speaker c's speech of Conv-TasNet, and λ is the Lagrange multiplier. $p = 0.75$, $q = 2.0$, $\alpha = 7e^{-7}$, $\beta = 3e^{-3}$, and $\eta = 0.1$ are hyperparameters used for guaranteed global minimum. $x_n, n = 1, 2, \cdots, N$ is the nth component of the vector $\boldsymbol{x} \in \mathbb{R}^N$, which is the input to the function $\Psi()$.

We examine the separation performance and sparsity of an ℓ_1-regularized model by varying $\lambda \geq 0$, i.e., 0.02, 0.03, 0.05, and 0.1. Note that the optimization is the minimization here due to ℓ_1-regularized setting. Fig. 9 shows the sparseness vs performance plots between ℓ_1-regularized models, the combined models, and baseline in SI-SNRi. The dots from five training runs for each λ are superimposed in Fig. 5(a). Almost all dots from ℓ_1-regularized models, including their averages over five training runs per λ, are situated at higher sparseness but at lower performance. The ℓ_1-regularized models achieve less performance than the baseline, leading to less performance than the proposed combined models.

VI. CONCLUSION

We explore the impact of sparse approximation on the separation performance of a deep-learning-based speech separation algorithm in speech mixtures without additional noise or reverberant interference.

We develop a combined model that embeds the sparse approximation ML-ISTA into the deep-learning-based Conv-TasNet speech separation. Adopting ML-ISTA as a sparse approximation algorithm is crucial for embedding as it avoids solving a bi-level optimization problem comprising sparse approximation and speech separation. ML-ISTA performs sparse approximation through forward calculations, thereby eliminating sparse approximation optimization. ML-ISTA's forward mechanism has a strong affinity for Conv-TasNet's convolution-based encoder structure and facilitates end-to-end computationally efficient model training.

We demonstrate that ML-ISTA's explicit integration as a sparse approximation influences Conv-TasNet's overall separation performance. We compare the separation performances of the combined models with Conv-TasNet. We create four combined models with different convergence levels using ML-ISTA's forward calculations with varying iteration counts. We observe that the sparseness of the combined models differs from that of the baseline model and that sparse approximation yields superior performance in all four criteria: SN-SNRi, SDRi, PESQ, and STOI. Improvements in peak performance of 1.1% to 4.7%, on average, over the baseline model are achieved in the four criteria, with increases in the model size (0.009%) and computational complexity (0.7% per iteration count). Furthermore, sparse approximation accelerates performance gain in the early stages of training, allowing all combined models to surpass the saturated performance of the baseline model even at epoch 30 for SI-SNRi.

We describe several directions in future work. One direction is to extend the ML-ISTA model to a noisy and reverberant environment. Noisy two-speaker mixtures, like Libri2mix and WSJ0 hipster ambient mixtures (WHAM!), or noisy reverberant two-speaker mixtures, like WHAM reverberant (WHAMR!), are the target speech corpora. Another direction is to apply ML-ISTA to the multilayered convolutional encoder case [38] to examine the similar level of ML-ISTA effect on separation performance in a multilayered convolutional case.

REFERENCES

[1] Y.-M. Qian, C. Weng, X.-K. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 40–63, Jan. 2018.

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

[3] S. Soni, R. N. Yadav, and L. Gupta, "State-of-the-art analysis of deep learning-based monaural speech source separation techniques," *IEEE Access*, vol. 11, pp. 4242–4269, 2023.

[4] J. Agrawal, M. Gupta, and H. Garg, "A review on speech separation in cocktail party environment: Challenges and approaches," *Multimedia Tools Appl.*, vol. 82, no. 20, pp. 31035–31067, Feb. 2023.

[5] J. Schnupp, I. Nelken, and A. King, "Auditory scene analysis," in *Auditory Neuroscience: Making Sense of Sound*. Cambridge, MA, USA: MIT Press, 2011, pp. 223–268.

