ISCA Archive
http://www.isca-speech.org/archive

Second Workshop on Child
Computer Interaction (WOCCI 2009)
Cambridge, MA, USA
November 5, 2009

Preliminary Study of Stress/Neutral Detection on Recordings of Children in the Natural Home Environment

Umit Yapanel, Dongxin Xu, John Hansen*, Sharmistha Gray, Jill Gilkerson and Jeffrey A. Richards
LENA Foundation, 5525 Central Avenue #100, Boulder, CO 80301, USA
dongxinxu,yapanel,sharmigray,jillgilkerson,jeffrichards@lenafoundation.org
*Center for Robust Speech Systems, The University of Texas at Dallas, Richardson, TX 75080, USA.
john.hansen@utdallas.edu

ABSTRACT
Emotion and stress/neutral detection based on an input audio stream has been a topic of interest in the literature with various applications. This paper reports on a preliminary study of stress/neutral detection based on naturalistic home environment recordings of children. One major motivation of the work is to add stress/neutral detection functionality into the LENA™ System [10]. The study started with an acted emotion database, and tested the acoustic feature of Mel-frequency cepstral coefficients and the Gaussian Mixture Model (GMM) for stress/neutral detection on this relatively simple database. The method was then applied to the adult speech segments automatically extracted from home recordings of children with the LENA System, achieving 72% accuracy for adult stress/neutral detection. The application of this new functionality to a large number of naturalistic home environment recordings of children reveals interesting and meaningful statistical differences among the families of typically developing children, language-delayed children, and children with Autism Spectrum Disorders (ASD). The result suggests the potential for stress/neutral detection, along with the LENA System, as an integrated solution for (i) quality assessment of the child language environment, (ii) monitoring language interventions for disordered children, or (iii) general psychological and behavioral research.

Categories and Subject Descriptors
J.4 [**Computer Applications**]: Social & Behavioral Science – *Psychology.* I.5.4 [**Computer Methodology**]: Pattern Recognition – *Applications.* J.3 [**Computer Applications**]: Life and Medical Science – *Health.*

General Terms
Algorithms, Measurement, Performance, Reliability, Languages.

Keywords
Stress/Neutral Detection, Emotion Detection, Autism, Autism Spectrum Disorder (ASD), Speech Signal Processing, Child Speech Signal Processing, Child Development, Pattern Recognition.

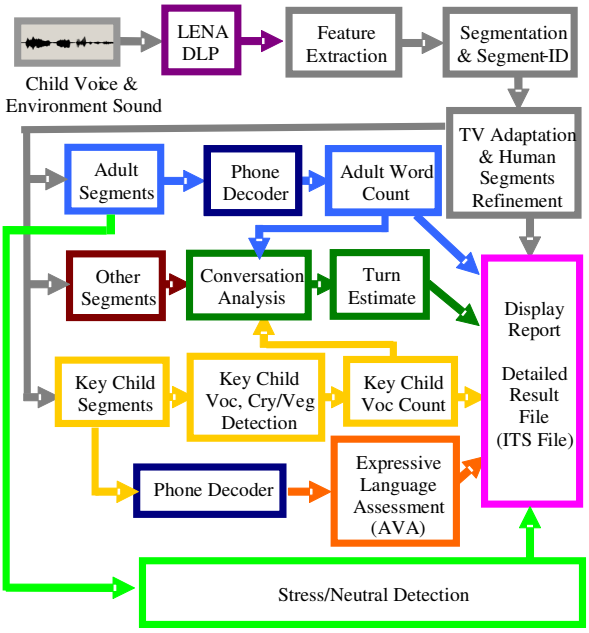

Figure 1: Diagram of the LENA System.

1. INTRODUCTION
The LENA™ (Language ENvironment Analysis) System as previously introduced in [10] utilizes speech signal processing technology to analyze and monitor a child's natural language environment and the vocalizations/speech of the child. As shown in Figure 1, the LENA System includes a digital language processor (DLP – essentially a small digital recorder) worn by a child in a pocket of specially designed clothing to record the sounds in the child's environment including his or her own voice. The software system provides estimates of adult word-counts, adult-child interactions (conversational turns) and child vocalizations, along with other language-related environment information such as the amount of audible TV/electronic media and an automatic vocalization assessment (AVA) of language development for the child. This study considers the prospect of adding a stress/neutral detection component to the LENA System.

