CHILDREN'S SPEECH RECOGNITION WITH APPLICATION TO INTERACTIVE BOOKS AND TUTORS

Andreas Hagen, Bryan Pellom, and Ronald Cole
Center for Spoken Language Research
University of Colorado at Boulder
http://cslr.colorado.edu

ABSTRACT

We present initial work towards development of a children's speech recognition system for use within an interactive reading and comprehension training system. We first describe the Colorado Literacy Tutor project and two corpora collected for children's speech recognition research. Next, baseline speech recognition experiments are performed to illustrate the degree of acoustic mismatch for children in grades K through 5. It is shown that an 11.2% relative reduction in word error rate can be achieved through vocal tract normalization applied to children's speech. Finally, we describe our baseline system for automatic recognition of spontaneously spoken story summaries. It is shown that a word error rate of 42.6% is achieved on the presented children's story summarization task after using unsupervised MAPLR adaptation and VTLN to compensate for inter-speaker acoustic variability. Based on this result, we point to promising directions for further study.

1. INTRODUCTION

Children's speech represents an extremely important yet poorly understood and little researched area in the field of computer speech recognition. Previous studies have considered an acoustic analysis of children's speech [1,2,3]. This work has shed light onto the challenges faced by systems that will be developed to automatically recognize children's speech. For example, it has been shown that children below the age of 10 exhibit a wider range of vowel durations relative to older children and adults, larger spectral and suprasegmental variations, and wider variability in formant locations and fundamental frequencies in the speech signal. In recent years, several studies have attempted to address the issue of children's speech recognition by adapting the acoustic features of children's speech to match that of acoustic models trained from adult speech [4,5,6,7]. Approaches of this sort have included vocal tract length normalization as well as spectral normalization. In [8], the number of tied-states of a speech recognizer were reduced to compensate for data sparcity. Each of these studies point to lack of children's acoustic data and resources to estimate speech recognition parameters relative to the over abundance of existing resources for adult speech recognition. More recently, corpora for children's speech recognition have begun to emerge. In [9] a small corpus of children's speech was collected for use in interactive reading tutors and led to a complete children's speech recognition system [10]. In [11], a more extensive corpus consisting of 1100 children in grades K through 10 was collected and used to develop a speech recognition system for isolated word and finite-state grammar vocabularies for U.S. English.

In this paper we describe our recent work on developing a speech recognition system tailored to children's speech. Our work is being conducted in the context of NSF ITR and IERI funded projects for improving foundational reading skills and comprehension in children. In Section 2 we describe the Colorado Literacy Tutor project and describe the manner in which speech recognition is used to assess pronunciation of words during reading out loud, and assessing comprehension of text through dialogue interaction and analysis of spoken summaries of the text. In Section 3, we present our baseline speech recognition system. In Section 4, we describe the audio corpora collected at CSLR for research on children's speech recognition. In Section 5 we describe recognition experiments and present initial results on those corpora. Here, we illustrate challenges in children's speech recognition by performing a series of experiments to illustrate the nature of acoustic mismatch in younger children (grades K through 5) and extend this work by presenting an initial system for spontaneous recognition of story summaries.

2. THE COLORADO LITERACY TUTOR

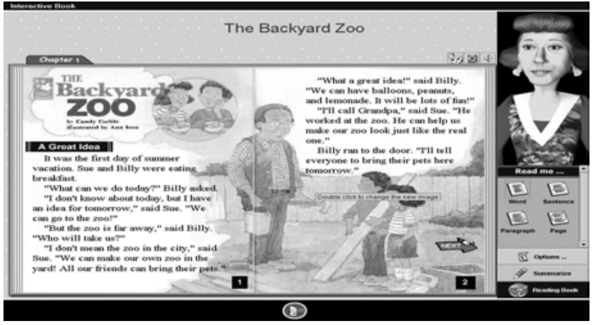
The Colorado Literacy Tutor (CLT) is a technology-based literacy program, based on cognitive theory and scientifically based reading research, which aims to improve literacy and student achievement in public schools [12,13]. The goal of the Colorado Literacy Tutor is to provide computer-based learning tools that will improve student achievement in any subject area by helping students learn to read fluently, to acquire new knowledge through deep understanding of what they read, to make connections to other knowledge and experiences, and to express their ideas concisely and creatively through writing. Therefore it shares a lot of goal with the LISTEN project [14]. A second goal is to scale up the program to both state and national levels in the U.S. by providing accessible, inexpensive and effective computer-based learning tools that are easy to use.