[6] P. Földiák, "Sparse coding in the primate cortex," in *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed., 2nd ed. Cambridge, MA, USA: MIT Press, 2002, pp. 1064–1067.

[7] A.-S. Sheikh, N. S. Harper, J. Drefs, Y. Singer, Z. Dai, R. E. Turner, and J. Lücke, "STRFs in primary auditory cortex emerge from masking-based statistics of natural sounds," *PLOS Comput. Biol.*, vol. 15, no. 1, Jan. 2019, Art. no. e1006595.

[8] A. A. Kressner, D. V. Anderson, and C. J. Rozell, "A novel binary mask estimator based on sparse approximation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 7497–7501.

[9] N. Zhao, X. Xu, and Y. Yang, "Sparse representations for speech enhancement," *Chin. J. Electron.*, vol. 19, no. 2, pp. 268–272, Apr. 2011.

[10] S. Mallat, "Sparsity in redundant dictionaries," in *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. Burlingtong, MA, USA: Elsevier, 2009, pp. 611–695.

[11] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, 2015.

[12] Q. Zhang, X. Hu, B. Hong, and B. Zhang, "A hierarchical sparse coding model predicts acoustic feature encoding in both auditory midbrain and cortex," *PLOS Comput. Biol.*, vol. 15, no. 2, Feb. 2019, Art. no. e1006766.

[13] J. Sulam, A. Aberdam, A. Beck, and M. Elad, "On multi-layer basis pursuit, efficient algorithms and convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1968–1980, Aug. 2020.

[14] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[15] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 1562–1566.

[16] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[17] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 31–35.

[18] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2840–2849, 2021.

[19] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 46–50.

[20] Y. Liu and D. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 12, pp. 2092–2102, Dec. 2019.

[21] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Ontario, ON, Canada, Jun. 2021, pp. 21–25.

[22] Z.-Q. Wang, G. Wichern, and J. L. Roux, "Convolutive prediction for monaural speech dereverberation and noisy-reverberant speaker separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3476–3490, 2021.

[23] R. Ullah, M. Shohidul Islam, M. Imran Hossain, F. E. Wahab, and Z. Ye, "Single channel speech dereverberation and separation using RPCA and SNMF," *Appl. Acoust.*, vol. 167, Oct. 2020, Art. no. 107406.

[24] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2351–2364, 2023.

[25] B. Olshausen and D. Field, "Sparse coding of sensory inputs," *Current Opinion Neurobiol.*, vol. 14, no. 4, pp. 481–487, Aug. 2004.

[26] P. Kloppenburg and M. P. Nawrot, "Neural coding: Sparse but on time," *Current Biol.*, vol. 24, no. 19, pp. R957–R959, Oct. 2014.

[27] S. J. Thorpe, "Localized versus distributed representations," in *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed., 2nd ed. Cambridge, MA, USA: MIT Press, 2002, pp. 643–646.

[28] M. Boos, J. Lücke, and J. W. Rieger, "Generalizable dimensions of human cortical auditory processing of speech in natural soundscapes: A data-driven ultra high field fMRI approach," *NeuroImage*, vol. 237, Aug. 2021, Art. no. 118106.

[29] W. Mlynarski and J. H. McDermott, "Learning midlevel auditory codes from natural sound statistics," *Neural Comput.*, vol. 30, no. 3, pp. 631–669, Mar. 2018.

[30] M. Edalatian, A. A. Soitani, and N. Faraji, "Sparse representation of human auditory system," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Washington, DC, USA, Dec. 2016, pp. 302–306.

[31] A. Wahid, A. W. U. Ullah, K. Kadir, and S. Saadain, "Multi-layer convolutional sparse coding framework for restoration of under-sampled MR images," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, May 2021, pp. 1–6.

[32] A. Wahid, J. A. Shah, A. U. Khan, M. Ahmed, and H. Razali, "Multi-layer basis pursuit for compressed sensing MR image reconstruction," *IEEE Access*, vol. 8, pp. 186222–186232, 2020.

