A STUDY OF CHILD SPEECH EXTRACTION USING JOINT SPEECH ENHANCEMENT AND SEPARATION IN REALISTIC CONDITIONS

Xin Wang[1]*, Jun Du*[1]*, Alejandrina Cristia*[2]*, Lei Sun*[1]*, Chin-Hui Lee*[3]

[1]University of Science and Technology of China, Hefei, Anhui, China
[2]Laboratoire de Sciences Cognitives et Psycholinguistique, ENS, Paris, France
[3]Georgia Institute of Technology, Atlanta, Georgia, USA

ABSTRACT

In this paper, we design a novel joint framework of speech enhancement and speech separation for child speech extraction in realistic conditions, targeting the problem of extracting child speech from daily conversations in BabyTrain mega corpus. To the best of our knowledge, it is the first discussion of a feasible method for child speech extraction in realistic conditions. First, we make detailed analysis of the BabyTrain mega corpus, which is recorded in adverse environments. We observe problems of background noises, reverberations and child speech that is partially obscured by adult speech (for instance due to speaker overlap but also imitation by the adult). Motivated by this, we conduct a joint framework of speech enhancement and speech separation for child speech extraction. To measure the extraction results in realistic conditions, we propose several objective measurements to evaluate the performance of the our system, which is different from those commonly used for simulation data. Compared with the unprocessed approach and classification approach, our proposed approach can yield the best performance among all subsets of BabyTrain.

Index Terms— Child Speech Extraction, Speech Separation, Measures, Speech Enhancement, Realistic Conditions

1. INTRODUCTION

Recent years have seen a literal explosion in the use of child-centered audio-recordings, gathered as infants and young children go about their day [1]. The resulting data are of interest to both a wide range of theories (e.g., developmental psychology, cognitive science) and numerous applications (e.g., the diagnosis of potential language disorders, the measurement of effects of an intervention). Despite the interest in these data, there are very few analysis algorithms that can cope with these data, which truly deserve the name of 'in the wild'. To begin with, much of the voice recorded belongs to the infant or child wearing the device, who produce non-speech vocalizations (such as crying as well as non-emotional, non-speech productions). Moreover, the other people recorded may vary in their closeness to the microphone, such that their voice alternates between near-field and far-field within the same recording. Finally, many people may be recorded; in our experience, children can come across 20 people over a normal day, with as many as 9 people in a 5-minute interval [2].

At present, to the best of our knowledge, there is only one algorithm that can be used for such data. The LENA Foundation developed an algorithm by training acoustic models on about 150h hand-annotated data [3]. In a nutshell, the algorithm extracts Mel-Frequency Cepstral Coefficients (MFCCs) in 10 ms windows, but then applies a minimal duration Gaussian Mixture Model (GMM) to segment the full recording (often 16h long) into segments that are at least 600 ms in length [4]. These segments correspond to broad types of speakers such as male adult and female adult. This software is widely used, with a recent LENA communication stating that over 100 publications had used it in the past 10 years, and interventions had employed it in the last year affecting over 10,000 children in USA [5]. And yet, there are several important disadvantages: The software is relatively expensive (at least 12,000 US$, and more depending on the volume of recordings to be analyzed), and it has not been updated since its inception in the early 2000s. Moreover, some of the key metrics that researchers and practitioners would like to draw from the recordings seems to have low reliability.

Recently, deep neural networks (DNNs) [6] have been utilized in many speech processing areas, such as speech enhancement [7], and speech separation [8]. In [9], long short-term memory recurrent neural network (LSTM-RNN) was used in speech separation. In [10], ideal ratio masks (IRMs) were used to make binary classification on time-frequency (T-F) units. However, all of these studies dealt only with adult speech separation while research in child speech separation is still quite limited in the literature. Moreover, these methods were mostly tested on simulation data, which is quite different from realistic adverse conditions. In our previous work [11], we proposed a progressive learning approach to separating child speech from signals with mixed adult speech in a speaker-independent manner based on a densely connected LSTM architecture [12]. However, the aforementioned method focused only on speech separation, without considering the simultaneous presence of noise and interference in realistic conditions.

