INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Frequency Normalization Technique for Kindergarten Speech Recognition Inspired by the Role of f_o in Vowel Perception

Gary Yeung[1], *Abeer Alwan*[1]

[1]Dept. of Electrical and Computer Engineering, University of California, Los Angeles, USA

garyyeung@g.ucla.edu, alwan@ee.ucla.edu

Abstract

Accurate automatic speech recognition (ASR) of kindergarten speech is particularly important as this age group may benefit the most from voice-based educational tools. Due to the lack of young child speech data, kindergarten ASR systems often are trained using older child or adult speech. This study proposes a fundamental frequency (f_o)-based normalization technique to reduce the spectral mismatch between kindergarten and older child speech. The technique is based on the tonotopic distances between formants and f_o developed to model vowel perception. This proposed procedure only relies on the computation of median f_o across an utterance. Tonotopic distances for vowel perception were reformulated as a linear relationship between formants and f_o to provide an effective approach for frequency normalization. This reformulation was verified by examining the formants and f_o of child vowel productions. A 208-word ASR experiment using older child speech for training and kindergarten speech for testing was performed to examine the effectiveness of the proposed technique against piecewise vocal tract length, $F3$-based, and subglottal resonance normalization techniques. Results suggest that the proposed technique either has performance advantages or requires the computation of fewer parameters.

Index Terms: child speech recognition, frequency normalization, fundamental frequency

1. Introduction

Over the past several years, speech has become one of the most desirable ways to interact with electronic devices. In fact, speech may be one of the only methods that young children have to interact with such devices as many children have not yet learned to read, write, or type. Additionally, child automatic speech recognition (ASR) can facilitate the development of automated educational and assessment tools [1, 2, 3, 4, 5]. Yet, while adult ASR systems have demonstrated significant improvements over the past years, child ASR systems continue to lag behind in performance.

A number of past studies has explored the performance of child ASR systems [6, 7, 8]. Notably, a recent study by Kennedy et al. explored the performance of child ASR in robotics applications [6]. The results revealed insufficient performance (15%-20% error rates) for even basic digit recognition tasks. Another study by Yeung and Alwan discovered the impact that a single year age difference can have on child ASR performance using deep neural network (DNN) hidden Markov model (HMM) ASR systems [7]. Those results highlighted the importance of targeting a specific age group (especially the kindergarten age group) when training and testing ASR systems, rather than grouping several age groups together.

One of the major hurdles that child ASR still faces is the scarcity of publicly-available child speech databases. As such, ASR systems for young children are often trained on speech from older children or adults. However, speech acoustics of children change dramatically as they grow [9, 10, 11, 12, 13, 14]. For example, formant frequencies and the fundamental frequency (f_o) are known to depend on the age of the child [9, 10, 11]. As the mismatch in age between training and testing data becomes larger, ASR performance degrades rapidly [7].

Several child ASR training techniques have been dedicated to reducing the acoustic mismatch between training and testing data. Some of these studies used a range of ages along with different vocal tract length normalization (VTLN) and maximum likelihood linear regression (MLLR) approaches with varying degrees of success [15, 16, 17, 18, 19, 20]. Variations of VTLN that replace the warping factor with acoustic parameters, such as the third formant ($F3$) [21] or subglottal resonances (SGR) [22], have also been shown to be effective for the normalization of child-to-adult speech. Additionally, some studies have shown success by using DNN architectures to learn from both child and adult speech [23, 24].

f_o has been used successfully in adult ASR. Several studies noted that the inclusion of f_o or voicing parameters as features resulted in increased performance for adult systems, even in atonal languages [25, 26, 27]. Faria and Gelbart argued that f_o could be used to predict vocal tract size and VTLN warping factors with a maximum-likelihood (ML) approach [28]. For children, Shahnawazuddin et al. used f_o to predict lifter sizes when extracting cepstral features [29].

