Spectral modification for recognition of children's speech under mismatched conditions

Hemant Kathania, Sudarsana Reddy Kadiri, Paavo Alku and Mikko Kurimo
Department of Signal Processing and Acoustics, Aalto University, Finland
(hemant.kathania, sudarsana.kadiri, paavo.alku, and mikko.kurimo)@aalto.fi

Abstract

In this paper, we propose spectral modification by sharpening formants and by reducing the spectral tilt to recognize children's speech by automatic speech recognition (ASR) systems developed using adult speech. In this type of mismatched condition, the ASR performance is degraded due to the acoustic and linguistic mismatch in the attributes between children and adult speakers. The proposed method is used to improve the speech intelligibility to enhance the children's speech recognition using an acoustic model trained on adult speech. In the experiments, WSJCAM0 and PFSTAR are used as databases for adults' and children's speech, respectively. The proposed technique gives a significant improvement in the context of the DNN-HMM-based ASR. Furthermore, we validate the robustness of the technique by showing that it performs well also in mismatched noise conditions.

Index Terms: Children speech recognition, Spectral sharpening, Spectral tilt, DNN.

1 Introduction

Recent advances in ASR have impacted many applications in various fields, such as education, entertainment, home automation, and medical assistance (Vajpai and Bora, 2016). These applications can benefit children in their daily life, in playing games, reading tutors (Mostow, 2012), and learning both native and foreign languages (Evanini and Wang, 2013; Yeung and Alwan, 2019).

The task of speech parameterization for the front-end aims at a compact representation that captures the relevant information in the speech signal by using short-time feature vectors. The two commonly used feature sets are Mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980) and the perceptual linear prediction cepstral coefficients (PLPCC) (Lee et al., 1999; Huber et al., 1999). Speech of adults and children have large acoustic and linguistic differences (Lee et al., 1999; Narayanan and Potamianos, 2002; Potaminaos and Narayanan, 2003; Gerosa et al., 2009). Both the Mel-filterbank and PLP coefficients are better suited for adults as they provide better resolution for low-frequency contents while a greater degree of averaging happens in the high-frequency range (Davis and Mermelstein, 1980; Hermansky, 1990a).

In the case of children's speech, more relevant information is available in the high-frequency range. Therefore, to enhance the system performance, a better resolution needs to be used for the high-frequency range. Previous studies have also shown that formant sharpening is helpful for increasing speech intelligibility (Chennupati et al., 2019; Zorila Tudor-Catalin and Yannis, 2012; Potaminaos and Narayanan, 2003; Kathania et al., 2014). Motivated by these observations, we suggest to modify the speech spectrum by formant sharpening and spectral tilt reduction.

In (Potamianos and Narayanan, 2003; Kathania et al., 2014, 2016) , it was shown that the word error rate (WER) in recognition of children's speech is much higher than that of adult speech and specifically under mismatched and noisy conditions. The problems are due to higher inter-speaker variance caused by the development of the vocal tract, leading to different formant locations and spectral distribution (Hermansky, 1990b), and due to the inaccuracy in pronunciation and grammar caused by language acquisition. Most importantly, the insufficient training data limits the performance because collecting large speech databases of children's speech is hard. Adult speech corpora normally contain hun-

dreds or thousands of hours of data, while most publicly available corpora for children's speech have less than 100 hours of data (Panayotov et al., 2015; Claus et al., 2013). Therefore, it is necessary that ASR systems built for children are robust for various mismatched conditions.

In this paper, a spectral sharpening and tilt reduction method is proposed to enhance the intelligibility of children's speech to boost the ASR system performance under mismatched conditions. Spectral sharpening and spectral tilt reduction have been used in enhancement of speech intelligibility in noise (Chennupati et al., 2019; Zorila Tudor-Catalin and Yannis, 2012). In this study, it is shown that the MFCC and PLPCC features computed after the spectral modification (referred to as SS-MFCC and SS-PLPCC) are found to outperform the conventional MFCC and PLPCC features. This is demonstrated by both the spectral analyses and experimental evaluations in this paper. The robustness of the technique is further validated by showing that it performs well in mismatched noise conditions also.