In this paper, we propose a novel joint framework of speech enhancement and speech separation for child speech extraction in realistic conditions. First, we make detailed analysis of the BabyTrain mega corpus [13–16] , which draws from recordings in a complex adverse environments. We observe problems of background noises, reverberations and child speech that is partially obscured by adult speech (for instance due to speaker overlap but also imitation by the adult). Motivated by this, we design a joint framework of speech enhancement and speech separation for child speech extraction. We use the state-of-the-art model as our enhancement model. For the separation model, we propose a SNR-Progressive Multi-Target Learning (PMT) based model to separate child speech from mixed noisy speech. To evaluate the extraction results in realistic conditions, we propose several objective measurements. Compared with the unprocessed results and results from the classification model, our proposed approach can yield the best performance.

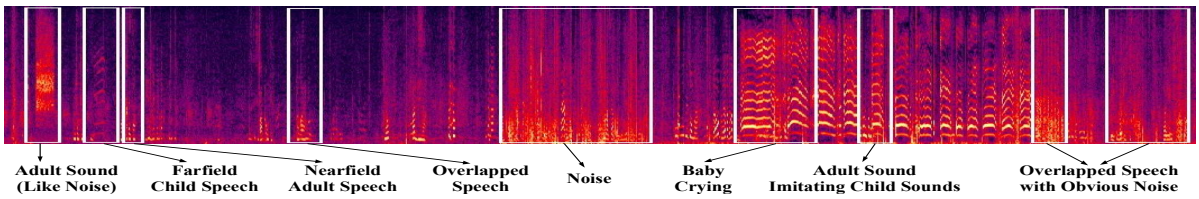

Fig. 1. An utterance example from the BabyTrain data set.

Table 1. A description of selected corpora present in the BabyTrain data set

Child Data	Language	Hours(h)	Farfield	Noise	Overlapped
aclew-starter	English Spanish French Tseltal	1.50			
LENA_Lyon	French	26.85	✓	✓	✓
Namibia	Yu'lhoan	23.73			
Tsimane	Tsimane'	4.53			
Vanuatu	Many	15.72			
Pretzer	English	0.83			

2. CORPUS AND CHALLENGE

The main data set that we used in our experiments was the BabyTrain mega corpora, which is an aggregation of 8 child-centered corpora. This results in 245 hours of recordings acquired in a wide range of conditions, including daily life, inside, outside, during parties and so on. Each recording is sampled at 44.1 kHz and comes with its human-made transcription files. For informational purposes, we give a description of main corpus in BabyTrain as shown in Table 1. First, BabyTrain data were recorded in an adverse environment, together with noise and reverberations, and speech of adults and children overlap because of the informal nature of interactions. Second, BabyTrain contains both farfield speech and nearfield speech. Third, BabyTrain consists of different languages, which vary in their phonetic characteristics. All of these effects make BabyTrain difficult to process in realistic adverse conditions.

To introduce the data set in a more intuitive way, an utterance selected from the subset of namibia in BabyTrain is presented in Fig.1. This utterance includes both nearfield speech and farfield speech with noise, a region of overlap between female adult speech and child speech, some crying sounds by the baby, and even an example of an adult who is imitating the child's sounds. These conditions are quite challenging to handle.

Motivated by those analyses, we propose a joint framework of speech enhancement and speech separation for child speech extraction in realistic conditions and come up with several objective measurements to evaluate the performance of our system.

3. THE PROPOSED OVERALL FLOW

As illustrated in Fig. 2, the overall flow of our proposed joint framework for child speech extraction in realistic conditions consists of three main modules, namely speech enhancement, speech separation and post-processing, which are elaborated in the following subsections.

