Speech Communication 151 (2023) 76–85

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Single-channel speech separation using soft-minimum permutation invariant training

Midia Yousefi, John H.L. Hansen [*,1]

Center for Robust Speech Systems (CRSS) Erik Jonsson School of Engineering & Computer Science The University of Texas at Dallas, Richardson, Texas, USA

ARTICLE INFO

Keywords:
Source separation
Speech separation
Cocktail party
Probabilistic permutation invariant training
PIT
Prob PIT
Soft-minimum PIT

ABSTRACT

The goal of speech separation is to extract multiple speech sources from a single microphone recording. Recently, with the advancement of deep learning and availability of large datasets, speech separation has been formulated as a supervised learning problem. These approaches aim to learn discriminative patterns of speech, speakers, and background noise using a supervised learning algorithm, typically a deep neural network. A long-lasting problem in supervised speech separation is finding the correct label for each separated speech signal, referred to as label permutation ambiguity. Permutation ambiguity refers to the problem of determining the output-label assignment between the separated sources and the available single-speaker speech labels. Finding the best output-label assignment is required for calculation of separation error, which is later used for updating parameters of the model. Recently, Permutation Invariant Training (PIT) has been shown to be a promising solution in handling the label ambiguity problem. However, the overconfident choice of the output-label assignment by PIT results in a sub-optimal trained model. In this work, we propose a probabilistic optimization framework to address the inefficiency of PIT in finding the best output-label assignment. Our proposed method entitled trainable Soft-minimum PIT is then employed on the same Long-Short Term Memory (LSTM) architecture used in Permutation Invariant Training (PIT) speech separation method. The results of our experiments show that the proposed method outperforms conventional PIT speech separation significantly (p-value < 0.01) by $+1$dB in Signal to Distortion Ratio (SDR) and $+1.5$dB in Signal to Interference Ratio (SIR).

1. Introduction

Extracting the underlying sources from a signal mixture is a general problem in many applications. A classical example for such an application is to recognize or isolate what is being said by an individual speaker in a cocktail-party scenario in which multiple speakers are talking simultaneously Yousefi and Hansen (2020a). The auditory system of the human brain encounters two main challenges in the cocktail party scenario. First, it carries out sound segregation, which is the act of deriving properties of the individual sources from the mixture Carlyon (1992). Second, it can switch attention between different sources when following distinct conversations Shinn-Cunningham (2008); Koch et al. (2011). Humans can accomplish this in part due to bilateral hearing, as well as learned effective neural decoding in the auditory cortex.

However, as shown in Fig. 1, listening and following one speaker in the presence of competing speakers is an easy task for the majority of people, a remarkable ability usually taken for granted. While extensive research has explored speaker recognition by machines Stöter et al. (2018), the current task requires expanded knowledge and capabilities. However, even for humans with normal hearing abilities, the capacity of the human auditory system to extract and separate simultaneous sources out of a mixture is severely compromised Stöter et al. (2018); Bronkhorst (2015); Yousefi and Hanse (2021). As reported in Kawashima and Sato (2015), humans are capable of detecting up to three simultaneous active speakers without using spatial information of the input mixture. Thus, solving the cocktail party problem for mixtures with more than three concurrent active speakers is a very challenging task in which even humans may not be able to address Kashino and Hirahara (1996).

The cocktail party problem can be viewed as addressing blind source separation Qian et al. (2018), which is the task of recovering a set of independent sources when only their mixtures with unknown coefficients are available. Source separation can be considered as the

* Corresponding author.
E-mail addresses: midia.yousefi@utdallas.edu (M. Yousefi), john.hansen@utdallas.edu (John H.L. Hansen).

[1] This project was funded in part by NSF CISE CyberLearning Award #1918032, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

https://doi.org/10.1016/j.specom.2023.05.005
Received 19 April 2022; Received in revised form 1 May 2023; Accepted 5 May 2023
Available online 18 May 2023

combination of many active research threads such as: speech enhancement Vincent et al. (2018), speech separation Makino et al. (2007); Yousefi and Hansen (2021), waveform preserving estimation Tong et al. (1993); Sidiropoulos et al. (1998), music separation Ozerov et al. (2007); Luo et al. (2017), etc. Each of these research sub-communities make specific assumptions on the structure and properties of the active sources in the mixture, which results in problem-specific solutions to the cocktail party problem. Given the central role of overlapping speech detection and separation in the cocktail party scenario, in this study, we focus on single-channel speaker-independent speech separation.

The pioneering work of separating audio signals from a mixture was represented by Bregman in Bregman (1990). He noted that the auditory system performs an Auditory Scene Analysis (ASA) on the mixture signal entering the ear. ASA is executed in two steps: *(i)* the acoustic signal is decomposed into a number of sensory components, and *(ii)* the components that are likely from the same source are combined into a single stream. Based on this study, Computational Auditory Scene Analysis (CASA) was proposed in Brown and Cooke (1994). CASA attempts to model different parts of the biological auditory system that encomposes outer ear, middle ear and inner ear. In this approach, the signal is passed through a set of filterbanks that mimics the sound transduction performed by the inner hair cells. Next, for each filterbank output, periodicity, frequency transition, oneset and offset are calculated. Using pitch tracking techniques and labeling T-F bins, CASA groups the voice and unvoiced segments likely belonging to the same source. Finally, each waveform is reconstructed based on the source-specific segments Brown and Cooke (1994).

Independent Component Analysis (ICA) is another main technique introduced to address source separation Comon (1994). The strength of ICA as a separation tool resides in a realistic assumption that different physical processes generate unrelated signals. Therefore, when given a mixture of multiple speech sources, ICA identifies those unrelated signals which are voice traits of different speakers Comon and Jutten (2010). In ICA, data vectors are represented using weighted a linear combination of basis functions. Higher order statistics are needed to derive the independent coefficients of each basis. ICA removes not only correlation but also higher order dependence between the estimated bases, which has contributed to its success in extracting individual sources Lee (1998). Several years later, a powerful decomposition method called Non-Negative Matrix Factorization (NMF) was introduced, which has been very effective in modeling latent structure of data Lee and Seung (1999). NMF finds the parts-based representation of non-negative data through matrix decomposition, and therefore it is capable of extracting underlying speech sources from a mixture Lee and Seung (2001). A number of techniques have been developed based on NMF Hoyer (2004); Ding et al. (2005). Sparse NMF and Convolutive NMF (CNMF) are among the most popular Yousefi and Savoji (2016). CNMF outperforms NMF by modeling the temporal continuity of the speech signal in a time span of several frames O'Grady and Pearlmutter (2006); Smaragdis (2006). Although both ICA and NMF are supervised machine learning approaches and can learn useful patterns from the input data, the linear structure of the trained model in both ICA and NMF prevents it from learning complex structure within the speech signal. Thus, non-linear machine learning approaches such as Deep Neural Networks have been of interest.

Recently, learning-based approaches have boosted the performance of speech separation dramatically Wang and Chen (2018); Hershey et al. (2016); Yousefi and Angkititrakul (2021). Deep Clustering (DPCL) Hershey et al. (2016) was among the first approaches that made significant progress in extracting speech signals out of a mixture without prior information concerning the number of speakers. DPCL converts the mixture speech into an embedding space using an Recurrent Neural Network (RNN), with hope that T-F bins belonging to the same speaker establish a cluster in the embedding space. K-means clustering is used in the embedding space to identify these clusters. Finally, another network is trained based on T-F bins grouped in each cluster to estimate source-specific masks in order to recover the individual speech signals from the mixture Hershey et al. (2016). Another related technique called Deep Attractor Network (DANet) was introduced in Chen et al. (2017). Similar to DPCL, DANet projects the T-F bins of the mixture into an embedding space. DANet uses Expectation-Maximization (EM) to represent each speaker in the embedding space using a vector called an attarctor point such that the T-F bins belonging to that speaker are pulled toward the corresponding attractor point. Finally, the speech signals are estimated based on the grouped T-F bins around each attarctor point Luo et al. (2018).