The remaining of this paper is presented as follows: In Section 2, the proposed spectral sharpening and tilt reduction technique is discussed. In Section 3, the speech corpora and ASR specifications are described. The results of the proposed method are presented in Section 4. In Section 5, the effects of noisy environment on the proposed method are discussed. Finally, the paper is concluded in Section 6.

2 The spectral modification method

The proposed spectral modification technique consists of formant sharpening and spectral tilt reduction as described below and depicted in the block diagram in Fig 1. From the spectral examples shown in Fig 2 and spectrograms shown in Fig 3, we can observe that the proposed method enhances formant peaks and the level of higher frequencies.

2.1 Adaptive spectral sharpening

The formant information is important for recognizing speech, and Adaptive Spectral Sharpening (ASS) is a method that emphasizes the formant information (Zorila Tudor-Catalin and Yannis, 2012). For sharpening of formants, an approach that was motivated in speech intelligibility is utilised (Zorila Tudor-Catalin and Yannis, 2012). In this method, the magnitude spectrum is extracted using the SEEVOC method (Paul, 1981) for the pre-emphasized voice speech frame. The adaptive spectral sharpening at frame t is given by

$$H_s(\omega, t) = \left(\frac{E(\omega, t)}{T(\omega, t)}\right)^{\beta}, \tag{1}$$

where E(ω,t) is the estimated spectral envelope computed using the SEEVOC method and T(ω,t) is the spectral tilt for frame t. Spectral tilt T(ω,t) is computed using cepstrum and is given by

$$\log T(\omega) = C_0 + 2C_1 \cos(\omega). \tag{2}$$

Here C_m is the mth cepstral coefficients and is given by

$$C_m = \frac{1}{(\frac{N}{2}+1)} \sum_{k=0}^{\frac{N}{2}} E(\omega_k) \cos(m\omega_k). \tag{3}$$

Formant sharpening is performed using Eq. (1) by varying β. Typically, the value of β is higher for low signal-to-noise ratio (SNR) values and lower for high SNR values. In this study, we have investigated the extent of spectral sharpening by varying the β parameter from 0.15 to 0.35. Note that spectral sharpening is performed only in voiced segments using probability of voicing as defined in (Zorila Tudor-Catalin and Yannis, 2012).

2.2 Spectral tilt modification

Apart from spectral sharpening, we also perform fixed spectral tilt modification ($H_r(\omega)$) to boost the region between 1 kHz and 4 kHz by 12 dB and to reduce the level of frequencies below 500 Hz (by 6 dB/octave). The resulting magnitude spectrum for a frame after the ASS and fixed spectrum tilt modification is given by

$$\hat{E}(\omega) = E(\omega) H_s(\omega) H_r(\omega) \tag{4}$$

The modified magnitude spectrum ($\hat{E}(\omega)$) is combined with the original phase spectrum for reconstructing the signal using IDFT and Overlap-and-Add (OLA) (Rabiner and Gold, 1975).

A schematic block diagram describing the steps involved in the proposed method is shown in Fig 1. Fig 2 illustrates the effect of spectral modification for a voiced child's speech segment. Here the blue curve is the spectrum of the original speech segment and the red curve is the modified speech spectrum. From the figure, it can be seen that

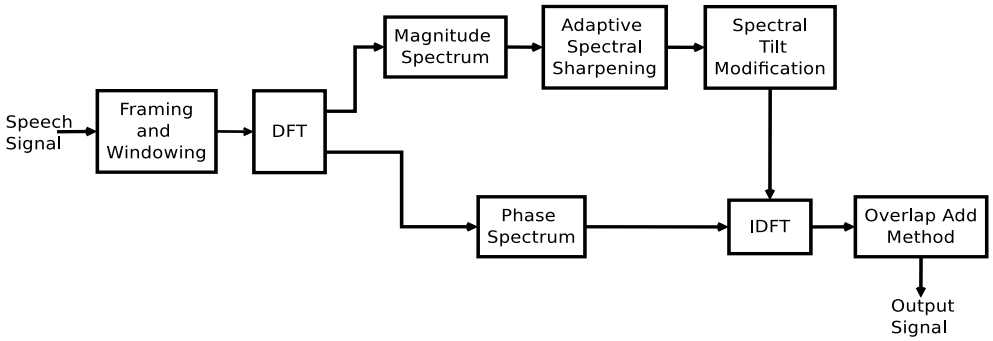

Figure 1: Block diagram of the spectral modification method.