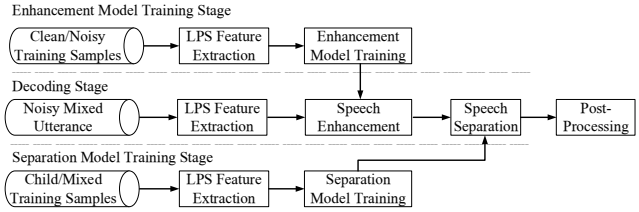

Fig. 2. The proposed overall flow for child speech signal processing.

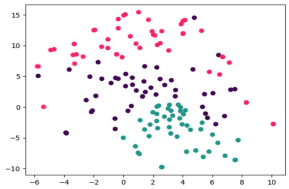

Fig. 3. T-SNE graph of the i-vector distances among Babies(Pink), Children(Purple) and Adults(Green).

3.1. Speech Separation

Firstly, we proved that distances between child and adult speech are large enough to warrant a possible separation. Based on our previous studies, speaker separability can be tied to distances between speaker groups by adopting i-vectors [17] to represent each speaker. To visualize the similarity of different individual objects in a low-dimensional space, each object can be represented by a point and the points are projected in order to approximate the distances between pairs of objects. We adopted t-SNE graph [18] to graphically describe the dissimilarity of different speaker groups. The t-SNE graphs of i-vector based distance matrices for 50 adult speakers, 50 child speakers and 50 baby speakers (age < 5) from our training set (introduced in Section 5) are shown in Fig.3. In this figure, the pink, purple, and green points represent the baby, child, and adult speakers, respectively. Fig.3 confirms that the both baby&child groups and the adult groups could be well separated in two clusters for most cases, which motivates our proposed separation approach in the next.

Our child speech separation framework is a progressive multi-targets LSTM network (LSTM_PMT) as shown in Fig.4. The LSTM_PMT can be divided into several successive stacked blocks and each block is made up with one LSTM layer and one fully connected layer via multi-targets learning. The fully connected layer in each block is also referred to as a target layer, which is designed to learn intermediate speech targets with a higher target-inference ratio (TIR) than the targets of previous target layers. A series of progressive ratio masks (PRM) are concatenated with the progressively separated log-power spectra (PLPS) features together as the learning targets.

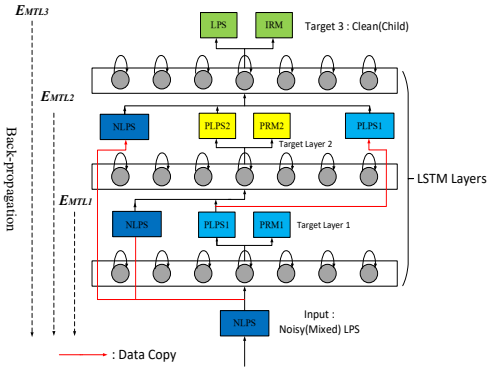

Fig. 4. The proposed LSTM_PMT architecture for child speech separation.

The PRM target is defined as follows:

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

where C(t,f) represents the power spectrum of the child speech signal at the time-frequency (T-F) unit (t,f), $A_{\text{T}}(t,f)$ and $A_{\text{I}}(t,f)$ represent the power spectrum of the adult in one PRM target and input signals at the T-F unit (t, f), respectively. When the numerator of Eq. (1) becomes the power spectrum of the child speech signal, $A_{\text{T}}(t,f)$ is zero and $z^{\text{PRM}}(t,f)$ is regressed to the traditional IRM. Hence, in practical use, the PRMs can serve as a progressively stronger separation ability. Here, the total number of target layers K is set to 3. Correspondingly, a weighed minimum mean squared error (MMSE) criterion is designed to optimize all network parameters randomly initialized with K target layers as follows:

$$E_{\text{MTL}}(k) = \sum_{m=1}^{K} E_{\text{PLPS}}(m) + E_{\text{PRM}}(m) \tag{2}$$

