Speech Communication 152 (2023) 102956

Contents lists available at ScienceDirect

Speech Communication

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

Using iterative adaptation and dynamic mask for child speech extraction under real-world multilingual conditions

Shi Cheng [a], Jun Du [a,*], Shutong Niu [a], Alejandrina Cristia [b], Xin Wang [a], Qing Wang [a], Chin-Hui Lee [c]

[a] University of Science and Technology of China, Hefei, Anhui, PR China
[b] Laboratoire de Sciences Cognitives et Psycholinguistique, ENS, Paris, France
[c] Georgia Institute of Technology, Atlanta, GA, USA

ARTICLE INFO

Keywords:
Speech extraction
Child speech process
Dynamic mask
Iterative adaptation based separation
Real-world multilingual condition

ABSTRACT

We develop two improvements over our previously-proposed joint enhancement and separation (JES) framework for child speech extraction in real-world multilingual scenarios. First, we introduce an iterative adaptation based separation (IAS) technique to iteratively fine-tune our pre-trained separation model in JES using data from real scenes to adapt the model. Second, to purify the training data, we propose a dynamic mask separation (DMS) technique with variable lengths in movable windows to locate meaningful speech segments using a scale-invariant signal-to-noise ratio (SI-SNR) objective. With DMS on top of IAS, called DMS+IAS, the combined technique can remove a large number of noise backgrounds and correctly locate speech regions in utterances recorded under real-world scenarios. Evaluated on the BabyTrain corpus, our proposed IAS system achieves consistent extraction performance improvements when compared to our previously-proposed JES framework. Moreover, experimental results also show that the proposed DMS+IAS technique can further improve the quality of separated child speech in real-world scenarios and obtain a relatively good extraction performance in difficult situations where adult speech is mixed with child speech.

1. Introduction

1.1. Child speech processing and challenges

Processing child speech is a crucial diagnostic tool for detecting early childhood diseases and understanding the intentions of children who lack language expression abilities (Gilkerson et al., 2008; Wang et al., 2018, 2020; Yeung et al., 2021). In recent decades, this topic has attracted attention from researchers in academic fields such as developmental psychology (Sattorovich, 2022) and cognitive science (Slobin, 2021), as well as in application domains such as diagnosing underlying language disorders and measuring intervention effects (Kohnert et al., 2020; Hus and Segal, 2021).

In 2008, the Language Environment Analysis (LENA) Foundation launched a software tool (Gilkerson et al., 2008) that was trained on a 150-h, hand-annotated dataset using Mel Frequency Cepstral Coefficients (MFCC) features (Godino-Llorente et al., 2006) and Gaussian Mixture Models (GMM) (Reynolds, 2009). The LENA tool analyzes children's language environments and has significantly impacted the field of child speech processing. However, its high starting cost of at least US $5000 makes large-scale implementation challenging. In recent years, there has been a surge of efforts in child-focused recordings and related research. A new corpus, called Tong, which documents language behaviors and monitors children's continuous development, was published in Xiangjun and Yip (2018). Although this corpus has attracted widespread attention and aims to investigate the influence of environmental factors on language behavior, access to the Tong corpus may be restricted to researchers affiliated with specific institutions, potentially limiting its widespread use within the research community. Tang et al. (2019) studied the challenges that language learning poses to children as the phonological environment changes. Despite the researchers providing open-source access, the study primarily focuses on the Mandarin tone sandhi process, limiting the generalizability of the findings to other phonological alternations or languages.

Open-source access and a relatively large volume of data are essential for research advancement. Established in 1984, the Children's Language Data Exchange System (CHILDES) and its database-formatted counterpart, childes-db (Sanchez et al., 2019; MacWhinney and Snow, 1985; MacWhinney, 1996, 2000, 2001, 2014) provide a large and

[*] Corresponding author.
E-mail addresses: chengshi@mail.ustc.edu.cn (S. Cheng), jundu@ustc.edu.cn (J. Du), niust@mail.ustc.edu.cn (S. Niu), alecristia@gmail.com (A. Cristia), wx0304@mail.ustc.edu.cn (X. Wang), qingwang2@ustc.edu.cn (Q. Wang), chinhui.lee@ece.gatech.edu (C.-H. Lee).

https://doi.org/10.1016/j.specom.2023.102956
Received 3 December 2022; Received in revised form 12 June 2023; Accepted 21 June 2023
Available online 27 June 2023

diverse collection of language data, making it an invaluable resource for researchers studying child language acquisition and development across various languages and sociocultural backgrounds. It was also used in the ADReSS challenge at INTERSPEECH 2020 (Luz et al., 2020) to explore automatic recognition methods of spontaneous speech for Alzheimer's patients at different ages, with childs-db included. Although CHILDES has been a pioneer in disseminating large-scale datasets about child speech behaviors, transcripts in CHILDES may not always be in a standardized format or may require additional processing, such as format conversion, cleaning, annotation, or parsing. This can be time-consuming for researchers and may create challenges in data analysis and comparison.

In Lyakso et al. (2019), a Russian phonetic database, AD-Child.Ru, was presented, which contains phonetic data of children aged 4 to 16 accompanied by detailed records. Researchers revealed the development of correlations between language uniqueness and differences in children with autism spectrum disorder on this dataset in Lyakso and Frolova (2020). AD-Child.Ru was also used to create an efficient tool for the semi-automatic detection of axial spondylopathy (axSpA) in patients with bone marrow edema lesions (Kucybała et al., 2020). However, the generalizability of the findings and applications developed using this corpus to other languages and cultures is uncertain, as the dataset is specifically a monolingual study of Russian-speaking children. Indeed, for various reasons, speech enhancement and separation have been proven to be more challenging in multilingual scenarios (Watanabe et al., 2017; Zhou et al., 2018). Some of the reasons include the increase in acoustic variability due to multiple languages and language-specific noise characteristics (Jansen and Van Durme, 2011; Delcroix et al., 2015), the complexity of model training (Hershey et al., 2016a), and the lack of resources for low-resource languages (Chen et al., 2020).

In Wang et al. (2020), we proposed a child speech extraction system using joint speech enhancement and speech separation (JES) in real-world conditions on BabyTrain (Lavechin et al., 2020), a multilingual real-scene large dataset. Using speech enhancement as the pre-processing for speech separation, the joint system leads to a preliminary performance improvement in child speech extraction compared to direct-mapped binary classification networks. Nevertheless, JES treats BabyTrain as a whole dataset in the real-world scenario and does not fully exploit its multilingual potential across different subsets. Therefore, we take a deeper look at it in this paper. Since the work of this paper is based on (Wang et al., 2020), we will provide a more detailed introduction to the scheme and the BabyTrain dataset in Section 2.

These children's datasets have attracted significant interest in various fields and led to some novel technical approaches. However, the processing of child speech faces numerous difficulties. Firstly, most of the recordings are collected from devices worn by children throughout the day. As a result, these recordings often contain a large number of non-speech vocalizations, such as crying, snoring, and screaming. Additionally, adults play crucial roles as participants and companions in children's living environments. Consequently, child speech is often accompanied by adult speech, and many adults even imitate children's voices. Such speech mixtures present considerable challenges for researchers. To address these challenges, more advanced voice signal front-end processing technology is needed.

1.2. Speech enhancement and separation techniques

Enhancement and source separation are two key front-end signal processing techniques. Speech enhancement can be used to suppress background noise, while speech separation aims to separate target speech from a speech mixture known as the "cocktail party" problem (Cherry, 1953; Arons, 1992; Haykin and Chen, 2005; Bee and Micheyl, 2008). These techniques have led to a series of cutting-edge applications in automatic speech recognition (ASR) (Demir et al., 2012; Kanda et al., 2019), sound event detection (SED) (Heittola et al., 2011; Kong et al., 2018; Turpault et al., 2020), and other areas such as call customer service channels (Rustamov et al., 2019), multi-speaker meeting minutes (Raj et al., 2021), and target instruction extraction of smart speakers in domestic settings (Ling et al., 2021). Speech enhancement and speech separation methods have undergone a long period of development. Before the advent of deep learning methods, the non-negative matrix factorization (NMF) method has been the mainstream speech separation method (Vincent et al., 2014; Virtanen et al., 2015; Wood et al., 2017) and a series of related technologies have been derived from NMF. In Wood et al. (2017), by combining unsupervised dictionary learning of non-negative matrix factorization with spatial localization using the generalized cross-correlation method, a flexible blind source separation algorithm called GCC-NMF was proposed and demonstrated. The deep NMF method was introduced in Le Roux et al. (2015) and shown to be competitive with deep neural networks on the 2nd CHIME Speech Separation and Recognition Challenge corpus (Vincent et al., 2013).

The above traditional algorithms have achieved significant results in speech enhancement and separation tasks. However, the advantages of deep learning have allowed it to quickly replace traditional algorithms in many artificial intelligence fields, including the speech signal processing area, leading to the emergence of various speech separation methods. A joint deep neural network and recurrent neural network optimization system was proposed in Huang et al. (2014) and evaluated on the TIMIT speech corpus (Garofolo et al., 1993b). Compared to NMF models, this system achieved approximately 3.8–4.9 dB signal-to-interference ratio (SIR) gain while maintaining better source-to-distortion ratio (SDR) and source-to-artifact ratio (SAR) (Greenberg et al., 1993; Vincent et al., 2006; Emiya et al., 2011; Le Roux et al., 2019). In recent years, transformers have become popular and achieved state-of-the-art (SOTA) performances in many artificial intelligence fields. In Subakan et al. (2021), the transformer-based SepFormer model was applied to the standard WSJ0-2MIX and WSJ0-3MIX datasets, obtaining 22.3 dB and 19.5 dB SI-SNR gains. Recent research on deep learning-based speech separation has also demonstrated that time-domain methods outperform traditional time–frequency-based methods on some simulated data. In Luo and Mesgarani (2019), a fully convolutional end-to-end temporal audio separation deep learning framework (Conv-TasNet) was proposed, which significantly outperformed previous time–frequency masking methods in separating two-speaker and three-speaker mixed speech (Kavalerov et al., 2019; Ditter and Gerkmann, 2020).

Despite the continuous emergence of new network structures, recurrent neural networks (such as RNN, LSTM, GRU, and Bi-LSTM), with their inherent advantages in time series modeling, have long dominated sequence-to-sequence tasks. Among them, the representative LSTM has been widely used in sequence modeling and has proven to be very effective on commonly used datasets such WSJ0 (Garofolo et al., 1993a; Hershey et al., 2016b), AISHELL corpus (Bu et al., 2017), TIMIT (Garofolo et al., 1993b), and CHIME series challenges' datasets (Vincent et al., 2016; Barker et al., 2018; Watanabe et al., 2020). However, the vast majority of researches focus on simulation data. In Wang et al. (2018), we validated the excellent performance of the progressive learning strategy for child speech separation on simulated data. Progressive learning can bring certain improvements to speech extraction, but it is challenging to alleviate some complex noises in the audio, especially in real-world scenarios with complex audio conditions. As for real data, the use of Conformer instead of recurrent neural networks for the separation model introduced significant performance gains in both word error rate (WER) and speaker-attributed WER in Chen et al. (2021). In Shi et al. (2020), a common strategy called Speaker-Conditional Chain Model was proposed to process complex speech recordings. With the predicted speaker information from the whole observation, the proposed model has been proven to help solve the problem of conventional speech separation and speaker extraction for multi-round long recordings under real scenarios. Despite some

Table 1
Description of data in the BabyTrain (Lavechin et al., 2020) corpus. As shown in the leftmost column, the child-centered subsets cover a wide range of conditions, including various scenarios in different languages. The "Tot. Dur." column gives the total duration of the corresponding subset. The next few columns denote the accumulated duration of different speakers, including KCHI: key children; OCH: other children; MAL: male adults; FEM: female adults; UNK: unknown speakers.

