Kathania, Hemant; Kadiri, Sudarsana; Alku, Paavo; Kurimo, Mikko

Using data augmentation and time-scale modification to improve ASR of children's speech in noisy environments

Published in:
Applied Sciences

DOI:
10.3390/app11188420

Published: 01/09/2021

Document Version
Publisher's PDF, also known as Version of record

applied sciences

MDPI

Article

Using Data Augmentation and Time-Scale Modification to Improve ASR of Children's Speech in Noisy Environments

Hemant Kumar Kathania [1,2,*], Sudarsana Reddy Kadiri [1], Paavo Alku [1] and Mikko Kurimo [1]

[1] Department of Signal Processing and Acoustics, Aalto University, Otakaari 3, FI-00076 Espoo, Finland; sudarsana.kadiri@aalto.fi (S.R.K.); paavo.alku@aalto.fi (P.A.); mikko.kurimo@aalto.fi (M.K.)
[2] Department of Electronics and Communication Engineering, National Institute of Technology Sikkim, Ravangla 737139, India
* Correspondence: hemant.ece@nitsikkim.ac.in or hemant.kathania@aalto.fi

Abstract: Current ASR systems show poor performance in recognition of children's speech in noisy environments because recognizers are typically trained with clean adults' speech and therefore there are two mismatches between training and testing phases (i.e., clean speech in training vs. noisy speech in testing and adult speech in training vs. child speech in testing). This article studies methods to tackle the effects of these two mismatches in recognition of noisy children's speech by investigating two techniques: data augmentation and time-scale modification. In the former, clean training data of adult speakers are corrupted with additive noise in order to obtain training data that better correspond to the noisy testing conditions. In the latter, the fundamental frequency (F_0) and speaking rate of children's speech are modified in the testing phase in order to reduce differences in the prosodic characteristics between the testing data of child speakers and the training data of adult speakers. A standard ASR system based on DNN–HMM was built and the effects of data augmentation, F_0 modification, and speaking rate modification on word error rate (WER) were evaluated first separately and then by combining all three techniques. The experiments were conducted using children's speech corrupted with additive noise of four different noise types in four different signal-to-noise (SNR) categories. The results show that the combination of all three techniques yielded the best ASR performance. As an example, the WER value averaged over all four noise types in the SNR category of 5 dB dropped from 32.30% to 12.09% when the baseline system, in which no data augmentation or time-scale modification were used, was replaced with a recognizer that was built using a combination of all three techniques. In summary, in recognizing noisy children's speech with ASR systems trained with clean adult speech, considerable improvements in the recognition performance can be achieved by combining data augmentation based on noise addition in the system training phase and time-scale modification based on modifying F_0 and speaking rate of children's speech in the testing phase.

Keywords: recognition of children's speech; data augmentation; time-scale modification; DNN

Citation: Kathania, H.K.; Kadiri, S.R.; Alku, P.; Kurimo, M. Using Data Augmentation and Time-Scale Modification to Improve ASR of Children's Speech in Noisy Environments. *Appl. Sci.* **2021**, *11*, 8420. https://doi.org/10.3390/app11188420

Academic Editor: Yoshinobu Kajikawa

Received: 8 August 2021
Accepted: 3 September 2021
Published: 10 September 2021

1. Introduction

Automatic speech recognition (ASR) has many potential applications for children in areas such as education (learning new languages and other skills), games, and entertainment. Building ASR systems for child users is, however, challenging for several reasons. First, ASR applications are typically used by children in noisy environments, and the data collection to cover different noise conditions is particularly difficult. Second, due to general problems in recording child speech (e.g., it is difficult to control recording conditions for children, and talkers are not always collaborative), it is difficult to collect enough training data to build ASR systems for children. Therefore, the performance of ASR systems in recognition of children's speech degrades due to the mismatch caused by training and testing under different noise conditions and due to the mismatch caused by training the system with adults' speech and testing with children's speech.

While the majority of publicly available ASR systems work effectively for adults' speech in noise-free environments, their performance degrades considerably when used in noisy environments and particularly when recognizing children's speech in noisy environments [1,2]. Classrooms are an example of environments where children are subject to noise exposure [3–6] and where ASR technology is increasingly used in education. The degradation of ASR systems in recognition of noisy children's speech depends on many issues such as noise type, signal-to-noise ratio (SNR), acoustic and linguistic differences in fundamental frequency (F_0), speaking rate, and formant frequencies between adult and child speech [7–16]. Decreased ASR performance of children's speech compared to adults' speech is also explained by the fact that the number of publicly available training data for children's speech (tens of hours) [17,18] is much smaller compared to that of adults' speech (thousands of hours) [19,20]. Overall, new research is needed to develop ASR systems capable of recognizing children's speech in noisy environments. In order to develop such a system, the following two techniques are taken advantage of in the current study: (1) using data augmentation based on noise addition to tackle the mismatch induced by having different noise conditions in training and testing and (2) using time-scale modification to tackle the mismatch induced by having adults' speech in training and children's speech in testing.

There are a few existing ASR studies that have addressed both the mismatch induced by having clean speech in training and noisy speech in testing as well as by having adult speech in training vs. child speech in testing. These research questions were studied, for example, in [21] by investigating the use of spectral moments and spectral smoothing-based features, and the results showed that the spectral moment time-frequency distribution augmented by low-order cepstral (SMAC) features were found to improve the recognition performance. In [22], variational mode decomposition (VMD)-based spectral smoothing was found to improve ASR for children's speech in noise. In [23–26], the feature-space maximum likelihood linear regression (fMLLR) transform, deep convolutional neural networks (CNNs), graph-based feature filtering, and Bayes methods were investigated to address the problems caused by different channel and noise conditions. The effect of feature learning using a CNN-based end-to-end acoustic modeling approach was studied in [27], and the method was shown to give a reduction in WER. Principal component analysis and heteroscedastic linear discriminant analysis based on a low-rank feature projection were explored in [28]. In [29], prosodic features including loudness, intensity, and voice probability were investigated, and it was found that combining the prosodic features with mel-frequency cepstral coefficients (MFCCs) improved ASR performance. The effect of the filter bank on ASR of children's speech was studied in [30], and it was found that the linear-frequency filter bank was better compared to the mel and inverse-mel filter banks.

The effects of acoustic and linguistic differences between adults' and children's speech have been investigated widely [22,29,31,32], and these differences have been observed to degrade ASR performance considerably. Modifying prosodic features (F_0 and speaking rate) was found to reduce the effect of mismatch induced by having adults' (clean) speech in training and children's (clean) speech in testing in [33–35]. The effect of modifying formants in ASR of children's speech in clean and noisy conditions was explored recently in [36], and the results showed a reduction in WER when formants of children's speech were modified towards those in adults' speech. The performance of the formant modification method proposed in [36] is, however, limited due to the use of all-pole spectral modeling methods whose accuracy deteriorates in low SNR levels below 5 dB. In [37], generative adversarial network (GAN)-based data augmentation was explored, and an improvement in WER was reported. A data augmentation strategy based on modifying prosody (F_0 and speaking rate) by changing glottal closure instants was studied in [38]. In [18], data augmentation using stochastic feature mapping (SFM) to transform out-of-domain adult data was found to improve recognition of children's speech.

This study investigates ASR of children's speech by focusing on two challenges: (1) recognition of speech in noisy conditions, which is a typical scenario for child users, and

(2) recognition of children's speech using an ASR system trained with adult speech due to the lack of training data from child speakers. In order to address these two challenges, the study combines data augmentation and time-scale modification. In the former, a straightforward data augmentation method is used by corrupting the training data of the ASR system with additive noise to obtain new speech data that correspond better with the testing data in noisy conditions. In the latter, two modification methods (F_0 modification and speaking rate modification) are used to modify the prosodic characteristics of the children's speech in the testing phase towards the prosodic characteristics of the adult speech that is used in the system training phase. F_0 and speaking rate were selected as prosodic features to be modified because their modification is easy to implement and because their modification has shown promising results in previous studies [33,34,39]. In addition, modification of these factors can be done in a more robust manner from noisy speech compared to factors such as formants, whose estimation deteriorates in noisy environments. Figure 1 shows a flow diagram demonstrating how data augmentation and time-scale modification are used in the current investigation. Note that it would also be possible to modify the prosodic structure of adult speech in the training phase using time-scale modification, but this should be done separately for each noise condition and is therefore not feasible.