$$E_{\text{PLPS}}(m) = \frac{1}{N}\sum_{n=1}^{N} \left\| \mathcal{F}_{\text{m}}(\hat{\boldsymbol{x}}_{\text{n}}^{0}, \hat{\boldsymbol{x}}_{\text{n}}^{1}, ..., \hat{\boldsymbol{x}}_{\text{n}}^{\text{m-1}}, \boldsymbol{\Lambda}_{\text{m}}) - \boldsymbol{x}_{\text{n}}^{\text{m}} \right\|_2^2 \tag{3}$$

$$E_{\text{PRM}}(m) = \frac{1}{N}\sum_{n=1}^{N} \left\| \mathcal{F}_{\text{PRM}}(\hat{\boldsymbol{x}}_{\text{n}}^{0}, \hat{\boldsymbol{x}}_{\text{n}}^{1}, ..., \hat{\boldsymbol{x}}_{\text{n}}^{\text{m-1}}, \boldsymbol{\Lambda}_{\text{PRM}}) - \boldsymbol{x}_{\text{n}}^{\text{PRM}} \right\|_2^2 \tag{4}$$

where $E_{\text{MTL}}(k)$ corresponds to the multi-target loss in k^{th} target layer. It's the sum of two kinds of losses, namely $E_{\text{PLPS}}(m)$ and $E_{\text{PRM}}(m)$ from all lower layers. $\hat{\boldsymbol{x}}_{\text{n}}^{\text{m}}$ and $\boldsymbol{x}_{\text{n}}^{\text{m}}$ are the n^{th} D-dimensional vectors of estimated and reference target PLPS feature vectors for m^{th} target layer, respectively ($m > 0$), with N representing the mini-batch size. $\hat{\boldsymbol{x}}_{\text{n}}^{0}$ denotes the n^{th} vector of input noisy LPS features with acoustic context. $\mathcal{F}_{\text{m}}(\hat{\boldsymbol{x}}_{\text{n}}^{0}, \hat{\boldsymbol{x}}_{\text{n}}^{1}, ..., \hat{\boldsymbol{x}}_{\text{n}}^{\text{m-1}}, \boldsymbol{\Lambda}_{\text{m}})$ is the neural network function for the m^{th} target with the dense structure using the previously learned intermediate targets from $\hat{\boldsymbol{x}}_{\text{n}}^{0}$ to $\hat{\boldsymbol{x}}_{\text{n}}^{\text{m-1}}$, and $\boldsymbol{\Lambda}_{\text{m}}$ represents the parameter set of the weight matrices and bias vectors before m^{th} target layer, which are optimized in the manner of BPTT with gradient descent. $\boldsymbol{x}_{\text{n}}^{\text{PRM}}$, $\mathcal{F}_{\text{PRM}}(\hat{\boldsymbol{x}}_{\text{n}}^{0}, \hat{\boldsymbol{x}}_{\text{n}}^{1}, ..., \hat{\boldsymbol{x}}_{\text{n}}^{\text{m-1}}, \boldsymbol{\Lambda}_{\text{PRM}})$, and $\boldsymbol{\Lambda}_{\text{PRM}}$ are corresponding versions to PRM targets.

Different from our previous work in [11], the proposed model in this paper is a progressive multi-targets network with progressive ratio masks (PRM), which is not considered in [11]. By using new structure, the separated speech has better intelligibility and less distortions.

3.2. Speech Enhancement

We use the state-of-the-art speech enhancement model proposed in [19] as our speech enhancement model. The deep model architecture is similar to that in Fig.4. The main difference is that the input and output of the child separation model is mixed speech and child speech, respectively, while the input and output of the enhancement model is noisy speech and clean speech, respectively. More details of the preprocessor can refer to [19].