Currently, there are only limited adult utterances from natural home recordings labeled with stress and neutral by human raters. The stress/neutral detector was built on these adult utterances and applied to adult speech as shown in Figure 1.

Emotion and stress/neutral detection based on audio signals has been a topic of interest in the literature [1,2,3,4,5,6,7,8,9], with many applications, such as: improving automatic speech recognition; enabling emotion speech synthesis; medical application; virtual teaching; application in psychological and behavioral research [2,3,5,9]. A number of research studies have been based on audio recordings of acted emotions [2]. Early studies based on the SUSAS (Speech Under Simulated and Actual Stress) corpus illustrated the differences in performance for stress/emotion detection as well as automatic speech recognition using simulated and actual stress/emotion data [11]. Alternative features and classifiers have also shown promise for emotion/stress versus neutral detection in actual speech under stress [12,13]. There have been emotion and stress/neutral detection studies on the cases such as TV talk-shows, human speech in certain extreme conditions. [2,7,9]. Due to the nature of the LENA System, we are interested in stress/neutral detection using naturalistic home environment audio recordings of the child. Potentially, such stress/neutral detection, incorporated into the LENA System, can be used for: quality assessment of child language environment with richer contents or as a tool for monitoring the language intervention of children with disorders such as ASD; it may also be used as a general tool for psychological and behavioral research such as those investigating how emotion/stress affects child development or how birth order could be related to emotion/stress behaviors. This task is unique with respect to its data and purpose.

In the following sections, the acoustic feature and the modeling method for stress/neutral detection in this study is described, along with the results on a simulated emotion database and the naturalistic data from the LENA Foundation Natural Language Corpus [14]. The application of the new stress/neutral detection functionality and the previous functionalities of the LENA System are applied to a large number of daylong recordings in the LENA Foundation Natural Language Corpus, revealing interesting statistical differences among the families of typically developing children, language-delayed children (i.e. children with language delays not including ASD) and children diagnosed with ASD.

2. Feature and Classification Method

As noted in the literature, the acoustic features correlated to emotion include pitch, vocal energy, various spectral features such as formant frequencies and bandwidths, and temporal features such as syllable rate, duration, etc. [11,6]. The Mel-frequency cepstral coefficients with high order of 36 (MFCC-36) contains information about vocal energy (first coefficient), excitation (related to high order coefficients) and spectral features. MFCC-36 and its temporal first order derivative were used as features for stress/neutral detection in this study, resulting in the acoustic feature of 72 dimensions for each frame for stress/neutral detection. Both first and second order derivatives of MFCC-36 were tested and the results showed no benefit with the second order derivative.

Cepstral features have previously been shown to perform well with Gaussian Mixture Models (GMM). GMM were used in this study as the classification modeling method. During the training phase, the k-means clustering method was first used to initialize the model, followed by the EM algorithm for model tuning. Various model sizes were explored for different tasks in the study.

With the limited amount of labeled data, testing was performed under a leave-one-out-cross-validation scheme to ensure the validity of the test. Various leave-one-out tests were considered depending on specific tasks. Speaker or entire audio stream recordings for a family were left out for cross-validation of different tasks.

3. Test on Simulated Emotion Database

A German database of acted emotion [1] was initially used in this study to verify the feature and method used. This database was downloaded from the Internet and has data from 5 male speakers and 5 female speakers. The same sentences were used for all emotions in order to reduce or eliminate content dependency. The utterances that were not clearly identified by human raters were eliminated. In total, there were 535 utterances of 7 emotion classes: anger (127 utterances); sadness (62); fear (69); disgust (46); happiness (71); boredom (81) and neutral (79). Since boredom was similar to neutral, they were merged as "neutral" in the stress/neutral classification task and the remaining 5 classes were merged as "stress" in this task. This rearrangement of emotion classes into 2 classes of stress/neutral is to enable comparison with the test on natural data from the LENA Foundation Natural Language Corpus, where only stress and neutral were labeled.

Average accuracy, which can be defined as the ratio of the number of correctly recognized utterances for all classes over total number of test utterances, is used as the overall performance measure. Leave-one-speaker-out-cross-validation was used for this task. Tables 1 and 2 show the average accuracies with different model sizes and different feature sizes. As can be seen, this is a relatively simple task. The 7-class emotion classification task achieves 73.4% accuracy and the 2-class stress/neutral classification task achieves 92.7% accuracy, which also indicates the effectiveness of MFCC-36 and its first derivative as the feature and the GMM modeling method for the stress/neutral detection. This is at least true for this particular database and provides the basis for us to further apply the method to the naturalistic data as shown in the next section.