Permutation Invariant Training (PIT) Yu et al. (2017); Kolbæk et al. (2017) is another effective solution which performs separation in two steps: first, it trains a neural network to separate the specific speech sources, and second, it finds the best output-label assignment to minimize the separation error. However, since the network generates unreliable outputs in the initial steps of training, the costs of different output-label permutations are close. The inefficiency of PIT in addressing permutation ambiguity has been considered in our previous study Yousefi et al. (2019). Therefore, we proposed Probabilistic PIT (Prob--PIT) which defines a log-likelihood function based on separation errors of all possible permutations. Unlike conventional PIT that uses one output-label permutation with the minimum cost, Prob-PIT uses all permutations by employing the soft-minimum function Yousefi et al. (2019). The effectiveness of Prob-PIT is achieved only when the parameters of the log-likelihood cost function are tuned well for each dataset. This can be very tedious and may require extensive time and computational resources to find the best hyperparameter value. To address this issue in Prob-PIT, in this study, we build on our previous work Yousefi et al. (2019) and propose a novel trainable Probabilistic PIT which we call Soft-minimum PIT to resolve the label permutation ambiguity challenge without requiring manual tuning of the parameters of the log-likelihood cost function. The contributions of this study are threefold:

- Proposing a novel trainable Probabilistic Permutation Invariant Training framework called soft-minimum PIT for single-channel supervised speech separation.

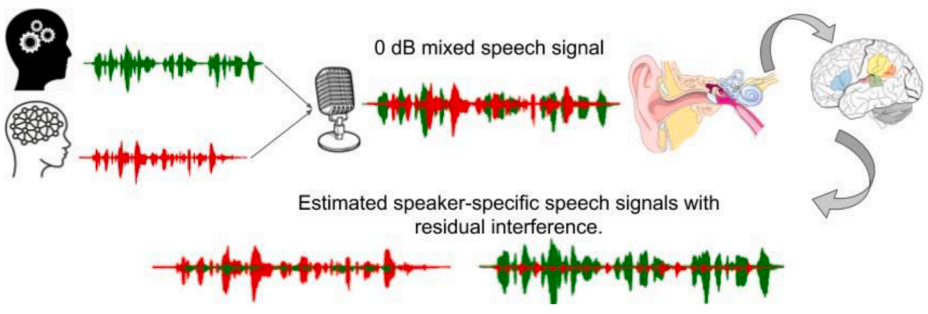

Fig. 1. Speech separation performed by human auditory system. The separated speech sources still contain residual speech from interfering talker.

- Comparing the results of both training and tuning of the hyper-parameters of the log-likelihood cost function.
- Comparing the results of our proposed system with the conventional PIT.

The remainder of the paper is organized as follows. We present the problem formulation, generating overlapping speech mixtures, and extracting spectral features in Section 2. Details of the proposed trainable Prob-PIT are explained in Section 3. We report on the experimental procedures and results in Section 4. The results are discussed in Section 5, and finally the conclusion is presented in the last section.

2. Problem formulation

Speech separation is extremely challenging under the single-microphone speaker-independent scenario, where no prior speaker information is available during evaluation. In supervised approaches, speech separation is formulated as a linear combination of single-speaker speech signals:

$$y[n] = \sum_{s=1}^{S} x_s[n] \tag{1}$$

in which S is the total number of speakers in the mixture signal y, and x_s is the speech signal corresponding to speaker s. While the number of total speakers S is assumed as available prior information, the goal of speech separation is to estimate the speaker-specific speech signals x_s from the mixture y. In this study, we consider the case of two-talker mixed speech separation. Therefore, Eq. 1 is modified as:

$$y[n] = x_1[n] + x_2[n] \tag{2}$$

with $x_1[n]$ and $x_2[n]$ represent the speaker-specific speech signal. Speech separation is generally performed in the frequency domain by transforming the signals using the Short-Time Fourier Transform (STFT). The main reason for this choice is that the speech structure such as harmonics, formants, and energy densities of are better represented in this domain. Therefore, speech separation is formulated as the task of recovering (STFT) of the source signals $X_s(t,f)$ for each time frame t and frequency bin f, given the mixed speech. However, since estimating phase information of this STFT representation is still an open problem Williamson et al. (2016), the phase information is acquired from the overlapping speech signal, and the separation is simplified to the task of estimating the magnitude spectrogram of the speaker-specific speech signals X_1 and X_2 from the mixture Y as:

$$Y(f,t) = X_1(f,t) + X_2(f,t), \tag{3}$$

where Y and X represent the magnitude spectra of y and x. The estimation of magnitude spectra of speech is usually achieved by training a model using supervised learning techniques. However, due to label permutation ambiguity, training a robust model for speech separation is challenging. Permutation ambiguity as depicted in Fig. 2, happens only in the model training stage and affects system performance in the testing phase. In Fig. 2, a single-channel mixed speech signal is processed by a DNN-based separation system. The goal of this separation is to extract speech signals corresponding to speakers A and B from the input mixture. Since there are only two speech sources in this mixture, the separation system has two outputs named o_1 and o_2. Each output contains a separated speech waveform corresponding to one of the speakers in the mixed speech. If the separation system is used for test, estimated speech signals at output 1 o_1 and output 2 o_2 are the separated final streams. However, in training, additional processing steps are required for updating the model parameters in each epoch. In supervised learning, updating parameters of the model is accomplished by comparing model's output with the desired ground truth signal, which in this application is the single-speaker speech waveform. The more the model output is similar to the desired label, the less the model's parameters are updated.

One important component in training, is backpropagation of the separation error which is the process of calculating the gradient of the error function with respect to the neural network's weights. Therefore, the effectiveness of backpropagation is highly dependent on the correct and precise value of the separation error. At this point, this question rises: "Which single-speaker speech waveform should be used as the desired form of each output?". Effectively, we should find the correct order between the outputs and the labels. There are two possible solutions:

Scenario 1:

- Output1: is the estimated speech related to Speaker A.
- Output2: is the estimated speech related to Speaker B.

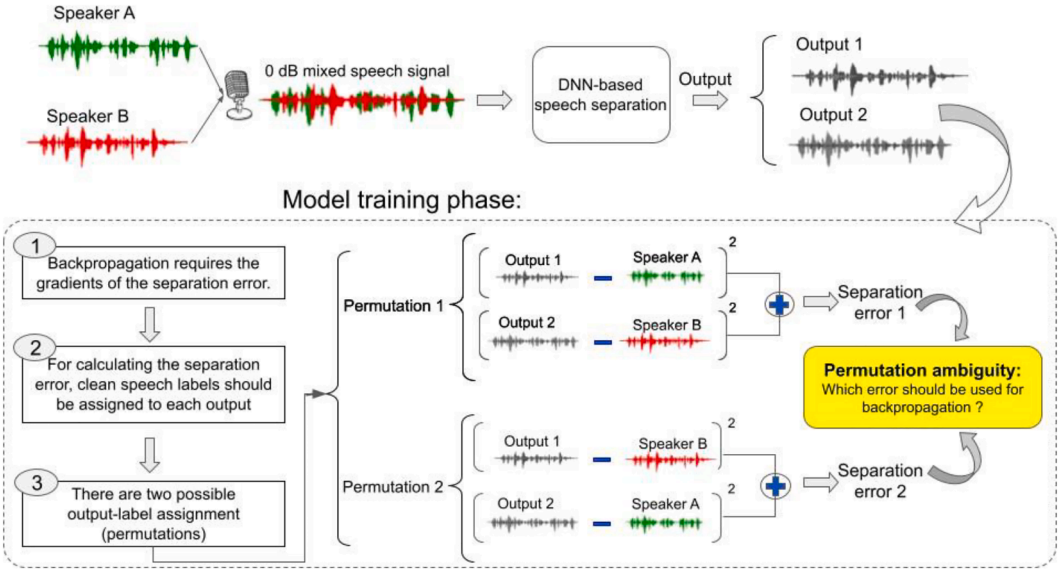

Fig. 2. Single-channel speaker-independent speech separation pipeline. Once the speaker-specific speech signals have been recovered at the output of the separation system, they should be assigned to their corresponding clean speech versions i.e. speaker A or speaker B for separation loss calculation. The permutation ambiguity caused by different possible output-label assignments leads to different separation errors causing gradient conflict in the training phase.