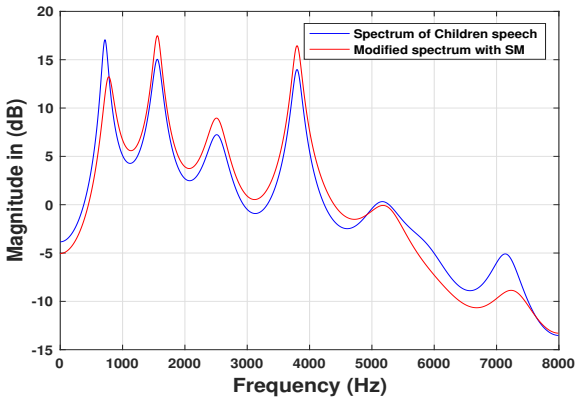

Figure 2: Spectrum for a segment of child's speech (blue) and the corresponding spectrum after the spectral modification (SM) (red).

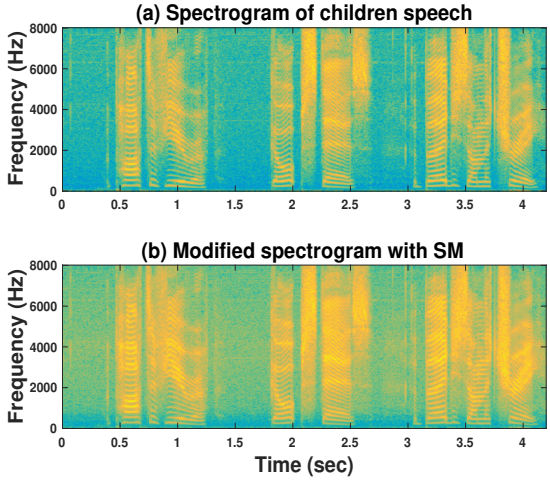

Figure 3: Spectrogram for a segment of child's speech shown in (a), and the corresponding spectrogram after spectral modification shown in (b).

formants are sharpened by the proposed method (red curve). Specifically, it can be clearly seen that formants are more prominent in the region of 1 kHz to 4 kHz for the proposed method (red curve), which is due to the spectral modification as described in Section 2.2. Furthermore, illustrations of the spectrograms are shown in Fig 3. Fig 3 (a) shows the child's original spectrogram before modifications and Fig 3 (b) shows the corresponding spectrogram after the proposed spectral modification (SM) method. Again it can be observed from Fig 3(b) that the spectrogram has a larger high-frequency emphasis compared to spectrogram in Fig 3(a), due to spectral modification in the proposed method.

3 Data and Experimental setup

This section describes the speech corpora (adult and children), front-end speech features and specifications of ASR system.

3.1 Speech Corpora

Adult speech data used in this work was obtained from WSJCAM0 (Robinson et al., 1995). Children's speech data was obtained from the PF-STAR corpus (Batliner et al., 2005) to simulate a mismatched ASR task. Both the WSJCAM0 and PF-STAR corpora are British English speech databases. Details of both corpora are given in Table 1

3.2 Front-end speech parameterization

The speech data was first pre-emphasized with a first order FIR high-pass filter (with zero at $z = 0.97$). For frame-blocking, overlapping Hamming windows with a length of 20 ms and an overlap of 50% were used. 13-dimensional MFCCs were extracted using 40 channels. The 13-dimensional base MFCC features were then spliced in time taking a context size of 9 frames. Time-splicing resulted in 117-dimensional features vectors. Linear discriminant analysis (LDA) and maximum-likelihood linear transformation (MLLT) were