3.3. Post-processing as a Classification Problem

After getting the child speech from our separation model, we post-processed separation results to classify child and adult frames. We used the output node of IRM to generate the final separation masks of each frame. Then we calculated the mean value of masks across all dimensions of each frame, and detected masks whose mean was below the threshold, which could be regarded as adult speech. Next, by comparing with oracle voice activity detection (VAD) files, the speech parts that were above the threshold could be regarded as speaker labels of child. Finally according to the above-mentioned principles, we generated the 2-class labels and calculated the measures to be introduced in the next session.

4. PROPOSED MEASURES IN REALISTIC CONDITIONS

For simulation data, there are many objective measures to evaluate the performances of separated speech, such as perceptual evaluation of speech quality (PESQ) [20] and short-time objective intelligibility (STOI) [21]. Developers using simulation data can calculate the aforementioned objective functions due to the existence of origin clean child speech. But in realistic conditions, there is no clean child speech corresponding to the noisy mixed speech. A simple idea is to use the backend applications to directly measure the goodness of separation methods.In this study, we focus on the evaluation without backend applications. We used post-processing approach mentioned in session 3.3 to convert separation results to classification labels and came up with several measures to compare with classification models.

The first measure is Jaccard error rate (JER), which is based on the Jaccard similarity index [22], a metric commonly used to evaluate the output of image segmentation systems. JER is defined as :

$$\text{JER} = \frac{\text{FA} + \text{Miss}}{\text{Total}} \tag{5}$$

where Total is the duration of the union of child and adult speaker segments, FA is the total child speaker time detected by system but not attributed to the reference child speaker, and Miss is the total reference child speaker time but not detected by the system. JER is commonly used in the diarization [23] task. But different from that the reference speaker in diarization task is determined due to the permutation problem, our JER directly regard child as the only reference speaker.

The second measure is child speech duration error rate (CSDER), inspired by the measure about child speech duration ratio in LENA Foundation [3], which is defined as:

$$\text{CSDER} = \frac{|\text{ECSD} - \text{OCSD}|}{\text{Total}} \tag{6}$$

where Total is the duration of the union of child and adult speaker segments, ESCD is the child speech duration time detected by system, OCSD is the oracle child speech duration time.

Table 2. Average performance comparison on different subsets and overall corpus among Unprocessed, Classification and Ours system.

Subset	Systems	JER	CSDER
namibia	Unprocessed	0.500	0.475
	Classification	0.495	0.468
	Ours	0.474	0.327
lena_lyon	Unprocessed	0.500	0.590
	Classification	0.498	0.378
	Ours	0.423	0.148
tsimane	Unprocessed	0.500	0.463
	Classification	0.499	0.426
	Ours	0.389	0.297
Overall	Unprocessed	0.500	0.505
	Classification	0.495	0.454
	Ours	0.449	0.272

5. EXPERIMENTS AND RESULT ANALYSIS

For BabyTrain mega corpus, the 245 hours of recordings have been splitted such that each speaker belongs to one and only one of train set, development set and test set. Each corpora is splitted such that around 60% of the key children go to the training set, 30% to the development set and 10% to the test set. The obtained proportion, in terms of cumulated duration, is 57.5% for the train set, 27% for the development set and 15.5% for the test set.

The configuration of our enhancement experiments was the same as [19]. In our separation experiments, the adult speech data was derived from four data sets, namely the BabyTrain mega corpus, WSJ0 corpus [24], part of AISHELL-1 corpus [25] and part of Librispeech corpus [26]. The child speech data was derived from two data sets, namely the BabyTrain mega corpus and the part with children age from kindergarden to grade 5 of CSLU Kids Corpus [27]. The whole 19562 utterances (about 55 hours speech) of child were mixed with the above mentioned 58686 adult utterances at three target-inference ratio (TIR) levels (-5dB, 0dB and 5dB) to build a 500-hour training set, consisting of pairs of child and mixed utterances. The BabyTrain development set was used for testing.

