A LONGITUDINAL STUDY OF INFANT SPEECH PRODUCTION PARAMETERS: A CASE STUDY

Hynek Bořil[1], *John H. L. Hansen*[1], *Dongxin Xu*[2], *Gill Gilkerson*[2], *Jeff Richards*[2]

[1]CRSS, Erik Jonsson School of Engineering & Computer Science, University of Texas at Dallas, USA
[2]LENA Foundation, Boulder, Colorado, USA

ABSTRACT

Recent studies demonstrate the potential and importance of children's speech processing in the detection of language delay and early communication disorders, automated reading tutoring, or emotional state assessment. In order to design and improve performance of such applications, good understanding of the children's speech structure and its development over time is necessary. This study tries to contribute to this domain by analyzing speech production parameters in longitudinal data acquired from female infant subject within the interval of 11 to 35 months of age, sampled with a step of 4 months. The analyzed parameters include fundamental frequency, formants, vocal tract length, average spectrum, spectral slope, duration of voiced segments, number of voiced segments in conversational turn, and recently proposed pitch micro-contour patterns. In addition, a simple automatic age classifier utilizing cepstral coding of speech and Gaussian mixture models (common techniques in automatic speech processing) is designed and shown to reach an accuracy of 70.6 % in the speaker-dependent open test task. The capability of the classifier to distinguish between the individual age models demonstrates that current speech acoustic modeling schemes used in automated systems are strongly age-dependent. It is believed that understanding the speech production changes during growing, combined together with a reliable age detection will strongly benefit the design of robust children speech processing systems.

***Index Terms*—** Infant speech production, longitudinal data, pitch patterns, age classification

1. INTRODUCTION

Speech-based interfaces represent an attractive way of human–machine interaction, with an application in automated online reservations, information retrieval, voice controlled home appliances, or automated dictation. Speech processing technology contributes also to the domains of security and forensics (speaker identification), and medical research (cochlear implants). While the major portion of speech research has been focused on adults, recent studies show the potential of children's speech processing in tasks such as the detection of language delay [1] and early communication disorders [2], automated reading tutoring (detection of reading miscues, text comprehension) [3], or emotional state assessment [4]. Current speech systems are prevalently targeted on adult speakers and exhibit a performance reduction when exposed to children due to the acoustic differences in the children and adult speech production [5]. To improve accuracy and robustness of such systems, it is necessary to understand and address the specifics of children speech and its development over time.

Analyses of children speech development over time utilize either *cross-sectional* or *longitudinal* data [6]. In the first case, samples are collected in a certain time instant from multiple subjects of different ages [7], [8]. In the latter case, multiple samples per each subject are collected over a certain time period [9], [10]. The *cross-sectional* approach allows for fast data acquisition at the cost of possible increased variance of the observed trends, as each time sample comes from different speaker. The *longitudinal* approach allows for carrying analysis for each individual speaker, eliminating the inter-speaker variance. On the other hand, collection of such data is very time consuming.

This study presents initial analysis of speech production parameters on a small subset of a children speech database acquired by the LENA Foundation. The database comprises more than 65.000 hours of recorded data. Subjects in the database are recorded starting from 2 months through 36 months of age [1]. The recordings are obtained in 16 hours sessions while the children carry the recording system in the pocket of a custom-made clothing [11]. Each recording is conducted during an 'ordinary' day in the child's natural environment. The subset analyzed in this study comprises longitudinal data from a healthy female infant acquired within the interval of 11 to 35 months of age, with a sampling step of 4 months. The analyzed parameters include fundamental frequency, formants, vocal tract length, average spectrum, spectral slope, duration of voiced segments, number of voiced segments in conversational turn, and recently proposed pitch micro-contour patterns [12]. In addition, simple age classifier that exploits the age dependency of speech parameter distributions is proposed and evaluated in the open test set task.

The remainder of the paper is organized as follows. First, results of the speech production analyses are presented are confronted with the literature. Second, an age classifier is proposed. Efficiency of several front-end processing schemes are evaluated and compared. Final part summarizes outcomes of the study.