Subset	Language	Tot. Dur. (h)	KCHI (h)	OCHI (h)	MAL (h)	FEM (h)	UNK (h)	Quality	Multi-scenarios	Overlapped
War2	English(US)	0.83	0.23	0.00	0.00	0.00	0.15	Fair	✗	✓
Paido	Greek, Eng., Jap.	40.13	10.93	0.00	0.00	0.00	0.00	Good	✗	✗
Vanuatu	Mixture	2.48	0.20	0.08	0.08	0.15	0.02	Poor	✓	✓
Tsimane	Tsimane	9.50	0.62	0.38	0.18	0.47	0.00	Poor	✓	✓
Namibia	Ju|'hoan	23.73	1.93	1.53	0.68	2.37	1.02	Fair	✓	✓
Lena_lyon	French	26.85	4.55	1.23	1.15	5.03	1.00	Poor	✓	✓
Aclew_starter	Mixture	1.50	0.17	0.08	0.10	0.33	0.00	Fair	✗	✓

completed and ongoing work, there is still a lack of relevant research in real-world scenarios.

In recent years, many unsupervised adaptation techniques that typically utilize the pseudo-labeling methods have shown promising results in various fields, including text classification (Liang et al., 2017; Xie et al., 2020), and natural language understanding (Liu et al., 2021). Pseudo-labeling methods also lead to various real-world speech processing applications, such as speaker diarization (Takashima et al., 2021), and speech recognition (Park et al., 2019; Laine and Aila, 2016).

Here, we first propose a new iterative adaptation based separation (IAS) framework based on pseudo-labeling methods. The proposed IAS framework is an application of pseudo-labeling and contains a speech enhancement model and a single-output pre-trained speech separation model which only extracts child speech. However, the pseudo labels often contain errors which will mislead the adapted model. In addition, as discussed above, child speech often tends to be present in extremely complex speech scenes, so simply adopting a fine-tuning strategy may not yield remarkable results. We find that during the iterative process, there was still a lot of noises that could not be effectively removed. Therefore, we hope to further explore some methods to locate children's speech in complex acoustic environments and adopt more stringent suppression methods for other parts. Accordingly, a dynamic mask based separation (DMS) framework following IAS (DMS+IAS) is further proposed. By adjusting SI-SNR in Luo and Mesgarani (2019), a variable length-position dynamic mask can be obtained and used to mask the noise regions, alleviating the interference of errors in pseudo-labels. The results in Section 4 show that IAS and DMS+IAS can achieve better results than JES, with DMS+IAS attaining the best results on all selected subsets.

The remainder of this paper is organized as follows. In Section 2, we describe the BabyTrain corpus and review our previous works. In Section 3, we elaborate on the proposed iterative adaptation and dynamic mask based separation techniques to improve JES. Experimental results with detailed analyses are presented in Section 4. Finally, we conclude our findings and discuss some future work in Section 5.

2. BabyTrain corpus and prior work

As mentioned in Section 1, there has only been a small collection of child-centric speech corpora in recent years. However, most of them are monolingual or single-speaker, and the corresponding acoustic scenes are usually narrow, making it difficult to address various complex scenes in the real world. In our previous study (Wang et al., 2020), we have proposed some front-end techniques based on the real-world multilingual BabyTrain corpus, which offers various complex scenes. The two techniques proposed in this paper are derived from our previous works, so we first introduce the BabyTrain dataset and review some previously-proposed algorithms next.

2.1. Data analysis

BabyTrain is a large corpus containing several child-centered subsets ranging in age from 1 month to 5 years old (Bergelson et al., 2017; Canault et al., 2016; VanDam et al., 2016; Pretzer et al., 2019). Each recording is sampled at 44.1 kHz with a human transcription. It contains 245-h recordings of various adverse environments in different languages. Its sufficient amount of comprehensive data makes its recording style cover almost all typical life scenes, including daily life, indoor, outdoor, party scenes, etc. Table 1 gives a broad description of the selected subsets of BabyTrain. The total duration of the selected part is 105.02 h, which are sufficient to demonstrate the effectiveness of our method in multilingual real-world scenarios. Some subsets even contain recordings belonging to different scenes, such as two-person conversations, multi-person gatherings, etc. Moreover, the recording equipment is worn by children with friction and obscuring by clothing, resulting in poor audio quality in most parts of each recording, making it difficult for conventional front-end processing. To visually demonstrate that the BabyTrain utterances cover a wide range of acoustic scenes, we plot spectrograms of three samples in Fig. 1. Fig. 1(a) indicates a recording of a sleeping child snoring, while Fig. 1(b) represents the audio of a father singing with his children at a family party. Finally, in Fig. 1(c), a complex dialogue in a family gathering. As shown in the figures, some recordings cover life scenarios with background noises and overlap speech segments accounting for a significant proportion of the recordings. Moreover, BabyTrain contains both far-field and near-field speech. All of these present significant challenges for our child speech extraction task to be discussed next.

2.2. Joint enhancement and separation

Due to the influence of unavoidable noises in real-world audio recordings, it is often necessary to add a speech enhancement front-end to remove the noises before the separation model. Based on this idea, our previously-proposed JES system combines speech enhancement and source separation (Wang et al., 2020) to deal with the complex scenes in the BabyTrain corpus.

The processing flow of our baseline JES system is shown in the middle part of Fig. 2. Feature extraction is first performed to extract log power spectrum (LPS) features from speech, followed by speech enhancement and source separation. To build the baseline JES system, we train a pair of enhancement and separation models, as shown in the left-hand and right-hand dotted boxes, respectively. Note that our enhancement network is directly adopted from Sun et al. (2020). It is a multi-target Bi-LSTM network that uses LPS and ideal ratio mask (IRM) as learning targets. LPS is a very commonly used feature type extracted from speech signal and IRM is defined as in Eq. (1), where $s_{t,f}$ and $n_{t,f}$ denote the power spectrum of child and adult speech signals at the time–frequency (T–F) unit (t, f), respectively.

$$z_{t,f}^{\mathrm{IRM}} = \frac{s_{t,f}}{s_{t,f} + n_{t,f}} \tag{1}$$

LPS and IRM are commonly used in speech separation, where the former has good preservation of the spectrogram but cannot completely remove interfering speech, and the latter can remove interfering speech more cleanly but can cause spectrogram loss (Bao and Abdulla, 2018). Learning both targets can effectively remove noise while preserving the target speech.

As for the separation network, we use a 3-layer Bi-LSTM network. In the definition of IRM, $s_{t,f}$ refers to the power spectrum of clean

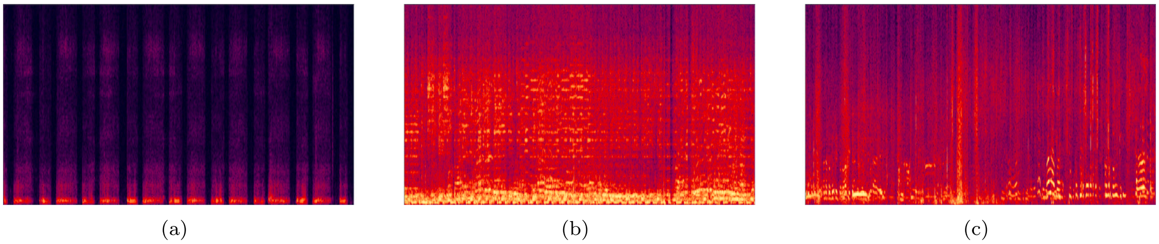


Fig. 1. Spectrograms of three recording samples from the BabyTrain corpus in different scenarios. They are all recorded by equipments worn on children for 24 h a day: (a) a clip of a child snoring while sleeping, (b) a pair of father and son singing to the music at a party, and (c) a family conversation scenario.

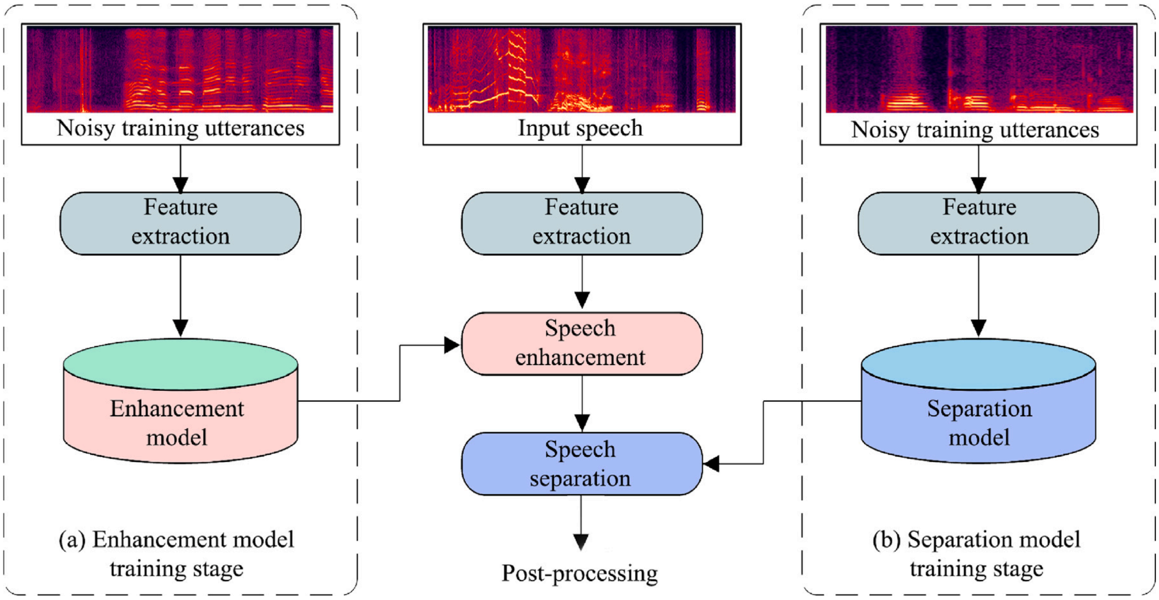


Fig. 2. Framework of our proposed joint enhancement and separation system, in which (a) and (b) represent the enhancement and separation model training stage, respectively.

child speech. However, the model does not directly learn clean child speech but rather learns progressively denoised speech. The targets in the model gradually reduce the noise present in the speech signal, ultimately approaching the target clean speech. Likewise, the LPS features extracted from the speech signal are also obtained through a step-by-step approximation process. Therefore, The learning targets actually are progressive log power spectrum feature (PLPS) and progressive ratio mask (PRM), denoted as $\mathbf{Z}^{\mathrm{PLPS}}$ and $\mathbf{Z}^{\mathrm{PRM}}$. The former represents the LPS target for each progressive layer, and the latter can be calculated by Eq. (2).

$$z_{t,f}^{\mathrm{PRM}} = \frac{s_{t,f} + n'_{t,f}}{s_{t,f} + n_{t,f}} \tag{2}$$

where $z^{\mathrm{PRM}}(t,f)$ is the value of $\mathbf{Z}^{\mathrm{PRM}}$ at the time–frequency (T–F) unit (t,f), and $n'_{t,f}$ is the power spectrum of the residual adult speech in the target speech. Note that the output of each layer is simultaneously concatenated to the LPS of the original speech as input to the next layer. Further detail can be found in Wang et al. (2020) and Sun et al. (2020).