For signal analysis, all of the speech was resampled at 16 kHz. A 512-point discrete Fourier transform (DFT) of each overlapping windowed frame was computed. Then 257-dimensional LPS vectors normalized by global mean and variance were used to train the proposed LSTM_PMT model. The Microsoft Computational Network Toolkit (CNTK) [28] was used for training. For our proposed LSTM_PMT separation systems, one LSTM layer was used to connect the input layer and target layers. Each target TIR gain was 10dB. The 7-frames input and the estimations of intermediate target are spliced together to learn next target. The number of LSTM memory cells in each layer was 1024, and the IRM output of final layer in LSTM_PMT model was used to test. As for the part of post-processing, oracle voice activity detection(VAD) information was used to get speech segments. As a comparison, a direct mapping classification LSTM network with the architecture 257-512-512-512-2, consisting of three LSTM layers and 512 memory cells for each LSTM layer, output being 2-class one-hot vector, was built as our baseline classification model.

Table 2 shows the average JER and CSDER on the three subsets of BabyTrain development set and the whole BabyTrain development set among unprocessed, classification and our system. Clearly, our system yielded consistent improvements on the measure of JER and CSDER over the classification approach and unprocessed approach in all different subsets and overall corpus. The reason for the poor results of classification model is that its generalization capability is quite weak by directly using noisy mixed speech as input for classifying child and adult frames in adverse environments. As a comparison, our system use enhancement model to purify the noisy speech and use separation model to further extract the child speech.

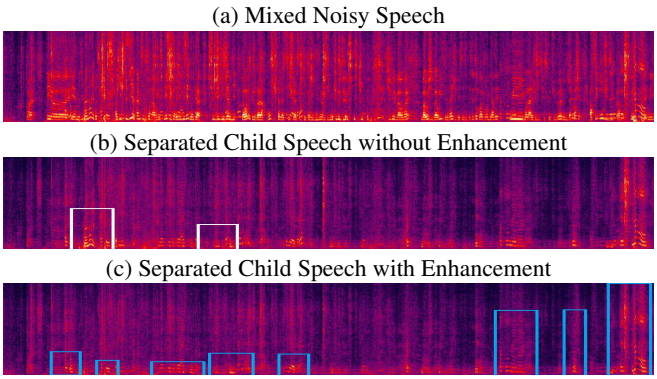

Fig. 5. Spectrograms of an utterance example in the subset of lena_lyon.

Figure 5 shows the spectrograms of an utterance example in the subset of lena_lyon. First of all, by comparing the enhanced separated speech with the origin mixed noisy speech, our separation system can generate the child speech with less speech distortion and suppress the female adult speech even if it's overlapped segment, as shown in the blue rectangles in subfigure(c). Moreover, by comparing the separated speech which is not enhanced with enhanced separated speech, our enhancement model can help the separation system to suppress adult speech because enhancement model can suppress babble noise, which is often generated by adults in daily conversations, as shown in the white rectangles in subfigure(b). So both speech enhancement and separation are important for extracting child speech in realistic environments.

6. CONCLUSION

In this study, we designed a novel joint framework of speech enhancement and speech separation for child speech extraction in realistic conditions, targeting the problem of separating child speech from daily conversations involving background noises, reverberations and overlapping speech. In a preliminary set of experiments, our approach could yield a relatively satisfied performance in child speech extraction even in the quite noisy and challenging realistic conditions.

7. ACKNOWLEDGEMENTS

This work was supported in part by the National Key R&D Program of China under contract No.2017YFB1002202, the National Natural Science Foundation of China under Grant Nos.61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005. This work was also funded by Huawei Noah's Ark Lab. The research reported here was conducted at the 2019 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies, hosted at L'École de Technologie Supérieure (Montreal, Canada) and sponsored by Johns Hopkins University with unrestricted gifts from Amazon, Facebook, Google, and Microsoft.

8. REFERENCES

[1] Marisa Casillas and Alejandrina Cristia, "A step-by-step guide to collecting and analyzing long-format speech environment (lfse) recordings," *Collabra*, vol. 5, no. 1, 2019.

