Dynamic Mask based Iterative Adapted Child Speech Separation under Multilingual Real Conditions

Shi Cheng, Jun Du, Shutong Niu, Chin-Hui Lee, Alejandrina Cristia, Xin Wang

Dynamic Mask based Iterative Adapted Child Speech Separation under Multilingual Real Conditions

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

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

Abstract

We propose two iterative adapted child speech separation systems in multilingual real-world scenarios. The newly proposed systems are developed from our previous joint enhancement and separation (JES) system. First, we introduce an iterative adapted separation (IAS) method based on the separation model in JES, which iteratively implements the pre-trained separation model on subsets of real scenes in different languages and styles to adapt the model. Subsequently, we further use the definition of scale-invariant signal-to-noise ratio (Si-SNR) to propose a variable length-position dynamic mask to purify the training data, which is called DM-IAS. The DM-IAS alleviates the problem of a large amount of noises in speech under real scenarios. The length of the activated segment in the dynamic mask is automatically determined by statistical features based on Si-SNR, and the onset of the active segment is positioned by a sliding window with a certain step size. Evaluated on BabyTrain corpus, our proposed IAS system achieves consistent and significant separation performance improvements when compared with the previously proposed system JES. In addition, the experiment results show that the proposed DM-IAS method can further improve the quality of separated children speech in real scenarios, and also obtains a relatively good separation performance in the common but very difficult situation where adults imitate children's speech.

Keywords: speech separation, child speech process, dynamic mask, iterative adapted separation, multilingual real scene

1. Introduction

Child speech processing has been paid more and more attention in these years. It's vital for the diagnosis of early childhood diseases and it helps to understand children's intentions due to their lack of language expression ability [1, 2, 3, 4]. It has been generally investigated in academic domain such as developmental psychology [5] and cognitive science [6], as well as in applied domain such as the diagnosis of underlying language disorders and the measurement of the effects of interventions [7, 8]. In earlier times, however, the extreme lack of data limited the development of children's speech signal processing. The lack of data also makes the associated research costs very high. The LENA Foundation launched their software [1] in 2008, which was trained on a 150 hours hand-annotated dataset using MFCC features [9] and GMM model [10]. In the more than ten years since the software was launched, tens of thousands of children have benefited from it, and it has made great contributions to the monitoring of early childhood diseases. But its prohibitive price tag (at least US$5000 strating costs) makes it difficult to get a large-scale rollout. Until recent years, there has been a wave of small outbreaks in child-focused recordings and related researches on

Email addresses: chengshi@mail.ustc.edu.cn (Shi Cheng), jundu@ustc.edu.cn (Jun Du), niust@mail.ustc.edu.cn (Shutong Niu), chl@ece.gatech.edu (Chin-Hui Lee), alecristia@gmail.com (Alejandrina Cristia), xinwang@mail.ustc.edu.cn (Xin Wang)

them. In [11], researchers published a new corpus that documents the language development of a Mandarin-speaking child from 1 year 7 months to 3 years 4 months, which surpassed existing published corpora in the amount of data and its monitoring of the continuous development of children has attracted widespread attention. The challenges that language learning poses to children as the phonological environment changes have been studied in [12]. In [13], a Russian phonetic database AD-Child.Ru was presented, which contains the phonetic data of children with atypical development aged 4-16. Researchers revealed the development of the correlation between language uniqueness and difference in children with autism spectrum disorder on this dataset in [14]. AD-Child.Ru was also used to create an efficient tool for semi-automatic bone detection to detect axial spondylopathy (axSpA) in patients with bone marrow edema lesions [15]. Founded in 1984, the Children's Language Data Exchange System (CHILDES) together with its database-formatted mirrors childes-db [16, 17, 18, 19, 20, 21] has been a pioneer in the dissemination of large-scale children's speech behavior datasets. It was also used in the ADReSS Challenge at INTERSPEECH 2020 [22], which aims to explore automatic recognition methods for Alzheimer's disease of spontaneous speech in terms of age, child-db included. In [4], researchers proposed a child speech recognition system based on frequency feature normalization and data augmentation on OGI Kids' Speech Corpus [23]. These children's datasets should have been of great interest in many fields and led to some novel technical approaches. However, due to various difficulties, there are not many effective front-end methods for speech processing that can well utilize these data. To begin with, most of the recordings are collected from devices worn by children for a whole day. Consequently, these corpora contain a large number of non-speech vocalizations, e.g., crying, snoring and screaming. Besides, in the process of children's lives, adults play the extremely important roles as participants and companions. Therefore, children's pronunciation is often accompanied by a large number of adult voices, and many of them even imitate children's voices, which brings great challenges for researchers.

Speech enhancement and speech separation are very important front-end technologies in speech signal processing, which are often used to denoise and extract speech in complex scenes. Speech enhancement can be used to suppress background noise beforehead while speech separation aims at separating target speech from a mixture, which is a key issue in the "Cocktail party problem" [24, 25, 26, 27]. Due to its importance in the front end of speech signal processing, speech separation has been an important research direction in academic and industry fields. It has derived a series of cutting-edge applications in ASR (Automatic Speech Recognition) [28, 29], SED (Sound Event Detection) [30, 31, 32] and other areas, such as call customer service channels [33], multi-speaker meeting minutes [34] and target instruction extraction of smart speakers in domestic scene [35]. Speech enhancement and speech separation methods have undergone a long development. Before the advent of deep learning methods, the non-negative matrix factorization (NMF) method has been the mainstream speech separation method [36, 37, 38] and a series of related technologies have been derived from NMF. In [38], through combining unsupervised dictionary learning of non-negative matrix factorization with spatial localization of generalized cross-correlation method, a blind source separation algorithm names GCC-NMF has been proposed and proven to be a flexible source separation. Deep NMF method has been introduced in [39] and proven to be competitive with deep neural networks on 2nd CHIME Speech Separation and Recognition Challenge corpus [40]. 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 and many speech separation methods have emerged. A joint deep neural network and recurrent neural network optimization system was proposed in [41] and evaluated on the TIMIT speech corpus [42]. Compared to NMF models, such a system achieved about 3.8-4.9dB signal to interference ratio (SIR) gain while maintaining better source-to-distortion ratio (SDR) and source-to-artifact ratio (SAR) [43, 44, 45, 46]. In recent years, transformers have become popular, and state-of-the-art (SOTA) performances have been achieved in a large number of artificial intelligence fields. In [47], the transformer-based SepFormer model was applied to the standard WSJ0-2 and WSJ0-3MIX datasets and obtained 22.3 dB and 19.5dB Si-SNR gain. 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 [48], 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 [49, 50]. Despite that new network structures

emerge in an endless stream, Recurrent neural network (such as RNN, LSTM, GRU and Bi-LSTM), with its inherent advantages in time series modeling, has long been dominating the sequence-to-sequence task modeling. Among them, the representative LSTM has been widely used in sequence modeling and has been proved to be very effective on commonly used datasets such WSJ0 [51, 52], AISHELL corpus [53] and TIMIT [42] and CHIME series challenges' datasets [54, 55, 56]. However, the vast majority of research focuses on simulation data. As for real data, [57] examined the use of Conformer in lieu of recurrent neural networks for separation model and introduced significant performance gains in both word error rate (WER) and speaker-attributed WER. In [58], a common strategy named Speaker-Conditional Chain Model to process complex speech recordings has been raised. With the predicted speaker information from the whole observation, the proposed model has been proven to be helpful in solving the problem of conventional speech separation and speaker extraction for multi-round long recordings under real scenarios. Despite some completed and ongoing work, there is still a lack of relevant research in real-world scenarios. In [2], we validated the excellent performance of the progressive learning strategy on the task of child speech separation on simulated data. Progressive learning can bring certain gains to speech extraction, but it is difficult to solve some complex noises in the audios, especially in real scenes with complex audio conditions. Actually, most of the released children's corpora then are monolingual or single-speaker and the corresponding acoustic scenes are narrow, making it difficult to approach various complex scenes in the real world. Hence, we further proposed a child speech extraction system using joint speech enhancement and speech separation in real-world conditions on BabyTrain [59], a multilingual real-scene large dataset. By using speech enhancement as the pre-processing for speech separation, the joint system leads to a preliminary performance improvement on child speech extraction compared to direct-mapped binary classification networks.