Scenario 2:

- Output1: is the estimated speech related to Speaker B.
- Output2: is the estimated speech related to Speaker A.

As shown in Fig. 2, each scenario is a possible permutation which results in alternate separation errors. For two-talker speech separation, there are two possible output-label assignments (permutation) and accordingly two possible separation errors. In general, for S sources in a mixture, there are $S!$ alternate possible permutations, which causes $S!$ different cost functions. In neural network training, it is necessary to find the correct separation error (correct cost function) and then perform back-propagation through the correct cost. The general consensus is to perform backpropagation through the gradients of the minimum separation error. This is the technique used in Permutation Invariant Training (PIT) speech separation method which has been shown to be effective in addressing the permutation ambiguity Yu et al. (2017). However, it has been discussed Yang et al. (2020); Yousefi et al. (2019) that the hard decision on choosing the minimum cost as the best solution results in training a sub-optimal separation model. To be more specific, the process of choosing the correct separation error is more challenging in the initial epochs of training, where the network is still naive and its outputs are not reliable. In those first epochs, costs of different permutations are very similar, and minimum separation error does not necessarily represent the correct output-label assignment. Therefore, the network will be trained based on a wrong decision epoch after epoch, which finally will contribute to a sub-optimal separation model.

To address this problem, it is important to optimize both model parameters and label assignments in the training phase. However, PIT uses a fixed label-assignment for every epoch and the benefit of using a flexible label assignment has not been explored sufficiently in the literature. The authors in Yang et al. (2020) have showed that label assignments chosen based on minimum separation error may be very random especially in initial epochs where network outputs show poor results. They studied the behavior of their network in selecting the output-label assignment, and discovered that the selected label assignments for a high percentage of training examples may be reversed in two consecutive epochs Yang et al. (2020). This rapid decision flip confuses both the network and the optimizer, which leads to updating the model parameters in opposite directions. These observations are detrimental in the training phase, which manifest the inadequacy of PIT.

Additionally, in our previous study, we explored the distributions of both separation errors associated with the possible permutations in a two-talker speech separation scenario Yousefi et al. (2019). The Kernel Distribution Estimation (KDE) of both separation error 1 and separation error 2 were plotted which revealed the inseparability of the separation errors in the first epoch of training. In the calculated KDE, error1 and error2 are more likely to be observed in regions where their values are very close Yousefi et al. (2019). We showed that choosing the minimum cost may lead to assigning the wrong label to the network output which affects quality of the trained model. Therefore, we proposed Probabilistic PIT (Prob-PIT) Yousefi et al. (2019) in which the output-label permutation was considered as a discrete latent random variable with a uniform prior distribution. Next, a log-likelihood function was defined based on prior distributions and separation errors of all possible permutations. We optimized the network parameters by maximizing the log-likelihood function. Unlike conventional PIT that enforces a hard decision by using one output-label permutation with the minimum cost, Prob-PIT uses all permutations by employing the soft-minimum function where leads to better overall separation results. To achieve a possible optimal possible model, in the cost function of Prob-PIT, a hyper-parameter was defined and manually tuned. However, manual tuning for the best value is a tedious process, which required extensive computational resources. Despite spending time and computational costs, the optimum value may not be found in challenging situations. Therefore, in this study we define a novel training framework in which the optimum value for the cost function parameter is learned from data during the training phase.

3. Trainable prob-PIT

We introduce the soft-minimum Permutation Invariant Training method in this section. As noted in Section 2, assume X_1 and X_2 contain magnitude spectra of speaker-specific clean speech signals shown as $X = [X_1, X_2]$ and Y is the magnitude spectra of the overlapping speech as introduced in Eq. 2. In supervised learning, we are give pairs of mixed speech signal and their associated single-speaker speech signals in the form of (Y, X) and our task is to train a model that estimates X based on the mixed speech Y observation. In the following subsections, we describe the proposed generative speech model, the network architecture, and the training framework in detail.

3.1. Model structure

In this proposed speech separation method, we define the magnitude spectra of single-speaker clean speech using a generative model as:

$$X = \widehat{X} + \epsilon, \tag{4}$$

in which $\widehat{X}$ is the estimated magnitude spectra by employing the separation system and ϵ is the separation error. For deriving $\widehat{X}$, we assume the neural network $D(\theta)$ with learnable parameters θ is employed. The network $D(\theta)$ takes overlapping speech observation Y as input, and estimates two speaker-specific speech signals in its outputs O_1 and O_2:

$$O_1, \mathrm{O}_2 = D(Y, \theta), \tag{5}$$

Once outputs O_1 and O_2 are derived, the goal of separation is accomplished. However, to build a reliable separation model, several epochs of training are required. In the training phase, backpropagation is performed to update the network parameters θ in order to optimize for the separation task. Backpropagation computes the gradients of the loss function with respect to the network parameters θ for each example in the dataset (i.e. each pair of (Y,X)). As noted in Section 2, for two-talker overlapping speech separation, there are two possible permutations between outputs O_1 and O_2 with clean speech labels X_1 and X_2. These two possible permutations lead to two alternate separation costs as depicted in Fig. 3. This situation causes a gradient conflict because it is not clear to the optimizer which separation cost should be used for the gradient calculation in backpropagation. In contrast to PIT, we define a one-to-one permutation function $Z(.)$ to solve the permutation ambiguity. The function $Z(.)$ permutes the order of the network outputs to match the correct order of the single-speaker speech labels. In a two-talker speech separation scenario, the permutation function, $Z(.)$, can take two forms as follow:

$$\widehat{X} = Z(O_1, \mathrm{O}_2), \tag{6}$$

As noted, the possible output-label assignment is considered as a latent variable with a uniform distribution. Therefore, in our task, the function $Z(.)$ can take two forms: $z_1(.)$ and $z_2(.)$ such that:

$$\begin{aligned} [O_1, O_2] &= z_1([O_1, O_2]), \\ [O_2, O_1] &= z_2([O_1, O_2]). \end{aligned} \tag{7}$$