Data augmentation and time-scale modification have been investigated separately in previous ASR studies (e.g., [40–42] for the former and [33,34,39,43] for the latter). However, the effect of *combining* these techniques has not been investigated in recognition of noisy children's speech before. Therefore, the main contribution of the current study is to investigate how the performance of a children's speech ASR system that suffers from the two challenges described in the beginning of this section is affected when using data augmentation, F_0 modification, and speaking rate modification either separately or by combining these techniques one by one. The study shows encouraging results, indicating that while none of the three previously studied methods alone gives an adequate improvement in the recognition performance, the combination of the three approaches as implemented in the current study results in a considerable improvement in recognition of children's speech in noisy conditions.

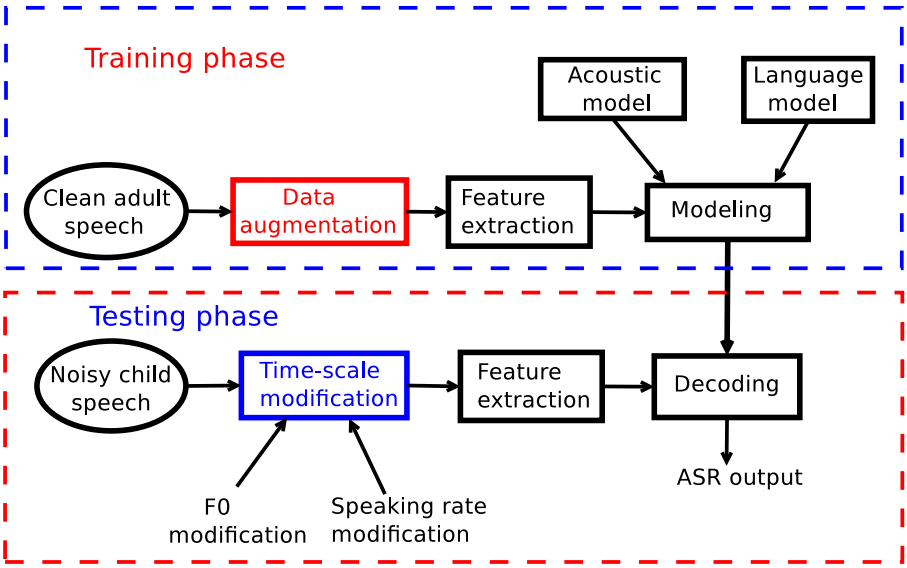

Figure 1. A flow diagram describing the ASR scenario studied in the current article to recognize noisy children's speech using training based on clean adults' speech. The two techniques studied, data augmentation and time-scale modification, are marked with red and blue color, respectively.

The remainder of the paper is organized as follows. The two main techniques studied, data augmentation and time-scale modification, are first described in Sections 2 and 3, respectively. Section 4 describes the speech databases and the ASR system used in the study.

The results of the ASR experiments are reported in Section 5 by describing in separate sub-sections how data augmentation, F_0 modification, speaking rate modification, and finally the combination of the three affect the recognition performance. The results are discussed in Section 6, and the conclusions of the study are drawn in Section 7.

The list of abbreviations used in this study are given in Table 1.

Table 1. List of abbreviations.

ASR	automatic speech recognition
TSM	time-scale modification
DA	data augmentation
F_0M	F_0 modification
SRM	speaking rate modification
RTISI-LA	real-time iterative spectrogram inversion with look-ahead
LP	linear prediction
WER	word error rate
DNN	deep neural network
TDNN	time delay neural network
VTLN	vocal tract length modification
SRA	speaking rate adaptation
MFCCs	mel-frequency cepstral coefficients
GMM	Gaussian mixture model
HMM	hidden Markov model
fMLLR	feature-space maximum likelihood linear regression
LM	language model
LDA	linear discriminant analysis
MLLT	maximum likelihood linear transform
SAT	speaker adaptive training
SNR	signal-to-noise ratio
VMD	variational mode decomposition
SMAC	spectral moment time-frequency distribution augmented by low-order cepstral
CNN	convolutional neural network
SFM	stochastic feature mapping
GAN	generative adversarial network
STFTM	short-time Fourier transform magnitude

2. Data Augmentation

In this study, we used noise addition as the data augmentation strategy. The main motivation for this type of data augmentation is to capture more acoustic variability of the data to improve the ASR system performance in noisy environments. The proposed approach is demonstrated in the block diagram shown in Figure 2. The input to the data augmentation procedure is clean adult speech taken from an existing large database, WSJCAM0 [44], which will be described in Section 4. Augmentation is conducted by corrupting the clean input signal with additive noise by varying the SNR from 0 dB to 15 dB with a step size of 5 dB and by using four different types of noise (babble, white,

factory, and volvo) extracted from the NOISEX-92 database [45]. It should be noted that the data augmentation approach is used in this study to generate new, noise-corrupted data for training, but the original clean adult speech taken from the WSJCAM0 database is not included in the generated training data.

Data augmentation is performed in the ASR experiments of the current study using three scenarios: the "same" scenario, the "different" scenario, and the "all" scenario. The "same" scenario refers to testing the ASR system in circumstances where the test speech is corrupted by one type of noise (e.g., babble), and system training is based on using the same noise type in data augmentation. The "different" scenario refers to corrupting the test speech with one noise type and using the other three types in augmenting the training data (e.g., babble noise in testing, and factory, volvo as well as white noise in augmentation). The "all" scenario refers to testing with speech corrupted by one noise type (e.g., white) but using all four noise types in data augmentation.

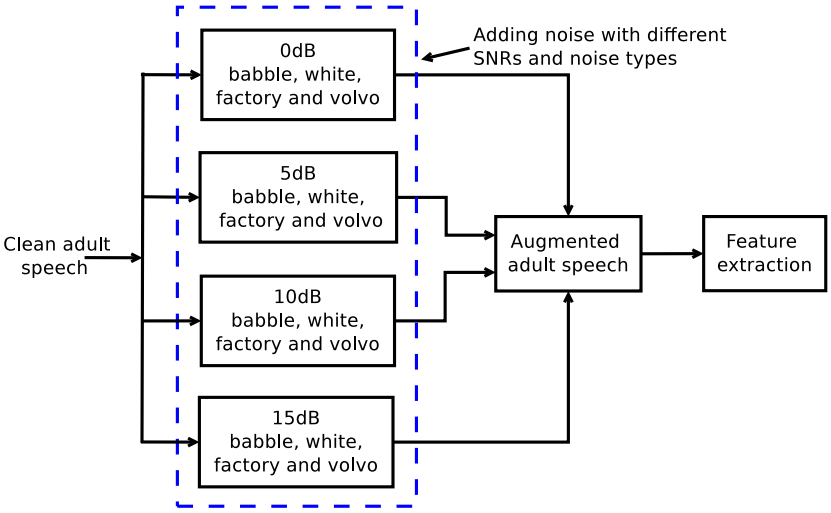

Figure 2. A block diagram of the data augmentation method.