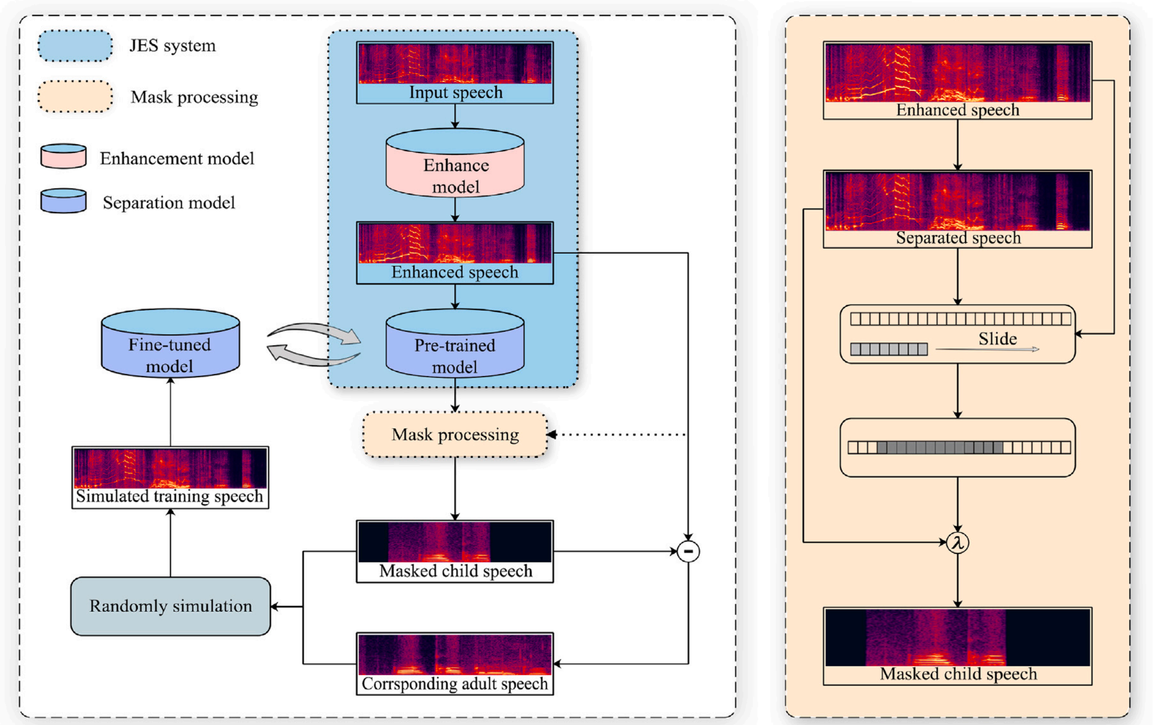
Due to the diverse nature of the BabyTrain corpus, which includes multi-language and multi-scenario subsets such as Namibia, Lena_lyon, War2, and Tsimane, the simple JES processing may not yield satisfactory results. To address this issue, we propose the iterative adaptation and dynamic mask-based separation methods to enhance the baseline separation model. This is achieved by fine-tuning the model with specific data from each subset, leading to improved performance in various scenarios.

3. Proposed iterative adaptation and dynamic mask based separation

Our proposed frameworks build upon the previously introduced JES system, as shown within the blue dashed box in Fig. 3(a). The separation model, which is trained on the entire BabyTrain corpus, serves as our pre-trained model and forms the foundation for our subsequent experiments. During the adaptive phase, we utilize the development set to construct the training data for each BabyTrain subset and evaluate the performance on the corresponding test set. In this process, our fine-tuned models and the subset category are one-to-one, which is equivalent to having this prior knowledge in the first place during inference.

3.1. Iterative adaptation based separation

First of all, as JES is trained and tested on the entire BabyTrain training set, we adopt JES to build a baseline system. Although JES yields satisfactory results on the entire dataset, we believe that the performance can be further improved for test data within each subset, leading to better overall results. Therefore, we propose an iterative adaptation based separation technique, denoted as IAS. In the IAS system, the separated speech obtained with the pre-trained model is treated as clean speech, and a new training set is created for each subset to fine-tune the pre-trained model. Specifically, we segment the separated speech into one-second intervals. Then, we calculate the corresponding adult speech using Eq. (3), where $\mathbf{x}_a$ represents adult



(a) Framework of our proposed systems

(b) Details of the mask processing

Fig. 3. Framework of our iterative adaptation based separation system with (DMS+IAS) and without (IAS) mask processing.

speech to be obtained, $\mathbf{x}_e$ and $\mathbf{x}_c$ denote the enhanced speech and child speech, respectively. Note that we only focus on child speech and simply remove other sources of audio. Therefore, our model adopts a single-output separation model to extract child speech.

$$\mathbf{x}_a = \mathbf{x}_e - \mathbf{x}_c \tag{3}$$

We randomly mix child and adult speech for each subset to construct a new training set. We then utilize the newly formed training set to fine-tune our pre-trained separation model, which has been updated in the previous iteration. Each fine-tuned model will help to generate the data to train the new pre-trained model in the next iteration. $\mathbf{Z}_m^n$ and $\hat{\mathbf{Z}}_m^n$ are the nth $2D$-dimensional reference and estimated splicing vectors of D-dimensional PLPS feature vector and D-dimensional PRM vector for mth target layer, respectively, with $m = 1, \ldots, M$ representing the layer index and $n = 1, \ldots, N$ representing the mini-batch index. The loss function in mth target layer of our network is shown in Eq. (4):

$$E(m) = \frac{1}{N}\sum_{n=1}^{N} \|\mathcal{F}_m(\hat{\mathbf{Z}}_0^n, \hat{\mathbf{Z}}_1^n, \ldots, \hat{\mathbf{Z}}_{m-1}^n, \Lambda_m) - \mathbf{Z}_m^n\|_2^2 \tag{4}$$

$E(m)$ represents the MSE loss between the prediction $\mathcal{F}_m$ and target $\mathbf{Z}_m^n$ at layer m. Here, we make it clear that $\mathcal{F}_m(\hat{\mathbf{Z}}_0^n, \hat{\mathbf{Z}}_1^n, \ldots, \hat{\mathbf{Z}}_{m-1}^n, \Lambda_m)$ is the mth layer's prediction based on the learned target $\hat{\mathbf{Z}}_0^n$ to $\hat{\mathbf{Z}}_{m-1}^n$ and the parameter set of the weight matrix and bias vector Λ_m before the mth target layer. Λ_m is optimized using gradient descent in a backpropagation through time (BPTT) manner. We stop the iteration when the error rate related metrics cease to decrease. Then, the corresponding separation model will be chosen as our separation model in the test stage. A diagram of this framework is shown in Fig. 3(a), with the dotted box part named *Mask processing* skipped. Algorithm 1 presents the specific operation flow of IAS, in which the SSfine-tuned and SSpre-trained represent the fine-tuned speech separation model and the pre-trained speech separation model, respectively. Considering the small amount of data in each development set, we only update the parameters of the fully connected layer of the network, effectively preventing the over-fitting problem. Simultaneously, the network can adapt to a specific subset without altering the information learned by the Bi-LSTM layers from large-scale data.

Algorithm 1 Iterative adaptation based separation.

```
JES results: Use the previous JES system to obtain the preliminary enhanced speech and separated speech, cut them into one-second segments, note as $\mathbf{x}_e$ and $\mathbf{x}_s$;
Initial inputs: Set $i = 1$, note that $\mathbf{x}_s^i$ represents separated speech in iteration $i$ and the JES system's output $\mathbf{x}_s = \mathbf{x}_s^0$;
while $i$-th iteration do
    $\mathbf{x}_c^i = \mathbf{x}_s^{i-1}$;
    $\mathbf{x}_a^i = \mathbf{x}_e - \mathbf{x}_s^{i-1}$;
    randomly mix $\mathbf{x}_c^i$ and $\mathbf{x}_a^i$ up to build a new training set;
    $\mathrm{SS}_{\text{pre-trained}} \rightarrow \mathrm{SS}_{\text{fine-tuned}}$;
    if the error rate on validation set stops decreasing then
        break;
    end if
    $i = i+1$;
end while
Choose the model with the lowest error rate and apply it to the test set.
```

In the post-processing stage, we utilize the fine-tuned model of each iteration corresponding to each subset for decoding and calculating performance metrics.