Here, we first propose a new iterative adapted speech separation framework which contains a speech enhancement model and a pre-trained speech separation model. The enhancement and separation models are used in turn to decode our input mixtures. We regard the separated speech as clean child speech and gain the noises by simply minus child speech from enhanced speech. For each subset, we use the corresponding development set to build a new training set and fine-tune the pre-trained model. The results in Section 3 show that such a method can obtain better results compared to our previously proposed joint speech enhancement and speech separation framework [3]. Moreover, as we discussed above, child speech corpus often tends to contain extremely complex speech scenes. As a result, separated children speech actually still contains some noises. So simply adopting fine-tune strategy on the corresponding subsets can not reach a remarkable result. Accordingly, while building the training set for updating the separation model, we want to discard the noises in separated children's speech. In this sense, a dynamic mask based iterative adaptive child speech separation framework is further proposed. By adjusting Si-SNR in [48], a length-position-variable mask can be obtained and used to mask the noise regions. Experimental results show that such a strategy could further improve the system performance.

The remainder of this paper is organized as follows. In Section 2, we describe the BabyTrain data set and give a review of our previous works. In Section 3, we elaborate on the proposed iterative adapted separation systems. Experimental results with detailed analyses are presented in Section 4. Finally, we conclude our findings and discuss some future works in Section 5.

2. Prior works on the BabyTrain

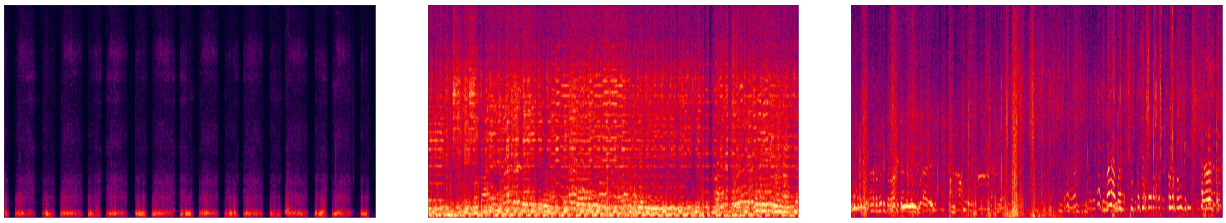
As mentioned above, there are few front-end methods that target real-world child speech. In our previous work, we have proposed some front-end methods based on the real multilingual BabyTrain dataset. By combining speech enhancement and speech separation, we obtained better results than single separate systems [3]. The systems proposed in this paper are based on these previous works. Accordingly, we first introduce the BabyTrain dataset and the previously proposed method here.

2.1. Data analysis

BabyTrain is a large corpus, it contains several child-centered corpora ranging in age from 1 month to 5 years [60, 61, 62, 63]. Each recording is sampled at 44.1 kHz in a common format and comes with

Table 1: Description of the subsets of BabyTrain [59] data set. Child-centered corpora included cover a wide range of conditions (including different languages and recording devices). Tot. Dur. represents total duration of the corresponding subset. Columns in italics denote accumulated duration of different speakers, including KCHI: key children; OCH: other children; MAL: male adults; FEM: female adults; UNK: speakers for whom age/sex is not known.

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	0.2	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	✗	✓



(a) (b) (c)

Figure 1: Spectrograms of recordings in different scenarios from the BabyTrain corpus, the audios are all recorded by the recording equipment worn by children for many hours in a day, (a) a clip of a child snoring while sleeping, (b) a father and son singing to the music at a party, (c) a family conversation scenario.

human transcription file. It contains 245-hour recordings with various adverse environments and different languages. Its huge amount of comprehensive data makes its recording style cover almost all common life scenes and some extreme ones, including daily life, indoor, outdoor, party scenes and so on. Table 1 gives a broad description of BabyTrain. It can be seen that the dataset covers a variety of languages and acoustic scenes, and even each recording in some subsets belongs to different scenes (such as two-person conversations, multi-person gatherings, etc.). Moreover, since the recording equipment is worn on children, friction and obscuring by clothing cause poor audio quality on most parts, which are difficult for front-end processing. In order to visually demonstrate that the BabyTrain dataset which covers a wide range of acoustic scenes, we further give the spectrograms of three samples in Figure 1. Figure 1(a) shows a recording of a sleeping child snoring while Figure 1(b) represents the audio of a father singing with his children in a family party. Figure 1(c) demonstrates a complex dialogue in a family scene. As is shown in the figures, the recordings cover a variety of life scenarios, with extreme scenes accounting for a significant proportion of the recordings. Additionally, BabyTrain contains both far-field and near-field speech. All of these present a big challenge for our separation task.

2.2. Prior work

As mentioned above, we proposed a joint speech enhancement and speech separation system. Due to the influence of unavoidable noises on the real dataset, it is necessary to add a speech enhancement model as a front-end processing to remove the noises before the separation model. Based on this idea, our previously proposed system takes the approach of combining the speech enhancement and speech separation [3], which mainly conducted on the BabyTrain mega corpus.

The flow of the joint system is illustrated in Figure 2. First we need to train the available speech enhancement and speech separation models as illustrated on either side of the dotted line in Figure 2. Secondly, in the decoding stage, we first extract Log Power Spectrum (LPS) features from the speech, and send them into the speech enhancement model, then transport the enhanced speech into the speech separation model

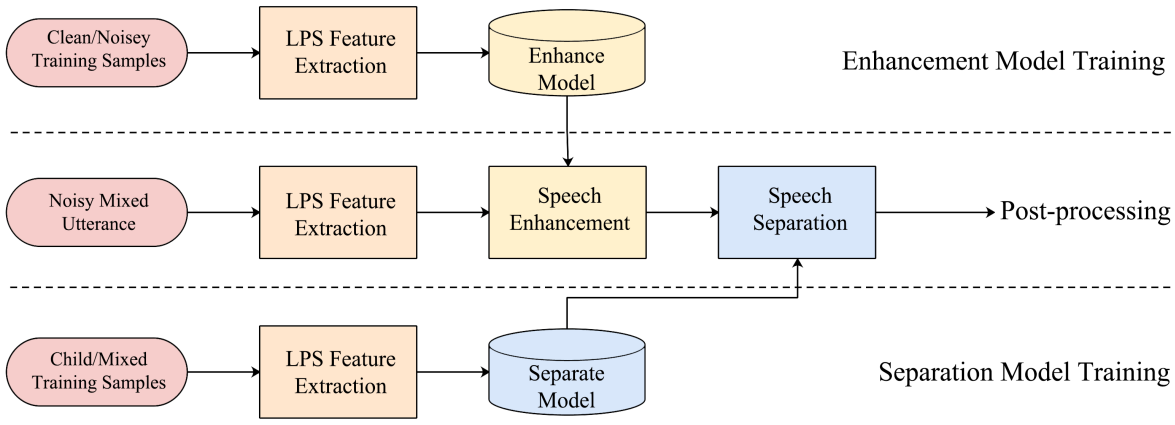


Figure 2: Framework of our proposed joint enhancement and separation system.

to obtain children's speech. Finally, after obtaining the children's speech from our separation model, we perform post-processing on the separation results to complete the classification of children's and adults' speech and calculate metrics. Note that our enhancement network directly adopts from [64]. As for separation network, we use a 3-layer Bi-LSTM network. Child speech, together with adult speech, are randomly mixed up at the signal to noise ratio (SNR) of -5dB, 0db and 5db as the inputs for the training of our separation model. Speech at SNR of 5db, 10db, 15db, 20db, 25db and clean speech are provided for the progressive learning. The learning targets are Progressive Log Power Spectrum feature (PLPS) and Progressive Ratio Mask (PRM). The former represents LPS target for each progressive layer and the latter can be calculated by Eq.(1):

