Assessing the Stress/Neutral Speech Environment in Adult/Child Interactions for Applications in Child Language Development

Sanjay A. Patil §
Center of Robust Speech Systems
The University of Texas at Dallas
Richardson, TX, USA
Sanjay.Patil@utdallas.edu

John HL Hansen §
Center of Robust Speech Systems
The University of Texas at Dallas
Richardson, TX, USA
John.Hansen@utdallas.edu

Gill Gilkerson†, Sharmi Gray†, Doungxin Xu†
Infoture Foundation.

ABSTRACT

It is known that for effective child language development, the number of adult words heard and adult-child exchanges in the early phase (8-20 months) is important. Language development can be represented in terms of adult word count (AWC) and conversational turns (CT) between the adult and child. The focus of this study is to investigate if perceived "stress" in the adult speech side of these exchanges impacts AWC or CTs, thus potentially impacting a child's language acquisition skills. We propose to develop a scheme to detect the presence of stress in the adult side of child-adult audio streams and relate this with metrics available for assessing language development. The proposed approach represents the first attempt to assess child-adult interactions from a stress/neutral assessment approach where recordings are monitored continuously for 12-hour periods of time. Here, a proposed speaking rate measure based on the utterance length (UL /AWC) shows a statistical correlation with stress levels, with male adults showing more significance as compared to female adults. Thus, adults increase speaking rate when under stress which impacts their ability to convey articulation details, and therefore a potential negatively impacting child language acquisition.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human factors, Human information processing.

General Terms

Algorithms, Measurement, Human Factors, Standardization.

Keywords

Stress Speech Detection, Adult Word Count (AWC), Conversational Turns (CT), Utterance length (UL), Child Language Development

1. INTRODUCTION

It is well known that a number of environmental factors impact effective speech communication between speakers. This is of great importance in child-adult communication since it influences child language development. Early child language acquisition requires a rich verbal and language environment. Greater conversational turns between adult and child will also have a positive influence on a child's language development. It has been shown that the adult stress state does affect the child's speech [4], [7], [11]. Wexler found that 2-year olds have a higher frequency of average oscillations (number of repetitions per instance of disfluency) and higher frequency of disrhythmic phonations as compared to other groups, under stressful conditions [11]. Hart and Risley conducted a longitudinal research study over a 10-year period [3]. Their study concluded that children who heard more words from birth to age 36 months had more sophisticated language skills than children who do not hear as many words. It is suggested that a rich adult-child language environment would have the adult produce somewhere between 15,000 to 30,000 words per day(WPD) in the presence of the child [3].

Since Hart and Risley emphasize the importance on the child language environment for the first 36 months, it is suggested that the stress component in adult-child interactions may also play a major role in child language acquisition and development. Though research with simulated stress conditions have been previously performed, this study is the first involving natural, spontaneous home environment based evaluations. As in many homes, much of the adult-child interaction occurs during routine events such as meal times, dressing and playtime. Hence, the test validity of the study rests on experimenting with real-life scenarios and stress-related situations. Therefore, having an extended recording, over a contiguous period of 12-hours, of the adult-child interactions at their homes provides a unique window into the home with minimal modifications of the actual environment. This study represents the first attempt to correlate the adult's stress level with adult word count and conversational turns in adult-child dialogs.

Due to a lack of clear verbal expressions of stress by infants, our study focuses on detecting the presence of adult stress. We define the adult stress state into two broad categories - neutral and stress; where neutral state includes plain interaction (interactions void of any stress), joy, happy state of emotion, and where the stress state includes all negative emotions - anger, frustration, sadness. In our study, the indicator of adult interaction with the child is defined in terms of AWC, and CT [6].

§ This Project was funded by The University of Texas at Dallas under project EMMITT.

† Collaborators from Infoture Foundation, JillGilkerson@infoture.org

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ICMI-MLMI'09 Workshop on Child, Computer and Interaction. November 5, 2009, Cambridge, MA, USA

Copyright 2009 ACM 978-1-600558-690-8/09/11 …$10.00

The paper is organized as follows; Sec. 2 describes details on child language acquisition. Sec. 3 deals with the speech database used for algorithm evaluation and algorithm details (Sec. 4), followed by results(Sec. 5). The last section suggests follow-up experiments and their implications.