The CLT project consists of four tightly integrated components: Managed Learning Environment, Foundational Reading Skills Tutors, Interactive Books, and Summary Street comprehension training [15,16,17]. A key feature of the project is the use of leading edge human communication technologies in learning tasks. The project has become a test bed for research and development of perceptive animated agents that integrate auditory and visual behaviors during face-to-face conversational interaction with human learners. The project enables us to evaluate component technologies with real users—students in classrooms—and to evaluate how the integration of these technologies affects learning using standardized assessment tools.

Within the CLT, Interactive Books are the main platform for research and development of natural language technologies and perceptive animated agents. Fig. 1 shows a page of an Interactive Book. Interactive Books incorporate speech recognition, spoken dialogue, natural language processing, computer vision and computer animation technologies to enable natural face-to-face conversational interaction with users. The integration of these technologies is performed using a client-server architecture that provides a platform-independent user interface for Web-based delivery of multimedia learning tools. Interactive Book authoring tools are designed for easy use by project staff, teachers and students to enable authors to design and format books by combining text, images, videos and animated characters. Once text and illustrations have been imported or input into the authoring environment, authors can orchestrate interactions between users, animated characters and media objects. Developers can populate illustrations (digital images) with animated characters, and cause them to converse with each other, with the user, or speak their parts in the stories using naturally recorded or synthetic speech. A mark up language enables authors to control characters' facial expressions and gestures while speaking. The authoring tools also enable authors to pre-record sentences and/or individual words in the text as well as utterances to be produced by animated characters during conversations.

Interactive Books enable a wide range of user and system behaviors. These include having the story narrated by one or more animated characters (while controlling their facial expressions and gestures), having conversations with animated characters in structured or mixed-initiative dialogues, having the student read out loud while words are highlighted, enabling the student to click on words to have them spoken by the agent or to have the agent interact with the student to sound out the word, having the student respond to questions posed by the agent either by clicking on objects in images or saying or typing responses, and having the student produce typed or spoken summaries which are analyzed for content using language processing techniques.

Read-aloud feedback involves following along as text is read, highlighting the read text, monitoring reading fluency and verifying pronunciation accuracy. Read-aloud feedback is obtained by building a statistical language model for the book, getting partial phrases from the speech recognizer as the user is reading and aligning the partial phrase with the book text using a Dynamic Programming search.



It was the first day of summer vacation. Sue and Billy were eating breakfast.

"What can we do today?" Billy asked.

"I don't know about today, but I have an idea for tomorrow," said Sue. "We can go to the zoo!"

"But the zoo is far away," said Billy. "Who will take us?"

"I don't mean the zoo in the city," said Sue. "We can make our own zoo in the yard! All our friends can bring their pets."

"What a great idea!" said Billy. "We can have balloons, peanuts, and lemonade. It will be lots of fun!"

"I'll call Grandpa," said Sue. "He worked at the zoo. He can help us make our zoo look just like the real one."

Billy ran to the door. "I'll tell everyone to bring their pets here tomorrow."

Figure 1: An example interactive book with animated agent (shown upper right)

Speech recognition technology plays a key role in the development of the Colorado Literacy tutor program. For example, within the Interactive Books speech recognition is used to enable student to provide spoken answers to questions about stories, to perform tracking and/or visual feedback about words that are pronounced correctly (based on confidence scores produced by the recognizer) while children read stories, and to provide pronunciation verification capabilities. In addition, speech recognition plays a key role in comprehension training. After reading stories within the interactive books, children will be asked to produce spoken summaries of the text they read, and the automatic transcription of these summaries can be used to provide feedback to the student about information that is missing from their summaries as well as the conciseness of their summaries, using an approach based on Latent Semantic Analysis (LSA) [18]. This leads to an interesting challenge for speech recognition in spoken story summarization tasks. We will consider this problem in more detail in Section 5.

3. BASELINE SPEECH RECOGNITION SYSTEM

The baseline speech recognition system used in this work is based on Sonic, the University of Colorado Speech Recognizer [19,20]. The recognizer uses Perceptual

Minimum Variance Distortionless Response (PMVDR) representation for speech features [21]. Acoustic modeling in Sonic is based on continuous density mixture-Gaussian HMMs. The acoustic models are decision-tree state-clustered where each state is associated with a gamma probability density function to model state durations. The recognizer implements an efficient time-synchronous, beam-pruned Viterbi token-passing search through a lexical prefix tree.