2. ANALYSIS OF SPEECH PRODUCTION PARAMETERS

The analyzed data set comprises 7 recordings acquired at the age of 11, 15, 19, . . ., 35 months. For each age sample, segments containing child's utterances were selected, yielding 5 minutes of speech per age. The recordings are stored in 16 kHz/16 bit format.

2.1. Fundamental Frequency

Past studies on average fundamental frequency (F_0) in children observed the following:

- Age 0–5 months: F_0 increases in both cry and non-cry utterances [13],
- Age 0–12 months: F_0 increases in hunger cries [14],
- Age 3, 6, 9 months: F_0 slightly increases at 6 months, followed by slight decrease at 9 months,

- Age 8–26 months: not significant changes of F_0 in monosyllables and bisyllables,
- Age 8 months–3.5 years: difficult to track any significant longitudinal trends in F_0 [7],
- Age 5–17 years: males – significant F_0 drop starting from 11 to 15 years, no significant change later; females – significant drop at 7–12 years, no significant change later [15],
- Age 8.5–11.5 years: F_0 in females decreases.

In our study, F_0 was extracted using cross-correlation RAPT algorithm [16]. As shown in Fig. 1, average F_0 displays an increasing trend in the interval of 11–35 months (slope $a = 1.64$ Hz/month, correlation coefficient $R^2 = 0.56$). Note that the vertical bars represent 95% confidence intervals.

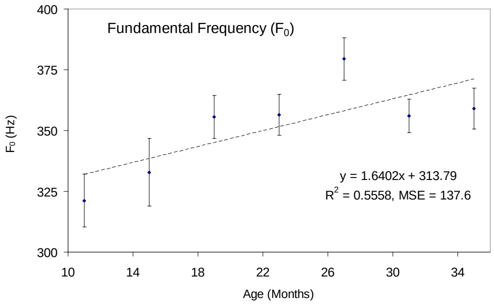

Fig. 1: Average fundamental frequency as a function of age. Vertical bars – 95% confidence intervals.

2.2. Formants

The following can be found in the literature on average formant frequencies:

- Age 4–60 months: F_1 decreases in females, F_1 decreases in males till approx. 30 months; F_2 decreases in both females and males in 4–18 months for some vowels, increase in other vowels [17],
- Age 8–18 months: F_1 decreases between 10–19 months in Canadian French speakers, no clear trend in Canadian English; F_2 decreases in Canadian English in 10–18 months, rather steady in Canadian French [18],
- Age 15–36 months: F_1, F_2 relatively unchanged till 24 months; significant decrease in 24–36 months [19],
- Age 5–16 years: F_1, F_2, F_3 decrease in most of the analyzed vowels [15],
- Age 8.5–11.5 years: F_1 mostly decrease between 10 and 11.5 years, F_2 mostly decreases between 8.5–10 years, F_3 consistently decreases in 8.5–11.5 years [7],

Formant center frequencies in our study were estimated using an algorithm that combines linear prediction of spectral envelope and dynamic programming [20]. As shown in Fig. 2, average F_2 displays descending trend with age. Similar was observed also for F_1 and F_3 (F_1: slope $a_{F1} = -2.21$ Hz/month, correlation coefficient $R^2 = 0.22$; F_2: $a_{F2} = -8.56$ Hz/month, $R^2 = 0.71$; F_3: $a_{F3} = -5.90$ Hz/month, $R^2 = 0.48$). Note that F_2 trend is steepest and most linear over time. It can be seen that the correlation coefficients of the F_2 and F_3 trends are considerably higher than in F_1. The reduction in average F_1–F_3 over time is somewhat intuitive since the formant frequencies (vocal tract resonances) are inversely proportional to the vocal tract length (which extends with age).