2. CHILD LANGUAGE ACQUISITION

An effective environment for child language learning for production of vocal sounds requires supportive / conducive conditions, having a playful scenario, along with a high degree of interaction with the care-giver [2], [9]. Some studies indicate that infants first respond to suprasegmental features of speech, and only later are able to discriminate segmental (e.g., vowel and consonant) aspects of speech [7], [12]. Studies show that children can improve verbal skills in an environment where adults talk with the child, as well as between themselves in the presence of the child [3]. Research elsewhere also indicates that the child learns to acquire language skills much earlier than before they are able to speak. During interaction with the child, the adult speech will contain their stress state. Most of the time, this is a positive state such as excitement, joy, or happiness when the child is happy, playful, or when the child is trying to repeat sounds/words, or when the child is well-behaved. On other occasions, it is likely that the adult will express a negative stressor, mostly when facing time-constraints, frustrated, angry, or when the child misbehaves. Hence, the positive stress states expressed are happiness, joy (termed as neutral), and the negative stress states are frustration, anger, irritation (termed as stress). It is believed these stress states will affect the adult-child interaction and hence child's language acquisition. Some of the terms that will be used are explained below:

Adult Word Count (AWC): The adult word count is the estimate of the number of words spoken in each adult segment after filtering portions of the signal containing non-speech sounds such as coughing or throat clearing. This represents the number of adult words a child hears per day, per hour or per utterance. For our study, it is words per utterance.

Conversationsal Turns (CT): Conversational turns are defined as alternating exchanges between the child and adult. Further, a conversation is defined as a contiguous region containing live human speech (recordings such as radio or TV excluded). A pause of 5 seconds or longer separates two conversation turns.

Utterance Length (UL): The time length (in sec.) for a single conversation is defined as the utterance length.

Utterance Length / Adult Word Count (UL/AWC): The ratio of utterance length to adult word count for every conversation defined in terms of sec./words is defined as UL/AWC. Though it can be thought of as reciprocal of AWC/UL but UL/AWC must be interpreted as time taken to speak a single word.

It is noted that in the field of speech signal processing, virtually all data collection is obtained in controlled laboratory conditions (for training data, and often for testing as well). This study employs a mobile recording unit (LENA[[6]), which is worn by the infant for up to 12 hour periods. The audio data is recovered on a daily basis, and all speech signal processing performed off-line. This is the first study employing such long-term adult-child interactions.

3. CORPUS

3.1 Infoture natural language corpus

The Infoture Natural Language Corpus is an extensive database of quantifiable adult-child speech produced in natural home environments [6]. The speech recordings contain adult-child interactions from families having a range of socioeconomic status, with infants spanning 2-36 months of age. There are over 3,000 12-16 hour recording sessions from over 460 families resulting in a total of over 61,000 hours of recordings. The corpus is a balanced for infant gender (male /female) and age (2-36 months).

Data collection is performed using an infant body-worn recording unit that contains an omnidirectional microphone with a flat 20-20kHz frequency response with a sample frequency of 16kHz, single-channel 16-bit PCM recording [6]. This represents 3.5 Trillion speech samples, occupying approximately 20TB of diskspace.

3.2 Core experiment corpus

We note that while a significantly larger amount of data is available, due to the human subjects IRB requirements, this study focused on a core database from 20 families, 12 hours per family, totaling 320 hours of data rich in speech content with all listener evaluations performed at Infoture, Inc. The focus here is on utterances addressed to the infant, hence utterances involving adult-adult conversations, and non-adult speech audio streams from recordings (e.g., TV and radio) were excluded via an automatic speech processing scheme and verified afterwards.

It should be noted that it is difficult to assess stress during adult-child interaction, since the adult can use a motherese tone while talking to their child (i.e., a speaking style which results in a 2-4 times increase in mean pitch f0, even though the adult is still in a neutral state - non-stress state). Thus, the pitch and other speech attributes are very different as compared to when two adults are conversing.

Table 1. Duration distribution for 20-recordings - most ([52,642/61,479]=85.62%) are less than 2 sec

Duration (sec)	**Number of Utterances**
< 1 sec	**8 077**
1 sec	**12 124**
(1-2 sec)	**32 441**
(2-5 sec)	8 225
(5-10 sec)	559
> 10 sec	53
Total	**61 479**

4. ALGORITHM

The first phase of algorithm processing is to establish the adult-word count (AWC) and conversational turns (CT). Here, each recording of over 12-hours is processed through the LENA (Language ENvironment Analysis) software to obtain female and male adult tags along-with the CT and AWC [6]. Each utterance within the stream which is more than 2 seconds long is perceptually scored by a human classifying it into one of the three categories – neutral (NX), stress (SX) or undecided (UX), where the undecided category being somewhere between the two extremes. As seen from Table 1, over 52,000 utterances of the total 61,479 are very short in duration (_ 2 sec). So, it was decided to have human experts classify the 2193 utterances (> 2 sec long) into one of three categories – NX (801 utterances), SX (730 utterances), or UX (662 utterances) [See Table 2].