Table 1: Speech corpora details for WSJCAM0 and PFSTAR used in ASR

Corpus	WSJCAM0		PF-STAR	
Language	British English		British English	
Purpose	Training	Testing	Training	Testing
Speaker group	Adult	Adult	Child	Child
No. of speakers	92	20	122	60
Speaker age	$>$ 18 years	$>$ 18 years	4-14 years	4-13 years
No. of words	132,778	5,608	46974	5067
Duration (hrs.)	15.50	0.60	8.3	1.1

used to reduce the feature vector dimension from 117 to 40. The 13-dimensional base PLPCC features were derived using 12^{th}-order linear prediction (LP) analysis. Cepstral mean and variance normalization (CMVN) as well as feature-space maximum-likelihood linear regression (fMLLR) were performed next to enhance robustness with respect to speaker-dependent variations. The required fMLLR transformations for the training and test data were generated through speaker adaptive training.

The MFCC and PLPCC features computed after the proposed spectral modification (i.e., spectral sharpening and tilting) are referred to as SS-MFCC and SS-PLPCC, respectively. ASR results are given for the baseline features (MFCC and PLPCC) and the proposed features (SS-MFCC and SS-PLPCC) for all the experiments conducted in this paper.

3.3 ASR system specifications

To build the ASR system on the adult speech data from the WSJCAM0 speech corpus, the Kaldi toolkit (Povey et al., 2011) was used. Context-dependent hidden Markov models (HMM) were used for modeling the cross-word triphones. Decision tree-based state tying was performed with the maximum number of tied-states (senones) being fixed at 2000. A deep neural network (DNN) was used in acoustic modeling. Prior to learning parameters of the DNN-HMM-based ASR system, the fMLLR-normalized feature vectors were time-spliced once again considering a context size of 9 frames. The number of hidden layers in the DNN was set to 5 with 1024 hidden nodes in each layer. The nonlinearity in the hidden layers was modeled using the *tanh* function. The initial learning rate for training the DNN-HMM parameters was set at 0.005 which was reduced to 0.0005 in 15 epochs. The minibatch size for neural net training was set to 512.

For decoding the test set for adults, the MIT-Lincoln 5k vocabulary Wall Street Journal bi-gram language model (LM) was used. The perplexity of this LM for the adult test set is 95.3 while there are no out-of-vocabulary (OOV) words. Furthermore, a lexicon consisting of 5850 words including pronunciation variants was used. While decoding the test set for children's speech, a 1.5k domain-specific bigram LM was used. This bigram LM was trained on the transcripts of speech data in PF-STAR after excluding those corresponding to the test set of children's speech. The domain-specific LM has an OOV rate of 1.20% and perplexity of 95.8 for the test set of children's speech. In total 1969 words used including pronunciation variations in lexicon for decoding the children's test set.

4 Results and discussion

The baseline WERs for children's test set in the DNN-HMM systems is 19.76% and 20.00% for the MFCC and PLPCC acoustic features respectively (see Table 2). In order to improve the recognition performance, the spectral sharpening technique is applied to mitigate the spectral differences between adults' and children's speech. The spectral sharpening algorithm includes the tunable β parameter according to Eq. (1), and this parameter was varied from 0.15 to 0.35 to sharpen the spectral peaks (formants). The WERs obtained with varying sharpening parameter are shown in Figure 4. From the figure, it can be observed that the best WER was obtained with $\beta = 0.25$. The remaining experiments are carried out using this value of β.

The baseline WERs for children's test set with respect to the DNN-HMM-based ASR systems trained using the MFCC and PLPCC features are given in Table 2. The MFCC and PLPCC features computed after the formant modification are denoted as SS-MFCC and SS-PLPCC, respectively in Table 2. A notable reduction in WER can be observed for both the features.

For further analysis, the children test data was divided into three different test sets based on age groups: $4-6$ years, $7-9$ years, and $10-13$ years. Table 3 shows the results for baseline and

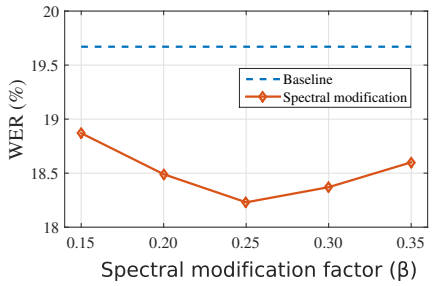

Figure 4: WER results depicting the effect of spectral modification (for varying the β parameter) on recognition of children's speech using an DNN-HMM system trained using adult speech.

proposed features for three age groups. It can be seen that the proposed approach improves the results in all the age groups for both of the proposed features, SS-MFCC and SS-PLPCC. We have also conducted significance test and notice that signed pair comparison found significant difference between the two approaches at level $p<0.01$.