Turning to the field of human speech perception provides additional insight into the usefulness of f_o. The tonotopic distances between formants and f_o, the distances between formants and f_o in some perceptual frequency scale, are a well-known set of features used to effectively normalize formant-based models of vowel perception [30, 31, 32]. The inclusion of f_o in the tonotopic distances suggests that f_o serves as a normalizing factor for formants, and consequently, the speech frequency spectrum. This is further supported by studies that show how f_o affects both vowel perception and the perception of voice naturalness when formants are fixed [33, 34].

In this study, we propose an f_o-based normalization procedure inspired by the tonotopic distances. Kindergarten is chosen as the target age group due to its relevance for educational and human-robot interface (HRI) applications [6], as well as its poor ASR performance [7]. We show that the tonotopic distances can be reformulated and generalized to motivate an f_o-based frequency normalization approach for the entire speech spectrum. Furthermore, this normalization technique is applied to a mismatched-grade ASR experiment using kindergarten speech as testing data and older grades as training data to simulate a scenario where kindergarten training data are not readily accessible.

The remainder of the paper is organized as follows. Section 2 reformulates the tonotopic distances and verifies the model-

http://dx.doi.org/10.21437/Interspeech.2019-1847

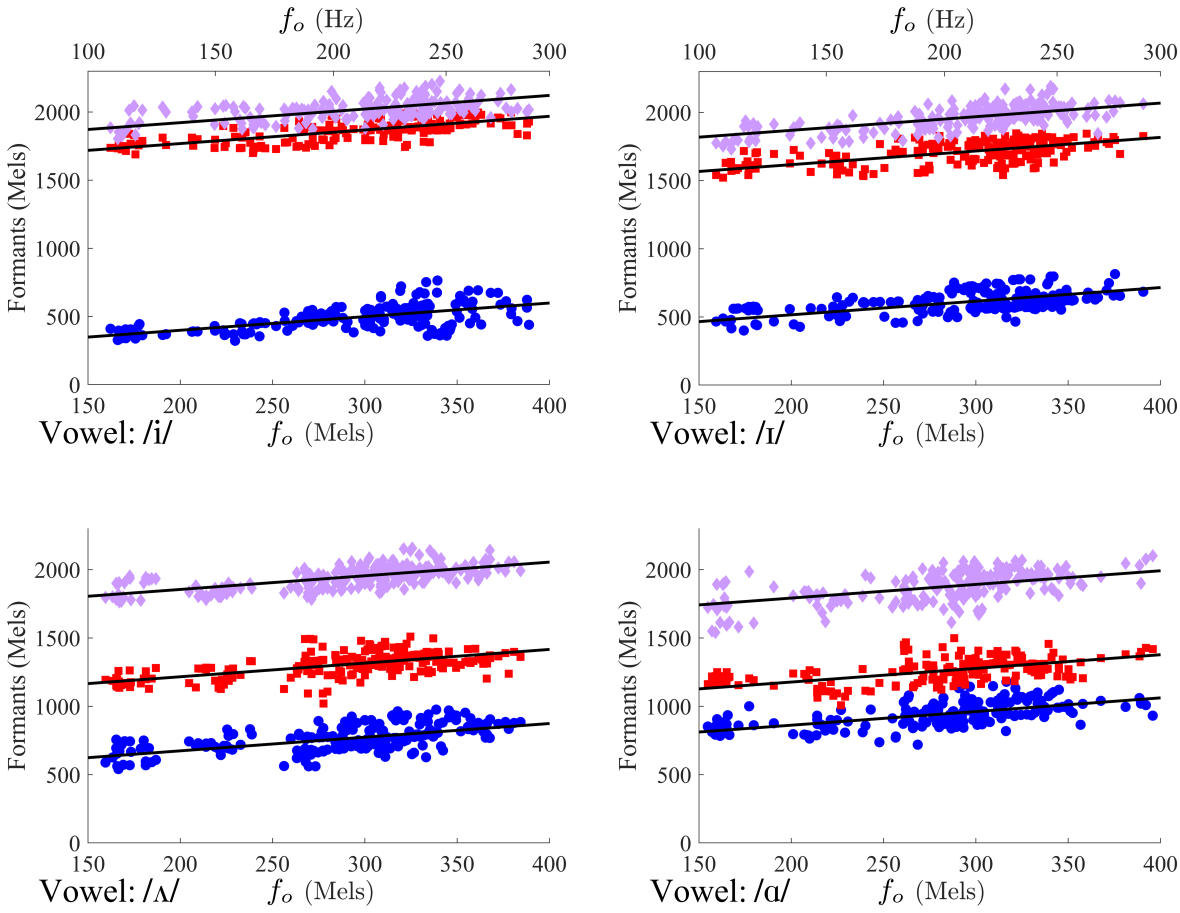

Figure 1: *$F1$ vs. f_o (blue), $F2$ vs. f_o (red), and $F3$ vs. f_o (purple) for /i/ (top-left), /ɪ/ (top-right), /ʌ/ (bottom-left), and /ɑ/ (bottom-right) from corresponding hVd words of children between the ages of 6 and 18 years. Also shown are the least-squares linear regression lines, fixed to have a slope of 1. The data seem to follow the linear relationship implied by the reformulation of the tonotopic distances.*

ing accuracy of the reformulation on child vowel productions. Section 3 describes the database and experimental setup of the mismatched-grade ASR experiment. Section 4 presents and discusses the results of the experiment. Section 5 concludes the paper with a brief summary and plans for future work.

2. Vowel Perception and Production

2.1. Tonotopic Distances for Vowel Perception