4. CU CHILDREN'S AUDIO SPEECH CORPORA

4.1 Read and Prompted Speech Corpus

During the 2000-2001 school year, we developed and tested data collection protocols, in collaboration with the Center for Spoken Language Understanding at the Oregon Graduate Institute (OGI). This work resulted in the collection of audio and video data from approximately 200 children in the Boulder Valley School District (BVSD). During the 2001-2002 school year, we further refined the data collection protocol, and captured audio and video data from an additional 580 students in schools within the Boulder Valley School District in Colorado. After data cleaning, a corpus consisting of 663 speakers from kindergarten through 5th grade was developed. The corpus consists of isolated words, sentences, and short spontaneous story telling. The protocol was developed and described in [11]. Table 1 provides the number of speakers per age level.

Grade	K	1	2	3	4	5
#Kids	84	136	150	92	91	110

Table 1: Number of children in the CU Children's audio speech corpus by grade level.

Each speaker has approximately 100 speech files, which vary in length depending on the protocol (the sequence of phrases read by the speaker). The recordings were made using one of three types of microphones: a commonly available head-mounted noise-canceling microphone (Labtec LVA-8450), an array microphone (CNnetcom-Voice Array Microphone VA-2000), and a commonly available desktop farfield microphone. The final corpus is sampled at 16kHz with 16 bits per sample. Each audio file has a corresponding word-level transcription. Information is provided about each speaker, such as age, sex, grade, and native language of speaker.

4.2 Read and Summarized Story Corpus

In 2003 we extended the Read and Prompted Speech Children's corpus by collecting data of children reading and summarizing stories. This corpus represents our test bed for research in the area of automatic recognition of children's spontaneous speech.

The CU Story Corpus currently consists of speech data and transcription of 106 children who were asked to read a story and provide a spontaneous spoken summary of the material. In addition, each child was asked to read 25 phonetically balanced sentences for future use in exploring strategies for speaker adaptation.

The data were collected from native English speaking children in the Boulder Valley School District. We have initially collected stories from children in grades 3, 4, and 5 (grade 3: 17 speakers, grade 4: 28 speakers, grade 5: 61 speakers). The data were originally collected in a quiet room using a commonly available Labtec Axis-502 microphone. The data were recorded at 44kHz and later resampled to 16kHz for the purposes of experimentation. The current corpus consists of 10 different stories. The number of speakers per story is shown in Table 2. Each story contained an average of 1054 words (min 532 words / max 1926 words) with an average of 413 unique words per story. The resulting summaries spoken by children contain an average of 168 words.

Story #	**(A)**	**(B)**	**(C)**	**(D)**
1	22	572	205	150
2	22	532	207	129
3	12	932	364	181
4	12	1668	606	224
5	11	828	329	185
6	9	1078	389	161
7	8	1926	631	276
8	5	1157	526	133
9	3	933	460	101
10	2	919	417	90

Table 2: Overview of the 10 stories used in the CU Read and Summarized Story corpus for (A) number of children who recorded the story, (B) number of words in the story, (C) number of unique words in the story, (D) average summary length in words

5. EXPERIMENTS USING CHILDREN'S SPEECH

The following sections describe experiments that highlight challenges faced in designing automatic speech recognition systems for children's speech. First, we consider the issue of inter-speaker variability and provide some insight as to the utility of state-of-the-art methods for coping with such variation between children of different ages. Then we consider the problem of recognition of spontaneously spoken story summaries.

5.1 Inter-Grade Level Acoustic Variability

Our initial work focused on achieving a better understanding of the nature of the acoustic variability that exists between children of various age/grade levels. In this experiment, we use the Sonic speech recognition system to compute a series of acoustic models, one per grade level (K through 5). Each acoustic model is trained by partitioning the CU Prompted and Read Speech Corpus (Section 4.1) into a training and testing set. The test set contained 25% of the overall corpus's speakers that were not in the training set. We then designed the test set to contain utterances in which only a single isolated word was recorded. We dynamically designed a language model for each test utterance such that the vocabulary size was fixed to 500 words and was guaranteed to contain the test-word (a closed-vocabulary experiment). Each word (N=500) was assumed to be equally likely during experimentation.

System performance was evaluated by presenting each mismatched condition (e.g., by using the acoustic model trained from children in grade 1 with audio data from children in grade 5). We compute the word error rate for each condition and summarize the results in Table 3.