Table 2. Count and Duration (sec) distribution of stress categories (neutral: NX, Stress:SX, undecided:UX) for male, female adults perceptually labeled by listeners at Infoture Foundation

	NX		SX		UX	
Gender	**Count**	**Durn**	**Count**	**Durn**	**Count**	**Durn**
Male	361	1136.8	173	512.1	190	571.9
Female	440	1297.7	557	1585.0	472	1393.2
Total	801	2424.5	730	2097.1	662	1965.1

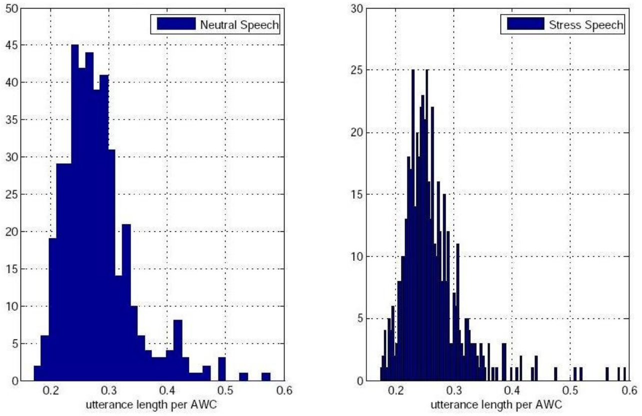

Figure 1. Histogram of (UL/Adult word count – sec/words) for female adults – NX and SX with mean value showing statistical significance to stress with 95% confidence.

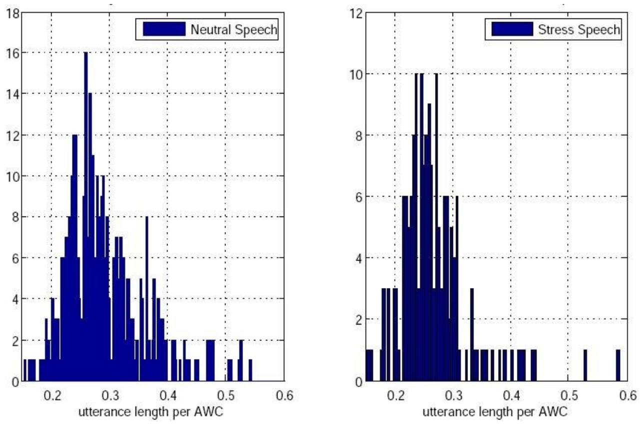

Figure 2. Histogram of (UL/Adult Word Coount - sec/words) for male adults - NX and SX with mean value showing statistical significance to stress with 99% confidence.

In the second phase of algorithm processing, we turn to stress/neutral speech detection. Here, 19-dimensional MFCC features are extracted using HTK with standard settings of 20ms window, 10ms skip, _ = 0.97, and a factor of 0.2 for normalized energy thresholding [5]. Using a five-fold cross validation scheme, 80% of the data under NX and SX category is used for building Gaussian mixture models (GMMs) with 128 mixtures (NX GMM model and SX GMM model), while the remaining 20% is used for test. The binary decision for the test utterance is either stress or neutral and is obtained by comparing the difference of log-likelihood scores with a threshold [10]. The results are averaged over the validation.

Table 3. Stress detection accuracy (%) for female, male adult and combined averaged over 5-fold cross validations - female change their style of speech with stress, higher stres detection accuracy as compared to males

Gender	Accuracy	Raw Results
	(%)	(corrects/Total)
Female	72.28	365/505
Male	67.53	260/385
Combined	70.22	625/890

5. ANALYSIS AND RESULTS

Using the 320 hour database, we focus on identifying stress in adults through their speech signals. The two models – neutral and stress are trained on the human-tagged utterances using the proposed GMM framework. A binary decision from the neutral/stressed classifier is bench-marked against tags from human transcription to evaluate algorithm accuracy. Table 3 shows that test results for female adult stress gave 72.27% accuracy, while that for male models gave 67.53% accuracy.

Fig 1 and 2 represent the difference in distribution for female and male utterances across neutral and stress conditions.