A number of past studies revealed that when using the bark scale, the tonotopic distances between formants ($F(x+1)-Fx$ for $x \in \{1, 2, 3, ...\}$), along with the tonotopic distance between the first formant and f_o ($F1 - f_o$), are effective features for modeling human vowel perception [30, 31, 32]. An equivalent set of features can be derived as the difference, in the bark scale, between formants and f_o ($Fx - f_o$ for $x \in \{1, 2, 3...\}$). This implies a linear relationship between formants and f_o in the bark scale. The formulation also implies that the linear relationship has a slope of 1. As the Mel scale is highly correlated with the bark scale, an equivalent relationship holds in the Mel domain, and we will use the Mel scale throughout this paper for consistency with Mel-frequency cepstral coefficients (MFCCs). The next section evaluates this reformulation for child vowel productions.

2.2. Child Vowel Production

To verify the relationship between formants and f_o for child vowel productions, we measured f_o, $F1$, $F2$, and $F3$ from the vowels of 4 hVd words in the WashU-UCLA Child Subglottal Resonances Database [35], which includes 43 children between the ages of 6 and 18 years. Two tense vowels, /i/ and /ɑ/, and two lax vowels, /ɪ/ and /ʌ/, were chosen for analysis. The least-squares linear regression lines for predicting $F1$, $F2$, and $F3$ from f_o were examined.

For all 12 linear regressions, the slopes of the least-squares regression lines were between 0.75 and 1.20. All slopes significantly contributed to the model ($p < 0.001$). The Pearson's correlation coefficients were always greater than $r > 0.51$. Figure 1 displays the data for the vowels, along with the least-squares regression lines, fixed to have a slope of 1. The linear relationships between the formants and f_o are visually obvious.

3. Database and Experimental Setup

3.1. Database

The OGI Kids' Speech Corpus [36] was used in this study. This corpus contains approximately 100 speakers per educational grade level, from kindergarten to 10th grade. Both scripted and spontaneous styles of speech were recorded from each speaker.

Utterances were recorded with a sampling rate of 16 kHz. In this study, only single-word scripted utterances were used as data for a word recognition task to avoid the usage of a child language model.

Each utterance contained one of 208 possible words. These words ranged from easy words such as "chair" to difficult words such as "organization." As some of the utterances did not contain the child saying the target word or were noisy, we only used files labeled as "1" in the OGI Kids' Speech Corpus verification files. This label indicated that the sound file both contained the child saying the word properly and had limited noise. Additionally, as the number of utterances of the words "push" and "spoons" was much larger than the other words, we only used a random subset of these words to remove a potential bias. To ensure a fair comparison between grades, a total of 1,654 word utterances were randomly chosen from each grade (K-10) as training or testing data.