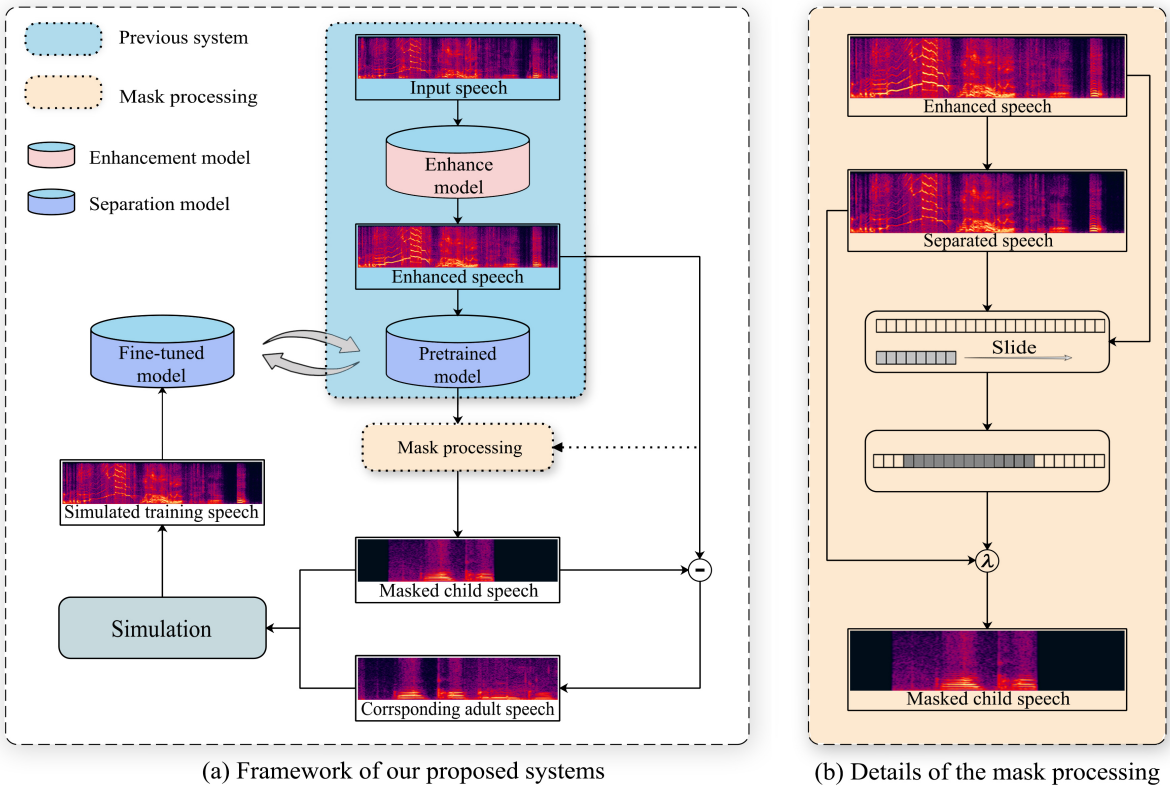
$$z^{l}_{PRM}(t,f) = \frac{C(t,f) + A^{l}_{T}(t,f)}{C(t,f) + A_{I}(t,f)} \tag{1}$$

where $l = 1, 2, 3$ represents the layer index and $C(t, f)$ stands for the power spectrum of the child speech signal at the time-frequency (T-F) unit (t, f). Likewise, $A^{l}_{T}(t, f)$ is the power spectrum of the noise at layer l and $A_I(t, f)$ represents that of noise in the input signals at time-frequency(T-F) unit (t, f). Note that the output of each layer is simultaneously concatenated to the features of the original speech as input to the next layer. More details can refer to [3] and [64].

The application of speech enhancement model can eliminate part of the environmental noises and make the inputs for the separation model in higher quality. Such a joint system is capable of obtaining a notable result in the overall data. However, the BabyTrain corpus contains multi-language and multi-scenario subsets, e.g., Namibia, Lena_lyon, War2 and Tsimane. The subsets are quite vary from each other as illustrated in Table 1. Simply joint speech enhancement and separation network can not solve the problems we mentioned above, such as the speech in which adults are imitating children and multi-style multilingual corpora. In this sense, it's natural to wonder whether the separation model can be further optimized for each subset to get targeted breakthroughs on different subsets. Therefore, we continue to explore on BabyTrain, and the proposed iterative adapted methods realize the adaptation on different style subsets.

3. Proposed frameworks

Our proposed frameworks are all based on the previously introduced JES system. We present its diagram in the blue dashed box in Figure 3 (a). The separation model trained on the whole BabyTrain corpus is regarded as our pre-trained model, which plays the cornerstone of our subsequent experiments. Then, in the adaptive phase, we use the development set to build different train set for each subset of BabyTrain and test on the corresponding test set.



(a) Framework of our proposed systems (b) Details of the mask processing

Figure 3: Framework of our iterative adapted with (yellow dashed box included) and without (yellow dashed box not included) mask systems.

3.1. Iterative adapted separation

First of all, as JES is trained and tested on the entire BabyTrain dataset, we employ the JES as a preliminary system. The system has a good results on the entire dataset, but for each subset, the information of current subset will be far more important than the rest. So in the first instance, we simply regard the separated speech from the pre-train model as clean speech and build a new training set for each subset to fine-tune the pre-trained model. In particular, we cut the separated speech into one-second segments. Then we calculate the corresponding adult speech via Eq.(2), where $x_a(t)$ represents adult speech to be obtained at frame t, $x_{se}(t)$ and $x_c(t)$ denote enhanced and separated child speech at frame t, respectively.

$$x_a(t) = x_{se}(t) - x_c(t) \tag{2}$$

For each subset, we randomly mix the child speech and adult speech at the SNR of -5dB, 0db, 5db, 10db, 15db, 20db and 25db to build a new training set. Among them, the segments at the SNR of -5db, 0db, 5db are used as the inputs of the model, the segments at the SNR of 5db, 10db, 15db are used as the first layer's targets of progressive learning, those of 15db, 20db and 25db are used as the second layer's targets, and the clean segments are the final learning objectives. We then use the newly constructed training set to fine-tune our pre-trained separation model updated in the previous iteration, and each fine-tuned model will become the new pre-trained model in the next iteration. The loss function of our network is shown in Eq. (3):

$$E = \frac{1}{N}\sum_{m=1}^{3}\sum_{n=1}^{N}||\mathcal{F}_m(\hat{\boldsymbol{x}}_n^0, \hat{\boldsymbol{x}}_n^1, , ..., \hat{\boldsymbol{x}}_n^{m-1}, \Lambda_m) - \boldsymbol{x}_n^m||_2^2 \tag{3}$$

where $m = 1, 2, 3$ denotes the layer index. Different from the time domain signals in Eq.(2), such as $x_a(t)$ and $x_{se}(t)$, $\boldsymbol{x}$ here represents the splicing vector of LPS and IRM features. $\mathcal{F}_m(\hat{\boldsymbol{x}}_n^0, \hat{\boldsymbol{x}}_n^1, , ..., \hat{\boldsymbol{x}}_n^{m-1}, \Lambda_m)$

Algorithm 1 Algorithm of the iterative adapted 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 $x_{se}(t)$ and $x_{ss}(t)$;
Initial inputs: Set $i = 1$, note that $x_{ss}^{i}(t)$ represents separated speech in iteration i and the JES system's output $x_{ss}(t) = x_{ss}^{0}(t)$;
while iteration i **do**
  $x_{c}^{i}(t) = x_{ss}^{i-1}(t)$;
  $x_{a}^{i}(t) = x_{se}(t) - x_{ss}^{i-1}(t)$;
  randomly mix $x_{c}^{i}(t)$ and $x_{a}^{i}(t)$ up to build a new training set;
  $SS_{pre-trained} \rightarrow SS_{fine-tuned}$;
  if the error rate on validation set stops decreasing **then**
    break;
  end if
  $i = i+1$;
end while
Choose the model with smallest error rate and apply it to the test set.

is the neural network function of the learned target $\hat{\boldsymbol{x}}_n^0$ to $\hat{\boldsymbol{x}}_n^{m-1}$ with densely connected structure. Λ_m represents the parameter set of the weight matrix and bias vector before the m-th target layer, which optimizes gradient descent in a Back Propagation Through Time (BPTT) manner. $\boldsymbol{x}_n^m$ is the splicing of LPS and PRM of the clean speech at m-th layer. When the error rate related metrics stop decreasing, we shut down the iteration. Then the corresponding separation model will be chosen as our separation model in the test stage. A diagram of this framework is shown in Figure 3 (a), with the dotted box part named *Mask processing* skipped. Algorithm 1 presents the specific operation flow of IAS, in which the $SS_{fine-tuned}$ and $SS_{pre-trained}$ represent fine-tuned speech separation model and 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, which effectively prevents the overfitting problem. Meanwhile, the network can adapt to a specific subset while not changing the amount of information learned by the Bi-LSTM layers from large-scale data.