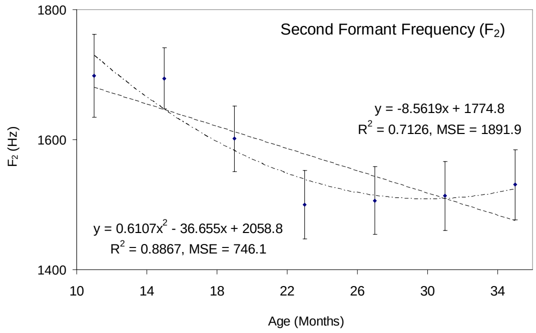

Fig. 2: Average second (F_2) formant frequency as a function of age. Vertical bars – 95% confidence intervals.

2.3. Vocal Tract Length

Studies employing resonance imaging (MRI) report typical values of vocal tract length (VTL) ranging from 7 cm in new-born babies to 13–18 cm in female and male adults, and an accelerated VTL growth between birth and 18 months [21]. The location of higher formant frequencies is believed to be more correlated with VTL compared to the lowest formant frequencies, which are strongly dependent on the rate of the jaw opening and vertical position of the tongue [22]. In the original version of the vocal tract length normalization (a popular technique used for normalizing inter-speaker vocal tract differences to improve accuracy of automatic speech recognition) [23], the warping parameter proportional to VTL differences is estimated from the inverse of higher formant frequencies; in particular, F_3 is typically used [24].

In our study, F_3 is chosen as the parameter for the estimation of the VTL changes. While VTL varies with the production of distinct phones and so does F_3, we assume that averaging F_3 over longer speech segments (utterances) will provide a reasonable estimation of the average, phone independent VTL. The inverse of average F_3 over time is shown in Fig. 3. It can be seen that $1/F_3$ displays fast

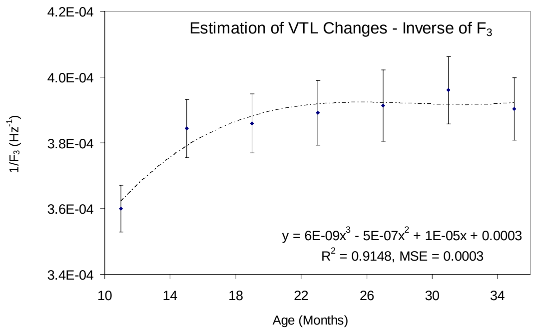

Fig. 3: Inverse of third formant.

increase between 11 and 15 months, while later, till 27 months, remains almost steady. Despite the obvious accuracy limitation of this technique compared to MRI, the observed trend seems to correlate well with the accelerated VTL growth reported for the early months after birth [21].

2.4. Average Spectrum and Spectral Slope

In this section, average spectra of the infant subject are compared with the average spectrum extracted from the spontaneous speech of 23 US native adult female subjects from the UTScope database [25] (average age $\mu = 22.2$ years, standard deviation $\sigma = 3.6$ years). For all sets, the average spectrum is extracted from a 25 ms window shifted with a skip rate of 10 ms. In addition, complementary average spectral slopes are extracted on the frame basis by fitting a straight line into the short term amplitude spectra in log frequency–log amplitude domain by means of linear regression. It can be observed that with increasing age, the number of local minima and maxima in the spectral envelope decreases (envelope smoothing) and the contours slowly approach those of adult speakers (see Fig. 4). Spectral slope in the infant (see Table 1; '*Duration*' stands for the total duration of voiced speech segments in the age sample data) is considerably flatter than the one usually seen in adult speakers (typically around -6 to -10 dB/oct), and its tilt increases with age, progressing towards adult values.