The decoding formula is shown in Eq. (5), in which $\hat{\mathbf{Z}}^{\text{LPS}}$ is the estimated decoded LPS of test speech, $\mathbf{Z}^{\text{LPS}}$ represents that of the noisy input speech, and $\hat{\mathbf{Z}}_M^{\text{PRM}}$ stands for the final layer's output PRM of our system.

$$\hat{\mathbf{Z}}^{\text{LPS}} = \mathbf{Z}^{\text{LPS}} + \ln(\hat{\mathbf{Z}}_M^{\text{PRM}}) \tag{5}$$

At the end of each iteration, we calculate the relevant metrics. Due to our focus on child speech separation in real-world scenarios, it is not possible to obtain clean references, making it difficult to calculate

Algorithm 2 Procedure of the post-processing.

```
Get masks: Take the fine-tuned model's output $\mathbf{Z}^{\mathrm{PRM}} \in \mathbb{R}^{T \times D}$ as the mask, where $T$ represents number of frames and $D = 257$ is the dimension of IRM;
Calculate the mean: For a given frame $t$, $\bar{z}^{\mathrm{PRM}}(t) = \frac{1}{D}\sum_{f=1}^{D} z^{\mathrm{PRM}}(t,f)$ is the mean of $\mathbf{Z}^{\mathrm{PRM}}$ over dimension $D$ at frame $t$;
Make decision: Set the threshold $th$, suppose $\hat{y}(t)$ is the predicted label at frame $t$;
  for t = 1 to $T$ do
    if frame $t$ is not silent frame then
      if $\bar{z}^{\mathrm{PRM}}(t) >= th$ then
        label $\hat{y}(t)$ as 'Child';
      else
        label $\hat{y}(t)$ as 'Adult';
      end if
    end if
  end for
Obtain labels: Compare obtained binary labels and calculate metrics.
```

commonly used separation metrics such as scale -invariant source-to-noise ratio (SI-SNR) (Luo and Mesgarani, 2018) and SDR (Jörnvall et al., 1995). Therefore, inspired by Gilkerson et al. (2008) and Hamers et al. (1989), we proposed Jaccard error rate (JER) and child speech duration error rate (CSDER) for child speech separation in real scenarios based on commonly used binary classification indicators in Wang et al. (2020), which are defined in Eqs. (6) and (7),

$$\mathrm{JER} = \frac{\mathrm{FA} + \mathrm{Miss}}{\mathrm{Total}} = \frac{\mathrm{FN} + \mathrm{FP}}{\mathrm{FN} + \mathrm{FP} + \mathrm{TP} + \mathrm{TN}} \tag{6}$$

$$\mathrm{CSDER} = \frac{|\mathrm{ECSD} - \mathrm{OCSD}|}{\mathrm{Total}} \tag{7}$$

$$\mathrm{BER} = \frac{\mathrm{FA}_{\mathrm{Rate}} + \mathrm{Miss}_{\mathrm{Rate}}}{2} = \frac{1}{2}\left(\frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} + \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}\right) \tag{8}$$

where ECSD represents the child speech duration time detected by our system and OCSD is the oracle child speech duration time. Total is the duration of the union of child and adult speaker segments, FA is the total child speaker time detected by our system but not attributed to the reference child speaker, and Miss is the total reference child speaker time but not detected by the system. It is worth mentioning that such FA and Miss are not our commonly used false alarm rate and missed alarm rate, but simply the number of wrong labels. In this paper, we switch them to the false and missed alarm rates, which are widely used in machine learning. The corresponding JER is also replaced by the balanced error rate (BER), which can more reasonably characterize the system's capabilities. The calculation of BER is shown in Eq. (8), and BER and CSDER are adopted as the evaluation metrics in this paper. The separated speech is subjected to our post-processing method to obtain the corresponding binary classification labels. Algorithm 2 demonstrates the post-processing procedure: we use the IRM nodes of the output layer to generate the separation masks for each frame. Then we calculate the mean of these masks over frequency domain dimensions for each frame, and check the frames whose measured mean is greater than a pre-defined threshold. Then these frames will be labeled as child speech segments and the rest of the non-silent segments are regarded as adult speech segments.

3.2. Dynamic mask based separation

As discussed earlier, due to the considerable complexity of the BabyTrain corpus, it is often insufficient to directly treat speech decoded from the pre-trained separation model as clean child speech. To address this issue, we propose a variable length-position dynamic mask based processing approach, denoted as DMS+IAS, as illustrated in the dotted box in Fig. 3(a).

The main difference between DMS+IAS and IAS (as described in Section 3.1) lies in the *Mask processing* step. Fig. 3(b) provides a detailed illustration of this module. After obtaining separated speech from the pre-trained model, we take it a step further by employing a dynamic mask generation process to eliminate residual noises instead of simply treating it as clean speech. As shown in Fig. 3(b), we use an activated window to slide over the speech segment to locate child speech and mask the remaining part.

Since our activated segment of the mask is continuous, and child speech and other interfering speech might be interlaced in longer speech segments, we divide the recordings into one-second segments to ensure that there is only one target in each segment, making it more convenient for the dynamic mask to locate it. The length and position of this mask are variable, and we will describe how to determine them next. Once the masked speech is treated as child speech, the subsequent operations remain the same as in Section 3.1.

3.2.1. Determining lengths of activated windows

As mentioned earlier, the pre-trained model cannot perfectly extract child speech. However, for specific speech segments, child speech does retain the main parts of the separated speech. With this in mind, we can locate the children's parts in the separated speech and remove the rest. We adopt a commonly used separation metric, SI-SNR (Luo and Mesgarani, 2018), to construct an indicator as a rule for calculating the active length of the dynamic mask:

$$s = \text{SI-SNR}(\mathbf{x}_s, \mathbf{x}_e) = 10\log_{10}\frac{\|\mathbf{x}_t\|^2}{\|\mathbf{x}_n\|^2} \tag{9}$$

where $\mathbf{x}_t = \frac{\langle \mathbf{x}_s, \mathbf{x}_e \rangle \mathbf{x}_e}{\|\mathbf{x}_e\|^2}$, $\mathbf{x}_n = \mathbf{x}_s - \mathbf{x}_t$ and SI-SNR is the function used to calculate SI-SNR. After gaining s, we want to use it to scale the active percentages of speech segments to [0, 1]. Inspired by the Sigmoid function (Han and Moraga, 1995; Yin et al., 2003), we construct a mapping function to achieve this goal, which is referred to as the variable length mapping function (VLM) in Eq. (10).

$$\mathrm{VLM}(s) = \max\left\{\frac{1}{1 + \exp(-\alpha \times s)}, 0.5\right\}, \quad \beta_1 < s < \beta_2 \tag{10}$$

When $s > \beta_2$, i.e., our pre-trained model retains most of the speech, we believe the separated speech can be considered clean speech, and we choose not to mask these segments, setting $\mathrm{VLM}(s) = 1$. As for $s < \beta_1$, which means these segments mostly contain noises, we fully mask them, i.e., $\mathrm{VLM}(s) = 0$. For each subset, we visualize the distribution of all speech and set β_1 = the lower 95% confidence interval boundary and β_2 = the median, where α is a hyperparameter. Fig. 4(a) displays the distribution of SI-SNR on the Tsimane subset and the corresponding VLM function. The two black dotted lines on the left and right represent β_1 and β_2, which can be automatically determined. The grey histograms represent the corresponding SI-SNR distributions on the Tsimane dataset before and after the pre-trained separation model. Fig. 4(b) shows the mapping relationship between different SI-SNR values and VLM values. Through this mapping relationship, we successfully establish the mapping of real values in the [0,1] range. Consequently, we can determine the active length l of speech through Eq. (11). Note that L is the frame number of the segment (one second), and $\lfloor * \rfloor$ represents the floor function.

$$l = \lfloor L \times \mathrm{VLM}(s) \rfloor \tag{11}$$

It is also worth noting that for $\beta_1 < s < \beta_2$, we set the minimum active length of the dynamic mask to $\frac{L}{2}$, as shown in Eqs. (10) and (11). Our experimental results demonstrate that the complex overlapping segments of child and adult speech may cause the SI-SNR value of speech (especially the overlapping segment speech) to fluctuate significantly. This constraint can help avoid generating too many fragments and preserve as much of the child speech as possible.

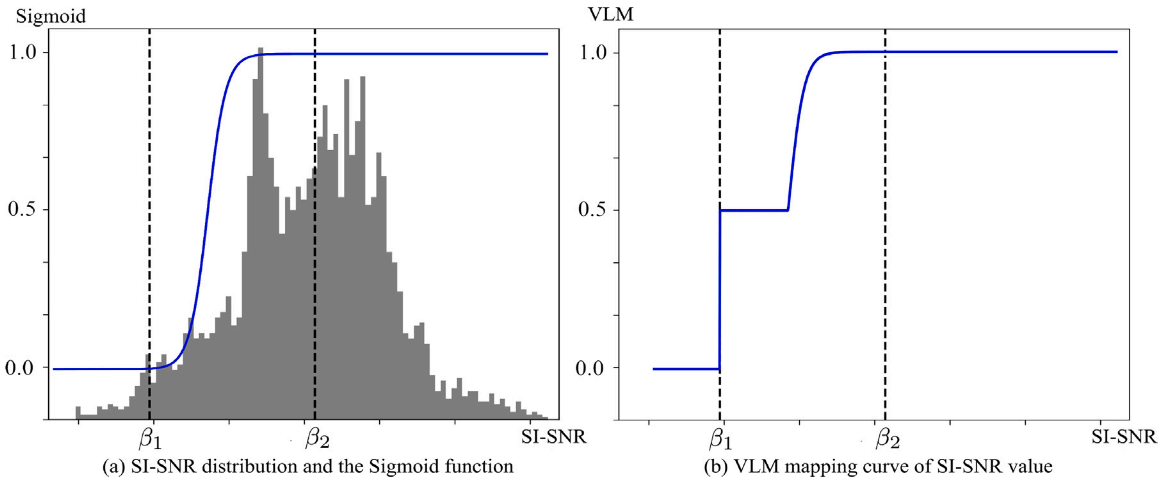


Fig. 4. An example of β_1, β_2 and corresponding SI-SNR distribution. The gray bars in (a) represent the distribution of SI-SNR value. The two blue curves represent the Sigmoid function and VLM function, respectively.

3.2.2. Determining locations of activated windows

As for the position determination problem, we employ a sliding window to find the start frame (SF) defined in Eq. (12). For t ranges from 0 to $L-l$, $\mathbf{x}_s^{t:t+l}$ and $\mathbf{x}_e^{t:t+l}$ represent separated and enhanced speech segments with length l and start frame t. We slide the sliding window with a $step = \frac{L}{1000}$ to find a position with the highest SI-SNR. The left endpoint t of such a window will be chosen as SF.

$$\mathrm{SF} = \operatorname*{argmax}_t(\text{SI-SNR}(\mathbf{x}_s^{t:t+l}, \mathbf{x}_e^{t:t+l})), \quad t \in [0, L-l] \tag{12}$$

In Fig. 3(b), λ represents our trust in the mask operation. As our confidence in the separation increases with each iteration, we can adjust the value of λ to reflect our current confidence in the separation results. In fact, our final clean speech $\mathbf{x}$ in each iteration is a weighted combination of the separated speech $\mathbf{x}_s$ and masked speech $\mathbf{x}_m$. In a sense, this can also be considered a method of model fusion. By adjusting the value of λ, we can control the contribution of the separated and masked speech to the final output, allowing for a more accurate and reliable separation result as the model's confidence increases.

$$\mathbf{x} = \lambda \times \mathbf{x}_m + (1-\lambda) \times \mathbf{x}_s \tag{13}$$

In Eqs. (14) and (15), $\mathbf{x}_m$ and λ are defined. It is important to note that the dynamic mask **dm** is actually a vector with the same dimension as the separated speech. The active frames are set to 1, and the rest are set to 0. This allows us to selectively apply the mask to the separated speech, preserving the desired child speech while suppressing the unwanted noise or adult speech components. The process of determining **dm** is illustrated in Fig. 3(b), which outlines the steps for calculating the variable length-position dynamic mask. By using this mask, the system can more accurately isolate child speech in each iteration, further improving the overall performance of the separation model.

$$\mathbf{x}_m = \mathbf{x}_s \odot \mathbf{dm} \tag{14}$$

$$\lambda = \begin{cases} 0.5 & \text{iter} = 1 \\ 1.0 & \text{iter} = 2, \dots \end{cases} \tag{15}$$

Indeed, as shown in Eq. (15), our confidence in the separation ability of the adapted model increases with each iteration. This is demonstrated in Fig. 3(a), where the process of updating the training data relies on a robust pre-trained model to assess speech quality and generate the appropriate dynamic mask. In turn, model optimization requires cleaner data to fine-tune the model parameters. As the pre-trained separation model becomes more capable of generating accurate dynamic masks, it can produce higher-quality adapted data. This improved data can then be used to further refine the adapted model, leading to better speech separation performance. In essence, the two modules dynamically form a closed loop, working together to enhance the overall separation ability of the system. Experimental results confirm that the dynamic mask can effectively address the challenges mentioned earlier in this section. By building upon the IAS method, the DMS+IAS approach offers even better speech separation performance, demonstrating the value of this joint optimization strategy.