In the decode stage, we utilize the fine-tuned model of each iteration corresponding to each subset for decoding and metrics calculating. The decoding formula is Eq.(4), in which $\hat{\boldsymbol{x}}_{\mathrm{LPS}}$ is the estimated decoded LPS of test speech and $\boldsymbol{x}_{\mathrm{LPS}}$ represents that of the noisy speech, $\hat{\boldsymbol{x}}_{\mathrm{PRM}^3}$ stands for the final output PRM of our system.

$$\hat{\boldsymbol{x}}_{\mathrm{LPS}} = \boldsymbol{x}_{\mathrm{LPS}} + \ln(\hat{\boldsymbol{x}}_{\mathrm{PRM}^3}) \tag{4}$$

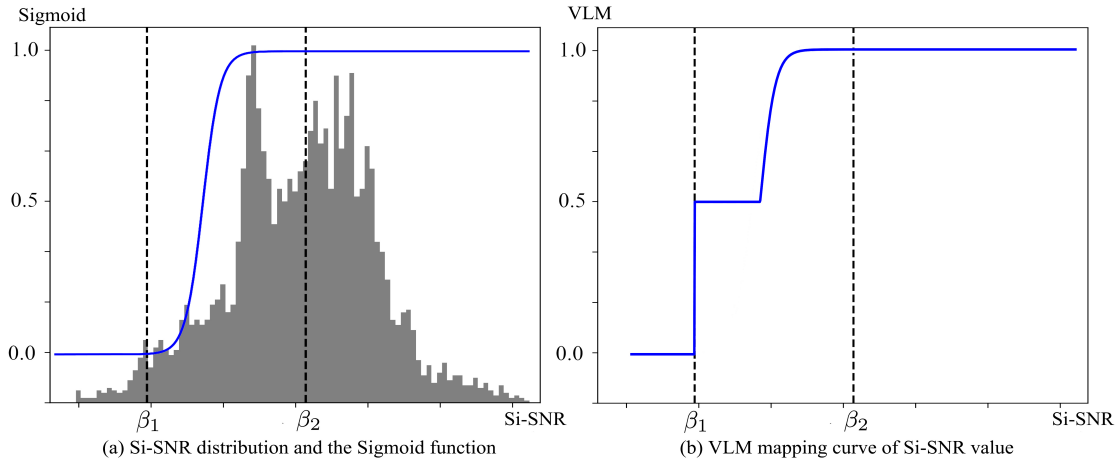
A series of post-processing operations are then performed on the decoded speech to obtain the corresponding labels, which can be used to calculate the corresponding metrics. Details of post-processing are described in Section 3.3.

This system is marked as IAS. Experimental results in Section 4 show that our system brings a better performance compared with the pre-trained model. Especially, the model performance will also improve as the number of iterations increases, which illustrates the effectiveness of our iterative strategy.

3.2. Dynamic mask based iterative adaptive separation

As we discussed above, due to the extreme complexity of BabyTrain, it is crude to directly treat the speech decoded from the pre-trained separation model as clean child speech. Hence we further propose a length-position-variable mask based processing as is shown in the dotted box part in Figure 3 (a), which is denoted as DM-IAS.

Apparently, the main difference compared with Section 3.1 is the *Mask processing* part. Figure 3 (b) shows the details of this module. After gaining separated speech from the pre-trained model, we take it one step further with a dynamic mask generation process to eliminate residual noises rather than simply treating it as clean data. As in shown in Figure.3 (b), we use an activated window to slide over the speech segment to



(a) Si-SNR distribution and the Sigmoid function (b) VLM mapping curve of Si-SNR value

Figure 4: An example of β_1, β_2 and corresponding Si-SNR distribution. The grey bars in (a) represents the distribution of Si-SNR value. The two blue curves represents Sigmoid function and VLM function, respectively.

locate child speech and mask the rest part. Since our activated segment of the mask is continuous, and the children's speech and other interfering speech are possibly interlaced in the longer speech segment, we divide the recordings into one-second segments to ensure that there is only one target in each segment, so that it is convenient for the dynamic mask to locate it. The length and position of such a mask are variable and we will describe how to determine them next. Masked speech is regarded as child speech and the subsequent operations are the same as in Section 3.1.

3.2.1. Determine the length of activated window

As previously mentioned, the pre-trained model can not extract the children's speech perfectly, but for specific speech segments, children's speech keep the main parts of separated speech. In this sense, it is possible for us to locate the children's parts of the separated speech and remove the rest. We refer to a commonly adopted metric for separation, namely scale-invariant source-to-noise ratio (Si-SNR) [48] to build an indicator as the rule to calculate the active length of the dynamic mask:

$$\mathbf{s}(x_{ss}(t), x_{se}(t)) = 10 \log_{10} \frac{||x_{ss}(t)||^2}{||x_{ss}(t) - x_{se}(t)||^2} \tag{5}$$

where $x_{ss}(t)$ represents the separated speech at t-th frame and $x_{se}(t)$ represents the enhanced speech. After gain $\mathbf{s}(x_{ss}(t), x_{se}(t))$, we want to use it to scale the active percentages of speech segments to [0, 1]. Inspired by the Sigmoid function [65, 66], we construct a mapping function to fulfill our goal, namely Variable Length Mapping function (VLM) in Eq.(6).

$$\text{VLM}(\mathbf{s}) = \max\{\frac{1}{1 + \exp(-\alpha \times \mathbf{s})}, 0.5\}, \quad \beta_1 < \mathbf{s} < \beta_2 \tag{6}$$

When $\mathbf{s} > \beta_2$, i.e., our pre-train model retains most of the speech, we believe the separated speech can be regarded as clean speech and choose not to mask these segments and set $\text{VLM}(\mathbf{s}) = 1$. As for $\mathbf{s} < \beta_1$, which means these segments contain mostly noises, we fully mask them, i.e., $\text{VLM}(\mathbf{s}) = 0$. For each subset, we visualize the distribution of all speech and set β_1 = lower 95% confidence interval boundary and β_2 = median. Figure 4 (a) shows the distribution of Si-SNR on subset Tsimane 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. Figure 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 a speech through Eq.(7). Note that L is the length of the segment (one second) and $\lfloor * \rfloor$ represents the floor function.

$$l = \lfloor L \times \mathrm{VLM}(\mathbf{s}) \rfloor \tag{7}$$

It is also noted that for $\beta_1 < \mathbf{s} < \beta_2$, we set the minimum active length of the dynamic mask to $\frac{L}{2}$, as shown in Eq.6 and Eq.7. Our experimental results show that the complex overlapping segments of adults and children's speech may cause the Si-SNR value of speech (especially the overlapping segment speech) to fluctuate widely, this constraint can avoid generating too many fragments and preserve as much of the child's voice as possible.

3.2.2. Determine the location of activated window

As for the position determination problem, we use a sliding window to find the start frame (SF) defined in Eq.(8). For t ranges from 0 to $L - l$, we slide the sliding window with $step = \frac{L}{1000}$ to find a position with the biggest Si-SNR. Correspondingly, the left endpoint of such a window will be chosen as SF.

$$\mathrm{SF} = \underset{t}{\mathrm{argmax}}(\mathbf{s}(x_{ss}(t), x_{se}(t))), \quad t \in [0, L - l] \tag{8}$$

In Figure.3 (b), λ represents our trust in the mask operation. Our confidence in the separation increases with each iteration, so we can adjust the value of the λ to reflect our current confidence in the separation results. In fact, our final masked speech $x(t)$ at frame t in each iteration is the weight of separated speech $x_{ss}(t)$ and masked speech $x_{mask}(t)$, in a sense, this is also a way of model fusion.

$$x(t) = \lambda \times x_{mask}(t) + (1 - \lambda) \times x_{ss}(t) \tag{9}$$

where $x_{mask}(t)$ and λ are defined as follows. Note that the dynamic mask is actually a one-hot vector with the same dimension as the separated speech, where the active frames are set to 1 and the rest are 0, and $dm(t)$ represents the value at frame t.

$$x_{mask}(t) = x_{ss}(t) \times dm(t) \tag{10}$$

$$\lambda = \begin{cases} 0.5 & \mathrm{iter} = 1 \\ 1.0 & \mathrm{iter} = 2, \ldots \end{cases} \tag{11}$$

Eq.11 shows that as the number of iterations increases, so does our confidence in the separation ability of the adapted model. As shown in the Figure.3.(a), the update of the training data requires an excellent pre-trained model to judge the speech quality, thereby generating the corresponding dynamic mask. Model optimization requires data updates to generate model-adapted data. The more the pre-trained separation model can generate accurate dynamic masks to simulate higher-quality adaptive data, the more the generated training data can further improve the adaptive model. Therefore, in general, the jointly optimized two modules dynamically form a closed loop to improve speech separation performance. Our experimental results also confirm that the use of DM can alleviate the problems pointed out at the beginning of this section, and further improve the separation performance based on the IAS method in the previous subsection.