In general, if there are S active speakers in the mixture audio signal, there are $S!$ possible permutations between the network outputs and the clean speech labels. This means that the permutation function $Z(.)$ will have $S!$ possible forms, with all having the same probability of $\frac{1}{S!}$. Only one permutation (either z_1 or z_2) is considered as the correct response. Therefore, by replacing $\widehat{X}$ in Eq. 4 with Eq. 5 and 6, the generative model can be reformulated as:

$$X = Z(D(Y, \theta)) + \epsilon, \tag{8}$$

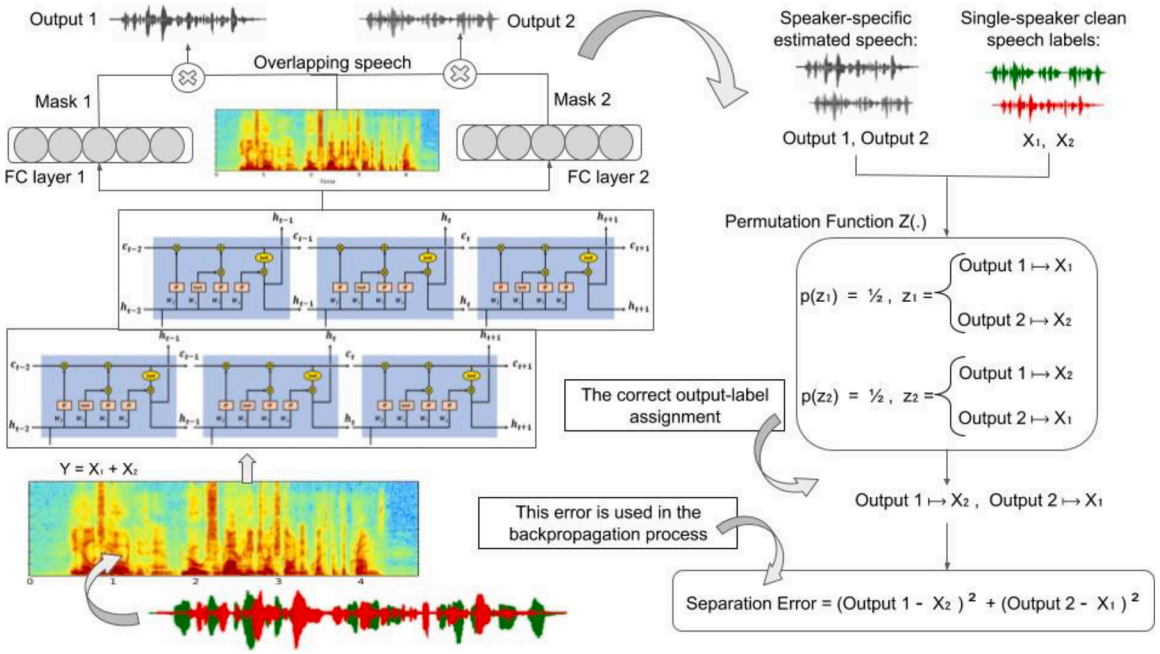

Fig. 3. Speech separation network architecture. First, the mixed speech is transformed to frequency domain using STFT. Next, the magnitude spectra of mixed speech is fed to a two-layer LSTM network followed by two Fully Connected (FC) layers for speaker-specific mask generation. Once the speaker-specific speech signal are estimated, they are passed to the permutation function $Z(.)$ for finding the best output-label assignment. Finally, the separation error is calculated based on the selected permutation.

Here, both ϵ and the permutation function $Z(.)$ are latent variables. ϵ is the estimation error which is typically modeled by a standard Gaussian distribution with mean zero and variance σ^2. Also, depending on the number of active speakers in the mixture, the permutation function $Z(.)$ could have different possible forms with uniform distribution. Eq. 8 is therefore our proposed generative model to represent single-speaker speech in this work.

3.2. Model training

In Section 3.1, we defined the proposed model pipeline. However, this model contains thousands of hyperparameters that must be trained based on extensive amounts of data. In contrast to conventional PIT, in our supervised training approach, we optimize the hyperparameters of the model by maximizing the log-likelihood function. The probability of estimating the speaker-specific speech signal X conditioned on the observation of the mixed speech signal Y, is the likelihood of the model which we aim to maximize. According to Bayes rule, the likelihood $P(X|Y)$ can be rewritten as:

$$P(X|Y) = \sum_{All\ possible\ Z} P(X|Z,Y)P(Z). \tag{9}$$

Here, $P(Z)$ is the probability of each possible permutation, which is set to $\frac{1}{S!}$ for S active speaker in a mixed speech signal. Therefore, $P(Z)$ is not dependent on z. $P(X|Z,Y)$ can be derived based on the Eq. 8. The distribution of X is determined by the distribution of ϵ with a new mean as:

$$P(X|Z,Y) = \mathcal{N}\big(Z(D(Y,\theta)),\ \sigma^2 I\big), \tag{10}$$

where, I is the identity matrix and $\mathcal{N}$ is the Gaussian distribution. By replacing $P(X|Z,Y)$ in Eq. 9, and inserting the Gaussian distribution equation, the following expression is obtained for the log-likelihood function :

$$\log P(X|Y) = \log\frac{1}{S!} + \log\frac{1}{\sqrt{\gamma\pi}} + \log\sum_{All\ Z} exp\left(\frac{-\parallel X - Z(D(Y,\theta))\parallel^2}{\gamma}\right), \tag{11}$$

where $\gamma = 2\sigma^2$. The log-likelihood function expressed in Eq. 11 is maximized in order to train the model. However, due to the logarithm function, this equation might be numerically unstable. Therefore, to ensure stability, we employ the log-sum-exp stabilization procedure: $\log\sum_i e^{x_i} = \max_i x_i + \log\sum_i e^{x_i - \max_i x_i}$. The following equations show the numerically stable form of Eq. (11):

$$\log P(X|Y) = \frac{-e(Z_{min},\theta)}{\gamma} + \log\left(1 + \sum_{Z\neq Z_{min}} \exp\left(\frac{e(Z_{min},\theta) - e(Z,\theta)}{\gamma}\right)\right) - \log\sqrt{\gamma\pi} + C, \tag{12}$$

$$e(Z,\theta) = \parallel X - Z(D(Y,\theta))\parallel^2,$$

$$Z_{min} = \underset{Z}{\operatorname{argmin}} e(Z,\theta).$$

Here, $e(Z,\theta)$ is the separation error of the permutation Z, and $Z_{\min}$ is the permutation that has the minimum separation error. Noted earlier, γ is equal to $2\sigma^2$, and σ^2 is the variance of the estimation error in the proposed model. Since $e(Z_{\min}) - e(z)$ is always negative, both exponential and logarithmic functions are numerically stable.

From Eq. 12, the model parameters θ are optimized to maximize the log-likelihood function $\log P(X|Y)$. This optimization is performed by applying the smooth minimum of the costs of all permutations with a smoothing factor of γ. The first part of this cost function is $-e(Z_{\min},\theta)$, which is the minimum error among all possible label permutations. This is the same cost that conventional PIT uses. The second part of the equation is the cost of all other possible permutations. Alternatively, by setting γ to larger values, a compromise is obtained between the cost of a minimum permutation versus the cost of all permutations (please refer to Yousefi et al. (2019) for more details).

4. Experiments

4.1. Dataset

As reported in von Neumann et al. (2019), real-world recording

datasets such as AMI only contain approximately 5–10% overlapping speech, which may not be sufficient for training a neural network model without overfitting the data. Therefore, we follow the same mixed speech generation process used in Yu et al. (2017); Hershey et al. (2016); Chen et al. (2017); Yousefi and Hansen (2020b). We generate overlapping speech utterances based on the GRID corpus, which is a multi-speaker, sentence-based corpus used in a monaural speech separation and recognition challenge Cooke et al. (2006). This corpus contains 34 speakers, 16 female and 18 male speakers, each providing 1000 sentences, which have been frequently used in several overlapping speech detection and separation studies Yousefi et al. (2018); Shokouhi and Hansen (2017); Tu et al. (2015); Yousefi et al. (2019). To generate mixed speech utterances, random speech recordings are selected from random speakers. The chosen utterances are first processed through a Speech Activity Detector (SAD) for removal of silence segments. Most recordings in the GRID corpus have almost the same duration. However, for utterances with different duration, longer utterances are cut so that their length matches the shorter utterance, then they are summed with a random Signal-to-Interference Ratio (SIR), which is uniformly distributed between 0 to 5 dB. Each data sample in our generated corpus contains three waveforms, the two selected random utterances and the output generated mixed speech. For each dataset, we have generated 10h of mixed data for the training set, 4h for development, and 2h mixtures for the test set[2] Also, speakers used for generating the test set are separate from those used in training and development sets.