3.2. Mel-Frequency Cepstral Coefficients

The default features were 13-dimensional MFCCs with a window size of 25 ms and a shift of 10 ms. MFCCs were extracted with a 1024-point discrete Fourier transform (DFT). When extracting MFCCs, a bandwith of 5.2 kHz was chosen such that $F3$ was contained in the signal for all children. The number of filters used during MFCC extraction differed depending on the normalization technique used and was based on preliminary experimentation. When not using any normalization, 19 filters were used. Cepstral mean normalization (CMN) was applied to all MFCCs.

3.3. Mismatched-Grade ASR Experiments

For the mismatched-grade ASR experiments, the 1st-10th grade data were used as training data. The data were separated by grade for a total of 10 sets of training data. For each grade, all 1,654 word utterances were used to train a DNN-HMM ASR system with 250 triphones. Various normalization strategies applied to MFCCs were used as input features. After feature extraction, a 7-frame linear discriminant analysis (LDA), along with feature space MLLR (fMLLR), was applied for a final 40-dimensional feature input. DNNs were trained on an additional 9-frame LDA, 2 hidden layers, and 2-norm non-linearities with an input dimension of 500 and output dimension of 100 [37]. All ASR systems were trained with the Kaldi Speech Recognition Toolkit [38]. This setup is based on our previous investigation of child ASR [7]. Each system was tested using all 1,654 word utterances from the kindergarten speech data.

3.4. Fundamental Frequency-Based Normalization

The reformulation of the tonotopic distances in Section 2.1 suggests that the formants of a vowel utterance can be normalized by computing f_o for the utterance and providing a default f_o value such that all speakers can be normalized to some default representation. To avoid the extraction of formants for children, which is generally unreliable, we simply warp the entire speech spectrum based on f_o. The DFT is warped as follows:

$$f_{norm} = f_{orig} - (f_{o,utt} - f_{o,def}) \quad (1)$$

where all parameters are in the Mel scale, f_{norm} is a normalized frequency corresponding to some DFT index, f_{orig} is the frequency from the original speech spectrum mapped to f_{norm}, $f_{o,utt}$ is the child's median f_o over an utterance, and $f_{o,def}$ is some chosen default value.

After initial testing for the value of $f_{o,def}$, we chose $f_{o,def} = 100$ Hz or 150.49 Mels for this experiment. To compute $f_{o,utt}$, f_o is computed per frame for each utterance using multi-band summary correlogram-based (MBSC) pitch detection [39], and the median was chosen as $f_{o,utt}$. For the utterances used in this experiment, the median f_o had a range of 78 Hz to 307 Hz. When using this f_o-based normalization method, the bandwidth of the MFCC extraction was chosen such that MFCC computation did not use a frequency higher than 5.2 kHz. Additionally, 15 filters were used.

3.5. Other Warping Strategies

To evaluate the effectiveness of the f_o-based normalization method, we also evaluated other standard normalization strategies on the mismatched-grade ASR task. These normalization strategies included piecewise VTLN, $F3$-based normalization [21], and SGR normalization [22]. Similar to the proposed f_o-based method, these methods were also applied at the utterance level. When extracting MFCCs with these normalization strategies, we used a bandwidth of 5.2 kHz with 19 filters. For the $F3$-based and SGR normalization, $F3$ and SGRs were computed as the median across the utterance.

4. Results and Discussion

The results of the mismatched-grade ASR experiment are displayed in Table 1. The first row shows word error rates (WERs) of the baseline systems with no normalization. The second row shows WERs of the systems trained using f_o-based normalization. The remaining three rows show the WERs of the systems trained using VTLN, $F3$-based, and SGR normalization.

Of the systems used, the system with the lowest WER was trained on 2nd grade speech with no normalization. For comparison, we performed this same task on the Google Cloud application programming interface (API) ASR with the 208 possible words included in the speech context. We also evaluated WER as whether the target word was contained in the resulting transcript (e.g., "athletes" contains "athlete"). However, the Google Cloud API ASR had a WER of 25.33%, which is worse than the system trained on 2nd grade speech. This demonstrates the importance of developing ASR systems that accommodate children with methods such as frequency normalization.