3. Time-Scale Modification

In order to address the second challenge of ASR for children's speech described in Section 1—the mismatch caused by the lack of adequate training data from child speakers—the present study investigates modifying the time-scale structure of children's speech in the system testing phase. The goal of the time-scale modification is to make the testing data of child speakers more similar to the training data of adult speakers by modifying two prosodic features of speech, F_0 and speaking rate. Both the F_0 modification and the speaking rate modification were conducted using the same method, the real-time iterative spectrogram inversion with look-ahead (RTISI-LA) algorithm [35,46,47]. RTISI-LA was originally developed as a method for estimating time-domain signals from overlapping magnitude spectra that have been computed frame by frame. RTISI-LA is a similar kind of phase recovery technique as the Griffin-Lim algorithm [46,48]. RTISI-LA is, however, much faster than the Griffin-Lim algorithm and therefore justified to be used in applications like the current study, where large numbers of speech data need to be processed. The RTISI-LA method consists of the following steps.

- The speech signal is processed in frames of L samples by computing the short-time Fourier transform magnitude (STFTM) spectrum using the FFT with the Hamming window. The frame shift (S) is selected as $S = L/4$ so that each frame overlaps with three previous and three following frames. In the following, the frame index is denoted by m and the window function by $w(n)$.
- To reconstruct the speech signal from its STFTM, an iterative frame-by-frame signal estimation process is applied. Let us suppose that the first $m - 1$ frames of the speech

signal have already been reconstructed from STFTM, and let us denote this signal by $x_{m-1}(n)$. The task is to synthesize $x_m(n)$.

- In order to estimate the mth frame, a partial analysis frame is created using overlap-adding (OLA) for the $(m-1)$th, $(m-2)$th, and $(m-3)$th frame of $x(n)$ considering an overlap of 75%. The fourth quarter of this partially filled frame is filled with zeros. Let the partial frame be denoted by $x_{m-1}(n)w(n-mS)$. In RTISI-LA, the future k frames influence the reconstruction of the mth frame. After the mth frame is generated, it is kept uncommitted until the $(m+k)$th frame is generated.
- Next, the Fourier transform of the partial frame is computed using a scaled Hamming window.
- The phase information computed from the Fourier transform of the partial frame is then combined with the STFTM for the mth frame.
- The inverse Fourier transform of the derived frequency-domain signal produces a new estimate for the mth frame. In each iteration, the estimation of $x(n)$ is updated.

As described in [46], the RTISI-LA algorithm can be used to modify both the F_0 and speaking rate of speech signals. By using the notations of [46], these two modifications were conducted in the current study as follows by using child speech as input.

F_0 Modification The F_0 modification was computed by first re-sampling the input speech signal in the time domain. To modify the F_0 of the input speech signal downwards by factor q, where $0 < q < 1$, the input frame of L' samples was re-sampled to obtain a longer frame of L samples (i.e., $L' = qL$), which was then used in the STFTM computation. We used simple linear interpolation in re-sampling because it is computationally inexpensive and was reported in [46] to provide reasonable sound quality. The value of L was fixed to 160 samples. The STFMs of the overlapping frames (i.e., the magnitude spectrogram) were then processed using the RTISI-LA algorithm (steps 1–6 above) to obtain the F_0-modified time domain output signal. The value of the factor q was selected by conducting ASR experiments and by searching for the value of q, which yielded the lowest WER. More details about this will be given in Section 5.2.

Speaking rate modification The modification of the speaking rate with RTISI-LA is based on the frequency-domain approach proposed in [46]. In this approach, the STFTM of the input signal is computed in the analysis stage using a frame shift of S_a, and the signal is transformed to the time domain in the synthesis stage with RTISI-LA using a different value of the frame shift (denoted by S_s). The amount of modification is defined by the factor α, which is defined such that $S_a = S_s/\alpha$. By using $0 < \alpha < 1.0$, the speaking rate of speech can be increased. In our experiments, the speaking rate modification was conducted using the frame size of $L = 256$ samples and by fixing $S_s = L/4$. The value of α was selected by searching for the value that yielded the lowest WER, as will be explained in Section 5.3.

4. Speech Databases and the ASR System

The ASR experiments of the study were conducted using two openly available speech databases. The adults' speech data that were used in training was taken from the WSJCAM0 British English speech corpus [44]. The children's speech data that were used for testing were taken from the PF-STAR British English speech corpus [49]. Detailed information about the two databases are given in Table 2. The children's speech data were split into two parts: the validation set and the test set. The former consisted of 2.5 h of speech from 62 speakers with an age range of 6–14 years. The latter consisted of 1.1 h of speech from 60 speakers with an age range of 4–13 years.

Table 2. Information on the two speech databases used in the current study.

Database	WSJCAM0	PF-STAR
Language	British English	British English
Use	training	testing
Speaker type	adult	child
No. of speakers (males / females)	92 (53/39)	60 (32/28)
Age	>18 years	4–13 years
No. of words	132,778	5067
Duration (h)	15.5	1.1

To build ASR systems, the Kaldi toolkit [50] was used. Deep neural network (DNN)-based context-dependent hidden Markov models (HMM) were used for acoustic modeling of the cross-word tri-phones. Decision-tree-based state tying was performed with the maximum number of tied-states (senones) fixed at 2000. Prior to learning the parameters of the DNN–HMM-based ASR system, the fMLLR-normalized feature vectors were time-spliced considering a context size of 9 frames. The number of hidden layers in the DNN was set to 5, with 1024 hidden units in each layer. The non-linearity in the hidden layers was modeled using the tanh function. The initial learning rate was set to 0.005, which was reduced to 0.0005 in 15 epochs for training the DNN–HMM system. The minibatch size of 512 was used in the DNN training. In decoding the test set for children's speech, a 1.5 k domain-specific bigram language model (LM) was used. This bigram LM was trained on the transcripts of the speech data in PF-STAR after excluding the test set. In total, 1969 words were used, including pronunciation variations in the lexicon for decoding the children's test set.

5. Results

As the first step in the series of experiments, we trained a baseline ASR system using original clean adult speech from the WSJCAM0 database and tested the system with original clean children's speech from the PF-STAR database (i.e., no data augmentation or time-scale modification was used). As expected, this system gave a poor performance (WER = 19.58%) due to the mismatch induced by having adults' speech in training and children's speech in testing. We then also included the other mismatch type discussed in the introduction by testing the same system with noise-corrupted children's speech. The results indicated, as expected, that the system performance deteriorated severely: in the noise condition with SNR = 5 dB, for example, the WER value rose to 82.67%, 87.40%, 92.32%, and 46.12% for babble, white, factory, and volvo noise, respectively. As also reported in previous studies [22,51], these experiments indicated that there is lots of room for improvement in recognition of children's speech in noisy conditions. In the following sub-sections, we report on the results from the experiments, which were conducted to improve the system performance step by step by first using data augmentation, then time-scale modification based on F_0 modification, then time-scale modification based on speaking rate modification, and finally all of these three methods combined.

5.1. Results Obtained by Using Data Augmentation

The effect of augmenting the training data (of adult speech) with the method described in Section 2 was studied by testing the ASR system with noise-corrupted children's speech. To generate noisy test data, the children's speech signals of the PF-STAR database were corrupted with additive noise using four noise types (babble, white, factory, volvo) and four SNR categories (between 0 dB and 15 dB). For each individual noise condition, we built a baseline ASR system, which was trained using solely the noise-corrupted adult

speech of the corresponding condition (i.e., no data augmentation was used in the training of the baseline systems). The results of the ASR experiments are reported in Table 3.

From Table 3, the following observations can be made. (1) The data augmentation based on the "same" condition improved the ASR performance compared to the baseline and the condition "different" worsened the performance in all noise types and SNR categories. (2) The data augmentation based on the "all" condition improved the performance compared to the baseline for babble and factory noise but decreased the performance for white and volvo noise. (3) Despite the fact that the data augmentation improved the performance compared to the baseline systems in many of the scenarios studied, overall the WER values were still unacceptably high.

Table 3. WERs obtained using data augmentation in recognizing noisy children's speech of four noise types in four SNR categories. Results obtained with the baseline system, which was trained without data augmentation using solely the noise-corrupted speech of the corresponding condition, are shown in the third column. WERs obtained using the three data augmentation scenarios described in Section 2 are shown in the fourth, fifth, and sixth column.