4.2. Evaluation metrics

Speech separation techniques are usually evaluated using the blind source separation evaluation (BSS-EVAL) toolbox Vincent et al. (2006); Wang and Chen (2018). Two of the widely used measures from this toolkit are Signal-to-Distortion (SDR), Signal-to-interference Ratio (SIR), and Signal-to-Artifact Ratio (SAR) in the estimated speech signals. SIR and SAR measure different types of residual noise in the estimated speaker-specific speech signal. SIR assesses the remaining noise due to the residual interference in the separated speech signal which is called mis-separation noise. Also, SAR measures the noise related to the reconstruction algorithm which includes, glitches due to the STFT phase estimation process. SDR measures the distortion introduced to the estimated signal by both mis-separation and the reconstruction algorithm.

Both SDR, SIR, and SAR have been shown to be well correlated with human assessments of signal quality Fox et al. (2007). From a mathematical point of view, SDR is defined as the ratio of the target signal power to the distortion introduced by the interference, reconstruction noise, and all other background noise. SIR is defined as the ratio of the target signal power to that of the interference signal still remained in the separated speech. SAR is defined as the ratio of the estimated signal power to the reconstruction noise. In this study, we evaluate system performance using these two metrics Vincent et al. (2006).

4.3. Experimental results

For network training, a 129-dim STFT magnitude spectra is employed, computed over a frame size of 32ms with a 50% frame shift. The network architecture consists of two Long Short-Term Memory (LSTM) Network with 128 neurons each. The output of the last LSTM layer is passed to the Softmax activation function which adds a non-linearity to the model. Next, two different fully connected layers are used to estimate speaker-specific masks. Finally, by multiplying the estimaed masks wih the overlapping speech magnitude spectra, two magnitude spectra with dimension [129 * frame-number] are generated for the two speech sources. This architecture is one of the effective structures employed by conventional PIT in Yu et al. (2017). Since the memory cells in LSTM keep track of the speaker-specific information, this architecture is a suitable choice for speech separation Chen and Wang (2017). However, since a different dataset has been used in this study, we tune the hyperparameters to ensure that network has a viable initial setup for the speech separation task prior to experiments. The network is trained for 50 epochs using Adam optimization algorithm. Our experiments show that a dropout rate of 20%, and a learning rate of 0.0005 reduced by 0.7 when the cross-validation loss improvement is less than 0.003 in two successive epochs, are the best choices for the hyperparameters.

Baseline PIT– In this study, we compare our proposed method with PIT introduced in Yu et al. (2017). PIT minimizes the mean-square-error of the estimated speech signal to train the network. To do so, PIT selects the permutation with minimum cost throughout backpropagation. The cost function used in PIT is as:

$$Cost_{(PIT)} = \| X - \widehat{X} \|^2 = \| X - Z_{min}(D(Y,\theta)) \|^2. \tag{13}$$

Constant γ Soft-minimum PIT– The results of our proposed approach is shown in Table 1. Evaluation metrics are reported for both speaker-specific estimated speech signals. For each reconstructed waveform, SDR, SIR, and SAR are reported. As mentioned in the previous section, PIT is considered as the baseline. As shown in Table 1, $1 \leq \gamma \leq 5$ improves the PIT baseline, however, the choice of $\gamma = 2$ seems to be the best choice with the best output performance in terms of SDR and SIR. Larger choices of γ result in under-performing PIT, which is expected as the optimizer ignores the permutation with the minimum cost.

Additionally, as mentioned in Section 3, γ is equal to $\gamma = 2 * \sigma^2$, in which σ is the variance of the separation error. Therefore, when $\gamma = 2$, the variance of the separation error is set to $\sigma = 1$ which is a reasonable choice for the separation cost.

It is worth mentioning that, the core role of γ is in the training phase. Our motivation in defining the log-likelihood function in Eq. 12 was to prevent the network parameters to be trained based on an unreliable cost function. Therefore, the log-likelihood function uses γ to replace the minimum cost by the soft-minimum cost function which results in a smoother optimization landscape and therefore the model is less likely to converge to a poor local minimum. Once the model is trained, the separation performance is evaluated by other metrics such as SDR, SIR, and SAR. Thus, in the testing phase, we follow the PIT testing procedure and set γ to zero.

During testing, we still need to find the correct output-label assignment for determining SDR, SIR, and SAR. In our proposed approach, there are two possible options in the testing phase: *(i)* choosing the output-label assignment with minimum separation loss, and *(ii)* choosing the output-label assignment with the permutation that maximizes the log-likelihood function with the same assigned γ in the training phase. These two approaches are evaluated and plotted in Fig. 4. Here, the blue bars represent the first case in which γ is set to zero during the testing phase. Alternatively, the red bars represent the case in which γ in the testing phase is the same value as in the training set. As depicted in the plot, system performance is much better when $\gamma = 0$ in terms of SDR and SIR, This is expected as our proposed method is aimed

Table 1
The results of the proposed Soft-minimum PIT speech separation in terms of Signal to Distortion ratio (SDR), Signal to Interference Ration (SIR), and Signal to Artifacts Ratio (SAR) for both speakers in the mixture. $\gamma = 2$ maximizes the log-likelihood of he separation cost and results in the best performance.

	Speaker 1			Speaker 2		
Constant γ	SDR	SIR	SAR	SDR	SIR	SAR
PIT	6.6693	8.1966	13.0914	2.9977	4.5250	10.3076
$\gamma = 1$	6.8775	8.4905	13.0699	3.2571	4.8916	10.2669
$\gamma = 2$	**7.1894**	**8.9429**	**13.0547**	**3.6061**	**5.3652**	**10.2844**
$\gamma = 3$	6.9817	8.6556	13.0409	3.3853	5.0869	10.2152
$\gamma = 4$	6.8852	8.4423	13.4668	3.1657	4.6821	10.9357

[2] The corpus generated here will be shared with the speech community.

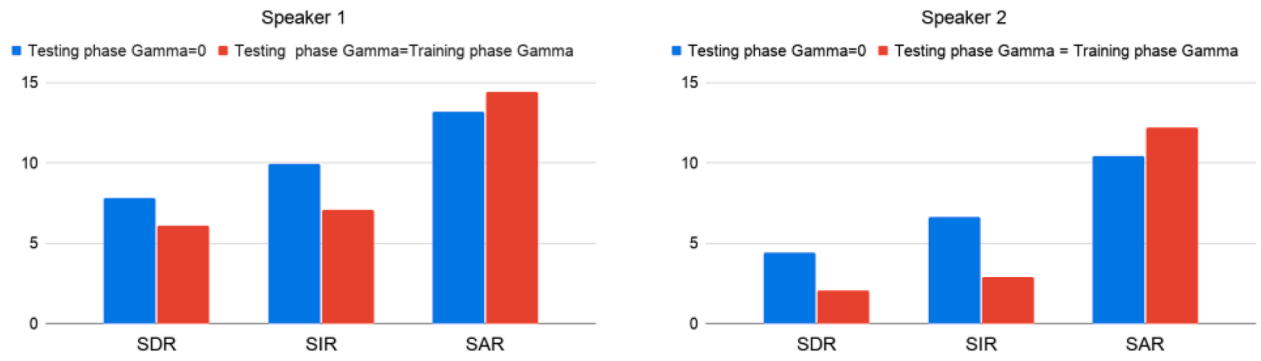

Fig. 4. The comparison of speech separation performance for different γ values in the testing phase. As depicted in the plots, once the model is trained with the Soft-minimum PIT method, γ should be set to Zero in the testing phase. This is because the trained model is reliable in the testing phase and the permutation with the minimum loss is the correct output-label assignment.

at modifying the training process to result in a higher quality separation model with improved performance. As mentioned in Section 2, choosing the permutation with a minimum cost in the training phase confuses the model and optimizer, and leads to updating the model parameters toward opposite directions in successive epochs. However, once the model is trained by maximizing the log-likelihood of all possible separation costs, the testing phase should be performed by selecting the permutation with the minimum cost.