4. Experiments and result analysis

4.1. Experimental settings

We focused on child speech extraction and took BabyTrain as the main dataset. We also introduced parts of some other datasets to improve our data diversity. The configurations of our enhancement experiments were the same as (Sun et al., 2020). In our pre-train stage, the adult speech data were derived from four data sets, namely the BabyTrain mega corpus, WSJ0 corpus (Garofolo et al., 1993a; Hershey et al., 2016b), part of AISHELL-1 corpus (Bu et al., 2017) and part of Librispeech corpus (Panayotov et al., 2015). The child speech segments were derived from two data sets: the BabyTrain mega corpus and the part with children aged from kindergarten to grade 5 of CSLU Kids Corpus (Shobaki et al., 2000). We utilized the ground-truth labels from the training set to obtain children's segments and adults' segments (ground-truth labels are divided into two categories: child and adult). Then, we randomly simulated child speech and adult speech to generate training data. 19562 children's utterances (about 55 h) were mixed with the above 58,686 adult utterances at seven target inference ratio (TIR) levels (−5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, and 25 dB) to construct a 500-h training set consisting of children's utterance pairs and mixed utterance pairs. Among them, the speech at the SNR of −5 dB, 0 dB, and 5 dB were used as the model's inputs. The segments at the SNR of 5 dB, 10 dB, and 15 dB were used as the first layers targets of progressive learning, those of 15 dB, 20 dB, and 25 dB were used as the second layer's targets, and the clean segments were the final learning targets. The BabyTrain development set was used for fine-tuning, and the BabyTrain test set was used for testing. All speech were resampled at 16 kHz, frame length was set to 32 ms and frameshift was 16 ms. The 512-point discrete Fourier transform (DFT) of each overlapping windowed frame is calculated. Then the pre-trained separation model was trained using the 257-dimensional LPS vectors with global mean and variance normalization. The outputs of each layer are the 257-dimensional PLPS and 257-dimensional PRM predicted by the model. It should be pointed out that the Paido data set is a child reading words at intervals in a tranquil environment. It does not contain any overlapping

Table 2
BER and CSDER values comparison. SS represents speech separation, and JES represents joint speech enhancement and speech separation. BER and CSDER denote balanced error rate and child speech duration error rate, respectively.

Systems	BER						CSDER					
	War2	Tsimane	Namibia	Vanuatu	Lena_lyon	Overall	War2	Tsimane	Namibia	Vanuatu	Lena_lyon	Overall
SS	0.396	0.409	0.452	0.435	0.461	0.443	0.163	0.389	0.328	0.346	0.243	0.327
JES (Wang et al., 2020)	0.373	0.408	0.437	0.426	0.448	0.432	0.123	0.383	0.306	0.342	0.243	0.313

segments, the scene is also very single, and the content is very clean. It is unsuitable for verifying the separation system in the real complex scene, so this dataset is not included in the subsequent experiments.

It is also worth mentioning that in the previous work, we used Microsoft CNTK (Seide and Agarwal, 2016) as our deep learning framework for model training and decoding, but here we migrated the prior work to the PyTorch (Paszke et al., 2019) framework due to applicability and cutting-edge issues. Our network structure is 3-layer progressive multi-target learning based Bi-LSTM (PMT-Bi-LSTM), M in Eq. (4) an Eq. (5) is 3. All experiments were conducted on GeForce RTX™ 3090. MSE was used as the optimization criterion in all pre-train and fine-tune stages. We used Adam as our optimizer, and the variable learning rate was set to 0.01 for the first 10 epochs and 0.005 for the rest epochs. Batchsize was 32 in the pre-train stage and 64 in the fine-tune stages. LPS features were used as our inputs in the training and decoding period. PLPS and PRM are adopted as our training targets for all Bi-LSTM layers. The input of the current layer and the estimations of the intermediate target are spliced together to learn the next target. The number of Bi-LSTM memory cells in each layer was 1024, and the PRM output of the final layer was used to decode the speech. α in Eq. (10) was set to 1.7, β_1 and β_2 were automatically decided according to the distribution of SI-SNR. In the post-processing part, oracle VAD information was used, and Kaldi[1] was applied to extract i-vectors (Dehak et al., 2011; Saon et al., 2013) to visualize the separated results. Note that our strategies and all parameter optimizations, including threshold setting, were done on the development sets and tested on test sets.

4.2. Ablation experiments for joint enhancement and separation

In Wang et al. (2020), we did not give the ablation experiments that quantitatively present the effectiveness of the enhancement model in a joint speech enhancement and separation system. So here we first present the ablation experiments of the enhancement model on the test sets in Table 2. From the table, we can see that the enhanced model improves the system's performance to a certain extent, but there is no noticeable improvement in the BER of the Tsimane set and the CSDER of the Lena_lyon set. Considering the fact that these two datasets are relatively more complex relative to the other datasets (as can be seen in Table 1), we believe that the quality of the data itself limits the improvement of system performance, which further increases the necessity in using dynamic masks to clean the data.

Fig. 5 compares the spectrograms before and after processing by the SS and the JES systems. The top rectangle represents the sound category, the blue segments represent the target child speech, the gray ones represent the non-negligible background noises, and the red ones represent adult speech. The circles on the spectrograms mark the target speech, and the boxes illustrate noises. It can be seen from the figure that the separation model performs better for relatively single target-speaker segments, such as the child speech in the first and third circles, as well as adult speech in the first boxes. However, for complex scenes with multiple overlaps, the enhancement model will inevitably cause some damage to the human voice while suppressing the noise. As a result, the voice distortion in the overlapping segment is relatively large (as shown in the second circle), which causes an inevitable loss of the target speech while suppressing the noises. In addition, the boxes on the far right show that the addition of the enhanced model can sufficiently suppress the low-frequency background noises, which can make the separated child speech purer.

4.3. Results using iterative adaptation

In general, the JES system can improve the performance of the separation model to a certain extent but is limited by factors such as the quality of speech. The separation performance is not significant. It has been further improved with the introduction of the proposed IAS system, and we will present the relevant results in this subsection. Since our adaptive training set is generated by cutting the development set and adding noise randomly, in order to avoid overfitting, we simply give the results of the development set and focus on the results of the test set. Moreover, the speakers in the test set are independent of the development set, which can better illustrate our method's effectiveness in different languages.

We selected the results of five representative subsets to draw a bar chart as shown in Fig. 6. The deep-blue bars represent the JES system, and the light-blue bars refer to the IAS system. It can be seen that compared with JES, IAS can further reduce the error rate on subsets of different languages, improve the separation accuracy, and realize a certain degree of adaptation. Table 3 presents an overall BER and CSDER comparison among different separation methods on selected subsets of the BabyTrain test set. Baseline represents the system trained only on the specific subset (i.e., no pre-trained JES). JES denotes joint speech enhancement and speech separation system. IAS stands for our newly proposed iterative adaptation based separation framework. JES+GT represents the JES system fine-tuned on supervised dev set using the ground-truth label. This would convey a clearer picture of how much of the errors can be recovered by our systems compared with using clean references. Since the development set (i.e., the training set for the adaptive stage) of each subset is relatively small, all the optimal values can be reached in the first few iterations, and then it begins to fluctuate. Moreover, different subsets achieve the optimal results in different iteration rounds, in order to avoid the results being too messy, we only give their respective optimal results here. It is worth mentioning that the results of other non-optimal iteration rounds are also generally better than JES. Our proposed IAS method achieves better results than the previous JES system, both on the subset with little data (e.g., War2 and Tsimane) and on the subset with relatively more data (e.g., Lena_lyon and Namibia). For example, the IAS method has an improvement of 2.3% and 4.3% in BER and CSDER on the Lena_lyon dataset compared with the JES method. By comparing the overall results of the baseline, JES+GT, and IAS, we find that the BERs of baseline and JES+GT are 0.491 and 0.400. In the BER of 0.091 that baseline is more than JES+GT, IAS can recover 74% of them and reach the BER of 0.423. Likewise, IAS can recover 63.0% of CSDER (0.378, 0.281, and 0.224 for baseline, IAS, and JES+GT). This shows that the IAS method can achieve good results in most cases. However, due to the constraints of voice quality, if the speech quality is very poor at the beginning, directly adopting an iterative strategy will make speech quality difficult to control or even get worse, such as in Vanuatu, which contains a significant portion of non-speech scenes, with only background noise. This phenomenon further enhances the confidence we have in improving the data quality. The dynamic mask can remove some noises well, making the voice quality controllable and bringing a lower error rate, reflected in Section 4.4.

[1] https://github.com/kaldi-asr/kaldi.