Noise Type	SNR (dB)	WER (%)			
		Baseline	Data Augmentation Scenario		
			Same	Different	All
Babble	0 dB	53.35	47.05	76.35	50.90
	5 dB	36.43	34.37	55.42	33.68
	10 dB	30.29	29.16	43.10	25.75
	15 dB	26.46	25.18	35.76	23.64
White	0 dB	43.22	40.18	54.71	46.48
	5 dB	30.15	28.59	42.90	37.35
	10 dB	26.11	24.86	36.05	31.83
	15 dB	24.30	23.04	33.66	30.77
Factory	0 dB	65.51	54.75	81.05	58.46
	5 dB	42.47	38.01	59.23	34.59
	10 dB	30.90	29.03	42.67	25.09
	15 dB	26.30	25.18	33.23	20.65
Volvo	0 dB	21.09	18.95	38.85	23.48
	5 dB	20.20	18.44	36.62	22.59
	10 dB	19.53	18.32	35.60	21.82
	15 dB	18.92	18.11	32.36	21.54

5.2. Results Obtained by Using F_0 Modification

Compared to adult speakers, children typically use higher F_0 values and they also vary the F_0 of their speech over a larger F_0 range [52,53]. This phenomenon is show in Figure 3, which shows F_0 histograms computed from adults' speech signals taken from the WSJCAM0 database (panel (a)) and from children's speech signals taken from the PFSTAR database (panel (b), gray color). For this analysis, 100,000 frames of speech were collected from adults' and children's speech. The figure shows that for the adults' speech signals, there are two peaks close to 100 Hz and 200 Hz corresponding to the average F_0 of male and female speakers, respectively. The peak in the children's F_0 histogram is higher (around 250 Hz) but the histogram also shows a clearly larger spread of the F_0 values compared to the adults' F_0 values. When the children's speech signals were processed with

the RTISI-LA method described in Section 3 using $q = 0.80$, the F_0 histogram shown in orange in panel (b) was obtained. By comparing panels (a) and (b), it can be clearly seen that RTISI-LA succeeded in converting the original F_0 histogram of children's speech much closer to that of adults' speech.

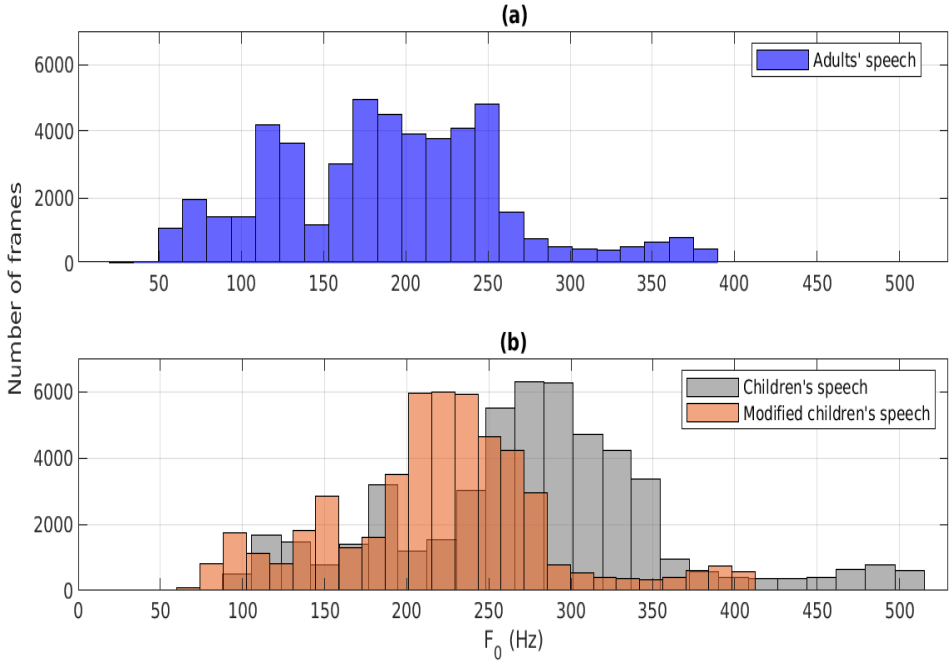

Figure 3. F_0 histograms for (**a**) adults' speech, (**b**) children's speech and modified children's speech processed by the F_0 modification method.

To evaluate the effect of F_0 modification on the ASR performance, the RTISI-LA method was used to modify all the validation data of children's speech that were noise-corrupted with four different noises with varying SNR values. In these tests, we used the same value for q for all the data in all the noise conditions. This q value was determined by varying the parameter between 0.95 and 0.75 (in steps of 0.05) for the different noise types and SNR categories and by searching for the q value that yielded the lowest WER value in each case. This optimization method was selected because it is straightforward and has low computational cost. These experiments, demonstrated in Figure 4, indicated that the lowest WER value was achieved in most cases with $q = 0.80$. Therefore, we fixed the F_0 modification parameter to $q = 0.80$ and processed the noisy children's test data using this parameter value with the RTISI-LA method. The WER values of these experiments are reported in Table 4. It can be seen that the F_0 modification improved the recognition performance in all the noise types and SNR categories studied.

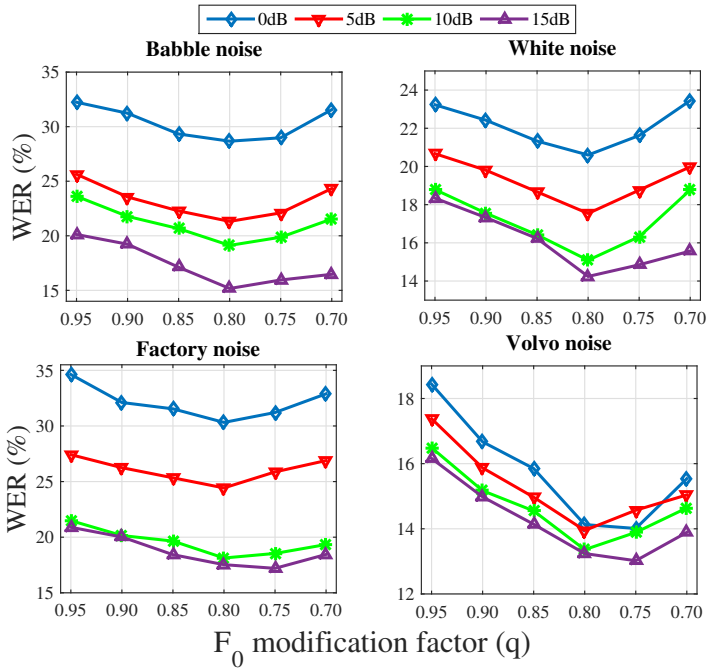

Figure 4. WERs as a function of the F_0 modification factor q for each of the four noise types and four SNR categories using the validation set. Note that the smaller the value of q (i.e., towards the right end of the x-axis), the larger the strength of the F_0 modification.

Table 4. WERs obtained in recognizing noisy children's speech of four noise types in four SNR categories using F_0 modification. The baseline system is the same as in Table 3.