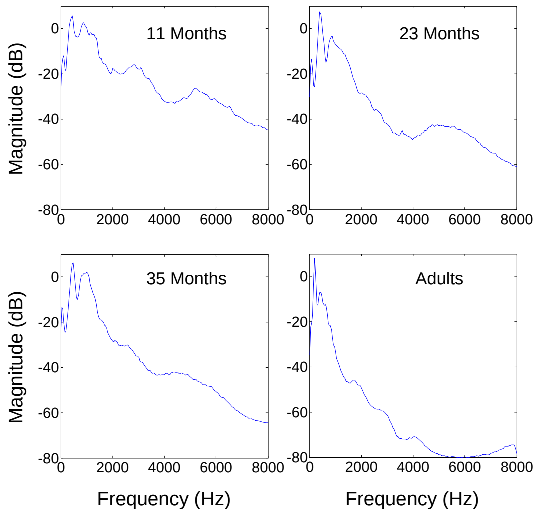

Fig. 4: Average spectra of 11, 23, and 35 months old subject versus average spectrum of adult female speakers extracted across 23 subjects.

Age (Months)	Duration (s)	Slope (dB/oct)	σ_{Slope} (dB/oct)
11	117.4	-2.0	1.9
15	89.6	-2.4	1.6
19	108.1	-2.4	1.6
23	99.1	-3.2	1.5
27	119.5	-3.9	1.6
31	152.4	-2.8	1.6
35	105.3	-3.5	1.3

Table 1: Average spectral slope in voiced speech segments.

2.5. Voiced Segment Durations and Frequencies

It has been observed that in children (typically 5 years and older), the duration of vowels, consonants [26], syllables [10], and sentences [15] tend to reduce with age and approach the one seen in adults. In our study, the duration of voiced speech segments is analyzed. Voiced segments are formed by a non-interrupted sequence of voiced frames. Voiced frame is identified using the output of the F_0 tracking algorithm; each frame that is assigned non-zero F_0 value is considered voiced. To obtain a further insight into the structure of the infant phrases, also the average number of voiced segments per conversational turn is counted (see Table 2). It can be seen that unlike reported for older children, the durations tend to increase for our subject, while the number of voiced segments per conversational turn either remains steady or reduces.

	Age (Months)						
Average	11	15	19	23	27	31	35
Segment Duration (s)	0.31	0.22	0.32	0.33	0.32	0.43	0.37
# Segments/Turn	2.3	2.5	2.0	1.8	1.8	2.4	2.1

Table 2: Average voiced segment duration and average number of voiced segments in conversation turn.

2.6. Pitch Micro-Contour Patterns

In [8], inspection of F_0 contours during vocalization in infants revealed frequently repeating intonation patterns. The authors defined 7 basic pattern shapes where each pattern was requested to have duration longer that 100 ms but shorter than 2 s. The patterns were subsequently used for analyzing pitch contours by means of counting frequency of the pattern occurrences. Recently, a novel pattern-

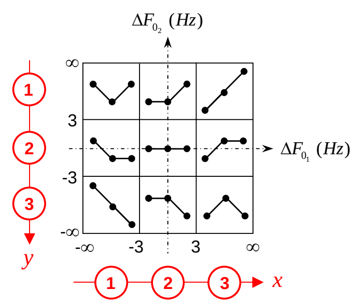

Fig. 5: Codebook of pitch micro-contour patterns.

based approach to pitch contour analysis and modeling was proposed in [12] for the purpose of dialect distance assessment. Similarly to [8], a small set of elementary patterns is defined, see Fig. 5. Unlike in [8], the patterns are not used to interpolate a large number of pitch samples but are rather fit into adjacent pitch sample values in order to describe the pitch contour micro-structure. By using common language modeling techniques, pattern bigram probabilities (a probability that a given pattern will be followed by another particular pattern) as well as higher order probabilities can be calculated and used to model complex pitch contour macrostructure from the elementary pattern units. In the present study, only the frequency of elementary patterns occurrence is analyzed and compared across age sets (see Fig. 6). Note that the x and y coordinates in Fig. 6

correspond to the x and y coordinates in circles in Fig. 5. For example, the center bar in Fig. 6 has coordinates [2; 2] and corresponds to the 'flat' pattern, while the bar in the bottom-right corner of the plots in Fig. 6 corresponds to the reversed 'V' pattern. It can be seen