		Train Grade						
		K	1	2	3	4	5	Avg
Test Grade	K	**47.9**	39.2	43.2	46.0	47.9	45.3	44.9
	1	38.1	**27.1**	28.3	30.1	32.9	29.4	31.0
	2	37.0	24.6	**24.4**	29.2	28.8	26.3	28.4
	3	34.5	22.5	21.9	**24.2**	24.9	22.3	25.1
	4	34.3	23.0	21.8	21.6	**23.9**	22.6	24.5
	5	34.7	23.5	20.3	23.3	23.5	**19.7**	24.2
	Avg	37.8	26.7	26.7	29.1	30.3	27.6	

Table 3: Word error rate (%) matrix showing matched and mismatched training / test conditions. Matched training and test conditions by grade-level are shown in bold.

From Table 3 it can be seen that the system performed worse the younger the children (grades K and 1). This can be seen especially well for acoustic model at grade level 5. Children at Kindergarten grade-level (K) had an average WER of 45.3% while the error rate drops drastically to 19.7% for children in grade-5. Some grade levels perform best if the training data for the acoustic model and the testing grade level match, as an example grade level one can be taken. But grade level 4 does not confirm this intuitive first idea. One assumption here is that children at a given grade-level may have similar growth characteristics. Certainly this is not the case as girls develop more rapidly in early grades (whereas we combined these data); children of both genders develop physically at differing rates; and the ages of children within each grade level may differ by as much as one year. Still a general rule of thumb seems to be that 1st graders and older children seem to provide more consistent (less variable) training data that leads to better results on ASR systems compared to children at Kindergarten grade-level. We hypothesize, consistent with analyses of children's speech in [1-3], that this may be due to clearer articulation and more well constructed sentences (fewer pauses, fewer filler words, better articulated training data). We will examine this issue further in a future study.

5.2 Coping with Inter-Speaker Acoustic Variability

As children's vocal tract lengths increase over time due to normal growth, and as children grow at different rates, it seems beneficial to apply vocal tract length normalization (VTLN) in order to compensate for these additional *inter-speaker* variations. This is in contrast to previous work such as [4,5,6,7] in which the use of VTLN was motivated to account for gross differences between adult and children's speech.

Using the VTLN method described in [22], we constructed a set of single-Gaussian triphone acoustic models and estimated the VTLN warping factor for each training speaker (see Sec. 5.1). Frequency warping factors ranging between 0.88 and 1.12 were estimated for each child. A single VTLN normalized acoustic model was then estimated using all data from children in grades K through 5.

Grade	K	1	2	3	4	5
(A)	47.9	27.1	24.4	24.2	23.9	19.7
(B)	42.9	26.8	24.0	18.4	19.5	16.9

Table 4: Word error rate (%) by child grade level for (A) grade-level dependent acoustic models and (B) a single VTLN normalized children's acoustic model.

Table 4 shows the average WERs for the different age-levels. The improvement can clearly be seen for each grade level. On average applying VTLN to normalize for differences in vocal tract length between children resulted in a relative error reduction of 11.2%.

5.3 Recognition of Spontaneous Spoken Summaries

An important goal of the Colorado Literacy Tutor project is to improve comprehension of text by students through automatic grading of spoken summaries. The Summary Street comprehension training program has been shown to improve student achievement by enabling students to improve their typed summaries of text following automatic grading of the text using Latent Semantic Analysis. Since most younger students cannot type well enough to benefit from this program, we are working to analyze spoken summaries of text, which requires achieving new levels of accuracy for children's speech. We are thus focusing our efforts on automatic transcription of spontaneously spoken story summaries by children from kindergarten through fifth grade.

This task is clearly challenging. First, it is well known that spontaneous speech is both disfluent, ungrammatical, and differs highly from read speech. Second, as mentioned in the introduction, the acoustic diversity of the speakers is especially a problem for young children at or below age 10 (e.g., formant frequency variability, vowel duration variability, etc.).

To illustrate the challenges of the spontaneous summary transcription problem, we constructed a language model consisting of the text from each of the 10 stories and ever increasing amounts of transcribed summaries from children (the summary from the story under test is always removed for experimentation). Using the recognition setup similar to that described in [20], we transcribed each summary spoken by each child in the summary corpus (106 summaries). The acoustic model for this simulation is based on a single non-VTLN normalized acoustic model.