3.3. Post-processing

At the end of each iteration, we calculate the relevant metrics. Since our test sets use data from real scenarios instead of simulation data, common metrics such as PESQ and STOI cannot be calculated. Inspired by [1, 67], we proposed the metrics Jaccard error rate (JER) and child speech duration error rate (CSDER) [3] for children's speech separation in real scenarios based on commonly used binary classification indicators. The separated speech are subjected to our post-processing method to obtain the corresponding binary classification labels. Algorithm 2 shows the post-processing algorithm: we use the IRM nodes of the

Algorithm 2 Algorithm of the post-processing.

Get Masks: Use the fine-tuned model's output IRM to get the Mask $\boldsymbol{z}^{T\times D}$, where T represents number of frames and $D = 257$ is the dimension of IRM;
Calculate the mean: $\bar{\boldsymbol{z}} = mean(\boldsymbol{z}^d), d = 1, ..., D$, $\bar{\boldsymbol{z}}^{T\times 1}$ is the mean of $\boldsymbol{z}$ over dimension D;
Detection and decision: Set the threshold $thres$, $\bar{z}(t)$ is the value of $\bar{\boldsymbol{z}}$ at frame t;
for t = 1 to T **do**
  if $x(t)$ is not silent frame **then**
    if $\bar{z}(t) >= thres$ **then**
      label $x(t)$ as 'Child';
    else
      label $x(t)$ as 'Adult';
    end if
  end if
end for
Obtain labels: Compare obtained binary labels and calculate metrics.

output layer to generate the separated masks for each frame, then we calculate the mean of these masks over all dimensions for each frame, and check the frames whose measured mean is lower than the threshold can be regarded as adult speech segments. Next, by comparing with the original Voice Activity Detection (VAD) file, the speech parts exceed the threshold can be labeled as child speech and the rest of the non-silent human speech were labeled as adult speech.

After obtaining the binary labels, we compute the corresponding metrics. In [3], we propose two binary label-based metrics JER and CSDER, which are defined in Eq.(12) and Eq.(13),

$$\text{JER} = \frac{\text{FA} + \text{Miss}}{\text{Total}} = \frac{\text{FN} + \text{FP}}{\text{FN} + \text{FP} + \text{TP} + \text{TN}} \tag{12}$$

$$\text{CSDER} = \frac{|\text{ECSD} - \text{OCSD}|}{\text{Total}} \tag{13}$$

$$\text{BER} = \frac{\text{FA}_{\text{rate}} + \text{Miss}_{\text{rate}}}{2} = \frac{1}{2}(\frac{\text{FP}}{\text{FP} + \text{TN}} + \frac{\text{FN}}{\text{FN} + \text{TP}}) \tag{14}$$

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 proportion of wrong labels in the total samples. In this paper, we adjust it to the false alarm rate and missed alarm rate widely used in machine learning, and the corresponding JER is also replaced by BER (Balanced Error Rate), which can more reasonably characterize the system's capabilities. The calculation of BER is shown in Eq.14, BER and CSDER are adopted as the evaluation metrics in this paper.

4. Experiment and result analysis

4.1. Experimental settings

We focused on child speech extraction and take BabyTrain as the main dataset. We also introduce parts of some other datasets to improve our data diversity. The configurations of our enhancement experiments were the same as [64]. In our pre-train stage, the adult speech data were derived from four data sets, namely the BabyTrain mega corpus, WSJ0 corpus [51, 52], part of AISHELL-1 corpus [53] and part of Librispeech corpus [68]. The child speech data were derived from two data sets, namely the BabyTrain mega corpus and the part with children aged from kindergarten to grade 5 of CSLU Kids Corpus [69]. 19562 children's

Table 2: Balanced Error Rate (BER) and CSDER values comparison. SS represents speech separation and SE represents speech enhancement.

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
SE+SS[3]	0.373	0.408	0.437	0.426	0.448	0.432	0.123	0.383	0.306	0.342	0.243	0.313

utterances (about 55 hours) were mixed with the above 58,686 adult utterances at 7 target inference ratio (TIR) levels (-5db, 0db, 5db, 10db, 15db, 20db and 25db) to construct a 500-hour training set consisting of children's utterance pairs and mixed utterance pairs. 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 32ms and frame shift was 16ms. 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 very quiet environment. It does not contain any overlapping segments, the scene is also very single, and the content is very clean which is not suitable for verifying the separation system in the complex real 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 [70] as our deep learning framework for model training and decoding, but here we migrated the prior work to the PyTorch [71] framework due to applicability and cutting-edge issues. Our network structure is Progressive Multi-target Learning based Bi-LSTM (PMT-Bi-LSTM). All experiments were conducted on GeForce RTX™ 3090. In all pre-train and fine-tune stages, MSE was used as optimization criterion. We used Adam as our optimizer, 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 feature were used as our inputs in the training and decoding period. Progressive Log Power Spectrum (PLPS) and Progressive Ratio Mask (PRM) are adopted as our training targets for all Bi-LSTM layers. For our proposed 3-layer Bi-LSTM based separation systems, each target TIR gain was 10dB. The input of current layer and the estimations of 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 final layer was used to decode the speech. α in Eq.(6) was set to 1.7, β_1 and β_2 were automatically decided according to the distribution of Si-SNR. All hyper-parameters were tuned with the development sets of BabyTrain subsets. In the post-processing part, oracle Voice Activity Detection (VAD) [72] information is used and kaldi [1] was applied to extract i-vectors [73, 74] to visualize the separated results. Note that our strategies and all parameter optimizations were done on the development sets and test on test sets.

4.2. Results and analysis

We conducted experiments on the test sets of BabyTrain. In [3], we didn't give the ablation experiments that quantitatively present the effectiveness of 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 performance of the system to a certain extent, but there is no obvious improvement in BER of the Tsimane set and the CSDER of the Lena_lyon set. Considering the fact that these two datasets are 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.

Figure 5 shows the comparison of the spectrograms before and after processing by the SS system and the SE+SS system. The top rectangle represents the sound category, the blue segments are the target child's