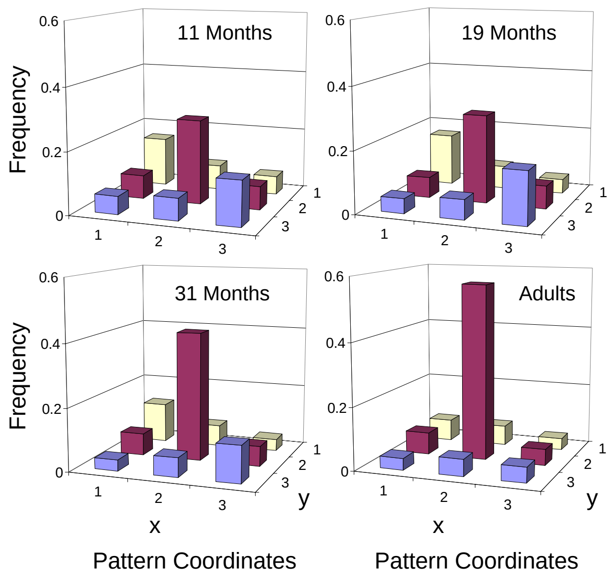

Fig. 6: Normalized histograms of children's pitch micro-contour patterns versus adult histogram extracted across 23 female subjects.

that while in the adult histogram the 'flat' pattern dominates, the pattern distribution in early infant speech is way more uniform. With increasing age, the 'flat' pattern becomes gradually more and more pronounced and the whole distribution approaches the one seen in adults.

3. AUTOMATIC AGE CLASSIFICATION

The previous sections have demonstrated the impact of age on the infant speech production. Intuitively, the observed production variations can be expected to have a direct impact on coding commonly used in speech systems. To evaluate this hypothesis, we train separate acoustic models for each age sample set, yielding 7 models representing 11-months speech, 15-months speech, . . ., 35-months speech. If the speech coding is age-sensitive, the models will capture age-dependent speech characteristics and can be used for speaker-dependent automatic age classification.

In the design of the age classifier, Gaussian mixture models (GMM's) are chosen to represent speech probability density functions (pdf's). The probability of speech observation vector $\mathbf{o}_t$ being generated by the j-th age GMM is calculated as:

$$b_j\left(\mathbf{o}_t\right) = \sum_{m=1}^{M} \frac{c_{jm}}{\sqrt{\left(2\pi\right)^n \left|\mathbf{\Sigma}_{jm}\right|}} \cdot \mathrm{e}^{-\frac{1}{2}\left(\mathbf{o}_t - \mu_{jm}\right)^T \mathbf{\Sigma}_{jm}^{-1}\left(\mathbf{o}_t - \mu_{jm}\right)}, \tag{1}$$

where m is the index of the Gaussian mixture component, M is the total number of mixtures, c_{jm} is the mixture weight such that:

$$\sum_{m=1}^{M} c_{jm} = 1, \tag{2}$$

n is the dimension of $\mathbf{o}_t$, $\mathbf{\Sigma}_{jm}$ is the mixture covariance matrix, and μ_{jm} is the mixture mean vector. During the age classification task, the sequence of acoustic observations extracted from the incoming utterance is scored against all acoustic models and the model that maximizes the probability in Eq. (1) is selected.

Efficiency of several speech coding strategies is compared. Common Mel frequency cepstral coefficients (MFCC) [27], perceptual linear prediction (PLP) cepstral coefficients [28], their alternatives with altered LPC–DCT stages denoted 'MFCC LPC' and 'PLP DCT', and Expolog-based features [29] (PLP front-end with Expolog filterbank – 'Expolog LPC'; MFCC front-end with Expolog filterbank – 'Expolog DCT') are evaluated. In each setup, static (c_0–c_{12}), dynamic, and acceleration coefficients form the feature vector. To allow for classifier training and evaluation on separate data sets, the available data for each age spot are divided into two halves, one assigned to the train set, other to the open test set.