VTLN showed significant improvement over the baseline results when training on 9th and 10th grade speech. However, f_o-based normalization provided a significant improvement over VTLN for these grades. Additionally, f_o-based normalization provided a significant improvement over $F3$-based normalization when training on 2nd and 3rd grade speech. There was no significant difference between the f_o-based and SGR normalization techniques.

When using VTLN in an ASR system, testing data must be passed through the system multiple times to compute the ML warping factor. However, ASR systems with f_o-based normalization only rely on the computation of f_o and can perform decoding in a single pass. As VTLN performance is also significantly worse than f_o-based normalization for the heavily mismatched systems (9th and 10th grade training data), the f_o-based technique should be used in favor of VTLN for kindergarten ASR given that the system is capable of computing f_o.

While the $F3$-based normalization only relies on a single parameter, the system seems to perform poorly in the kindergarten ASR systems that are more closely matched. This could be caused by a number of possible complications, such as dif-

Table 1: *Word error rates (WERs) (%) of DNN-HMM ASR systems for the mismatched-grade experiments. Each ASR system was trained on a single grade level (1^{st}-10^{th} grade) and tested on kindergarten speech. MFCCs were extracted with no normalization, f_o-based normalization, VTLN, $F3$-based normalization, and SGR normalization. Features were extracted with a bandwidth of 5.2 kHz. All WERs that are not significantly different ($p > 0.05$) from the f_o-based normalization are in **bold**.*

Feature Normalization	**Training Grade** 1	2	3	4	5	6	7	8	9	10
None	**23.58**	**22.49**	**25.03**	**25.03**	**26.42**	**30.17**	**33.43**	37.91	41.05	47.28
f_o	**25.70**	**24.55**	**25.76**	**26.96**	**26.06**	**29.63**	**32.59**	**34.28**	**34.76**	**37.76**
VTLN	**24.43**	**26.60**	**27.45**	**25.94**	**25.09**	**29.32**	**31.38**	**36.15**	38.27	41.72
$F3$	**26.90**	27.45	28.90	**26.48**	**26.48**	**30.53**	**30.59**	**31.80**	**34.95**	**37.36**
SGR	**25.63**	**23.10**	**28.05**	**26.48**	**25.70**	**28.05**	**29.99**	**32.41**	**34.82**	**38.75**

ficulties in estimating formants or problems with performing frequency normalization when all $F3$ values are similar. As such, the f_o-based method performed significantly better than the $F3$-based method for 2nd and 3rd grade training data and is preferable for the kindergarten ASR system.

While f_o-based and SGR normalization methods had similar performance, the f_o-based method may have computational benefits. The computation of SGRs requires estimation of formants [22], which are difficult to estimate for young children with high f_o. Additionally, the SGR estimation algorithm may further require an estimate of age [22]. The usage of f_o-based normalization may save kindergarten ASR systems computational resources, especially for systems that already compute f_o for other processes.

5. Conclusion

This study proposed an f_o-based frequency normalization method based on the tonotopic distances between formants and f_o, commonly used to model human vowel perception. The tonotopic distances were reformulated to be represented as a linear relationship (in the Mel scale) between f_o and formants. This reformulation was verified on child vowel productions. For the vowels /i/, /ɪ/, /ʌ/, and /ɑ/, least-squares regression lines between f_o and formants were shown to have slopes close to 1, and the data had a clear linear relationship.

The reformulation was further generalized by applying the linear relationship with f_o to the entire spectrum rather than just the formant locations. As such, the normalization method relied only on f_o as an additional parameter when extracting MFCCs. An ASR experiment with 1st-10th grade speech data for training and kindergarten speech data for testing was performed to compare the f_o-based normalization method to VTLN, $F3$-based, and SGR normalization. The f_o-based method had advantages in performance or computation over each of the other methods.