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

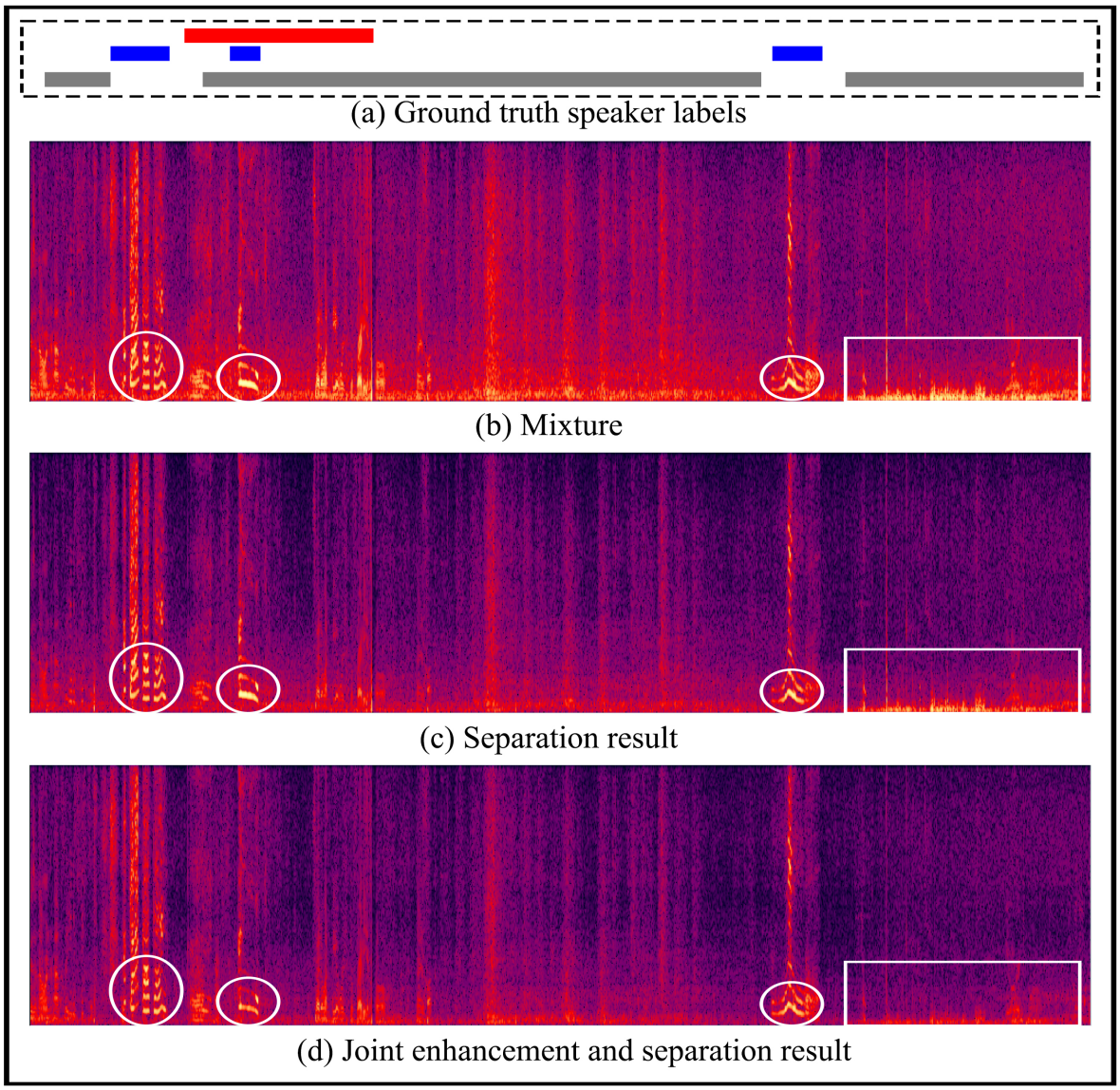


Figure 5: Spectrograms comparison of an utterance from test set. In (a), the red bar represents speech regions of adult, while the blue bar represents the target child speech and the grey bar denotes environmental noises. (b) gives the original spectrum. (c) and (d) show the results processed by SS and SE+SS, respectively.

speech, the gray ones are the non-negligible background noises, and the red ones stand for adult speech. The circles on the spectrogram 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 children's speech in the first and third circles. However, the distortion of voice in the overlapping segment is relatively large, which causes a certain loss to the target speech while suppressing the noises. In addition, the boxes on the far right show that the addition of the enhanced model can well suppress the low-frequency background noises, which can make the separated children's speech purer.

For IAS and DM-IAS, we have the following results. 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 on the test set. Since the speakers in the test set

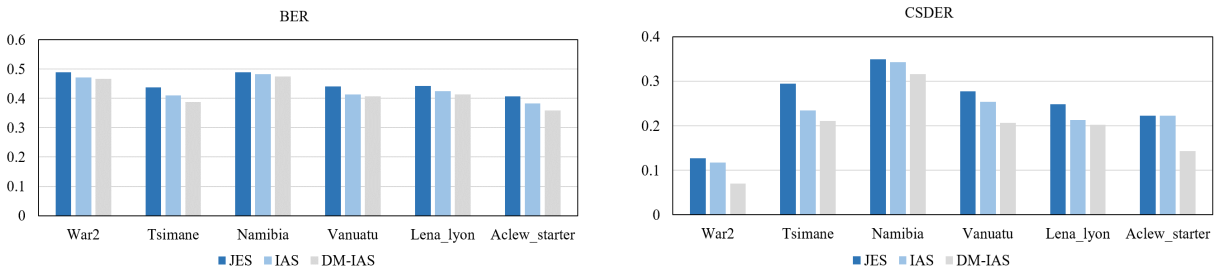


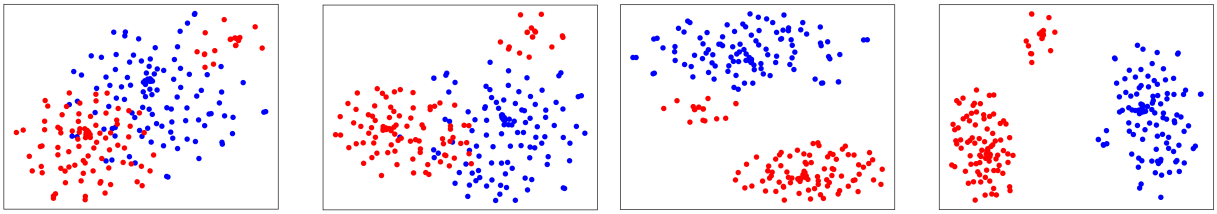
Figure 6: Bar charts of BER and CSDER results on several subsets of the development set

Table 3: BER and CSDER values on test sets, JES denotes joint separation and enhancement system. IAS and DM-IAS stand for our proposed iterative adaptive separation framework without and with dynamic mask operation.

Subset	BER					CSDER				
	JES	IAS		DM-IAS		JES	IAS		DM-IAS	
		Iter1	Iter2	Iter1	Iter2		Iter1	Iter2	Iter1	Iter2
War2	0.373	**0.366**	0.368	**0.354**	0.387	0.123	0.081	**0.078**	**0.066**	0.086
Tsimane	0.408	0.399	**0.395**	0.405	**0.395**	0.383	0.289	**0.267**	0.340	**0.170**
Namibia	0.437	0.437	**0.429**	0.436	**0.424**	0.306	0.313	**0.297**	0.293	**0.283**
Vanuatu	0.426	0.447	**0.445**	0.429	**0.381**	0.342	**0.313**	0.380	0.344	**0.243**
Lena_lyon	0.448	**0.425**	0.426	**0.421**	0.426	0.243	0.207	**0.200**	**0.192**	0.200
Aclew_starter	0.454	**0.449**	0.469	**0.447**	0.448	0.206	**0.182**	0.263	**0.147**	0.160
Overall	0.432	0.428	**0.423**	0.427	**0.415**	0.312	0.290	**0.281**	0.288	**0.244**

are independent but belong to the same language as the development set, this can better illustrate the effectiveness of our method in different languages. We selected the results of four representative subsets to draw a bar chart as shown in Figure 6. The deep-blue bars represent the JES system, and the light-blue and gray bars refer to the IAS and DM-IAS systems, respectively. It can be found that our DM-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).

Table 3 presents an overall BER and CSDER comparison among different separation methods on selected subsets of BabyTrain corpus. JES denotes joint speech enhancement and speech separation system. IAS and DM-IAS stand for our newly proposed iterative adaptive separation framework without and with dynamic mask operation. We mentioned earlier that the separation model used in the testing phase is the best model among all iterations selected according to the development set. Here, in order to better demonstrate the effectiveness of the iterative strategy, we give the results of each iteration round in IAS and DM-IAS columns. It can be seen that in a relatively simple acoustic scene, our system can be optimal in one iteration. But as the complexity of the scene increases, more iterations are needed to gradually clean out the residual noises in the separated speech. 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. It can be seen that 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 more data (e.g., Lena_lyon and Namibia), except for the BER on Vanuatu. As a result, the introduction of DM 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 IAS method has an improvement of 0.7% and 4.2% in BER and CSDER on the War2 dataset, respectively. compared with the JES method, and after the introduction of the DM method, an increase of 1.2% and 1.5% was obtained in these two indicators.



(a) Origin (b) JES (c) IAS (d) DM-IAS

Figure 7: T-SNE graph comparison between adult and child speech

Compared with the system that only uses the pre-trained model (SS in Table 2), our proposed DM-IAS achieved an improvement of 4.2% and 9.7% absolutely. It is worth noting that the optimal values of BER and CSDER for the same subset may be reached in different iterations. But the general trend is that the results of DM-IAS are better than IAS and JES. Overall results show that both BER and CSDER decrease with the increase of iterations, and the results of DM-IAS are better than those of IAS in the corresponding iteration, which are both better than JES.

Figure 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's speech and the blue ones represent the child's speech [75]. Figure 7(a) represents the mapping of adult speech and child speech distance information in two-dimensional space in the original utterance, as shown in the figure, although there is an inherent gap between adult speech and child voice, there is still lots of mixed regions. 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 this parts and found that these deviated parts are mainly the overlapping segments of adult-child speech and some adults imitate the child's voices to tease the child. As previously introduced, the recording device is worn on the children, so the energy of the child's voices are stronger and the adult voices are far-field. Hence, this parts of the speech will be more inclined to the child's parts in terms of spatial distance. From the figure, this parts are at the edge of the child's speech but there are still some overlap regions. In Figure 7(b), the original audio was processed by our JES approach. It can be found that compared with Figure 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 have no overlap regions, but it is still close to the blue edge. This indicates that the JES system plays a certain role, but it is difficult to separate the mixed speech of adults and children in some extreme cases. By observing Figure 7(b) and Figure 7(c), it can be found that after adopting our proposed IAS system, the separation results are further improved. The main parts of adult and child speech are separated by a larger distance in Figure 7(c). This indicates that our IAS method is able to obtain targeted breakthroughs on each subset. However, IAS does not deal well with the speech segments in which adults imitate children. Although this parts of speech can be separated from children's speech after using dynamic mask (as is shown in Figure 7(d)), the distance between them is still relatively close in general, which also limits our separation performance to a certain extent.