In the initial classification trial with the MFCC front-end, the age classification accuracy reached $Acc = 52.2\,\%$, proving the age dependency of speech coding and acoustic models (chance = 1/7 x 100 = 14.3 %). Subsequently, in an effort to obtain more meaningful system, the task was modified from 'pick 1 age from 7 possibilities' to 'select 2 adjacent neighbors from 6 possible neighbor pairs'. In this case, the classification is successful if one of the selected model pairs matches the actual age in the test data. In this task, the MFCC system performance increased to 68.8 %. Performance of all tested setups in the '2 neighbors' task is compared in Table 3. It can be seen that the best performance is reached by the 'PLP DCT' system ($Acc = 70.6\,\%$).

Front-End					
MFCC	MFCC LPC	PLP	PLP DCT	Expolog LPC	Expolog DCT
68.8	66.0	64.6	**70.6**	67.1	64.1

Table 3: Comparing front-end performance in age classification.

4. CONCLUSIONS

This study has presented an initial analysis of speech production development in a healthy female infant subject conducted on the longitudinal data spanning 11–35 months of age. Trends in the reduction of formant frequencies and extending vocal tract length with age confirm our intuition and observations presented in the literature. On the other hand, the observed gradual increase of F_0 after the 12 months of age, together with the increase of average voiced segment duration and reduction of voiced segments in conversational turns are somewhat surprising. Novel approach to children pitch contour analysis and modeling utilizing a codebook of pitch micro-contour patterns has been presented and shown to capture additional aspects of speech production maturing. Finally, age dependency of common speech coding strategies has been exploited in the design of a speaker-dependent automatic age classifier.

Some of the observed differences between the child's and adult speech production can be directly utilized in improving current speech processing techniques. For example, linear prediction-based (LP) spectral analysis typically incorporates a pre-emphasis of +6 dB/oct in order to compensate for the average spectral tilt in neutral adult speech. As shown in this paper, LP analysis conducted on children should use pre-emphasis that will compensate for the actual, flatter spectral tilt.

In the next step, the gradual development in the infant's speech sounds' articulation will be studied. Groups of syllable-like and phone-like sounds will be searched using stochastic modeling and unsupervised clustering techniques. It is expected that the analysis may bring a further insight into the early language acquisition process.

References

[1] D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, and J. H. L. Hansen, "Signal processing for young child speech language development," in *1^{st} Workshop on Child, Computer, and Interaction*, Chania, Greece, Oct. 2008.

[2] P. Zlatník and R. Čmejla, "Disordered speech assesment using different speech parameterizations," in *19th International Congress on Acoustics*, Madrid, Spain, 2007, pp. 1–4.

[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. ICSLP'06*, Pittsburgh, PA, USA, 2006, pp. 845–848.

[4] C. Lee S. Yildirim, S. Lee, A. Potamianos, and S. Narayanan, "Detecting politeness and frustration state of a child in a conversational computer game," in *Proc. EUROSPEECH'05*, Lisbon, Portugal, 2005, pp. 2209–2212.

[5] A. Hagen, B. Pellom, and R. Cole, "Highly accurate children's speech recognition for interactive reading tutors using subword units," *Speech Commun.*, vol. 49, no. 12, pp. 861–873, 2007.

[6] Shigeaki Amano, Tadahisa Kondo, Kazumi Kato, and Tomohiro Nakatani, "Development of japanese infant speech database from longitudinal recordings," *Speech Communication*, vol. 51, no. 6, pp. 510 – 520, 2009.

[7] P. Keating and R. Buhr, "Fundamental frequency in the speech of infants and children," *The Journal of the Acoust. Soc. of America*, vol. 63, no. 2, pp. 567–571, 1978.

[8] R. D. Kent and A. D. Murray, "Acoustic features of infant vocalic utterances at 3, 6, and 9 months," *The Journal of the Acoust. Soc. of America*, vol. 72, no. 2, pp. 353–365, 1982.