To further validate the effectiveness of the proposed modification method, another DNN-HMM-based ASR system was developed by pooling together speech data from training sets of both adults and children. For children's speech, the training set derived from PF-STAR consisted of 8.3 hours of speech by 122 speakers. The total number of utterances in this training set was equal to 856 with a total of 46974 words. The training set of adult speakers consisted of 15.5 hours of speech from 92 speakers (both male and female). Further, the training set comprised 132, 778 words and the total number of utterances was 7852. The developed ASR system exhibits a lower degree of acoustic/linguistic mismatch due to the pooling of children's speech into training. As a result, the baseline WERs for the developed system (given in Table 2) are significantly lower when compared to those obtained with respect to the ones trained on adult speech only. Still, further reductions in WERs are achieved when the spectral modification technique is applied to enhance the speech intelligibility as shown in Table 2.

5 Experiments in Noisy conditions

To further validate the proposed technique, noise robustness of the spectral modification technique was studied. Four different noises (babble, white, factory and volvo noise) extracted from NOISEX-92 (Varga and Steeneken, 1993) were added to the test data under varying SNR levels. The noisy test sets were then decoded using the acoustic models trained with clean speech. WERs in the case of adult/child mismatched testing are given in Table 4 for SNR values of 5 dB, 10 dB, and 15 dB. While the MFCC features seem slightly more robust to additive noise than the PLPCC features, the spectral modification reduces WER clearly for both of the acoustic features (denoted as SS-MFCC and SS-PLPCC) at the three different SNR levels. Hence, it can be concluded that the spectral sharpening of formant peaks improves the ASR performance also in various noisy conditions.

Table 2: WERs of the baseline and proposed spectral modification method for children's ASR. The performance evaluation is done separately using two ASR systems: a system trained with only adult speech from WSJCAM0 and a system trained by pooling also children's speech.

Training Data	Testing Data	WER in (%)			
		DNN-HMM (Acoustic Model)			
		PLPCC	**SS-PLPCC**	MFCC	**SS-MFCC**
Adult speech	Children's speech	20.00	19.38	19.76	18.23
Adult + children's speech	Children's speech	12.89	12.43	12.26	11.70

Table 3: WERs for the age-wise grouped children speech test sets with respect to adults data trained ASR systems demonstrating the effect of the proposed spectral modification.

Age wise setup	WER (in %)			
	PLPCC	**SS-PLPCC**	MFCC	**SS-MFCC**
4 - 6	72.36	70.18	70.48	68.18
7 - 9	20.11	17.24	19.38	16.20
10 - 13	12.35	11.72	11.78	10.53

6 Conclusion

This work explores spectral modification (sharpening of formants and reduction of spectral tilt) to achieve robust recognition of children's speech under mismatched conditions. The explored spectral modification technique is observed to enhance ASR of children's speech for both the MFCC and PLPCC features. Also, ASR results are analyzed for different age-groups and it was found that for all the age-groups there exists an improvement

Table 4: WERs of the proposed spectral modification method for children's speech test set under varying additive noise conditions.

Noise Type	SNR (dB)	WER in (%)			
		PLPCC	**SS-PLPCC**	MFCC	**SS-MFCC**
Babble	5dB	83.69	82.67	79.70	80.35
	10dB	64.62	58.36	59.7	56.41
	15dB	48.47	42.61	40.34	38.08
White	5dB	86.54	83.61	87.40	86.25
	10dB	79.01	77.26	73.78	72.62
	15dB	66.79	63.58	54.00	53.46
Factory	5dB	86.54	83.61	92.32	90.86
	10dB	67.13	65.96	68.96	66.95
	15dB	49.32	48.65	45.33	43.55
Volvo	5dB	34.71	26.22	26.12	24.70
	10dB	29.16	24.58	23.10	22.03
	15dB	25.61	22.89	21.64	20.75

with the proposed approach compared to baseline. Further, improvements were also observed in mismatch conditions caused by additive noise.