		WER (%)	
Noise Type	**SNR (dB)**	**Baseline**	**With F_0 Modification**
Babble	0 dB	53.35	38.36
	5 dB	36.43	23.69
	10 dB	30.29	18.28
	15 dB	26.46	16.25
White	0 dB	43.22	33..85
	5 dB	30.15	20.34
	10 dB	26.11	17.77
	15 dB	24.30	16.38
Factory	0 dB	65.51	47.50
	5 dB	42.47	31.59
	10 dB	30.90	17.89
	15 dB	26.30	15.87
Volvo	0 dB	21.09	13.83
	5 dB	20.20	13.16
	10 dB	19.53	12.82
	15 dB	18.92	12.61

5.3. Results Obtained by Using Speaking Rate Modification

In addition to F_0, speaking rate is another major factor for the mismatch between children's and adults' speech. Children typically speak at slower rates compared to

adults [35,52,53]. The effect of the speaking rate modification algorithm described in Section 3 is demonstrated in Figure 5, which shows the histograms of the number of phonemes per second. For this analysis, 500 utterances of adult speech and 500 utterances of child speech were taken from the WSJCAMO and PFSTAR databases, respectively, and the speaking rate of the latter was modified with the RTISI–LA algorithm described in Section 3 by using $\alpha = 0.74$. The difference in the histograms between the adults and children can be noted easily in Figure 5: the children's utterances show clearly smaller numbers of phonemes per time unit compared to the adults' utterances, that is, the children speak at a lower rate. The lower panel of Figure 5 shows the effect of the RTISI–LA algorithm when the speaking rate of the children was modified. By comparing the histograms shown in this panel with the histogram shown in the upper panel of Figure 5, it can be observed that the histogram of the rate-modified child utterances has become closer to that of the adults' utterances compared to the histogram computed from the original children's utterances.

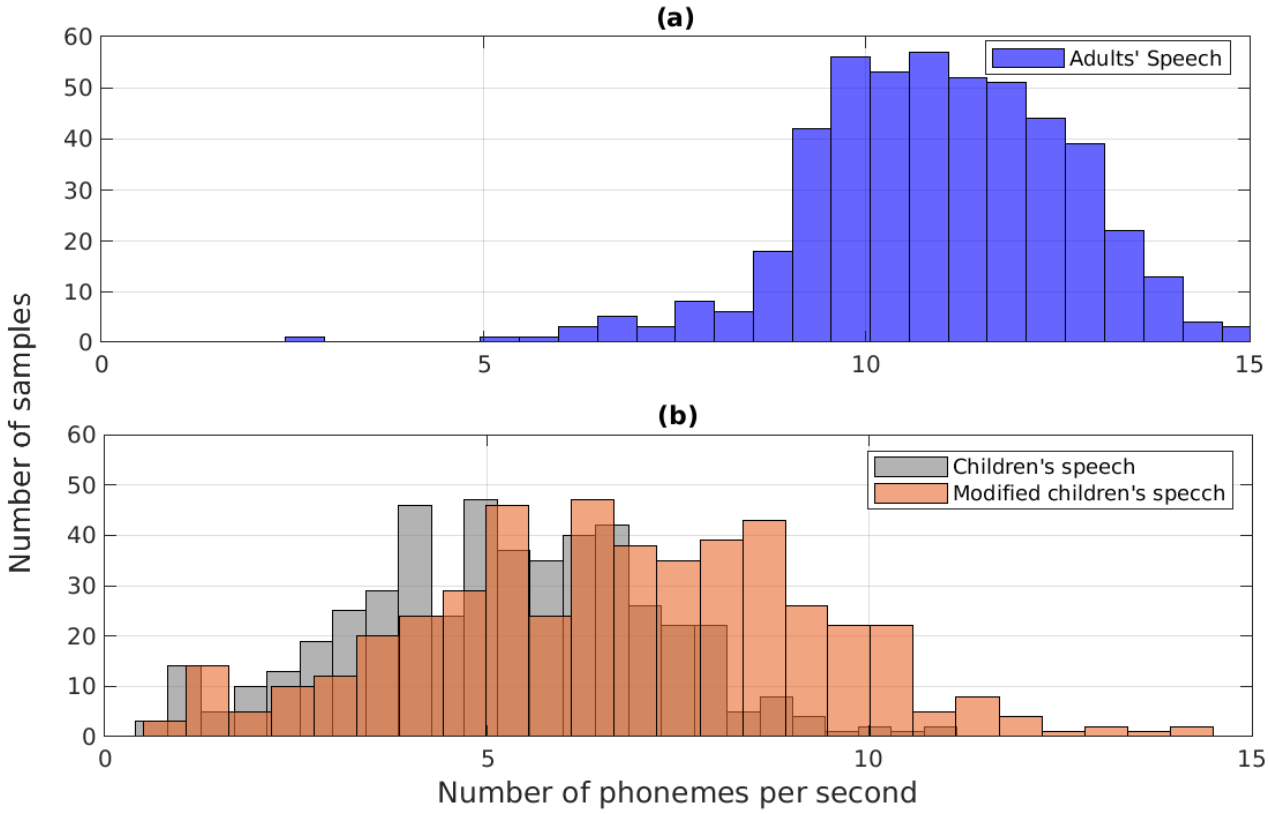

Figure 5. Number of phonemes per second as histograms for (**a**) adults' speech, (**b**) children's speech and modified children's speech processed by the speaking rate modification method.

To evaluate the effect of the speaking rate modification on the ASR performance, the RTISI-LA algorithm was applied to the noise-corrupted children's speech of the entire validation set. We first varied the speaking rate modification factor α from 0.83 and 0.68 (in steps of 0.03) in the same noise conditions and evaluated the corresponding WER value for all the data. As demonstrated in Figure 6, the lowest WER value was obtained using $\alpha = 0.74$ for most of the noise types and SNR categories. Therefore, we fixed $\alpha = 0.74$ and modified the speaking rate of the children's speech test data in all noise conditions. The results of these ASR experiments are shown in Table 5. From Table 5, it can be noted that the speaking rate modification algorithm improved the system performance over the baseline system for all the noises of varying SNR levels.

Table 5. WERs obtained in recognizing noisy children's speech of four noise types in four SNR categories using speaking rate modification. The baseline system is the same as in Table 3.

		WER (%)	
Noise Type	**SNR (dB)**	**Baseline**	**With Speaking Rate Modification**
Babble	0 dB	53.35	40.32
	5 dB	36.43	28.64
	10 dB	30.29	25.73
	15 dB	26.46	23.32
White	0 dB	43.22	38.65
	5 dB	30.15	26.53
	10 dB	26.11	23.12
	15 dB	24.30	22.76
Factory	0 dB	65.51	51.34
	5 dB	42.47	26.37
	10 dB	30.90	25.65
	15 dB	26.30	24.87
Volvo	0 dB	21.09	18.65
	5 dB	20.20	17.43
	10 dB	19.53	17.17
	15 dB	18.92	16.34

5.4. Results Obtained by Using the Combined System

As the last step, we evaluated the recognition of noisy children's speech by combining the data augmentation scene described in Section 2 to the two time-scale modification methods described in Section 3. The evaluation was computed in a similar manner to that in Section 5.1 by using the four noise types and the four SNR categories and by including the three different data augmentation scenarios ("same", "different", and "all"). In each of the three scenarios, the recognizer trained in the corresponding data augmentation scenario was tested with noisy children's speech, which was time-scale-processed either using F_0 modification or speaking rate modification and using both of them. The WER results obtained are reported in Table 6 by referring to data augmentation, F_0 modification, and speaking rate modification by DA, F_0M, and SRM, respectively. From this table, the following main observations can be made. First, the system that combined all the three studied components (i.e., DA + F_0M + SRM) performed considerably better than the baseline in all the cases studied. In addition, combining *both* of the two time-scale modification methods with DA yielded the best combined system in all cases. Second, by comparing the best combined systems (i.e., DA + F_0M + SRM) between the three noise augmentation scenarios, it can be seen that the WER values were best in the "all" scenario for all the noise types and SNR categories.