Trainable γ Soft-minimum PIT– The main limitation of the soft-minimum PIT is finding the right value for γ. This can be exhausting since it requires time and computational resources which makes it a sub-optimal tuning process. Therefore, instead of tuning this parameter, in this section, γ is trained by the optimizer.

The separation performance for the trained γ is reported in Table 2. The reported values for γ in the first column are the initial values for this trainable parameter. As reported in the table, the initial value of $\gamma = 1$ leads to the best performance in terms of SDR and SIR. However, in terms of SAR, some degradation is observed for which the root cause and potential methods to address this bimodal performance change is still under investigation. In our experiments, after several epochs, γ converged to a number in the range $[1.5 \leq \gamma \leq 2]$, and it was very close to the manually selected value reported in the previous table. Since the optimal value of γ for the data used is in the range of [1,2], then setting the initial value of γ higher than 1 makes it more challenging for the optimizer to converge to the optimal γ, causing an overall lower performance.

Furthermore, in order to minimize the effect of parameter initialization on our final separation metrics, we train each network five times with different initial parameters. The distributions of the final evaluation metrics of those five experiments are depicted in Fig. 5. In these plots, the boxplot and kernel distribution estimation of those five experiments are depicted. The boxplot is a standardized way of displaying the distribution of experiments based on a five number summary, which are the minimum, first quartile (Q1), median, third quartile (Q3), and maximum. Additionally, each violin-shaped object represents the kernel distribution estimation of the results. The more the result points are in a specific range, the larger the violin is for that range. Also, each blue circle on the blue line represents the mean of the evaluation metrics in the five experiments for each separation system. The black solid line is the mean of the performance metrics for $\gamma = 0$ as the PIT baseline.

The first row in Fig. 5 represents the distribution estimation for the fist separated speaker in terms of SDR (first row, first column), and SIR (first row, second column). Likewise, the distribution estimation plots are depicted for the second speaker in the second row. As shown, the PIT separation method has a long and narrow violin response with high variance of the results. This system behavior is not desired because the output performance is not reliable and may tend to vary significantly. However, in our proposed soft-minimum PIT for both constant γ and Trained γ the violin plots tend to be wide and short which confirms a small variance in the output results. However, as mentioned before, since manual tuning of γ does not guarantee finding the optimal value, this approach is sub-optimal compared to learning γ in the training phase. Once the optimal value for γ is learned by the optimizer, the SDR of the first speaker is improved by almost +1dB compared to the baseline PIT. This is almost +13% relative improvement with lower variance, revealing the superiority and reliability of the proposed approach compared to PIT. In terms of SIR, training γ in the proposed soft-minimum PIT achieves the best results with +1.5dB improvement compared to PIT. Similar to SDR, the variance of the output SIR is lower in the Trained γ scenario. Nevertheless, as depicted in the second row of Fig. 5, the pattern of the results is repeated for the second speaker as well. The output SDR and SIR for the second speaker have the same relative improvement in the proposed approach compared to the baseline. This is a very important accomplishment, because the second speaker has lower energy in the mixture (due to the lower input SDR we chose in the mixing process), therefore, recovering such a degraded signal is very challenging. Additionally, most separation systems are only effective in recovering the target speaker speech from the mixture at the expense of the remaining speaker-specific speech signals. Therefore, an effective speech separation solution should be capable of recovering both speakers from the mixture with the same level of quality. Similar to the first speaker, once soft-minimum PIT is employed in the training phase, the variance of the output SDR and SIR for the second speaker is also lowered compared to PIT.

Table 2
The results of the proposed Soft-minimum PIT with trained γ in terms of Signal to Distortion ratio (SDR), Signal to Interference Ration (SIR), and Signal to Artifacts Ratio (SAR) for both speakers in the mixture. The initial value of $\gamma = 1$ is the best option for learning its optimal value.

	Speaker 1			Speaker 2		
Initial value	SDR	SIR	SAR	SDR	SIR	SAR
$\gamma = 1$	**7.6471**	**9.8187**	**12.6322**	**4.2793**	**6.6073**	**9.6374**
$\gamma = 2$	7.2199	9.0586	12.8717	3.6579	5.5009	10.1288
$\gamma = 3$	7.1882	8.9946	12.8792	3.6487	5.4935	10.0569
$\gamma = 4$	7.0982	8.8117	13.1499	3.4787	5.1941	10.4676

5. Discussion

In general, speech separation is usually performed in two steps *(i)* separating the specific speech sources, and *(ii)* determining the best output-label assignment, allowing for assessment of separation error via the evaluation metrics. The second step called *label permutation ambiguity* has been a long standing challenge in training neural networks for speech separation. Recently proposed *Permutation Invariant Training (PIT)* addressed this problem by determining the output-label