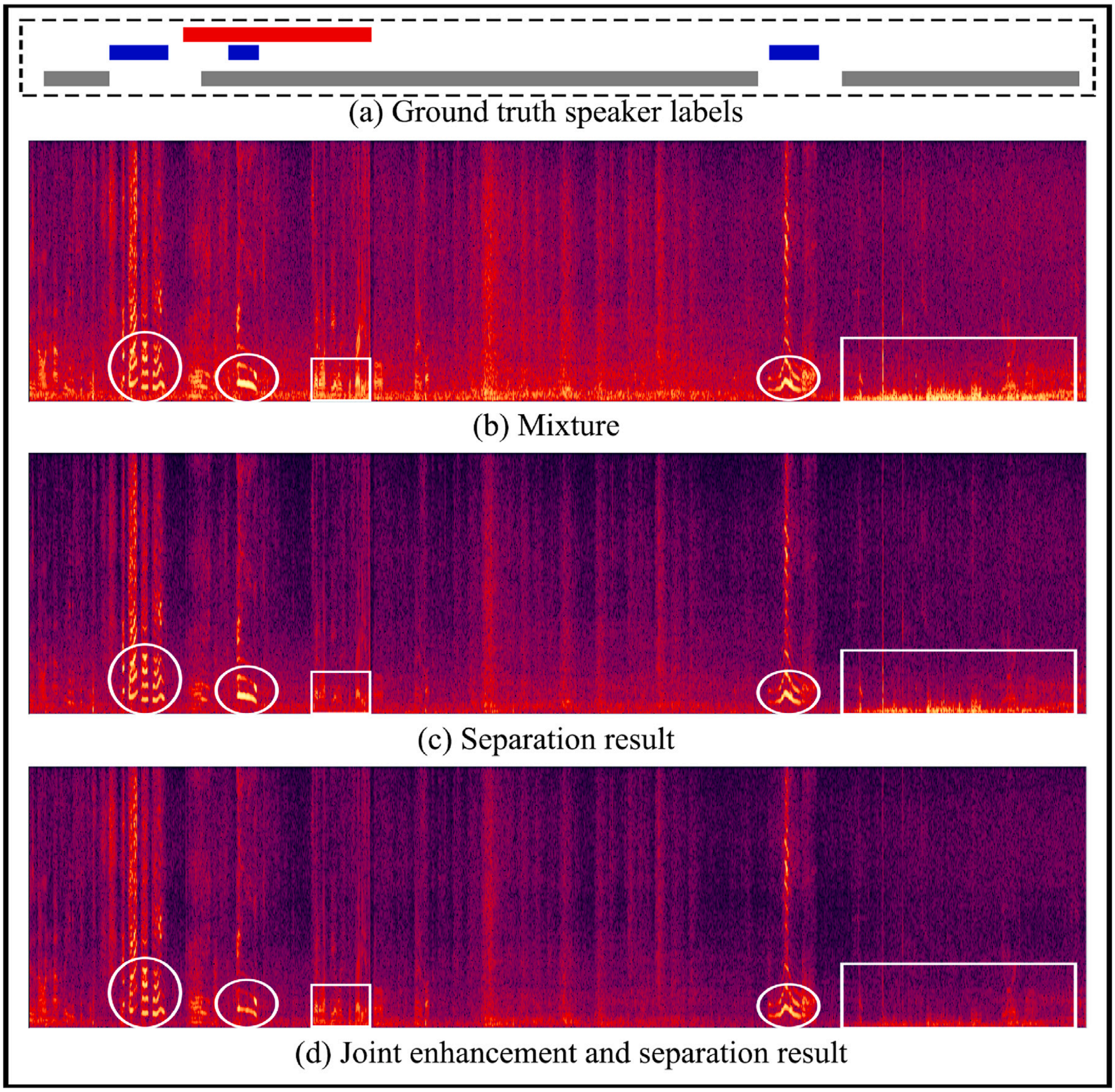


Fig. 5. Spectrograms comparison of an utterance from the test set. In (a), the red bars represent the speech regions of adults, while the blue bars represent the target child speech, and the gray bar denotes environmental noises. (b) gives the original spectrum. (c) and (d) show the results processed by SS and JES, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

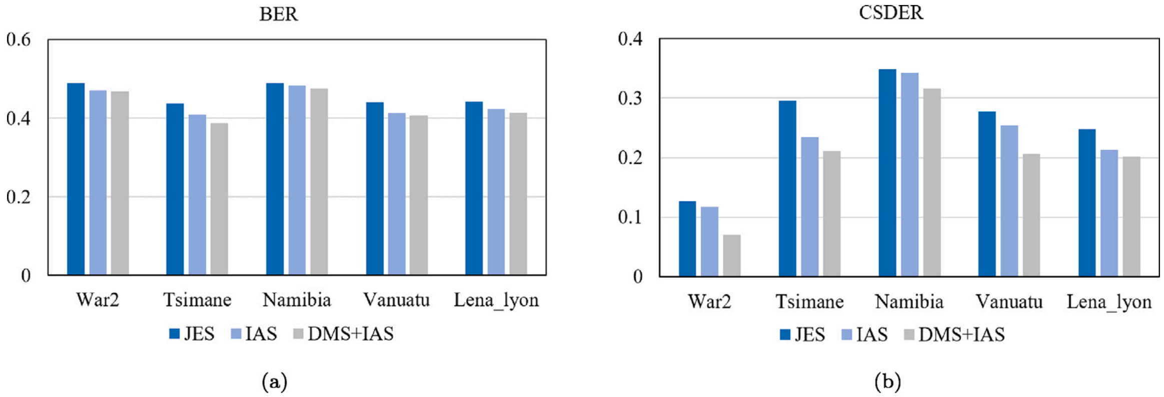
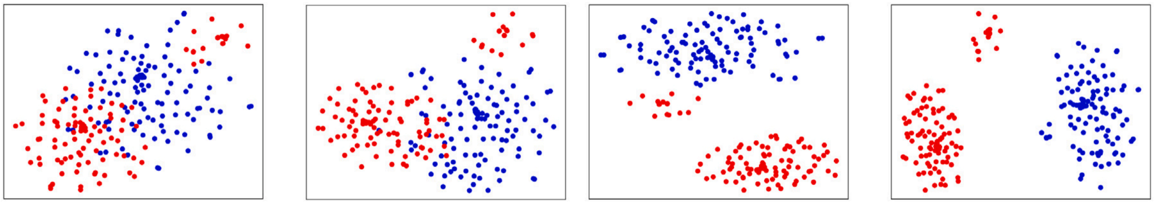


Fig. 6. Bar charts of BER and CSDER results on several subsets of the development set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Origin (b) JES (c) IAS (d) DMS+IAS

Fig. 7. T-SNE graph comparison between adult and child speech on Namibia test set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
BER and CSDER values on test sets. Baseline represents the system trained only on the specific subset (i.e., no pre-trained JES). JES denotes joint separation and enhancement systems. IAS and DMS+IAS stand for our proposed iterative adaptation based separation framework without and with dynamic mask operation. JES+GT represents the JES system fine-tuned on supervised dev set using the ground-truth labels.

Subset	BER					CSDER				
	Baseline	JES	IAS	DMS+IAS	JES+GT	Baseline	JES	IAS	DMS+IAS	JES+GT
War2	0.450	0.373	0.366	**0.354**	0.343	0.134	0.123	0.078	**0.066**	0.053
Tsimane	0.480	0.408	0.395	**0.395**	0.390	0.435	0.383	0.267	**0.170**	0.158
Namibia	0.499	0.437	0.429	**0.424**	0.415	0.355	0.306	0.297	**0.283**	0.269
Vanuatu	0.469	0.426	0.445	**0.381**	0.356	0.479	0.342	0.313	**0.243**	0.207
Lena_lyon	0.498	0.448	0.425	**0.421**	0.398	0.323	0.243	0.200	**0.192**	0.167
Aclew_starter	0.502	0.454	0.449	**0.447**	0.435	0.249	0.206	0.182	**0.147**	0.139
Overall	0.491	0.432	0.423	**0.415**	0.400	0.378	0.312	0.281	**0.244**	0.224

Fig. 7 shows the comparison of the spatial distance between adult and child speech before and after different separation strategies, where the red dots represent the adult speech and the blue ones represent the child speech (Van der Maaten and Hinton, 2008). Fig. 7(a) represents the mapping of adult speech and child speech distance information in two-dimensional space in the original utterances. Although there is an inherent gap between adult speech and child voice, many mixed regions still exist. Note that parts of the adult speech deviate from the main parts and are closer to the child parts, as shown in the upper right corner of the figure. We did a point-to-audio mapping of these parts and found that these deviated parts are mainly the overlapping segments of adult–child speech and some adults imitate the children's voices to tease the child, which is exactly the significant challenge for child speech processing that we mentioned earlier. As previously introduced, the children wear the recording device, so the energy of the children's voices is stronger, and the adults' voices are far-field. Hence, these parts of the speech will be more inclined to the child's parts in terms of spatial distance. From the figure, these parts are at the edge of the child speech, but there are still some overlap regions. In Fig. 7(b), the original audio was processed by our JES approach. It can be found that compared with Fig. 7(a), the red and blue points start to separate. The overlap parts in the lower left corner become less, and the upper right corner almost has no overlap regions, but it is still close to the blue edge. This indicates that the JES system plays a vital role, but it is not easy to separate the mixed speech of adults and children in some extreme cases. By observing Figs. 7(b) and 7(c), the separation performances are further improved after adopting our proposed IAS system. The main parts of adult and child speech are separated by a more considerable distance in Fig. 7(c). This indicates that our IAS method can obtain targeted breakthroughs on each subset. However, IAS does not deal well with the speech segments in which adults imitate children and the corresponding parts are still quite close to the child parts.

Fig. 8 shows the ground truth labels and the spectrograms comparison of different methods and the mixtures. The top rectangular box is the ground-truth label. The red, blue, and gray rectangular segments represent adult speech, child speech, and the existence of noises, respectively. Note that the two dark colors on the gray bar represent the sharp noises generated by the collision of household products. From top to bottom are the original speech without any processing, the speech processed by JES, IAS, and DMS+IAS. The white circles on the spectrograms represent the child speech, and the white boxes represent the adult speech. By comparison, we can find that the proposed IAS system is better than the JES method in extracting child speech, as shown in the white circles. However, IAS is not much better than JES for suppressing adult speech, as shown in the white boxes.

The above results show that, compared with the JES system, both qualitatively and quantitatively, our IAS system can achieve better results on subsets of different languages and styles to achieve the goal of adaptation.

4.4. Results using dynamic mask

Due to the complexity and difficulty of controlling speech quality in real scenarios, we believe that the introduction of DMS can further improve the performance of the IAS system. In this section, we discuss the experimental results of the DMS+IAS system following the discussion in Section 4.3.

In Fig. 6, the gray bars refer to the DMS+IAS systems. We can find that our proposed DMS+IAS system is not only able to achieve optimal results on single-scene subsets (such as War2 & Aclew_starter) but also brings improvements on complex scene datasets (such as Namibia & Vanuatu). In Table 3, we can see that the introduction of DMS can further improve the model performance and achieve the best results in both BER and CSDER on all subsets. For example, compared with the JES method, the DMS+IAS method has an improvement of 4.5% and 9.9% in BER and CSDER on the Vanuatu dataset, where IAS fails to work. By comparing the overall results of the baseline, JES+GT, and DMS+IAS, we find that the BERs of baseline and JES+GT are 0.491 and 0.400. In the BER of 0.091 that baseline is more than JES+GT, DMS+IAS can recover 83.5% of them and reach the BER of 0.415. Likewise, DMS+IAS can recover 98.7% of CSDER (0.378, 0.244, and 0.224 for baseline, DMS+IAS, and JES+GT). Overall results show that the results of DMS+IAS are better than those of IAS, which are both better than JES.