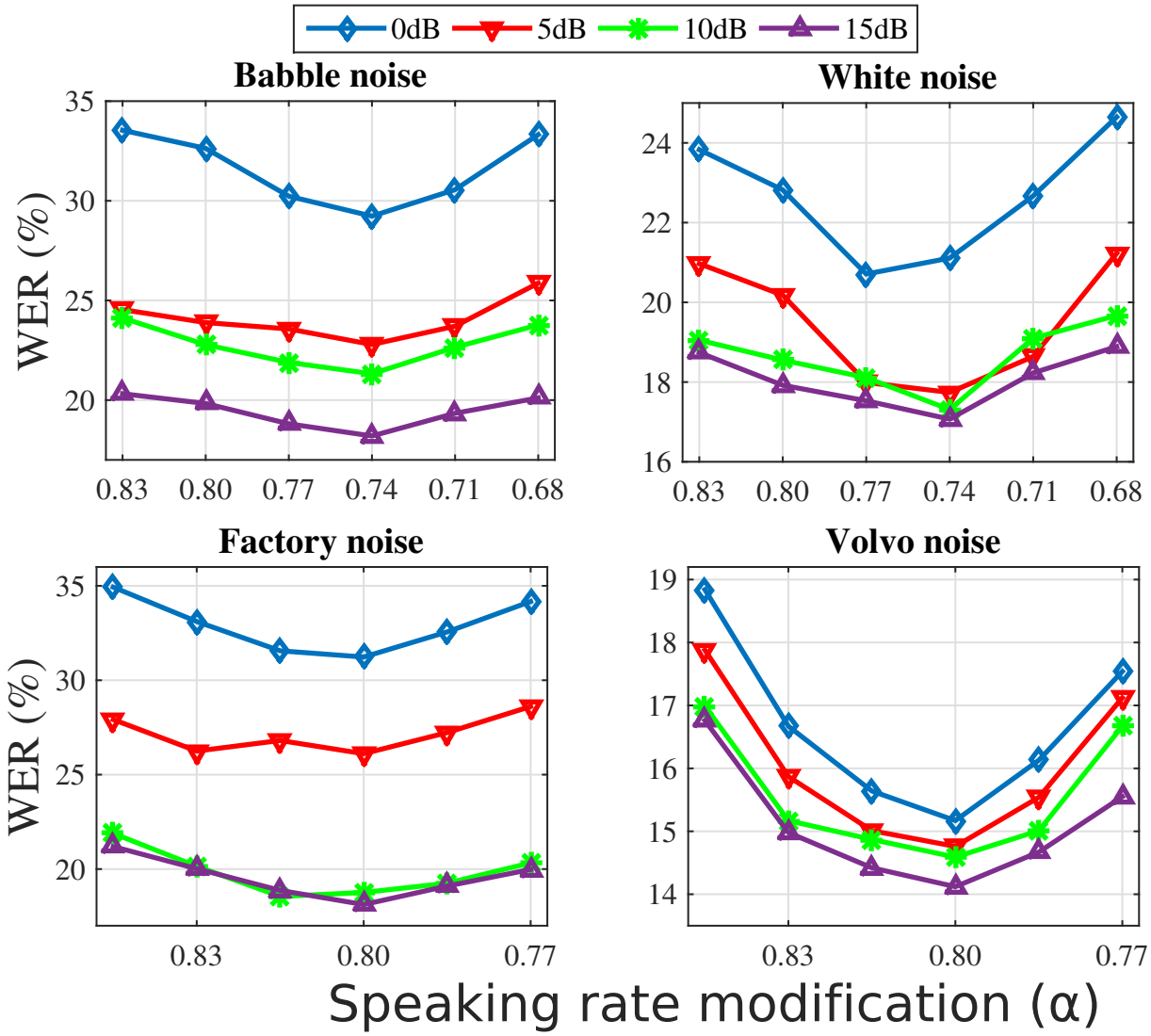

Figure 6. WERs as a function of the speaking rate modification factor α for each of the four noise types and four SNR categories using the validation set. Note that the smaller the value of α (i.e., towards the right end of the x-axis), the larger the strength of the speaking rate modification.

Table 6. WERs obtained in recognizing noisy children's speech of four noise types in four SNR categories using different combinations of data augmentation (DA), F_0 modification (F_0M), and speaking rate modification (SRM). The baseline system is the same as in Table 3. WERs are reported separately for each of the three data augmentation scenarios described in Section 2.

Noise Type	SNR (dB)	WER (%)									
		Baseline	Combined System								
			Same			Different			All		
			DA + F_0M	DA + SRM	DA + F_0M + SRM	DA + F_0M	DA + SRM	DA + F_0M + SRM	DA + F_0M	DA + SRM	DA + F_0M + SRM
Babble	0 dB	53.35	28.47	37.82	25.21	47.36	66.95	38.34	26.59	40.73	21.15
	5 dB	36.43	19.57	26.97	17.83	29.44	31.14	21.52	16.25	25.11	12.85
	10 dB	30.29	16.80	23.26	15.44	23.32	28.14	15.54	13.45	17.33	10.29
	15 dB	26.46	14.81	21.62	13.76	18.56	24.66	13.84	12.66	15.74	9.65
White	0 dB	43.22	32.54	34.27	28.69	29.60	45.92	23.53	31.22	34.83	17.95
	5 dB	30.15	19.55	23.67	18.80	17.87	32.77	14.12	23.71	23.79	12.34
	10 dB	26.11	15.10	19.98	14.99	14.12	26.63	10.94	15.66	20.02	10.54
	15 dB	24.30	13.95	19.27	13.70	13.94	21.80	10.56	13.41	18.56	10.27
Factory	0 dB	65.51	35.64	46.28	33.25	69.65	75.81	60.81	34.33	46.14	26.59
	5 dB	42.47	22.49	29.18	20.75	42.17	49.73	32.30	18.80	22.65	13.94
	10 dB	30.90	16.84	23.32	16.13	27.13	30.42	16.13	13.07	15.76	10.07
	15 dB	26.30	14.99	22.90	14.37	19.57	23.47	14.20	12.20	13.39	8.92
Volvo	0 dB	21.09	13.11	16.59	11.87	15.52	22.11	11.92	12.24	14.92	9.38
	5 dB	20.20	12.62	15.08	11.47	15.30	21.87	11.33	12.18	14.67	9.24
	10 dB	19.53	12.28	15.00	11.45	15.12	21.32	11.12	12.04	14.32	9.18
	15 dB	18.92	12.10	14.89	11.43	14.89	20.96	10.94	11.86	14.13	8.76

6. Discussion

Achieving high accuracy in recognition of children's speech is difficult using state-of-the-art ASR systems because of two types of mismatch between the system training and testing. First, children typically use ASR applications in noisy environments such as when playing games and when taking part in education with other children. Therefore, when the system is trained with clean speech, there is mismatch between the testing and training stages. Second, due to practical problems in recording young child speakers, few training data are available from child speakers, and current ASR systems are mostly trained using adults' speech only. Therefore, when these ASR systems are used to recognize children' speech, another mismatch will be brought about between the system training and testing stages. The severity of these two mismatches was first demonstrated in the current study using a standard ASR system: a poor WER value (of about 20%) was obtained in recognizing clean children's speech using the system trained using adults' speech. Furthermore, when the children's speech was contaminated with noise in order to simulate the use of children's ASR in realistic environments, the recognition performance deteriorated severely to WER values larger than 80% in some noise conditions.

In order to tackle the effects caused by the two mismatches described above in recognition of children's speech in noisy conditions, the current study investigated the utilization of data augmentation and time-scale modification. Furthermore, the time-scale modification technique consisted of two parts, modification of F_0 and modification of speaking rate, which were used to convert the prosodic structure of the children's speech test data to become closer to that of the adults' speech used in the system training. The experiments of the study were planned in order to first investigate how the recognition performance was affected when each of the modification techniques was utilized alone in building the recognizer, after which all the studied techniques were combined aiming at the best system. The experiments were conducted using an existing deep neural network (DNN)-based recognizer. Although CNNs are currently increasingly used in ASR, they call for larger numbers of training data compared to DNN-based systems. Therefore, choosing a DNN-based architecture was justified for the current investigation, and we leave the verification of the studied approach with big and complex systems as future work. The data augmentation involved corrupting the original clean training data of adults' speech using additive noise of different types and SNR categories. Three augmentation scenarios ("same", "different", "all") were generated, and these scenarios differ in the way the noise type in testing is seen by the data augmentation procedure. The recognition experiments indicated that data augmentation yielded a consistent improvement in WER only in the case when the noise type was the same in the augmentation and testing. However, when the noise type in testing was different from that used in data augmentation, the performance compared to the baseline decreased considerably, and this happened in all the noise conditions studied. Hence, the utilization of the straightforward data augmentation approach based on noise-corrupting the adult speech in the system training stage did not give an adequate improvement in recognition of noisy children's speech. As the next steps, ASR experiments were conducted by time-scale modifying the prosodic structure of the children's speech in the test stage. The experiments showed that both F_0 modification and speaking rate modification improved WER values compared to the baseline system and that this happened for all the noise conditions studied. From the two modification methods, F_0 modification yielded smaller WER values in all noise conditions. As the final step in our experiments, we combined data augmentation with F_0 modification, with speaking rate modification, and with both of them. The results indicated that combining data augmentation with both of the time-scale modification methods yielded the lowest WER in all noise conditions studied. For this combination, the WER values obtained in the three data augmentation scenarios were lowest in "all", second lowest in "same", and highest in "different".