Figure 8 shows the ground truth labels and the spectrograms comparision of different methods and the mixtures. The top rectangular box is the ground-truth label, in which the red rectangular segments represent adults' speech, the middle blue rectangular segments represent children's speech, and the bottom gray bar denotes the existence of a lot of noises in the whole speech. Note that there are two dark colors on the gray bar, which 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 and the speech processed by IAS and DM-IAS, respectively. The white boxes on the spectrograms are the child's speech, and the whilte ellipsoids are the recognition speech. By comparison, we can find that the proposed DM-IAS system is significantly better than the JES method and the pure IAS framework in retaining the child's speech. Our proposed IAS and DM-IAS systems also suppress the adult speech to some extent, and even include the adult parts in the overlapping segments. The above results also prove the effectiveness of our proposed method. It is worth mentioning that the development sets of War2 and Aclew_starter only contain

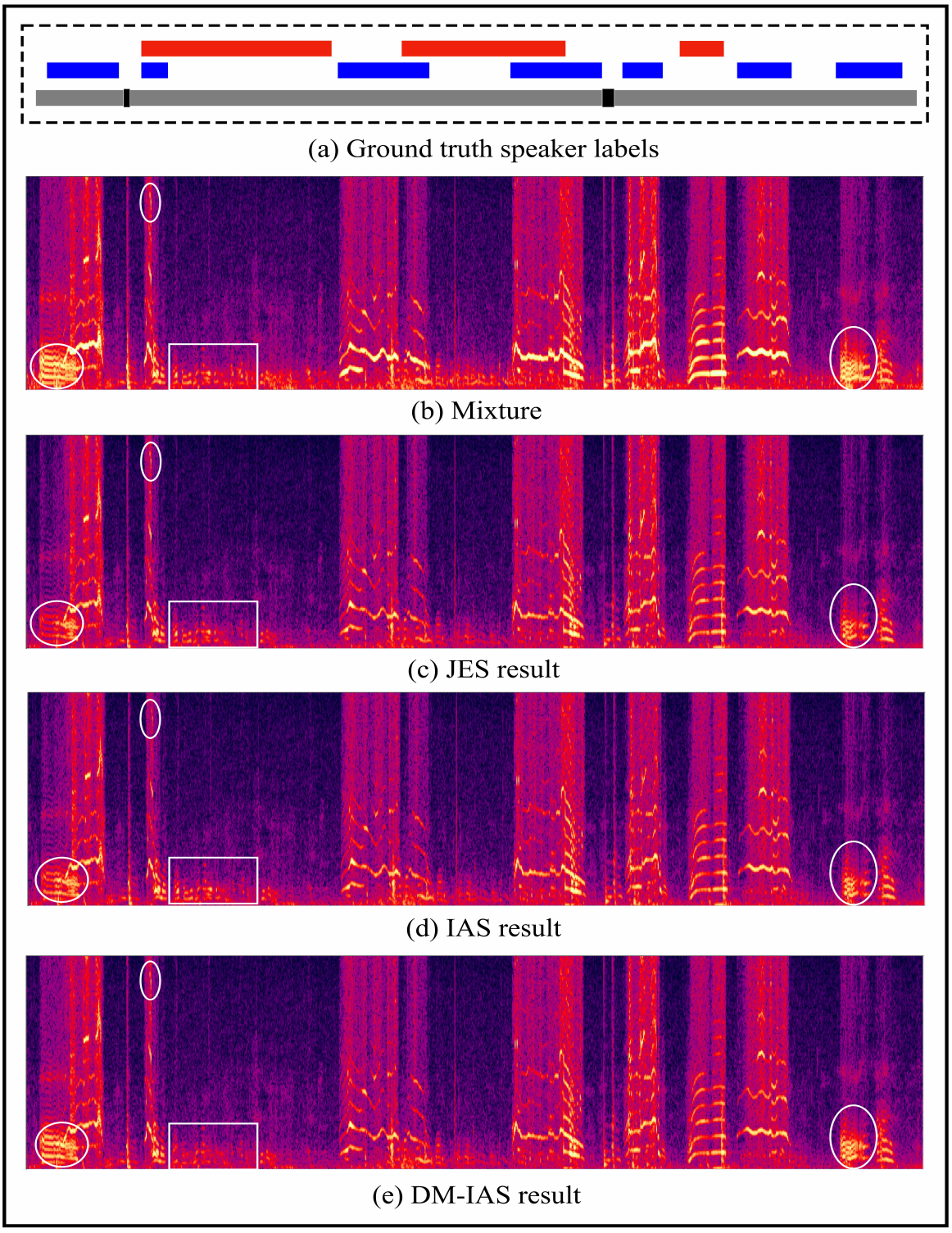


Figure 8: Spectrograms comparison of an utterance from test set. In (a), the red bar represents speech regions of adult, while the blue bar represents thetarget child speech and the grey bar denotes environmental noises. (b) gives the origin spectrum. (c)-(e) show the results processed by JES, IAS and DM-IAS, respectively. Black parts on grey bars represent high frequency sharp noises.

3 and 5 segments (each piece is about 5 minutes long) respectively, and the results of the development set and the test set both show that when we get a completely 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-training separation model first. Afterwards, a new training set is constructed based on the separation results of this recording and the original sentence, and an adaptive separation model can be obtained by finetuning the pre-trained model with little computational cost and time.

5. Conclusions

In this paper, we first propose the IAS framework on the complex real data of multi-speaker and multilingualism. After proving that it is more effective than the joint system JES we proposed earlier, we further introduce a length-position variable dynamic mask to further perform the data purification. Experimental results show that the proposed DM-IAS framework is valid on both BER and CSDER indicators. However, how to better retain children's speech while removing adult speech as much as possible is still a problem that needs to be studied in depth. In the future, we will further expand our research to more complex and severe overlapping conditions, as well as more children of different ages who speak multiple languages under real-world scenarios and continue to explore the effects of some newer deep learning networks on this task.

References

[1] J. Gilkerson, K. K. Coulter, J. A. Richards, Transcriptional analyses of the lena natural language corpus, LENA Foundation (2008).

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

[3] X. Wang, J. Du, A. Cristia, L. Sun, C.-H. Lee, 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, 2020, pp. 7304–7308.

[4] G. Yeung, R. Fan, A. Alwan, 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), 2021, pp. 6993–6997. doi:10.1109/ICASSP39728.2021.9413801.

[5] E. Z. Sattorovich, Psychological influence of speech disorders and the causes that cause them on the child's psyche, Academicia globe: inderscience research 3 (01) (2022) 39–42.

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

[7] K. Kohnert, K. D. Ebert, G. T. Pham, Language disorders in bilingual children and adults, Plural Publishing, 2020.

[8] Y. Hus, O. Segal, Challenges surrounding the diagnosis of autism in children, Neuropsychiatric Disease and Treatment 17 (2021) 3509.

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

[10] D. A. Reynolds, Gaussian mixture models., Encyclopedia of biometrics 741 (659-663) (2009).

[11] D. Xiangjun, V. Yip, A multimedia corpus of child mandarin: The tong corpus, Journal of Chinese Linguistics 46 (1) (2018) 69–92.

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

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

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

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

[16] A. Sanchez, S. C. Meylan, M. Braginsky, K. E. MacDonald, D. Yurovsky, M. C. Frank, childes-db: A flexible and reproducible interface to the child language data exchange system, Behavior research methods 51 (4) (2019) 1928–1941.

[17] B. MacWhinney, C. Snow, The child language data exchange system, Journal of Child Language 12 (2) (1985) 271–295. doi:10.1017/S0305000900006449.

[18] B. MacWhinney, The childes system, American Journal of Speech-Language Pathology 5 (1) (1996) 5–14.

[19] B. MacWhinney, The CHILDES project: The database, Vol. 2, Psychology Press, 2000.

[20] B. MacWhinney, From childes to talkbank, Department of Psychology (2001) 182.

[21] B. MacWhinney, The CHILDES project: Tools for analyzing talk, Volume II: The database, Psychology Press, 2014.