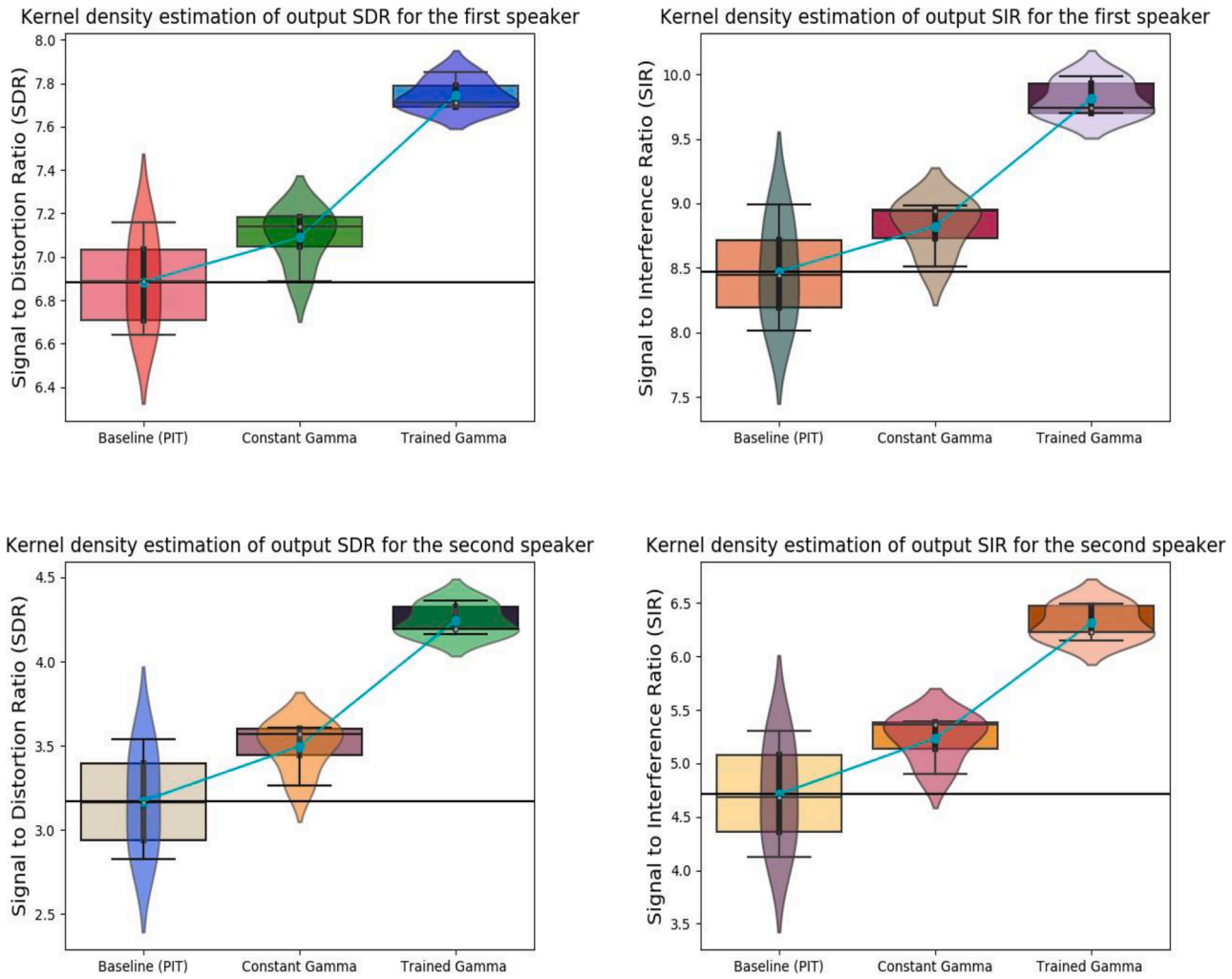

Fig. 5. The boxplot and kernel distribution of the Baseline (PIT), Soft-minimum PIT with Constant γ, and Soft-minimum PIT with trained γ is depicted. For each separation system, 5 experiments have been performed. Each violin-shaped object represents the boxplot and the kernel distribution estimation of those five experiments. The black solid line represents the mean of the results for the PIT baseline. The blue circles on the blue line are the mean of the output evaluation metrics for other two separation method: Constant γ in Soft-minimum PIT and Trained γ in the Soft-minimum PIT. As shown in the figure, the proposed soft-minimum PIT in both scenarios outperforms PIT baseline. Additionally, the output SDR and SIR of the proposed method have a lower variance for both separated speakers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

assignment which minimizes the separation error. In PIT, a neural network is trained that separates the speaker-specific speech signals. In the training phase, PIT determines the best output-label assignment which minimizes the separation error. Next, backpropagation is performed based on the minimum separation error. However, studies Yousefi et al. (2019); Yang et al. (2020) have shown that choosing the minimum separation error is a hard decision imposed on the optimizer, especially in the initial epochs of training where network is still naive. Each possible output-label assignment result in a different cost function. In the initial epochs of the training, the value of these costs are very close. Therefore, the minimum separation error does not necessarily represent the correct output-label assignment. Additionally, in the beginning of the training phase, the selected label assignments may be reversed in two consecutive epochs which confirms the unreliability of the network output. If backpropagation is performed based on only the minimum separation error, then the rapid decision flip confuses both the network and optimizer, which leads to updating the model parameters toward opposite directions. Therefore, updating network parameters based on the cost of one single permutation is not an optimal solution, and leads to an inefficient training of the network. These observations are detrimental in the training phase, which manifests the inadequacy of PIT.

In contrast to PIT, we propose the Soft-minimum PIT which considers the output-label assignment as a latent variable with uniform distribution. In Soft-minimum PIT, the network is trained by maximizing the log-likelihood of the prior distributions and the separation errors of all possible permutations. Since in the proposed method, all possible output-label assignment are taken into consideration in the backpropagation, the optimization landscape becomes smoother, which is in contrast to the hard decision of minimizing the Mean-Square-Error of the minimum separation cost performed in PIT. In the Soft-minimum PIT, the smoothness of the cost function is controlled by γ, which is $2 * \sigma^2$ with σ being the variance of the separation error. In this work, we have explored both tuning and training γ in the proposed method to evaluate the separation performance. The results of our experiments on the simulated two-talker overlapping speech dataset shows that Soft-minimum PIT outperforms PIT significantly (p-value < 0.01). Also, the greatest improvement is achieved by training γ with other parameters of the network. Trained γ in the Soft-minimum PIT results in improved output SDR and SIR by $+1$dB and $+1.5$dB with lower variance during multiple repeated experimental runs with different initialization.

The effectiveness of the proposed Soft-minimum PIT can be attributed to several reasons. The core strength of Soft-minimum PIT is the

incorporation of all possible output-label assignments in training the model parameters. This is in contrast to PIT, which uses a hard decision in assigning the output-label permutation that minimizes the total separation error. During training, the network is not able to estimate the speaker-specific speech signals correctly, therefore, its decision in assigning the correct output-label permutation is not reliable. Also, since in the initial epochs of training, all model parameters are randomly selected, so the separated speech signals at the output are far from the desired speech signals. Consequently, the separation error of different possible permutations are very similar and the correct output-label assignment does not necessarily have a minimum separation loss. Therefore, if the selected output-label permutation in PIT is not correct, then the model parameters are updated based on a wrong decision, resulting in deteriorating the training process.

In addition, it has been shown Yang et al. (2020) that the output-label assignment selected in PIT tends to change in two successive epochs for most of the data samples in the corpus. This uncertainty in finding the correct permutation causes a disorientation in the optimizer because the model parameters are updated in opposite directions for most of the initial epochs. Hence, in our Soft-minimum PIT, we consider the costs of all possible permutations for training the network in a probabilistic framework.

Another reason for the success of our proposed approach is that the minimum cost function used in PIT is replaced by a soft-minimum function. In several applications of machine learning, it has been shown that replacing the minimum by the soft-minimum results in a smoother optimization landscape and therefore it is less likely to converge to a poor local minima. This can also be explained in terms of the decision flips that PIT experiences during training. Since in Soft-minimum PIT the decisions are reliable and comprehensive, then the optimization landscape does not have many poor local minimums. Two core observations in this study confirms this finding for speech separation as well. First, SDR and SIR values of the soft-minimum are better than PIT significantly (p-value < 0.01); (2) the variance of SDR and SIR values are lower for both constant and trained γ. A lower variance in the results show a more stable system, which may be caused by a smoother optimization landscape.

6. Conclusion

In this study, we proposed Soft-minimum PIT to address label permutation ambiguity in speech separation. For Training single-channel speaker-independent speech separation models, two steps are required: first, estimating the speaker-specific speech signal; second, finding the correct output-label assignment for calculating the separation error. The second step known as label permutation ambiguity has been a long-standing challenge in training neural networks for the task of speech separation. One general solution introduced in PIT proposes to train a neural network based on the output-label assignment with minimum separation cost. Unfortunately, the hard choice of minimum cost permutation is not the best technique, especially in initial epochs of training where the network is still not strong enough to effectively separate the speech signals. In contrast to PIT, in our proposed Soft-minimum PIT, we consider all possible permutations as a discrete latent variable with a uniform prior distribution. Next, we trained the network by maximizing the log-likelihood function defined based on prior distributions and separation errors of all possible permutations. In our proposed approach the smoothness of the decision was controlled by a variable parameter that can be either tuned or trained. In this study, we explored both cases and results based on GRID datasets show that the proposed Soft-minimum PIT significantly outperforms PIT in terms of SDR and SIR. This solution therefore offers a viable option to effectively separate overlap/mixed speaker audio streams, especially in naturalistic audio scenarios.