Table 1: Model size and average accuracy of acted German database based on MFCC with order of 36

#GAUSSIAN/MODEL	32	64	128	256
7 EMOTION CLASSES	60.8%	64.2%	**64.6%**	60.7%
STRESS/NEUTRAL	88.1%	88.8%	**89.8%**	88.7%

Table 2: MFCC feature size and average accuracy with 128 Gaussians/model for acted German database

FEATURE SIZE	18	22	24	26
7 EMOTION CLASSES	69.5%	72.9%	**73.4%**	69.7%
STRESS/NEUTRAL	90.3%	92.7%	**92.7%**	91.3%

4. Test on Labeled LENA Data

Currently, among the large amount of data available through the LENA Foundation Natural Language Corpus, there are a very

limited number of adult segments labeled with stress and neutral. These data were drawn from 20 natural home environment recordings of 20 families with one full-day recording from each family. The adult segments (utterances) were automatically extracted by the LENA System software for the 20 daylong recordings. The machine-generated adult segments were selected and labeled with stress and neutral by a human rater. There were 801 neutral segments, 731 stress segments and 662 undecided (uncertain) segments. For simplicity and clarity, undecided segments were not included in this study. There were in total 1532 segments (utterances) with stress/neutral labeled. Moreover, different model and feature sizes were evaluated for this test. The leave-one-family-out-cross-validation scheme was used for this test to eliminate the potential confounding effect of the same speaker and/or family speakers.

Table 3 shows the average accuracies versus different choices of the MFCC-size and the GMM size. For this particular data set, MFCC-36 and its first derivative and 256 Gaussians for each class gave better performance. A detailed confusion matrix for MFCC-36 feature with 256 Gaussians for each class is shown in Table 4. An overall stress/neutral detection rate of 72% was achieved for this particular task, suggesting the possibility of applying the obtained model to the unlabeled larger quantity of LENA data to test if any interesting, meaningful results could be obtained which may be helpful and inspirational to child development research.

There are currently no child segments in the LENA database labeled with stress and neutral. The obtained adult stress/neutral model was also directly applied to child segments, resulting in more than 90% of child segments recognized as "stress," which obviously indicates the failure of this direct application of the mis-matched model. The reason is simple: The key-child who wears the recorder has high pitch and high volume, confusing the adult stress/neutral model.

Table 3: Average accuracy for Labeled Naturalistic Data Set using different model size and feature size

MFCC-SIZE/#GAUSS	36 / 128	36 / 256	36 / 512	24 / 256
STRESS/NEUTRAL	70.8%	**71.9%**	70.8%	70.6%

Table 4: Confusion matrix of MFCC-36 / 256 Gaussians for Labeled Naturalistic Data Set

		MACHINE RESULT		
		NEUTRAL	STRESS	TOTAL
HUMAN LABEL	NEUTRAL	**579**	222	801
	STRESS	208	**523**	731
	TOTAL	787	745	1532

Table 5: Mean and STD of the Stress-ratio for Adult Female, Adult Male and Adult Overall. All are normalized with the mean and standard-deviation (std) of the typical group

MEAN / STD	FEMALE STRESS-RATIO	MALE STRESS-RATIO	ADULT STRESS-RATIO
TYPICAL	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00
DELAYED	-0.14 / 0.87	-0.11 / 0.89	-0.05 / 0.93
ASD-NO-T	-0.01 / 0.99	0.00 / 1.09	0.00 / 0.97
ASD-T	**0.57 / 0.89**	**0.47 / 1.24**	**0.87 / 0.87**
ASD-ALL	0.18 / 0.99	0.15 / 1.16	0.28 / 1.02

Table 6: Word Counts Mean and STD for Adult Female, Adult Male and Adult overall. All are normalized with the mean and standard deviation (std) of the typical group

MEAN / STD	FEMALE WC	MALE WC	ADULT WC
TYPICAL	0.00 / 1.00	0.00 / 1.00	0.00 / 1.00
DELAYED	-0.14 / 1.01	-0.20 / 0.79	-0.20 / 0.92
ASD-NO-T	-0.23 / 1.19	-0.19 / 0.80	-0.26 / 1.09
ASD-T	**0.83 / 1.08**	**-0.36 / 0.73**	**0.37 / 0.83**
ASD-ALL	0.11 / 1.26	-0.24 / 0.78	-0.06 / 1.06

5. Application to Natural Home Environment

As mentioned above, the major purpose of this study is to establish a new module in the LENA System for stress/neutral detection. This could potentially increase the capacity of the LENA System, helping researchers, teachers, and parents to better understand child development and its relation to the stress/neutral environment. It could also help them understand a child's emotion/stress status or monitor a therapy session, a class, nanny time, obtaining useful feedback.