We elaborate in Section 4.4 that IAS does not deal well with the speech segments in which adults imitate children, as shown in Fig. 7. In contrast, the introduction of DMS alleviates this problem and separates the speech of adults and children better than IAS in this brutal scene (as is shown in Fig. 7(d)). But the distance between them

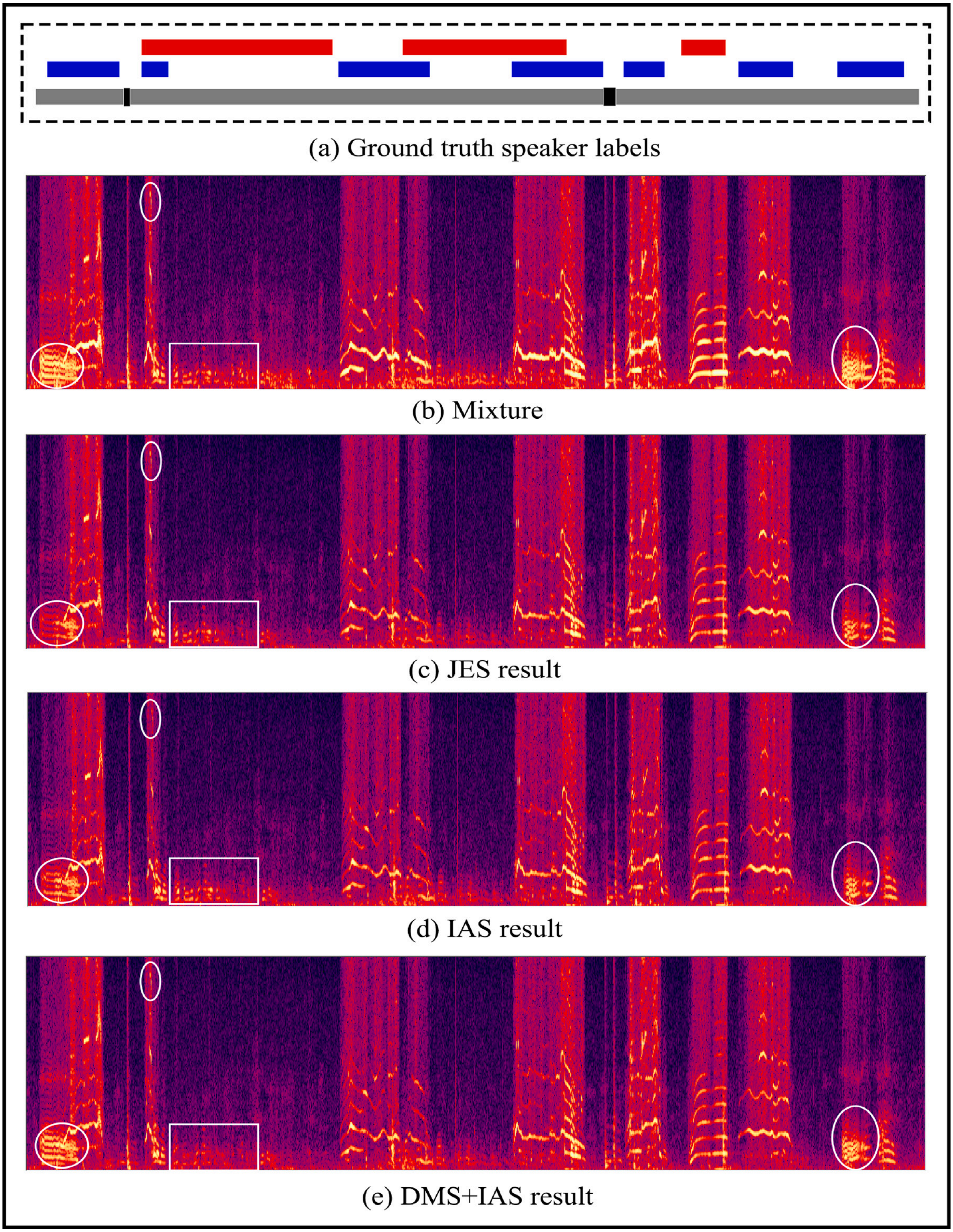


Fig. 8. Spectrograms comparison of an utterance from the test set. In (a), the red bar represents the speech regions of adults, while the blue bar represents the target child speech, and the gray bar denotes environmental noises. (b) gives the original spectrum. (c)–(e) show the results processed by JES, IAS, and DMS+IAS, respectively. Black parts on gray bars represent sharp high-frequency noises. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is still relatively close, which limits our extraction performance to a certain extent. The bottom spectrogram in Fig. 8 represents the speech processed by DMS+IAS. It can be seen that the proposed DMS+IAS system is significantly better than the JES method and the pure IAS framework in retaining the child speech, as shown in white circles. Our proposed DMS+IAS systems also suppress adult speech to some extent, as is shown in the white boxes. These experimental results above demonstrate the effectiveness of our method.

It is worth mentioning that the development sets of War2 and Aclew_starter only remain 13.4 and 24.3 min long, respectively, and the results of the development sets and the test sets both show that when we get a complete unknown test recording, we can also use it as a separate subset for processing. The separation result of this recording is obtained through the pre-trained separation model first. Afterward, a new training set is constructed based on the separation results of this recording and the original sentence, and an adaptation based separation model can be obtained by finetuning the pre-trained model with little computational cost and time.

5. Conclusion

In this paper, we first propose an IAS framework to improve over our previously-proposed JES system to deal with real-world recordings of child speech in multi-speaker and multi-lingual environments. To purify the data used in fine-tuning adaptation, we further propose a DMS framework to correctly obtain variable length-position dynamic masks that match well with the meaningful speech segments needed. Experimental results show that the proposed DMS+IAS framework is valid on both BER and CSDER metrics.

The DMS+IAS approach demonstrated its potential for practical applications in real-world scenarios. However, under real-world conditions, how to better extract children's speech while removing adult speech as much as possible is still a problem that needs to be studied in the future. In our future work, we will further extend our research to even more complex and severe speech-overlap conditions, and also explore its application in other fields, such as diarization and egocentric speech separation (Hershey et al., 2016a; Yang et al., 2019; Sell and Garcia-Romero, 2014).

CRediT authorship contribution statement

Shi Cheng: Conception and design of study, Analysis and/or interpretation of data, Writing – original draft. **Jun Du:** Conception and design of study, Analysis and/or interpretation of data, Writing – review & editing. **Shutong Niu:** Conception and design of study, Writing – original draft. **Alejandrina Cristia:** Acquisition of data, Writing – review & editing. **Xin Wang:** Conception and design of study, Acquisition of data, Writing – review & editing. **Qing Wang:** Analysis and/or interpretation of data, Writing – original draft. **Chin-Hui Lee:** Conception and design of study, Writing – review & editing.

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

The authors do not have permission to share data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427. All authors approved the version of the manuscript to be published.

References

Arons, B., 1992. A review of the cocktail party effect. J. Am. Voice I/O Soc. 12 (7), 35–50.

Bao, F., Abdulla, W.H., 2018. Noise masking method based on an effective ratio mask estimation in gammatone channels. APSIPA Trans. Signal Inf. Process. 7, e5. http://dx.doi.org/10.1017/ATSIP.2018.7.

Barker, J., Watanabe, S., Vincent, E., Trmal, J., 2018. The fifth'chime'speech separation and recognition challenge: dataset, task and baselines. arXiv preprint arXiv:1803.10609.

Bee, M.A., Micheyl, C., 2008. The cocktail party problem: what is it? how can it be solved? and why should animal behaviorists study it? J. Comp. Physiol. 122 (3), 235.

Bergelson, E., Warlaumont, A., Cristia, A., Casillas, M., Rosemberg, C., Soderstrom, M., Rowland, C., Durrant, S., Bunce, J., 2017. Starter-ACLEW. Databrary, Retrieved November 9 (2017) 2018.

Bu, H., Du, J., Na, X., Wu, B., Zheng, H., 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In: 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). IEEE, pp. 1–5.

Canault, M., Le Normand, M.-T., Foudil, S., Loundon, N., Thai-Van, H., 2016. Reliability of the language environment analysis system (lena™) in european french. Behav. Res. Methods 48 (3), 1109–1124.

Chen, Y., Wang, Y., Li, D., Chen, Z., Guo, F., 2020. Bilingualism in speech enhancement. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 7609–7613.

Chen, S., Wu, Y., Chen, Z., Wu, J., Li, J., Yoshioka, T., Wang, C., Liu, S., Zhou, M., 2021. Continuous speech separation with conformer. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 5749–5753.

Cherry, E.C., 1953. Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. 975–979.

Dehak, N., Torres-Carrasquillo, P.A., Reynolds, D., Dehak, R., 2011. Language recognition via i-vectors and dimensionality reduction. In: Twelfth Annual Conference of the International Speech Communication Association. Citeseer, p. None.

Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Ito, N., Kinoshita, K., Espi, M., Araki, S., Hori, T., et al., 2015. Strategies for distant speech recognitionin reverberant environments. EURASIP J. Adv. Signal Process. 2015, 1–15.

Demir, C., Saraclar, M., Cemgil, A.T., 2012. Single-channel speech-music separation for robust asr with mixture models. IEEE Trans. Audio Speech Lang. Process. 21 (4), 725–736.

Ditter, D., Gerkmann, T., 2020. A multi-phase gammatone filterbank for speech separation via tasnet. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 36–40.

Emiya, V., Vincent, E., Harlander, N., Hohmann, V., 2011. Subjective and objective quality assessment of audio source separation. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2046–2057.

Garofolo, J., Graff, D., Paul, D., Pallett, D., 1993a. Csr-I (Wsj0) Complete Ldc93s6a, Web Download, Vol. 83. Linguistic Data Consortium, Philadelphia.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., 1993b. Darpa Timit Acoustic-Phonetic Continous Speech Corpus Cd-Rom. Nist Speech Disc 1-1.1. NASA STI/Recon technical report n 93, p. 27403.

Gilkerson, J., Coulter, K.K., Richards, J.A., 2008. Transcriptional Analyses of the Lena Natural Language Corpus. LENA Foundation.

Godino-Llorente, J.I., Gomez-Vilda, P., Blanco-Velasco, M., 2006. Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters. IEEE Trans. Biomed. Eng. 53 (10), 1943–1953.

Greenberg, J., Peterson, P., Zurek, P., 1993. Intelligibility-weighted measures of speech-to-interference ratio and speech system performance. J. Acoust. Soc. Am. 94 (5), 3009–3010.

Hamers, L., et al., 1989. Similarity measures in scientometric research: The jaccard index versus salton's cosine formula. Inf. Process. Manage. 25 (3), 315–318.

Han, J., Moraga, C., 1995. The influence of the sigmoid function parameters on the speed of backpropagation learning. In: International Workshop on Artificial Neural Networks. Springer, pp. 195–201.

Haykin, S., Chen, Z., 2005. The cocktail party problem. Neural Comput. 17 (9), 1875–1902.

Heittola, T., Mesaros, A., Virtanen, T., Eronen, A., 2011. Sound event detection in multisource environments using source separation. In: Machine Listening in Multisource Environments. p. None.

Hershey, J.R., Chen, Z., Le Roux, J., Watanabe, S., 2016a. Deep clustering: Discriminative embeddings for segmentation and separation. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 31–35. http://dx.doi.org/10.1109/ICASSP.2016.7471631.

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

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P., 2014. Deep learning for monaural speech separation. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1562–1566. http://dx.doi.org/10.1109/ICASSP.2014.6853860.

Hus, Y., Segal, O., 2021. Challenges surrounding the diagnosis of autism in children. Neuropsychiatr. Dis. Treat. 17, 3509.

Jansen, A., Van Durme, B., 2011. Efficient spoken term discovery using randomized algorithms. In: 2011 IEEE Workshop on Automatic Speech Recognition and Understanding. pp. 401–406. http://dx.doi.org/10.1109/ASRU.2011.6163965.