7. Conclusions

Poor accuracy is obtained by DNN-HMM -based standard ASR systems in recognition of children's speech in noisy conditions due to two mismatches between the system training and testing. The study indicated that an effective way to tackle the effects of these mismatches is to combine data augmentation (adding different types of noise to adults' speech in the training phase) to the modification of both the F_0 and speaking rate structure of children's test speech to make the children's speech in testing become closer to the adults' speech in training. The study showed that compared to the baseline ASR system, the recognizer, which combined the data augmentation and time-scale modification, yielded substantial improvements in WER in all the noise conditions studied. For example, for the most severe noise type (factory noise), the WER values obtained using the baseline system were very poor in all SNR categories (the average WER = 41.29%), but the WER values obtained by the combined system in the "all" scenario dropped to clearly lower levels (the average WER = 14.88%). For the least severe noise type (volvo noise), the corresponding improvement in the averaged WER given by the combined system was from 19.93% to 9.14%. Despite the fact that the study showed promising results in recognition of noisy children's speech, new research is needed to understand, for example, how the reported results are affected when the amount of adult speech is increased in training. Moreover, studying how the proposed ASR system functions in the presence of adversarial attacks is another topic of future investigations.

Author Contributions: Conceptualization, H.K.K. and S.R.K.; methodology, H.K.K. and S.R.K.; software, H.K.K.; validation, H.K.K., S.R.K., P.A. and M.K.; formal analysis, H.K.K. and S.R.K.; investigation, H.K.K. and S.R.K.; resources, H.K.K.; data curation, H.K.K.; writing—original draft preparation, H.K.K. and S.R.K.; writing—review and editing, H.K.K., S.R.K., P.A. and M.K.; visualization, H.K.K.; supervision, P.A. and M.K.; project administration, P.A. and M.K.; funding acquisition, P.A. and M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Academy of Finland grant numbers 329267 and 330139.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: http://www.thespeechark.com/pf-star-page.html and https://catalog.ldc.upenn.edu/LDC95S24 (accessed on 10 June 2021).

Acknowledgments: The computational resources were provided by Aalto ScienceIT.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schalkwyk, J.; Beeferman, D.; Beaufays, F.; Byrne, B.; Chelba, C.; Cohen, M.; Kamvar, M.; Strope, B. Your Word is my Command: Google Search by Voice: A Case Study. In *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*; Springer: Boston, MA, USA, 2010; Chapter 4, pp. 61–90.
2. Li, J.; Deng, L.; Gong, Y.; Haeb-Umbach, R. An Overview of Noise-Robust Automatic Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 745–777. [CrossRef]
3. Chetoni, M.; Ascari, E.; Bianco, F.; Fredianelli, L.; Licitra, G.; Cori, L. Global Noise Score Indicator for Classroom Evaluation of Acoustic Performances in LIFE GIOCONDA Project. *Noise Mapp.* **2016**, *3*, 157–171. [CrossRef]
4. Zacarías, F.; Molina, R.H.; Ancela, J.L.C.; López, S.L.; Ojembarrena, A. Noise Exposure in Preterm Infants Treated with Respiratory Support Using Neonatal Helmets. *Acta Acust. United Acust.* **2013**, *99*, 590–597. [CrossRef]
5. Minichilli, F.; Gorini, F.; Ascari, E.; Bianchi, F.; Coi, A.; Fredianelli, L.; Licitra, G.; Manzoli, F.; Mezzasalma, L.; Cori, L. Annoyance Judgment and Measurements of Environmental Noise: A Focus on Italian Secondary Schools. *Int. J. Environ. Res. Public Health* **2018**, *15*, 208. [CrossRef] [PubMed]
6. Erickson, L.; Newman, R. Influences of Background Noise on Infants and Children. *Curr. Dir. Psychol. Sci.* **2017**, *26*, 096372141770908. [CrossRef]
7. Potamianos, A.; Narayanan, S. Robust Recognition of Children's Speech. *IEEE Trans. Speech Audio Process.* **2003**, *11*, 603–616. [CrossRef]
8. Cosi, P. On the Development of Matched and Mismatched Italian Children's Speech Recognition System. In Proceedings of the Interspeech, Brighton, UK, 6–10 September 2009; pp. 540–543.