Based on the adult stress/neutral detection result in Section-4, the best GMM model size with 256 Gaussian for either stress or neutral and MFCC-36 with its first derivative were used for the new module of stress/neutral detection in the LENA System shown in Figure 1. The system in Figure 1 with the new module is applied to 1227 natural home environment recordings of which all are longer than 12 hours. The data set includes 712 recordings from 76 typically developing children, 290 recordings from 30 language-delayed children and 225 recordings from 34 children diagnosed with ASD. The ASD sub-set can be further divided into two groups: the recordings with/without therapy time (intervention sessions). The ASD sub-group without therapy time includes 153 recordings and the sub-group with therapy time has 72 recordings. These make for a total of 5 groups. The LENA System can produce several estimates for each recording, including adult word count, child vocalization count, conversational turn count, and AVA score. Here, we focus on stress/neutral detection results and also use adult word count estimates to help interpret the results.

The initial goal of this application is to test whether the families of children with ASD tend to be more stressed than others and whether there is any stress-related difference associated with the families of language-delayed children, compared with the families of typically developing children. These comparisons check the usefulness of the new module and also provide a good example of how the new module could be used in other research or clinical practice.

We are interested in finding how much of the total amount of speech is stressed speech, which can be defined as the ratio of the number of stressed segments to the total number of segments. This stress-ratio can be applied to adult female segments, adult male segments and the overall adult segments (with no gender distinction.) This will give the stress-ratio for adult female, adult male and adult overall. The mean and standard deviation of such ratios can be calculated for the 5 groups mentioned above, for example: typically developing (annotated as Typical in Table 5 and Table 6); language delayed (Delayed); ASD without therapy (ASD-No-T); ASD with therapy (ASD-T) and overall ASD (ASD-All). For easy comparison, the stress ratio of any recording was

normalized with the mean and standard deviation of the typical group: $normalized\_ratio = (ratio - mean_{Typical}) / std_{Typical}$. Using this convention, the typically developing group always has 0-mean and unit-variance which can be used as a standard for comparison. Since the ratio is applied to adult female segments, adult male segments and overall adult segments individually, the normalization of the ratio is also performed individually on each unique segment type. Similarly, word counts were examined in the same way for adult female segments, adult male segments and overall adult segments, and the method of normalization is also the same.

As can be seen from Table 5, most of the statistics of stress-ratio are similar for the 3 groups of Typical, Delayed and ASD-No-Therapy. The Delayed group shows a very slight lower average stress-ratio and very slight lower variance. The group of ASD-With-Therapy shows a significant higher stress-ratio on average. This is the case for both female and male adult segments. It is likely that the higher stress-ratio actually reflects real life: Therapy days are times when a focused effort is made to engage with children with ASD. Thus, everybody is impacted, independent of whether the speaker is (i) the therapist, (ii) the mother or farther, or (iii) female or male. Table 6 provides further information and better illustrates the situation. As shown, the female word count of the ASD-With-Therapy group is significantly increased compared with all other groups. Meanwhile, the male word count of the same group decreases noticeably. This could be due to the fact that most therapists are female and the therapy time takes away from the original opportunities for males to speak during therapy days. Independent of whether the word count increases or decreases, the stress-ratio of both female and male all increase during the therapy day for children with ASD. This is an interesting finding using the stress/neutral detector and the LENA System. This result appears to be reasonable with respect to expected child/adult interactions for cases where ASD is present.

It is somewhat surprising to see similar stress-ratios for the group of ASD-without-therapy compared to that of delayed and typically developing groups. Perhaps, in the regular family environment, parents remain in fairly consistent stress states a majority of the time, regardless of whether there are language-delayed children or children with ASD. After all, people can't stay in a stressed state all the time. Everything will be regressed toward its normal status. However, the adult word-counts for the delayed group and ASD-without-therapy group did show the lower average compared with the typical group, which is consistent with our common sense.