Several directions can be considered for future work. As most studies verified the relationships between formants and f_o without considering within-speaker effects, an investigation of speakers, both children and adults, producing vowels at varying f_o values may prove interesting. An initial pilot study on 3 adults (2 males, 1 female) suggests that speakers follow this relationship across several f_o values. Additionally, we only used the median f_o value of an utterance in this study. A further investigation of whether the tonotopic distances hold across a single speaker's range of f_o values could lead to a more effective per-frame normalization method.

Finally, we plan to apply these more effective child ASR systems to educational and clinical applications to investigate how they perform in realistic conversational situations. This may include the use of ASR to perform speech pathology assessments, assist in the teaching of reading and language skills, or identify social cues during interactions with children.

6. Acknowledgements

This work was supported in part by NSF Grant #1734380.

7. References

[1] D. Kewley-Port, C. S. Watson, M. Elbert, D. Maki, and D. Reed, "The Indiana Speech Training Aid (ISTRA) II: Training Curriculum and Selected Case Studies," *Clinical Linguistics and Phonetics*, vol. 5, no. 1, pp. 13–38, 1991.

[2] H. T. Bunnell, D. M. Yarrington, and J. B. Polikoff, "STAR: Articulation Training for Young Children," in *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, 2000, pp. 85–88.

[3] J. Tepperman, J. Silva, A. Kazemzadeh, H. You, S. Lee, A. Alwan, and S. Narayanan, "Pronunciation Verification of Children's Speech for Automatic Literacy Assessment," in *Proc. of INTERSPEECH*, 2006, pp. 845–848.

[4] G. Yeung, A. Afshan, K. E. Ozgun, K. Kaewtip, S. M. Lulich, and A. Alwan, "Predicting Clinical Evaluations of Children's Speech with Limited Data Using Exemplar Word Template References," in *Proc. of the Workshop on Speech and Language Technology in Education (SLaTE)*, 2017, pp. 161–166.

[5] R. Sadeghian and S. A. Zahorian, "Towards an Automated Screening Tool for Pediatric Speech Delay," in *Proc. of INTERSPEECH*, 2015, pp. 1650–1654.

[6] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations," in *Proc. of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017, pp. 82–90.

[7] G. Yeung and A. Alwan, "On the Difficulties of Automatic Speech Recognition for Kindergarten-Aged Children," in *Proc. of INTERSPEECH*, 2018, pp. 1661–1665.

[8] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A Review of ASR Technologies for Children's Speech," in *Proc. of the Workshop on Child, Computer and Interaction (WOCCI)*, 2009, pp. 7.1–7.8.

[9] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of Children's Speech: Developmental Changes of Temporal and Spectral Parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[10] ——, "Analysis of Children's Speech: Duration, Pitch and Formants," in *Proc. of EUROSPEECH*, 1997, pp. 473–476.

[11] H. K. Vorperian and R. D. Kent, "Vowel Acoustic Space Development in Children: A Synthesis of Acoustic and Anatomic Data," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 6, pp. 1510–1545, 2007.

[12] L. L. Koenig, J. C. Lucero, and E. Perlman, "Speech Production Variability in Fricatives of Children and Adults: Results of Functional Data Analysis," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3158–3170, 2008.

[13] L. L. Koenig and J. C. Lucero, "Stop Consonant Voicing and Intraoral Pressure Contours in Women and Children," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 1077–1088, 2008.

[14] B. L. Smith, "Relationships Between Duration and Temporal Variability in Children's Speech," *The Journal of the Acoustical Society of America*, vol. 91, no. 4, pp. 2165–2174, 1992.

[15] P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving Speech Recognition for Children Using Acoustic Adaptation and Pronunciation Modeling," in *Proc. of the Workshop on Child Computer Interaction (WOCCI)*, 2014, pp. 15–19.

[16] G. Stemmer, C. Hacker, S. Steidl, and E. Nöth, "Acoustic Normalization of Children's Speech," in *Proc. of EUROSPEECH*, 2003, pp. 1313–1316.

[17] X. Cui and A. Alwan, "MLLR-Like Speaker Adaptation Based on Linearization of VTLN with MFCC Features," in *Proc. of INTERSPEECH*, 2005, pp. 273–276.