9. Narayanan, S.; Potamianos, A. Creating Conversational Interfaces for Children. *IEEE Trans. Speech Audio Process.* **2002**, *10*, 65–78. [CrossRef]
10. Sunil, Y.; Prasanna, S.; Sinha, R. Children's Speech Recognition Under Mismatched Condition: A Review. *IETE J. Educ.* **2016**, *57*, 96–108. [CrossRef]
11. Kathania, H.K.; Kadiri, S.R.; Alku, P.; Kurimo, M. Spectral Modification for Recognition of Children's Speech Under Mismatched Conditions. In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), Reykjavik, Iceland, 31 May–2 June 2021; Linköping University Electronic Press: Linköping, Sweden, 2021; pp. 94–100.
12. Gowda, D.; Kadiri, S.R.; Story, B.; Alku, P. Time-varying Quasi-closed-phase Analysis for Accurate Formant Tracking in Speech Signals. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1901–1914.
13. Chavan, K.; Gawande, U. Speech Recognition in Noisy Environment, Issues and Challenges: A Review. In Proceedings of the 2015 International Conference on Soft-Computing and Networks Security (ICSNS), Coimbatore, India, 25–27 February 2015; pp. 1–5. [CrossRef]
14. Fernando, S.; Moore, R.K.; Cameron, D.; Collins, E.C.; Millings, A.; Sharkey, A.J.C.; Prescott, T.J. Automatic Recognition of Child Speech for Robotic Applications in Noisy Environments. *arXiv* **2016**, arXiv:1611.02695.
15. Martinek, R.; Vanus, J.; Nedoma, J.; Fridrich, M.; Frnda, J.; Kawala-Sterniuk, A. Voice Communication in Noisy Environments in a Smart House Using Hybrid LMS + ICA Algorithm. *Sensors* **2020**, *20*, 6022. [CrossRef]
16. Walker, E.A.; Sapp, C.; Oleson, J.J.; McCreery, R.W. Longitudinal Speech Recognition in Noise in Children: Effects of Hearing Status and Vocabulary. *Front. Psychol.* **2019**, *10*, 2421. [CrossRef]
17. Claus, F.; Gamboa-Rosales, H.; Petrick, R.; Hain, H.U.; Hoffmann, R. A Survey About Databases of Children's Speech. In Proceedings of the 14th Annual Conference of the International Speech Communication, Lyon, France, 25–29 August 2013; pp. 2410–2414.
18. Fainberg, J.; Bell, P.; Lincoln, M.; Renals, S. Improving Children's Speech Recognition Through Out-of-Domain Data Augmentation. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016; pp. 1598–1602. [CrossRef]
19. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR Corpus Based on Public Domain Audio Books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5206–5210.
20. Battenberg, E.; Chen, J.; Child, R.; Coates, A.; Gaur, Y.; Li, Y.; Liu, H.; Satheesh, S.; Seetapun, D.; Sriram, A.; et al. Exploring Neural Transducers for End-to-End Speech Recognition. *arXiv* **2017**, arXiv:1707.07413.
21. Shahnawazuddin, S.; Deepak, K.T.; Pradhan, G.; Sinha, R. Enhancing Noise and Pitch Robustness of Children's ASR. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5225–5229.
22. Yadav, I.C.; Shahnawazuddin, S.; Govind, D.; Pradhan, G. Spectral Smoothing by Variationalmode Decomposition and its Effect on Noise and Pitch Robustness of ASR System. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5629–5633.
23. Mitra, V.; Franco, H.; Bartels, C.; van Hout, J.; Graciarena, M.; Vergyri, D. Speech Recognition in Unseen and Noisy Channel Conditions. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5215–5219.
24. Roffo, G.; Melzi, S.; Castellani, U.; Vinciarelli, A.; Cristani, M. Infinite Feature Selection: A Graph-based Feature Filtering Approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, 1. [CrossRef] [PubMed]
25. Xia, S.; Chen, B.; Wang, G.; Zheng, Y.; Gao, X.; Giem, E.; Chen, Z. mCRF and mRD: Two Classification Methods Based on a Novel Multiclass Label Noise Filtering Learning Framework. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–15. [CrossRef]
26. Zhang, C.; Zhang, S. Bayesian Joint Matrix Decomposition for Data Integration with Heterogeneous Noise. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1184–1196. [CrossRef] [PubMed]
27. Dubagunta, S.P.; Kabil, S.H.; Doss, M.M. Improving Children Speech Recognition through Feature Learning from Raw Speech Signal. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5736–5740.
28. Shahnawazuddin, S.; Kathania, H.; Dey, A.; Sinha, R. Improving Children's Mismatched ASR Using Structured Low-rank Feature Projection. *Speech Commun.* **2018**, *105*, 103–113. [CrossRef]
29. Kathania, H.K.; Shahnawazuddin, S.; Adiga, N.; Ahmad, W. Role of Prosodic Features on Children's Speech Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5519–5523.
30. Kathania, H.K.; Shahnawazuddin, S.; Ahmad, W.; Adiga, N. Role of Linear, Mel and Inverse-Mel Filterbanks in Automatic Recognition of Speech from High-Pitched Speakers. *Circuits Syst. Signal Process.* **2019**, *38*, 4667–4682. [CrossRef]
31. Shahnawazuddin, S.; Dey, A.; Sinha, R. Pitch-Adaptive Front-End Features for Robust Children's ASR. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016.
32. Gurunath Shivakumar, P.; Georgiou, P. Transfer Learning From Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations. *Comput. Speech Lang.* **2020**, *63*, 101077. [CrossRef] [PubMed]

33. Ahmad, W.; Shahnawazuddin, S.; Kathania, H.; Pradhan, G.; Samaddar, A. Improving Children's Speech Recognition Through Explicit Pitch Scaling Based on Iterative Spectrogram Inversion. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 2391–2395. [CrossRef]
34. Shahnawazuddin, S.; Adiga, N.; Kathania, H.K. Effect of Prosody Modification on Children's ASR. *IEEE Signal Process. Lett.* **2017**, *24*, 1749–1753. [CrossRef]
35. Kathania, H.K.; Shahnawazuddin, S.; Ahmad, W.; Adiga, N.; Jana, S.K.; Samaddar, A.B. Improving Children's Speech Recognition Through Time Scale Modification Based Speaking Rate Adaptation. In Proceedings of the 2018 International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 16–19 July 2018.
36. Kathania, H.K.; Kadiri, S.R.; Alku, P.; Kurimo, M. Study of Formant Modification for Children ASR. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7429–7433.
37. Sheng, P.; Yang, Z.; Qian, Y. GANs for Children: A Generative Data Augmentation Strategy for Children Speech Recognition. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 129–135.
38. Shahnawazuddin, S.; Adiga, N.; Kathania, H.K.; Sai, B.T. Creating Speaker Independent ASR System Through Prosody Modification Based Data Augmentation. *Pattern Recognit. Lett.* **2020**, *131*, 213–218. [CrossRef]
39. Knill, K.; Gales, M.; Kyriakopoulos, K.; Malinin, A.; Ragni, A.; Wang, Y.; Caines, A. Impact of ASR Performance on Free Speaking Language Assessment. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 1641–1645. [CrossRef]
40. Siegler, M.A.; Stern, R.M. On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems. In Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing, Detroit, MI, USA, 9–12 May 1995; Volume 1, pp. 612–615. [CrossRef]
41. Fosler-Lussier, E.; Morgan, N. Effects of Speaking Rate and Word Frequency on Pronunciations in Convertional Speech. *Speech Commun.* **1999**, *29*, 137–158. [CrossRef]
42. Stollman, M.H.P.; Kapteyn, T.S.; Sleeswijk, B.W. Effect of Time-Scale Modification of Speech on the Speech Recognition Threshold in Noise for Hearing-Impaired and Language-Impaired Children. *Scand. Audiol.* **1994**, *23*, 39–46. [CrossRef]
43. Yadav, I.C.; Pradhan, G. Significance of Pitch-Based Spectral Normalization for Children's Speech Recognition. *IEEE Signal Process. Lett.* **2019**, *26*, 1822–1826. [CrossRef]
44. Robinson, T.; Fransen, J.; Pye, D.; Foote, J.; Renals, S. WSJCAM0: A British English Speech Corpus For Large Vocabulary Continuous Speech Recognition. In Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing, Detroit, MI, USA, 9–12 May 1995; Volume 1, pp. 81–84.
45. Varga, A.; Steeneken, H.J. Assessment for automatic speech recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems. *Speech Commun.* **1993**, *12*, 247–251. [CrossRef]
46. Zhu, X.; Beauregard, G.T.; Wyse, L.L. Real-time Signal Estimation from Modified Short-time Fourier Transform Magnitude Spectra. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 1645–1653. [CrossRef]
47. Beauregard, G.T.; Zhu, X.; Wyse, L. An Efficient Algorithm for Real-Time Spectrogram Inversion. In Proceedings of the 8th International Conference on Digital Audio Effects, Madrid, Spain, 20–22 September 2005; pp. 116–118.
48. Griffin, D.; Lim, J. Signal Estimation from Modified Short-time Fourier Transform. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 236–243. [CrossRef]
49. Batliner, A.; Blomberg, M.; D'Arcy, S.; Elenius, D.; Giuliani, D.; Gerosa, M.; Hacker, C.; Russell, M.; Wong, M. The PF_STAR Children's Speech Corpus. In Proceedings of the Interspeech, Lisbon, Portugal, 4–8 September 2005; pp. 2761–2764.
50. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi Speech Recognition Toolkit. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Big Island, HI, USA, 11–15 December 2011.
51. Kathania, H.K.; Shahnawazuddin, S.; Pradhan, G.; Samaddar, A.B. Experiments on Children's Speech Recognition Under Acoustically Mismatched Conditions. In Proceedings of the 2016 IEEE Region 10 Conference (TENCON), Singapore, 22–25 November 2016; pp. 3014–3017. [CrossRef]
52. Yildirim, S.; Narayanan, S.; Byrd, D.; Khurana, S. Acoustic Analysis of Preschool Children's Speech. In Proceedings of the International Congresses of Phonetic Sciences (ICPhS), Barcelona, Spain, 3–9 August 2003.
53. Tavares, E.L.M.; de Labio, R.B.; Martins, R.H.G. Normative Study of Vocal Acoustic Parameters From Children From 4 to 12 Years of Age Without Vocal Symptoms. A Pilot Study. *Braz. J. Otorhinolaryngol.* **2010**, *76*, 485–490. [CrossRef] [PubMed]