Our transcription system for spoken summaries is based on multi-pass speech recognition with iterative unsupervised speaker adaptation (Fig. 2).

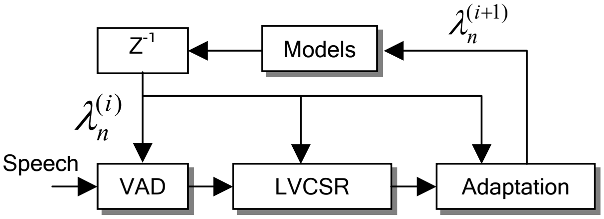


Figure 2: Diagram of multi-pass search used for recognition of spontaneous children's story summaries.

During each recognition pass, a voice activity detector (VAD) is dynamically constructed from the current adapted system acoustic models [20]. The VAD generates a segmentation of the audio into utterance units and LVCSR is performed on each detected speech region. The resulting output (a confidence tagged lattice or word string) is then used to adapt the acoustic model means and variances in an unsupervised fashion. Here we use the MAPLR adaptation approach described in [23]. The adapted acoustic models are then reapplied to obtain an improved segmentation, recognition hypothesis, and new set of adapted system parameters. The integrated adaptation procedure can be repeated several times resulting in sequential improvements to both segmentation and recognition hypotheses. For this work we found that one iteration of MAPLR adaptation was sufficient to achieve convergence.

Results are shown in Table 5, which shows clearly that the summarization task has a relatively low vocabulary size (~ 2000 words) and a low out-of-vocabulary rate (~ 1.5%). However, the task language model perplexity is high (~ 175) and the error rates of the current system, relative to adult systems, clearly demonstrate the challenges faced in spontaneous children's speech recognition (error rates approach 40-50% for a state-of-the-art transcription system). Adding more transcribed summaries aids in reducing the error rate from 51.7% with only 20 transcribed summaries to 46.1% with 100 transcribed summaries. As expected from the simulation results in section 5.2 additional vocal tract length normalization reduced the WER to 42.6% on the recognition task utilizing 100 transcribed summaries. So the additional performance gain through VTLN on the spoken summary recognition task accounts to about 8% relative.

# Sum	(A)	(B)	(C)	(D)	(E)
20	1819	213	2.4%	55.7%	51.7%
40	1872	235	1.5%	55.1%	50.5%
60	1958	194	1.5%	52.7%	48.3%
80	2070	179	1.4%	51.4%	47.0%
100	2143	159	1.2%	51.0%	46.1%

Table 5: ASR performance on spoken summaries as a function of the number of transcribed summaries used for language model training. Results are shown for (A) vocabulary size, (B) LM Perplexity, (C) OOV rate, (D) First-pass word error rate, and (E) word error rate after MAPLR adaptation (no VTLN).

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented our initial work towards development of a robust large vocabulary continuous speech recognition system for children's speech recognition. The system developed as a result of this work is now integrated into the Colorado Literacy Tutor project's Interactive Books, enabling a number of speech recognition services, including tracking the performance of students who are reading aloud, recognizing spoken responses to questions about stories that have been narrated or read, and phonetic-alignment for synchronization of audio and visible speech production by an animated character. In the near future, integration of continuous speech recognition will be incorporated into Interactive Books so we can begin research on the effectiveness of Latent Semantic Analysis techniques to grade summaries of text based on transcriptions provided by the recognizer.

We illustrated the importance of normalizing for vocal tract length *within* children's speech and demonstrated an 11.2% error reduction on an isolated word task trained on speech from children in grades K through 5. Our story summarization component, which involves spontaneous recognition of children's speech, exhibits error rates that likely prevent effective use in grading of summaries within Interactive Books in the near-term. However, we will be conducting basic research to improve recognition performance, and are also collecting more summaries from children in the Fall of 2004 and plan to incorporate data both from typed and spoken summaries as well as age-appropriate children's text to improve language modeling for this new and challenging task-domain.

7. ACKNOWLEDGEMENTS

The authors would like to thank Linda Corson, David Cole, Scott Schwartz and Taylor Struemph for collection and transcription of the CU children's speech corpora.
This work was supported in part by grants from the National Science Foundation's Information Technology Research (ITR) Program and the Interagency Educational Research Initiative (IERI) under grants NSF/ITR: REC-0115419, NSF/IERI: EIA-0121201, NSF/ITR: IIS-0086107, NSF/IERI: 1R01HD-44276.01; California Digital Media Innovation Grant 01-10130; and the Coleman Institute for Cognitive Disabilities. The views expressed in this paper do not necessarily represent the views of the NSF.