Jörnvall, H., Persson, B., Krook, M., Atrian, S., Gonzalez-Duarte, R., Jeffery, J., Ghosh, D., 1995. Short-chain dehydrogenases/reductases (sdr). Biochemistry 34 (18), 6003–6013. http://dx.doi.org/10.1021/bi00018a001, pMID: 7742302. arXiv:https://doi.org/10.1021/bi00018a001.

Kanda, N., Boeddeker, C., Heitkaemper, J., Fujita, Y., Horiguchi, S., Nagamatsu, K., Haeb-Umbach, R., 2019. Guided source separation meets a strong asr backend: Hitachi/paderborn university joint investigation for dinner party asr. arXiv preprint arXiv:1905.12230.

Kavalerov, I., Wisdom, S., Erdogan, H., Patton, B., Wilson, K., Le Roux, J., Hershey, J.R., 2019. Universal sound separation. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA, IEEE, pp. 175–179.

Kohnert, K., Ebert, K.D., Pham, G.T., 2020. Language Disorders in Bilingual Children and Adults. Plural Publishing.

Kong, Q., Xu, Y., Wang, W., Plumbley, M.D., 2018. A joint separation-classification model for sound event detection of weakly labelled data. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 321–325.

Kucybała, I., Tabor, Z., Polak, J., Urbanik, A., Wojciechowski, W., 2020. The semi-automated algorithm for the detection of bone marrow oedema lesions in patients with axial spondyloarthritis. Rheumatol. Int. 40 (4), 625–633.

Laine, S., Aila, T., 2016. Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242.

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., Cristia, A., 2020. An open-source voice type classifier for child-centered daylong recordings. arXiv preprint arXiv:2005.12656.

Le Roux, J., Hershey, J.R., Weninger, F., 2015. Deep nmf for speech separation. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 66–70. http://dx.doi.org/10.1109/ICASSP.2015.7177933.

Le Roux, J., Wisdom, S., Erdogan, H., Hershey, J.R., 2019. Sdr–half-baked or well done? In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 626–630.

Liang, B., Li, H., Su, M., Bian, P., Li, X., Shi, W., 2017. Deep text classification can be fooled. arXiv preprint arXiv:1704.08006.

Ling, H., Han, P., Qiu, J., Peng, L., Liu, D., Luo, K., 2021. A method of speech separation between teachers and students in smart classrooms based on speaker diarization. In: 2021 13th International Conference on Education Technology and Computers. pp. 53–61.

Liu, Z., Lin, W., Shi, Y., Zhao, J., 2021. A robustly optimized bert pre-training approach with post-training. In: Chinese Computational Linguistics: 20th China National Conference, CCL 2021, Hohhot, China, August 13–15, 2021, Proceedings. Springer, pp. 471–484.

Luo, Y., Mesgarani, N., 2018. Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 696–700. http://dx.doi.org/10.1109/ICASSP.2018.8462116.

Luo, Y., Mesgarani, N., 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (8), 1256–1266. http://dx.doi.org/10.1109/TASLP.2019.2915167.

Luz, S., Haider, F., de la Fuente, S., Fromm, D., MacWhinney, B., 2020. Alzheimer's dementia recognition through spontaneous speech: the adress challenge. arXiv preprint arXiv:2004.06833.

Lyakso, E.E., Frolova, O.V., 2020. Early development indicators predict speech features of autistic children. In: Companion Publication of the 2020 International Conference on Multimodal Interaction. pp. 514–521.

Lyakso, E., Frolova, O., Kaliyev, A., Gorodnyi, V., Grigorev, A., Matveev, Y., 2019. Ad-child. ru: Speech corpus for russian children with atypical development. In: International Conference on Speech and Computer. Springer, pp. 299–308.

MacWhinney, B., 1996. The childes system. Am. J. Speech-Lang. Pathol. 5 (1), 5–14.

MacWhinney, B., 2000. The CHILDES Project: The Database, Vol. 2. Psychology Press.

MacWhinney, B., 2001. From Childes to Talkbank. Department of Psychology, p. 182.

MacWhinney, B., 2014. The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database. Psychology Press.

MacWhinney, B., Snow, C., 1985. The child language data exchange system. J. Child Lang. 12 (2), 271–295. http://dx.doi.org/10.1017/S0305000900006449.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 5206–5210.

Park, D.S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E.D., Le, Q.V., 2019. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, Vol. 32. p. None.

Pretzer, G.M., Lopez, L.D., Walle, E.A., Warlaumont, A.S., 2019. Infant-adult vocal interaction dynamics depend on infant vocal type, child-directedness of adult speech, and timeframe. Infant Behav. Dev. 57, 101325.

Raj, D., Denisov, P., Chen, Z., Erdogan, H., Huang, Z., He, M., Watanabe, S., Du, J., Yoshioka, T., Luo, Y., et al., 2021. Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis. In: 2021 IEEE Spoken Language Technology Workshop. SLT, IEEE, pp. 897–904.

Reynolds, D.A., 2009. Gaussian mixture models. In: Encyclopedia of Biometrics, Vol. 741. pp. 659–663.

Rustamov, S., Akhundova, N., Valizada, A., 2019. Automatic speech recognition in taxi call service systems. In: International Conference for Emerging Technologies in Computing. Springer, pp. 243–253.

Sanchez, A., Meylan, S.C., Braginsky, M., MacDonald, K.E., Yurovsky, D., Frank, M.C., 2019. Childes-db: A flexible and reproducible interface to the child language data exchange system. Behav. Res. Methods 51 (4), 1928–1941.

Saon, G., Soltau, H., Nahamoo, D., Picheny, M., 2013. Speaker adaptation of neural network acoustic models using i-vectors. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE, pp. 55–59.

Sattorovich, E.Z., 2022. Psychological influence of speech disorders and the causes that cause them on the child's psyche. Acad. Globe: Indersci. Res. 3 (01), 39–42.

Seide, F., Agarwal, A., 2016. Cntk: Microsoft's open-source deep-learning toolkit. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 2135.

Sell, G., Garcia-Romero, D., 2014. Speaker diarization with plda i-vector scoring and unsupervised calibration. In: 2014 IEEE Spoken Language Technology Workshop. SLT, pp. 413–417. http://dx.doi.org/10.1109/SLT.2014.7078610.

Shi, J., Xu, J., Fujita, Y., Watanabe, S., Xu, B., 2020. Speaker-conditional chain model for speech separation and extraction. arXiv preprint arXiv:2006.14149.

Shobaki, K., Hosom, J.-P., Cole, R., 2000. The ogi kids' speech corpus and recognizers. In: Proc. of ICSLP. pp. 564–567.

Slobin, D.I., 2021. Imitation and grammatical development in children. In: Psychological Modeling. Routledge, pp. 166–177.

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., Zhong, J., 2021. Attention is all you need in speech separation. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 21–25.

Sun, L., Du, J., Zhang, X., Gao, T., Fang, X., Lee, C.-H., 2020. Progressive multi-target network based speech enhancement with snr-preselection for robust speaker diarization. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 7099–7103.

Takashima, Y., Fujita, Y., Horiguchi, S., Watanabe, S., García, P., Nagamatsu, K., 2021. Semi-supervised training with pseudo-labeling for end-to-end neural diarization. arXiv:2106.04764.

Tang, P., Yuen, I., Rattanasone, N.X., Gao, L., Demuth, K., 2019. The acquisition of phonological alternations: The case of the mandarin tone sandhi process. Appl. Psycholinguist. 40 (6), 1495–1526.

Turpault, N., Wisdom, S., Erdogan, H., Hershey, J., Serizel, R., Fonseca, E., Seetharaman, P., Salamon, J., 2020. Improving sound event detection in domestic environments using sound separation. arXiv preprint arXiv:2007.03932.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. J. Mach. Learn. Res. 9 (11).

VanDam, M., Warlaumont, A.S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., MacWhinney, B., 2016. Homebank: An online repository of daylong child-centered audio recordings. In: Seminars in Speech and Language, Vol. 37. Thieme Medical Publishers, pp. 128–142.

Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., Matassoni, M., 2013. The second 'chime' speech separation and recognition challenge: Datasets, tasks and baselines. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 126–130. http://dx.doi.org/10.1109/ICASSP.2013.6637622.

Vincent, E., Bertin, N., Gribonval, R., Bimbot, F., 2014. From blind to guided audio source separation: How models and side information can improve the separation of sound. IEEE Signal Process. Mag. 31 (3), 107–115.

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., Watanabe, S., Barker, J., Marxer, R., 2016. The 4th chime speech separation and recognition challenge. URL: http://spandh.dcs.shef.ac.uk/chime_challenge/ (last accessed on 1 August, 2018).

Virtanen, T., Gemmeke, J.F., Raj, B., Smaragdis, P., 2015. Compositional models for audio processing: Uncovering the structure of sound mixtures. IEEE Signal Process. Mag. 32 (2), 125–144.

Wang, X., Du, J., Cristia, A., Sun, L., Lee, C.-H., 2020. A study of child speech extraction using joint speech enhancement and separation in realistic conditions. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 7304–7308.

Wang, X., Du, J., Sun, L., Wang, Q., Lee, C.-H., 2018. A progressive deep learning approach to child speech separation. In: 2018 11th International Symposium on Chinese Spoken Language Processing. ISCSLP, IEEE, pp. 76–80.

Watanabe, S., Hori, T., Hershey, J.R., 2017. Language independent end-to-end architecture for joint language identification and speech recognition. In: 2017 IEEE Automatic Speech Recognition and Understanding Workshop. ASRU, pp. 265–271. http://dx.doi.org/10.1109/ASRU.2017.8268945.

Watanabe, S., Mandel, M., Barker, J., Vincent, E., Arora, A., Chang, X., Khudanpur, S., Manohar, V., Povey, D., Raj, D., et al., 2020. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. arXiv preprint arXiv:2004.09249.

Wood, S.U., Rouat, J., Dupont, S., Pironkov, G., 2017. Blind speech separation and enhancement with gcc-nmf. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (4), 745–755.

Xiangjun, D., Yip, V., 2018. A multimedia corpus of child mandarin: The tong corpus. J. Chin. Linguist. 46 (1), 69–92.

Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q., 2020. Unsupervised data augmentation for consistency training. Adv. Neural Inf. Process. Syst. 33, 6256–6268.

Yang, X., Deng, C., Zheng, F., Yan, J., Liu, W., 2019. Deep spectral clustering using dual autoencoder network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Yeung, G., Fan, R., Alwan, A., 2021. Fundamental frequency feature normalization and data augmentation for child speech recognition. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 6993–6997. http://dx.doi.org/10.1109/ICASSP39728.2021.9413801.

Yin, X., Goudriaan, J., Lantinga, E.A., Vos, J., Spiertz, H.J., 2003. A flexible sigmoid function of determinate growth. Ann. Botany 91 (3), 361–371.

Zhou, S., Xu, S., Xu, B., 2018. Multilingual end-to-end speech recognition with a single transformer on low-resource languages. arXiv:1806.05059.