CRediT authorship contribution statement

Midia Yousefi: Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft. **John H.L. Hansen:** Supervision, Conceptualization, Resources, Writing – review & editing, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Bregman, A., 1990. Auditory scene analysis: the perceptual organization of sound. Cambridge, MA, US.

Bronkhorst, A.W., 2015. The cocktail-party problem revisited: early processing and selection of multi-talker speech. Attent. Percept. Psychophys. 77 (5), 1465–1487.

Brown, G.J., Cooke, M., 1994. Computational auditory scene analysis. Comput. Speech Lang. 8 (4), 297–336.

Carlyon, R.P., 1992. The psychophysics of concurrent sound segregation. Philos. Trans. R. Soc. Lond. Ser. B: Biol. Sci. 336 (1278), 347–355.

Chen, J., Wang, D., 2017. Long short-term memory for speaker generalization in supervised speech separation. J. Acoust. Soc. Am. 141 (6), 4705–4714.

Chen, Z., Luo, Y., Mesgarani, N., 2017. Deep attractor network for single-microphone speaker separation. Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, pp. 246–250.

Comon, P., 1994. Independent component analysis, a new concept? Signal Process. 36 (3), 287–314.

Comon, P., Jutten, C., 2010. Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic press.

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. 120 (5), 2421–2424.

Ding, C., He, X., Simon, H.D., 2005. On the equivalence of nonnegative matrix factorization and spectral clustering. Proceedings of the 2005 SIAM international conference on data mining. SIAM, pp. 606–610.

Fox, B., Sabin, A., Pardo, B., Zopf, A., 2007. Modeling perceptual similarity of audio signals for blind source separation evaluation. International Conference on Independent Component Analysis and Signal Separation. Springer, pp. 454–461.

Hershey, J.R., Chen, Z., Le Roux, J., Watanabe, S., 2016. Deep clustering: discriminative embeddings for segmentation and separation. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 31–35.

Hoyer, P.O., 2004. Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. 5 (Nov), 1457–1469.

Kashino, M., Hirahara, T., 1996. One, two, many-judging the number of concurrent talkers. J. Acoust. Soc. Am. 99 (4), 2596–2603.

Kawashima, T., Sato, T., 2015. Perceptual limits in a simulated “cocktail party”. Attent. Percept. Psychophys. 77 (6), 2108–2120.

Koch, I., Lawo, V., Fels, J., Vorländer, M., 2011. Switching in the cocktail party: exploring intentional control of auditory selective attention. J. Exp. Psychol.: Hum. Percept. Perform. 37 (4), 1140.

Kolbæk, M., Yu, D., Tan, Z.-H., Jensen, J., 2017. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (10), 1901–1913.

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401 (6755), 788.

Lee, D.D., Seung, H.S., 2001. Algorithms for non-negative matrix factorization. Advances in neural information processing systems, pp. 556–562.

Lee, T.-W., 1998. Independent component analysis. Independent component analysis. Springer, pp. 27–66.

Luo, Y., Chen, Z., Hershey, J.R., Le Roux, J., Mesgarani, N., 2017. Deep clustering and conventional networks for music separation: stronger together. 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp. 61–65.

Luo, Y., Chen, Z., Mesgarani, N., 2018. Speaker-independent speech separation with deep attractor network. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (4), 787–796.

Makino, S., Lee, T.-W., Sawada, H., 2007. Blind Speech Separation, Vol. 615. Springer.

von Neumann, T., Kinoshita, K., Delcroix, M., Araki, S., Nakatani, T., Haeb-Umbach, R., 2019. All-neural online source separation, counting, and diarization for meeting analysis. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 91–95.

O’Grady, P.D., Pearlmutter, B.A., 2006. Convolutive non-negative matrix factorisation with a sparseness constraint. 2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing. IEEE, pp. 427–432.

Ozerov, A., Philippe, P., Bimbot, F., Gribonval, R., 2007. Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs. IEEE Trans. Audio Speech Lang. Process. 15 (5), 1564–1578.
Qian, Y.-m., Weng, C., Chang, X.-k., Wang, S., Yu, D., 2018. Past review, current progress, and challenges ahead on the cocktail party problem. Front. Inf. Technol. Electron. Eng. 19 (1), 40–63.
Shinn-Cunningham, B.G., 2008. Object-based auditory and visual attention. Trends Cogn. Sci. (Regul. Ed.) 12 (5), 182–186.
Shokouhi, N., Hansen, J.H., 2017. Teager–kaiser energy operators for overlapped speech detection. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (5), 1035–1047.
Sidiropoulos, N.D., Giannakis, G.B., Bro, R., 1998. Deterministic waveform-preserving blind separation of ds-cdma signals using an antenna array. Ninth IEEE Signal Processing Workshop on Statistical Signal and Array Processing (Cat. No. 98TH8381). IEEE, pp. 304–307.
Smaragdis, P., 2006. Convolutive speech bases and their application to supervised speech separation. IEEE Trans. Audio Speech Lang. Process. 15 (1), 1–12.
Stöter, F.-R., Chakrabarty, S., Edler, B., Habets, E.A., 2018. Countnet: estimating the number of concurrent speakers using supervised learning. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (2), 268–282.
Tong, L., Inouye, Y., Liu, R.-W., 1993. Waveform-preserving blind estimation of multiple independent sources. IEEE Trans. Signal Process. 41 (7), 2461–2470.
Tu, Y.-H., Du, J., Dai, L.-R., Lee, C.-H., 2015. Speech separation based on signal-noise-dependent deep neural networks for robust speech recognition. 2015 ICASSP. IEEE, pp. 61–65.
Vincent, E., Gribonval, R., Févotte, C., 2006. Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. 14 (4), 1462–1469.
Vincent, E., Virtanen, T., Gannot, S., 2018. Audio Source Separation and Speech Enhancement. John Wiley & Sons.
Wang, D., Chen, J., 2018. Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (10), 1702–1726.
Williamson, D.S., Wang, Y., Wang, D., 2016. Complex ratio masking for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) 24 (3), 483–492.
Yang, G.-P., Wu, S.-L., Mao, Y.-W., Lee, H.-y., Lee, L.-s., 2020. Interrupted and cascaded permutation invariant training for speech separation. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6369–6373.
Yousefi, M., Angkititrakul, P., 2021. System for end-to-end speech separation using squeeze and excitation dilated convolutional neural networks. US Patent App. 16/805,716.
Yousefi, M., Hanse, J.H., 2021. Speaker conditioning of acoustic models using affine transformation for multi-speaker speech recognition. arXiv preprint arXiv:2111.00320.
Yousefi, M., Hansen, J.H., 2020. Block-based high performance cnn architectures for frame-level overlapping speech detection. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 28–40.
Yousefi, M., Hansen, J.H., 2020. Frame-based overlapping speech detection using convolutional neural networks. arXiv preprint arXiv:2001.09937.
Yousefi, M., Hansen, J.H., 2021. Real-time speaker counting in a cocktail party scenario using attention-guided convolutional neural network. arXiv preprint arXiv:2111.00316.
Yousefi, M., Khorram, S., Hansen, J.H., 2019. Probabilistic permutation invariant training for speech separation. arXiv preprint arXiv:1908.01768.
Yousefi, M., Savoji, M.H., 2016. Supervised speech enhancement using online group-sparse convolutive nmf. 2016 8th International Symposium on Telecommunications (IST). IEEE, pp. 494–499.
Yousefi, M., Shokouhi, N., Hansen, J.H., 2018. Assessing speaker engagement in 2-person debates: overlap detection in united states presidential debates. Interspeech, pp. 2117–2121.
Yu, D., Kolbæk, M., Tan, Z.-H., Jensen, J., 2017. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 241–245.