8. REFERENCES

[1] S. Lee, A. Potamianos and S. Narayanan, "Analysis of Children's Speech: Duration, Pitch and Formants," Proc. EUROSPEECH 97, Rhodes, Greece, Sept. 1997.

[2] S. Lee, A. Potamianos and S. Narayanan, "Acoustics of Children's Speech: Vowels", Journal of the Acoustical Society of America, pp. 1455-1468, March 1999.

[3] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of Children's speech: Developmental changes of temporal and spectral parameters," Journal of the Acoustical Society of America, Vol. 105, pp. 1455-1468, March 1999.

[4] Potamianos, S. Narayanan and S. Lee, "Automatic Speech Recognition for Children," Proc. EUROSPEECH 97, Rhodes, Greece, Sept. 1997

[5] S. Das, D. Nix, M. Picheny, "Improvements in Children's Speech Recognition Performance," Proc. ICASSP, Seattle, WA, May, 1998

[6] Potamianos and S. Narayanan, "Robust Recognition of children's speech", IEEE Transactions on Speech and Audio Processing, 2003.

[7] D. Giuliani and M. Gerosa, "Investigating Recognition of Children's Speech", Proc. ICASSP, April 2003.

[8] M. Eskenazi, G. Pelton, "Pinpointing Pronunciation Errors in Children's Speech: Examining the Role of the Speech Recognizer", Proceedings of the PMLA Workshop, 2002.

[9] Eskanazi, M., "KIDS: A Database of Children's Speech", Journal of the Acoustical Society of America, Vol. 100, No. 4, Part 2, December 1996.

[10] G. Aist, P. Chan, X. Huang, L. Jiang, R. Kennedy, D. Latimer, J. Mostow, C. Yeung, "How Effective is Unsupervised Data Collection for Children's Speech Recognition?", Proc. ICSLP, Sydney, Australia, 1998.

[11] Khaldoun Shobaki, John-Paul Hosom, and Ronald Cole, "The OGI Kids' Speech Corpus and Recognizers", Proc. ICSLP, Beijing, China, 2000.

[12] CSLR Reading Tutor Project (2002). http://cslr.colorado.edu/beginweb/reading/reading.html.

[13] Colorado Literacy Tutor (2002). http://www.colit.org

[14] S. Banerjee, J. Beck, and J. Mostow, "Evaluating the Effect of Predicting Oral Reading Miscues", Proc. EUROSPEECH 03, Geneva, Switzerland, 2003.

[15] D. Steinhart, "Summary Street: An intelligent tutoring system for improving student writing through the use of Latent Semantic Analysis", Ph.D. Dissertation, Dept. Psychology, Univ. of Colorado, Boulder, CO, 2001.

[16] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by Latent Semantic Analysis", *Journal of the Society for Information Science,* vol. 41, no. 6, pp. 391-407.

[17] T. Landauer and S. Dumais, "A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge", *Psychological Review*, vol. 104, pp. 211-240, 1997.

[18] E. Kintsch, D. Steinhart, G. Stahl, C. Matthews, R. Lamb, and LRG, "Developing summarization skills through the use of LSA-based feedback", *Interactive Learning Environments*, vol. 8, pp. 87-109, 2000.

[19] B. Pellom, "Sonic: The University of Colorado Continuous Speech Recognizer", Technical Report TR-CSLR-2001-01, University of Colorado, March 2001.

[20] Bryan Pellom, Kadri Hacioglu, "Recent Improvements in the CU Sonic ASR System for Noisy Speech: The SPINE Task", Proc. ICASSP, Hong Kong, April 2003.

[21] Umit H. Yapanel, John H.L. Hansen, "A new perspective on Feature Extraction for Robust In-vehicle Speech Recognition", Proc. EUROSPEECH, Geneva, Sept. 2003.

[22] L. Welling, S. Kanthak, H. Ney, "Improved Methods for Vocal Tract Length Normalization", Proc. ICASSP, Phoenix Arizona, 1999.

[23] O. Siohan, T. Myrvoll, and C.-H. Lee, "Structural Maximum a Posteriori Linear Regression for Fast HMM Adaptation", Computer, Speech and Language, 16, pp. 5-24, January 2002.