[22] S. Luz, F. Haider, S. de la Fuente, D. Fromm, B. MacWhinney, Alzheimer's dementia recognition through spontaneous speech: the adress challenge, arXiv preprint arXiv:2004.06833 (2020).
[23] K. Shobaki, J.-P. Hosom, R. Cole, The ogi kids' speech corpus and recognizers, in: Proc. of ICSLP, 2000, pp. 564–567.
[24] C. E. C, Some experiments on the recognition of speech, with one and with two ears, The Journal of the acoustical society of America (1953) 975–979.
[25] B. Arons, A review of the cocktail party effect, Journal of the American Voice I/O Society 12 (7) (1992) 35–50.
[26] S. Haykin, Z. Chen, The cocktail party problem, Neural computation 17 (9) (2005) 1875–1902.
[27] M. A. Bee, C. Micheyl, The cocktail party problem: what is it? how can it be solved? and why should animal behaviorists study it?, Journal of comparative psychology 122 (3) (2008) 235.
[28] C. Demir, M. Saraclar, A. T. Cemgil, Single-channel speech-music separation for robust asr with mixture models, IEEE Transactions on Audio, Speech, and Language Processing 21 (4) (2012) 725–736.
[29] N. Kanda, C. Boeddeker, J. Heitkaemper, Y. Fujita, S. Horiguchi, K. Nagamatsu, R. Haeb-Umbach, Guided source separation meets a strong asr backend: Hitachi/paderborn university joint investigation for dinner party asr, arXiv preprint arXiv:1905.12230 (2019).
[30] T. Heittola, A. Mesaros, T. Virtanen, A. Eronen, Sound event detection in multisource environments using source separation, in: Machine Listening in Multisource Environments, 2011, p. None.
[31] Q. Kong, Y. Xu, W. Wang, M. D. Plumbley, 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, 2018, pp. 321–325.
[32] N. Turpault, S. Wisdom, H. Erdogan, J. Hershey, R. Serizel, E. Fonseca, P. Seetharaman, J. Salamon, Improving sound event detection in domestic environments using sound separation, arXiv preprint arXiv:2007.03932 (2020).
[33] S. Rustamov, N. Akhundova, A. Valizada, Automatic speech recognition in taxi call service systems, in: International Conference for Emerging Technologies in Computing, Springer, 2019, pp. 243–253.
[34] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo, et al., 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, 2021, pp. 897–904.
[35] H. Ling, P. Han, J. Qiu, L. Peng, D. Liu, K. Luo, 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, 2021, pp. 53–61.
[36] E. Vincent, N. Bertin, R. Gribonval, F. Bimbot, From blind to guided audio source separation: How models and side information can improve the separation of sound, IEEE Signal Processing Magazine 31 (3) (2014) 107–115.
[37] T. Virtanen, J. F. Gemmeke, B. Raj, P. Smaragdis, Compositional models for audio processing: Uncovering the structure of sound mixtures, IEEE Signal Processing Magazine 32 (2) (2015) 125–144.
[38] S. U. Wood, J. Rouat, S. Dupont, G. Pironkov, Blind speech separation and enhancement with gcc-nmf, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (4) (2017) 745–755.
[39] J. Le Roux, J. R. Hershey, F. Weninger, Deep nmf for speech separation, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 66–70. doi:10.1109/ICASSP.2015.7177933.
[40] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, M. Matassoni, The second 'chime' speech separation and recognition challenge: Datasets, tasks and baselines, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 126–130. doi:10.1109/ICASSP.2013.6637622.
[41] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smaragdis, Deep learning for monaural speech separation, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 1562–1566. doi:10.1109/ICASSP.2014.6853860.
[42] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1, NASA STI/Recon technical report n 93 (1993) 27403.
[43] J. Greenberg, P. Peterson, P. Zurek, Intelligibility-weighted measures of speech-to-interference ratio and speech system performance, The Journal of the Acoustical Society of America 94 (5) (1993) 3009–3010.
[44] E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation, IEEE transactions on audio, speech, and language processing 14 (4) (2006) 1462–1469.
[45] V. Emiya, E. Vincent, N. Harlander, V. Hohmann, Subjective and objective quality assessment of audio source separation, IEEE Transactions on Audio, Speech, and Language Processing 19 (7) (2011) 2046–2057.
[46] J. Le Roux, S. Wisdom, H. Erdogan, J. R. Hershey, Sdr–half-baked or well done?, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 626–630.
[47] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, J. Zhong, Attention is all you need in speech separation, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 21–25.
[48] Y. Luo, N. Mesgarani, Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (8) (2019) 1256–1266. doi:10.1109/TASLP.2019.2915167.
[49] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, J. R. Hershey, Universal sound separation, in: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), IEEE, 2019, pp. 175–179.
[50] D. Ditter, T. Gerkmann, 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, 2020, pp. 36–40.
[51] J. Garofolo, D. Graff, D. Paul, D. Pallett, Csr-i (wsj0) complete ldc93s6a, Web Download. Philadelphia: Linguistic Data Consortium 83 (1993).
[52] J. R. Hershey, Z. Chen, J. Le Roux, S. Watanabe, Deep clustering: Discriminative embeddings for segmentation and

separation, in: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2016, pp. 31–35.
[53] H. Bu, J. Du, X. Na, B. Wu, H. Zheng, 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, 2017, pp. 1–5.
[54] E. Vincent, S. Watanabe, J. Barker, R. Marxer, The 4th chime speech separation and recognition challenge, URL: http://spandh. dcs. shef. ac. uk/chime_challenge/(last accessed on 1 August, 2018) (2016).
[55] J. Barker, S. Watanabe, E. Vincent, J. Trmal, The fifth'chime'speech separation and recognition challenge: dataset, task and baselines, arXiv preprint arXiv:1803.10609 (2018).
[56] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, et al., Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings, arXiv preprint arXiv:2004.09249 (2020).
[57] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, M. Zhou, Continuous speech separation with conformer, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 5749–5753.
[58] J. Shi, J. Xu, Y. Fujita, S. Watanabe, B. Xu, Speaker-conditional chain model for speech separation and extraction, arXiv preprint arXiv:2006.14149 (2020).
[59] M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, A. Cristia, An open-source voice type classifier for child-centered daylong recordings, arXiv preprint arXiv:2005.12656 (2020).
[60] E. Bergelson, A. Warlaumont, A. Cristia, M. Casillas, C. Rosemberg, M. Soderstrom, C. Rowland, S. Durrant, J. Bunce, Starter-aclew, Databrary Retrieved November 9 (2017) 2018.
[61] M. Canault, M.-T. Le Normand, S. Foudil, N. Loundon, H. Thai-Van, Reliability of the language environment analysis system (lena™) in european french, Behavior research methods 48 (3) (2016) 1109–1124.
[62] M. VanDam, A. S. Warlaumont, E. Bergelson, A. Cristia, M. Soderstrom, P. De Palma, B. MacWhinney, Homebank: An online repository of daylong child-centered audio recordings, in: Seminars in speech and language, Vol. 37, Thieme Medical Publishers, 2016, pp. 128–142.
[63] G. M. Pretzer, L. D. Lopez, E. A. Walle, A. S. Warlaumont, Infant-adult vocal interaction dynamics depend on infant vocal type, child-directedness of adult speech, and timeframe, Infant Behavior and Development 57 (2019) 101325.
[64] L. Sun, J. Du, X. Zhang, T. Gao, X. Fang, C.-H. Lee, 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, 2020, pp. 7099–7103.
[65] J. Han, C. Moraga, The influence of the sigmoid function parameters on the speed of backpropagation learning, in: International workshop on artificial neural networks, Springer, 1995, pp. 195–201.
[66] X. Yin, J. Goudriaan, E. A. Lantinga, J. Vos, H. J. Spiertz, A flexible sigmoid function of determinate growth, Annals of botany 91 (3) (2003) 361–371.
[67] L. Hamers, et al., Similarity measures in scientometric research: The jaccard index versus salton's cosine formula., Information Processing and Management 25 (3) (1989) 315–18.
[68] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: an asr corpus based on public domain audio books, in: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2015, pp. 5206–5210.
[69] K. Shobaki, J.-P. Hosom, R. Cole, The ogi kids' speech corpus and recognizers, in: Proc. of ICSLP, 2000, pp. 564–567.
[70] F. Seide, A. Agarwal, Cntk: Microsoft's open-source deep-learning toolkit, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 2135–2135.
[71] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems 32 (2019) None.
[72] J. Sohn, N. S. Kim, W. Sung, A statistical model-based voice activity detection, IEEE signal processing letters 6 (1) (1999) 1–3.
[73] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, R. Dehak, Language recognition via i-vectors and dimensionality reduction, in: Twelfth annual conference of the international speech communication association, Citeseer, 2011, p. None.
[74] G. Saon, H. Soltau, D. Nahamoo, M. Picheny, Speaker adaptation of neural network acoustic models using i-vectors, in: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, IEEE, 2013, pp. 55–59.
[75] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (11) (2008).