[18] R. Serizel and D. Giuliani, "Vocal Tract Length Normalization Approaches to DNN-Based Children's and Adults' Speech Recognition," in *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 135–140.

[19] S. Panchapagesan and A. Alwan, "Multi-Parameter Frequency Warping for VTLN by Gradient Search," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006, pp. 1181–1184.

[20] S. S. Gray, D. Willett, J. Lu, J. Pinto, P. Maergner, and N. Bodenstab, "Child Automatic Speech Recognition for US English: Child Interaction with Living-Room-Electronic-Devices," in *Proc. of the Workshop on Child Computer Interaction (WOCCI)*, 2014, pp. 21–26.

[21] X. Cui and A. Alwan, "Adaptation of Children's Speech with Limited Data Based on Formant-Like Peak Alignment," *Computer Speech and Language*, vol. 20, no. 4, pp. 400–419, 2006.

[22] J. Guo, R. Paturi, G. Yeung, S. M. Lulich, H. Arsikere, and A. Alwan, "Age-Dependent Height Estimation and Speaker Normalization for Children's Speech Using the First Three Subglottal Resonances," in *Proc. of INTERSPEECH*, 2015, pp. 1665–1669.

[23] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large Vocabulary Automatic Speech Recognition for Children," in *Proc. of INTERSPEECH*, 2015, pp. 1611–1615.

[24] R. Serizel and D. Giuliani, "Deep Neural Network Adaptation for Children's and Adults' Speech Recognition," in *Proc. of the First Italian Computational Linguistics Conference (CLiC-it)*, 2014.

[25] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-Regression Hidden Markov Model," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001, pp. 513–516.

[26] A. Ljolje, "Speech Recognition Using Fundamental Frequency and Voicing in Acoustic Modeling," in *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 2137–2140.

[27] M. Magimai-Doss, T. A. Stephenson, and H. Bourlard, "Using Pitch Frequency Information in Speech Recognition," in *Proc. of EUROSPEECH*, 2003, pp. 2525–2528.

[28] A. Faria and D. Gelbart, "Efficient Pitch-Based Estimation of VTLN Warp Factors," in *Proc. of INTERSPEECH*, 2005, pp. 213–216.

[29] S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-Adaptive Front-End Features for Robust Children's ASR," in *Proc. of INTERSPEECH*, 2016, pp. 3459–3463.

[30] A. K. Syrdal and H. S. Gopal, "A Perceptual Model of Vowel Recognition Based on the Auditory Representation of American English Vowels," *The Journal of the Acoustical Society of America*, vol. 79, no. 4, pp. 1086–1100, 1986.

[31] H. Traunmüller, "Perceptual Dimension of Openness in Vowels," *The Journal of the Acoustical Society of America*, vol. 69, no. 5, pp. 1465–1475, 1981.

[32] R. P. Fahey, R. L. Diehl, and H. Traunmüller, "Perception of Back Vowels: Effects of Varying F1-F0 Bark Distance," *The Journal of the Acoustical Society of America*, vol. 99, no. 4, pp. 2350–2357, 1996.

[33] S. Barreda and T. M. Nearey, "The Direct and Indirect Roles of Fundamental Frequency in Vowel Perception," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 466–477, 2012.

[34] P. F. Assmann and T. M. Nearey, "Relationship Between Fundamental and Formant Frequencies in Voice Preference," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. EL35–EL43, 2007.

[35] G. Yeung, S. M. Lulich, J. Guo, M. S. Sommers, and A. Alwan, "Subglottal Resonances of American English Speaking Children," *The Journal of the Acoustical Society of America*, vol. 144, no. 6, pp. 3437–3449, 2018.

[36] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The OGI Kids' Speech Corpus and Recognizers," in *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, 2000, pp. 258–261.

[37] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving Deep Neural Network Acoustic Models Using Generalized Maxout Networks," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 215–219.

[38] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi Speech Recognition Toolkit," in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[39] L. N. Tan and A. Alwan, "Multi-Band Summary Correlogram-Based Pitch Detection for Noisy Speech," *Speech Communication*, vol. 55, no. 7-8, pp. 841–856, 2013.