7 Acknowledgements

This work was supported by the Academy of Finland (grant 329267). The computational resources were provided by Aalto ScienceIT.

References

A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, and M. Wong. 2005. The PF_STAR children's speech corpus. In *Proc. INTERSPEECH*, pages 2761–2764.

Nivedita Chennupati, Sudarsana Reddy Kadiri, and B. Yegnanarayana. 2019. Spectral and temporal manipulations of sff envelopes for enhancement of speech intelligibility in. *Computer Speech Language*, 54:86 – 105.

Felix Claus, Hamurabi Gamboa-Rosales, Rico Petrick, Horst-Udo Hain, and Rüdiger Hoffmann. 2013. A survey about databases of children's speech. In *14th Annual Conference of the International Speech Communication Association At: Lyon, France*, pages 2410–2414.

S. Davis and P. Mermelstein. 1980. https://doi.org/10.1109/TASSP.1980.1163420 Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 28(4):357–366.

K. Evanini and X. Wang. 2013. Automated speech scoring for non-native middle school students with multiple task types. In *Proc. INTERSPEECH*, pages 2435–2439.

Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos. 2009. A review of ASR technologies for children's speech. In *Proc. Workshop on Child, Computer and Interaction*.

H. Hermansky. 1990a. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 57(4):1738–52.

Hynek Hermansky. 1990b. Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752.

Jessica Huber, Elaine Stathopoulos, Gina Curione, Theresa Ash, and Kenneth Johnson. 1999. https://doi.org/10.1121/1.427150 Formants of children, women, and men: The effects of vocal intensity variation. *The Journal of the Acoustical Society of America*, 106:1532–42.

H. K. Kathania, S. Shahnawazuddin, G. Pradhan, and A. B. Samaddar. 2016. Experiments on children's speech recognition under acoustically mismatched conditions. In *2016 IEEE Region 10 Conference (TENCON)*, pages 3014–3017.

H. K. Kathania, S. Shahnawazuddin, and R. Sinha. 2014. Exploring hlda based transformation for reducing acoustic mismatch in context of children speech recognition. In *2014 International Conference on Signal Processing and Communications (SPCOM)*, pages 1–5.

Sungbok Lee, Alexandros Potamianos, and Shrikanth S. Narayanan. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468.

J. Mostow. 2012. Why and how our automated reading tutor listens. In *Proc. INTERSPEECH*, 4.

S. Narayanan and A. Potamianos. 2002. Creating conversational interfaces for children. *IEEE Transactions on Speech and Audio Processing*, 10(2):65–78.

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

D Paul. 1981. The spectral envelope estimation vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(4):786–794.

A. Potamianos and S. Narayanan. 2003. Robust recognition of children's speech. *IEEE Transactions on Speech and Audio Processing*, 11(6):603–616.

A. Potaminaos and S. Narayanan. 2003. Robust Recognition of Children Speech. *IEEE Transactions on Speech and Audio Processing*, 11(6):603–616.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *Proc. ASRU*.

Lawrence R Rabiner and Bernard Gold. 1975. Theory and application of digital signal processing. *Englewood Cliffs, NJ, Prentice-Hall, Inc.*

T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals. 1995. WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In *Proc. ICASSP*, volume 1, pages 81–84.

J. Vajpai and A. Bora. 2016. Industrial applications of automatic speech recognition. *International Journal of Engineering Research and Applications*, 6(3):88–95.

Andrew Varga and Herman J.M. Steeneken. 1993. 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 Communication*, 12(3):247–251.

Gary Yeung and Abeer Alwan. 2019. A Frequency Normalization Technique for Kindergarten Speech Recognition Inspired by the Role of fo in Vowel Perception. In *Proc. INTERSPEECH*, pages 6–10.

Kandia Varvara Zorila Tudor-Catalin and Stylianou Yannis. 2012. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. *INTERSPEECH*, pages 635 – 638.