[2] Alejandrina Cristia, Shobhana Ganesh, Marisa Casillas, and Sriram Ganapathy, "Talker diarization in the wild: The case of child-centered daylong audio-recordings," in *Interspeech 2018*, 2018, pp. 2583–2587.

[3] Jill Gilkerson, Kimberly K Coulter, and Jeffrey A Richards, "Transcriptional analyses of the LENA natural language corpus," 2008.

[4] Dongzin Xu, Umit Yapanel, and Sharmi Gray, "Reliability of the LENATM Language Environment Analysis System in young children's natural home environment," 2009, LENA Technical Report LTR-05-2.

[5] Alex Paul, "Thoughts on 15 years and the staying power of LENA," 2019, Blog entry available from `https://www.lena.org/15-years/`.

[6] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[7] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[8] Yanhui Tu, Jun Du, Yong Xu, Lirong Dai, and Chin-Hui Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *The 9th International Symposium on Chinese Spoken Language Processing*. IEEE, 2014, pp. 250–254.

[9] Felix Weninger, John R Hershey, Jonathan Le Roux, and Björn Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2014, pp. 577–581.

[10] Arun Narayanan and DeLiang Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7092–7096.

[11] Xin Wang, Jun Du, Lei Sun, Qing Wang, and Chin-Hui 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.

[12] Tian Gao, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Snr-based progressive learning of deep neural network for speech enhancement.," in *INTERSPEECH*, 2016, pp. 3713–3717.

[13] Elika Bergelson, Anne Warlaumont, Alejandrina Cristia, Marisa Casillas, Celia Rosemberg, Melanie Soderstrom, Florian Metze, Emmanuel Dupoux, Okko Rasanen, Caroline Rowland, and Samantha Durrant, "Starter-aclew," 2017, Accessed: 2018-06-17.

[14] M. Canault, M.-T. Le Normand, S. Foudil, N. Loundon, and H Thai-Van, "Reliability of the language environment analysis system (lenatm) in european french. behavior research methods," 2015.

[15] A. Cristia, G. Yetish, H. Colleran, C. Scaff, J. M. Siegel, and J Stieglitz, "Excerpts from daylong recordings of young children growing up in namibia, vanuatu, and bolivia. homebank," 2019.

[16] Gina M Pretzer, Lukas D Lopez, Eric A Walle, and Anne S Warlaumont, "Infant-adult vocal interaction dynamics depend on infant vocal type, child-directedness of adult speech, and timeframe," *Infant Behavior and Development*, vol. 57, pp. 101325, 2019.

[17] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[18] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[19] Lei Sun, Jun Du, Xueyang Zhang, Tian Gao, Xin Fang, and Chin-Hui Lee, "A progressive multi-target network based speech enhancement model for robust speaker diarization.," in *Submitted to ICASSP, 2020*.

[20] ITU-T Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[21] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[22] Lieve Hamers et al., "Similarity measures in scientometric research: The jaccard index versus salton's cosine formula.," *Information Processing and Management*, vol. 25, no. 3, pp. 315–18, 1989.

[23] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman, "The second dihard diarization challenge: Dataset, task, and baselines," *arXiv preprint arXiv:1906.07839*, 2019.

[24] John Garofalo, David Graff, Doug Paul, and David Pallett, "Csr-i (wsj0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[25] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao 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.

[26] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev 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.

[27] Khaldoun Shobaki, John-Paul Hosom, and Ronald Cole, "C-slu: Kids speech version 1.1," *Linguistic Data Consortium*, 2007.

[28] Dong Yu, Adam Eversole, Mike Seltzer, Kaisheng Yao, Zhiheng Huang, Brian Guenter, Oleksii Kuchaiev, Yu Zhang, Frank Seide, Huaming Wang, et al., "An introduction to computational networks and the computational network toolkit," *Microsoft Technical Report MSR-TR-2014–112*, 2014.