6. DISCUSSION

This study establishes a new module in the LENA System for adult stress/neutral detection. This preliminary effort achieves an accuracy rate of 72% with the simple method of using MFCC-36 and its first derivative as the feature and the GMM as the modeling method. The space to further improve the performance is open, including more acoustic features such as speaking rate, duration, explicit pitch estimation, etc [5,6]. The discriminant large margin modeling methods such as support vector machine and AdaBoosting [3] have high potential.

In a natural home environment, adult speech may frequently be broken into shorter segments due to other types of sound sources in the environment. This phenomenon suggests using a general smoothing method over neighboring adult segments rather than using many short adult segments for stress/neutral detection separately. This process may involve speaker clustering for the audio stream of a recording, ensuring that smoothing is applied to the same person.

The failure when applying the new adult-based module to child segments reinforces the need for labeling child segments with stress/neutral in the LENA database. Even for adult speech, more stress/neutral labeled segments are needed for more rigorous testing and model improvement. More human raters are also needed for cross-checking and to improve the quality of the labeled data. The failure of the new module to classify child segments accurately is due to the different characteristics of child pitch, etc. This suggests the need for gender-dependent modeling even for adult stress/neutral detection to take into account gender differences that may interfere with stress/neutral detection.

An interesting application of the new module has been shown in this study. More applications will be tried in the future, possibly for ASD research, research with deaf and hard of hearing children, other child development research and more general psychological research for children and/or adults.

7. ACKNOWLEDGEMENTS

We gratefully acknowledge Terry Paul for his conception of the LENA System and for personally funding and directing its development as well as the development of the LENA Foundation Natural Language Corpus.

8. REFERENCES

[1] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, "A Database of German Emotional Speech", Proceedings of the InterSpeech, 2005

[2] D. Ververidis, C. Kotropoulos, "A State of the Art Review on Emotional Speech Databases", Proc. Of 1st Rich Media Conference, Laussane, Switzerland, pp. 109-119, Oct, 2003

[3] M. Shami, W. Verhelst, "An Evaluation of the Robustness of Existing Supervised Machine Learning Approaches to the Classification of Emotions in Speech", Speech Communication, 2007

[4] A. Liberman, "Apparatus and Methods for Detecting Emotions in the Human Voice", US Patent No: 7,165,033 B1.

[5] J. Courtright, I. Courtright, "The perception of nonverbal vocal cues of emotional meaning by language-disordered and normal children", Journal of speech and hearing research, vol. 26 412-417, Sept. 1983

[6] M. Forsell, "Acoustic Correlates of Perceived Emotions in Speech", Mater's Thesis in Speech Communication, School of Computer Science and Communication, Royal Institute of Technology, Stockholm, Sweden, 2007

[7] M. Grimm, K. Kroschel, S. Narayanan, "Support Vector Regression for automatic recognition of spontaneous emotions in speech", Proc. Of International Conference on Acoustics, Speech and Signal Processing, 2007

[8] G. Zhou, J. Hansen and J. Kaiser, "Methods for stress classification: nonlinear TEO and linear speech based features", Proc. Of International Conference on Acoustics, Speech And Signal Processing, 1999

[9] J. Hansen, S. Bou-Ghazale, “Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database”, EUROSPEECH-97, vol.4, pp. 1743-1746

[10] D. Xu, U. Yapanel, S. Gray, J. Gilkerson, J. Richards, J. Hansen “Signal Processing for Young Child Speech Language Development” 1st Workshop on Child, Computer and Interaction, Oct. 2008, Chania, Crete, Greece. Also available:http://www.lenafoundation.org/DownloadFile.aspx/pdf/SignalProcessing_ChildSpeech

[11] J.H.L. Hansen, "Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition," *Speech Communications*, Special Issue on Speech Under Stress, vol. 20(2), pp. 151-170, November 1996

[12] B.D. Womack, J.H.L. Hansen, "Classification of Speech Under Stress using Target Driven Features," *Speech Communications*, Special Issue on Speech Under Stress, vol. 20(1-2), pp. 131-150, November 1996.

[13] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Nonlinear Feature Based Classification of Speech under Stress," *IEEE Transactions on Speech & Audio Processing*, vol. 9, no. 2, pp. 201-216, March 2001.

[14] http://www.lenafoundation.org/DataServices/Database.aspx?sub=true