With this, we consider UL with AWC and CT for stress/neutral assessment. UL, AWC, CT, and a derived parameter, (UL/AWC) ratio were analyzed for their dependence on stress in adult speech. Table 4 shows larger variations in female speech across stress conditions. Mothers show larger variations while speaking in neutral state (_ = 0.4327) as compared to speaking under stress state (_ = 0.0561). This implies that mothers uses more motherese speech under neutral conditions but tend to speak in a more adult-like speech while under stress. Males show more variations when under neutral as compared to under stress, but variations are reduced if compared to female speech. As mean, median and std. dev. Values under stress for both females and males are similar when under stress, this indicates that both genders when under stress, tend to speak to the child in a manner similar to conversing with an adult. Using Welch t-test for means and F-test for standard deviation, we found that UL/AWC showed a statistically significant dependence on stress. Female speech showed statistical significance with 97% confidence and male speech showed statistical significance with 99% confidence [See Table 4]. The dependence is more prominent for male adults compared to females

6. CONCLUSION / FUTURE STUDY

In this study, we employed a real-life scenario with contiguous recordings of 12-hours per family, for 20 families. Though it is difficult to perform stress assessment on adult speech within an adult-child framework because adults will speak in a motherese style (wherein the pitch f0 is about 2 to 4 times the normal pitch f0 ,) shifting their fundamental frequencies, it is very encouraging to see a GMM based binary stress/neutral classification rate of 72.27% and 67.53% accuracy for female and male adults respectively. Most important, is that the stress levels correlated with the change in (Utterance length to adult word count) ratio. The results represent that adults use much shorter duration of utterances for the same amount of words while speaking under stress with a child; thus, saying the words fast and affecting individual word articulation. This increases the chance of mispronunciation and reduces intelligibility for the infant, and hence impact language acquisition skills. The study thus indicates that adult stress does have an impact on the child's word count exposure, and therefore is expected to impact language development. An extensive speech corpus was employed with this study, yet other signal processing strategies can be employed to improve the system performance. Also, we plan to conduct an extensive longitudinal study on the impact of adult stress, word usage under stress, and vocabulary usage by the child which will provide further validation of these initial results

Future studies will extend our experiments on a much larger database, and analyze the multiple-session recordings from the same family over a wide time-span.

7. ACKNOWLEDGMENTS

Special thanks to Elisabeth Bownik of Infoture foundation for assessing the stress levels on the speech files. The first author expresses thanks to H. Boril, K. Godin, P. Boyraz for their helpful discussions.

8. REFERENCES

[1] S. Barnes, M. Gutfreund, D. Satterly, G. Wells, "Characteristics of Adult Speech Which Predict Children's Language Development," J. Child Lang., (10):65-84, 1983.

[2] L. Bloom and C. Margulis and E. Tinker and N. Fujita, "Early Conversations and Word Learning: Contributions from Child and Adult," Society for Research in Child Development, vol. 67. no. 6, pp. 3154-3175, December 1996.

[3] B. Hart and T. Risley, Meaningful Differences in the Everyday Experiences of Young American Children. Baltimore, Maryland: Brookers Publication, 1995.

[4] S.M.Higman, "The Influence of Individual and Family Characteristics on the Relationship between a Child's Language and Development and Adult Social Interaction and Life Satisfaction," Doctoral Thesis, The John Hopkins University, USA, 2004.

[5] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valchev, and P. Woodland, "The HTK book," Microsoft Corporation, July 2000.

[6] Technical Reports [Online]. Available: DOI=http://www.infoture.org/

[7] H.S.Kihm, Who Said 'words can never hurt?' An Investigation of Child Weight Status, Childhood Physiological Variables, and Latter Adult Quality of Life, Doctoral Thesis, Louisiana State University, USA, 2006.

[8] McRoberts and C. Best "Accommodation in Mean F0 During Mother-Infant and Father-Infant Vocal Interactions: a Longitudinal Case Study," J. Child Language, vol. 24, pp. 719-736, 1997.

[9] E.L.Moerk, Environmental Factors in Early Language Acquisition.

[10] G.J.WhiteHurst ed., Annals of Child Development, vol.3, Greenwich, JAI Press, 1986.

[11] S.A.Patil and John H.L. Hansen, Detection of Speech Under Physical Stress: Model Development, Sensor Selection, and Feature Fusion, Interspeech 2008, Brisbane, Australia.

[12] K.B.Wexler, "Developmental Disfluency in 2-, 4-, and 6-Year-Old Boys in Neutral and Stress Situations," Journal of Speech and Hearing Research, vol. 25, pp. 229-234, June 1982.

[13] -, "Mother-Child Interaction, Private Speech and Task Performance in Preschool Children with Behavior Problems," Journal of Child Psychiatry, vol. 40, no. 6, pp. 891-904, 1999.