[33] Z. Wen, H. Wang, Y. Gong, and J. Wang, "Denoising convolutional neural network inspired via multi-layer convolutional sparse coding," *J. Electron. Imag.*, vol. 30, no. 2, pp. 1–20, Mar. 2021.

[34] B. Willmore and D. J. Tolhurst, "Characterizing the sparseness of neural codes," *Netw., Comput. Neural Syst.*, vol. 12, no. 3, pp. 255–270, Jan. 2001.

[35] B. D. B. Willmore, J. A. Mazer, and J. L. Gallant, "Sparse coding in striate and extrastriate visual cortex," *J. Neurophysiol.*, vol. 105, no. 6, pp. 2907–2919, Jun. 2011.

[36] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Mahwah, NJ, USA: Lawrence Erlbaum, 1988.

[37] A. Cherni, E. Chouzenoux, L. Duval, and J.-C. Pesquet, "SPOQ ℓ_p-over-ℓ_q regularization for sparse signal recovery applied to mass spectrometry," *IEEE Trans. Signal Process.*, vol. 68, pp. 6070–6084, 2020.

[38] B. Kadioglu, M. Horgan, X. Liu, J. Pons, D. Darcy, and V. Kumar, "An empirical study of Conv-TasNet," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 7264–7268.

HIROSHI SEKIGUCHI received the B.E. degree in applied physics from The University of Tokyo, Japan, in 1978, and the M.S. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, in 1984. He is currently pursuing the Ph.D. degree in advanced interdisciplinary studies with The University of Tokyo. He worked for Semiconductor Group, Toshiba Corporation, where he designed large-scale integrated circuits and developed applications in digital signal processing. His research interest includes speech signal processing with machine learning. He is a member of IEICE.

YOSHIAKI NARUSUE (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan, in 2012, 2014, and 2017, respectively. He is currently an Associate Professor with the Department of Electrical Engineering and Information Systems, Graduate School of Engineering, The University of Tokyo. His research interests include wireless power transfer, next-generation wireless communication systems, and the Internet of Things. He is a member of The Institute of Electronics, Information and Communication Engineers (IEICE) and IPSJ. He was a recipient of the Second-Best Student Paper Award from the IEEE Radio and Wireless Symposium, in 2013, the Hiroshi Harashima Academic Encouragement Award, in 2013, the Best Paper Award from the IEEE Consumer Communications and Networking Conference, in 2018, and the ACM IMWUT Distinguished Paper Award, in 2020.

HIROYUKI MORIKAWA (Member, IEEE) received the B.E., M.E., and Dr.Eng. degrees in electrical engineering from The University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1992, respectively. Since 1992, he has been with The University of Tokyo, where he is currently a Professor. His research interests include the Internet of Things, wireless communications, digital transformation, cloud robotics, and digital society design. He has authored or coauthored the books titled *Data-Driven Economy and 5G*. His research interests include the Internet of Things, wireless communications, digital transformation, cloud robotics, and digital society design. He was a recipient of more than 100 honors and awards, including The Institute of Electronics, Information and Communication Engineers (IEICE) Best Paper Award (thrice), the IPSJ Outstanding Paper Award, the NTT DoCoMo Mobile Science Award, the Radio Day Ministerial Commendation, the Rinzaburo Shida Award, and the Okawa Publications Prize. He is the Chairperson of the Communications and Information Network Association of Japan, the Beyond 5G New Business Strategy Center, the 5G-Driven Social Design Consortium, the Smart Resilience Network, the Digital Society Design Council, and the Information and Communication Council Working Group of the Ministry of Internal Affairs and Communications, and the President of IEICE, Tokyo. He was the Vice Chair of the OECD Committee on Digital Economy Policy. He serves on numerous advisory committees and frequently serves as a consultant for governments and companies.