[9] M. P. Robb, J. H. Saxman, and A. A. Grant, "Vocal fundamental frequency characteristics during the first two years of life," *The Journal of the Acoust. Soc. of America*, vol. 85, no. 4, pp. 1708–1717, 1989.

[10] B. L. Smith and M. K. Kenney, "A longitudinal study of the development of temporal properties of speech production: Data from 4 children," *Phonetica*, vol. 56, no. 1-2, pp. 73–102, 1999.

[11] S. S. Gray, C. T. Baer, D. Xu, and U. Yapanel, "The LENA^{TM} language environment analysis system: The Infoture time segment (ITS) file," ITR-04-1, Infoture, Inc., Boulder, CO, 2007.

[12] M. Mehrabani and J. H. L. Hansen, "Automatic dialect separation assessment," *submitted to IEEE Trans. on Audio, Speech, and Language Processing*, 2009.

[13] W. C. Sheppard and H. L. Lane, "Development of the Prosodic Features of Infant Vocalizing," *J. Speech Hear. Res.*, vol. 11, no. 1, pp. 94–108, 1968.

[14] H. R. Gilbert and M. P. Robb, "Vocal fundamental frequency characteristics of infant hunger cries: birth to 12 months," *Int. Jour. of Pediat. O-R-L*, vol. 34, no. 3, pp. 237 – 243, 1996.

[15] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoust. Soc. of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[16] D. Talkin, *Speech Coding and Synthesis*, chapter A Robust Algorithm for Pitch Tracking (RAPT). W.B. Kleijn and K.K. Paliwal (Eds.), pp. 495–518, Elsevier, Amsterdam, Netherlands, 1995.

[17] K. Ishizuka, R. Mugitani, H. Kato, and S. Amano, "Longitudinal developmental changes in spectral peaks of vowels produced by Japanese infants," *The Journal of the Acoust. Soc. of America*, vol. 121, no. 4, pp. 2272–2282, 2007.

[18] L. Polka, S. Rvachew, and K. Mattoc, *Blackwell Handbook of Language Development*, chapter Experiential Influences on Speech Perception and Speech Production in Infancy. E. Hoff and M. Shatz (Eds.), pp. 153–172, Blackwell Pub., 2006.

[19] H. R. Gilbert, M. P. Robb, and Y. Chen, "Formant frequency development: 15 to 36 months.," *Journal of Voice*, vol. 11, no. 3, pp. 260–266, 1997.

[20] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *JASA*, vol. 82, no. S1, pp. S55–S55, 1987.

[21] H. K. Vorperian, R. D. Kent, M. J. Lindstrom, C. M. Kalina, L. R. Gentry, and B. S. Yandell, "Development of vocal tract length during early childhood: A magnetic resonance imaging study," *The Journal of the Acoustical Society of America*, vol. 117, no. 1, pp. 338–350, 2005.

[22] R. D. Kent and C. Read, *The Acoustic Analysis of Speech*, Whurr Publishers, San Diego, 1992.

[23] H. Wakita, "Normalization of vowels by vocal-tract length and its application to vowel identification," *IEEE Trans. on Ac., Speech, and Sig. Processing*, vol. 25, no. 2, pp. 183–192, Apr 1977.

[24] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. of ICASSP'96*, Los Alamitos, CA, USA, 1996, vol. 1, pp. 346–348, IEEE Computer Society.

[25] V. Varadarajan and J. H.L. Hansen, "Analysis of Lombard effect under different types and levels of noise with application to in-set speaker ID systems," in *Proc. of Interspeech'06*, Pittsburgh, Pennsylvania, 2006, pp. 937–940.

[26] M. Gerosa, S. Lee, D. Giuliani, and S. S. Narayanan, "Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition.," in *IEEE ICASSP'06*, 2006, vol. 1, pp. 393–396.

[27] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. 28, no. 4, pp. 357–366, 1980.

[28] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoust. Soc. of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[29] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on Speech & Audio Processing*, vol. 8, no. 4, pp. 429–442, 2000.