HOW EFFECTIVE IS UNSUPERVISED DATA COLLECTION FOR CHILDREN'S SPEECH RECOGNITION?

G. Aist, P. Chan*, X. Huang**, L. Jiang**, R. Kennedy*, D. Latimer*, J. Mostow*, C. Yeung**

*Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3720 USA
http://www.cs.cmu.edu/~listen

**Microsoft Research, One Microsoft Way, Redmond, WA 98052 USA
http://research.microsoft.com/stg

ABSTRACT

Children present a unique challenge to automatic speech recognition. Today's state-of-the-art speech recognition systems still have problems handling children's speech because acoustic models are trained on data collected from adult speech. In this paper we describe an inexpensive way to mend this problem. We collected children's speech when they interact with an automated reading tutor. These data are subsequently transcribed by a speech recognition system and automatically filtered. We studied how to use these automatically collected data to improve children's speech recognition system's performance. Experiments indicate that automatically collected data can reduce the error rate significantly on children's speech.

1. INTRODUCTION

Children present a unique challenge to automatic speech recognition. There is a dramatic difference between the acoustic characteristics of children and adults. In addition, previous study shows that children exhibit wider dynamic range of vowel duration, longer suprasegmental duration, and larger temporal and spectral variations (Lee et al. 1997). As such, most of the state-of-the-art speech recognition systems break down when tested with children's speech.

To improve the performance of speech recognition system on children's speech, it is clear that we need to use children's data extensively to adapt the existing acoustic model. However, extensive speech data collection of children is not trivial. Depending on the age group, children may be very uncooperative, easy to be bored, or unable to read correctly. Therefore, the quality control is much harder compared to the normal data collection for adults. Even if we can collect a huge amount of children's speech, it remains very difficult to transcribe these data accurately for the purpose of acoustic training. For example, the pronunciation by children is often atypical when compared with the standard dictionary. To be accurate, transcription into phonetic levels is often required. Consequently, the cost for supervised data collection and manual transcription could be several times higher than the adult speech collection and transcription.

The KIDS corpus, published by Linguistic Data Consortium, was described by Eskenazi (1996). Briefly, it was collected by trained research assistants on a NeXT™ workstation using a Sennheiser noise-canceling headset microphone. Data was recorded in fairly quite rooms with just one child speaker present at a time, supervised individually by a research assistant. A standard program for collecting speech data displayed one sentence to read at a time, taken from *Weekly Reader* (a newsmagazine for children). The recorded speech was transcribed manually, including noises and phonetic transcriptions of oral reading disfluencies (Eskenazi 1996).

The cost of such high-quality data collection will limit the amount of data we can get. In addition, the data collected in highly supervised style does not always match the real scenario where children interact with speech recognition systems. Therefore, finding a low-cost way to collect natural speech from children is a challenge for successful children's speech recognition. In this paper we introduce a corpus collected and automatically transcribed during children's in-school use of Reading Tutor developed at Carnegie Mellon University (Mostow and Aist 1997, Aist and Mostow 1997). It is well known that unsupervised or lightly supervised data collection and automatic transcription might have quality problem (Zavaliagkos et al. 1998). To compensate the problem, we developed some techniques to improve the quality of these data.

The goal of data collection is to improve the performance of speech recognition on children's speech. There are a number of studies focusing on children's speech. Wilpon and Jacobsen studied speech recognition for children and the elderly, investigating various training data compositions and their effects on performance. They also compared LPCC and MFCC feature for children's speech (Wilpon and Jacobsen 1996). Potamianos et al performed research on speaker normalization and adaptation for children's speech. They found frequency warping to be very effective for improved speech recognition. They also discovered that age-dependent modeling could further reduce the error rate (Potamianos et al. 1997). Das et al had the same conclusion regarding frequency warping. They also experimented with children's language models that substantially improved the recognition performance (Das et al. 1998).

In this paper, we conducted a comparative study between the performance of manually collected/transcribed KIDS corpus and automatically collected/transcribed RT corpus. We also studied if gender-dependent models for children could help speech recognition.

This paper is organized as follows. In section 2, we describe in details how Reading Tutor corpus is developed, including how we filter the raw data to ensure relatively high quality. In section 3 we discuss the acoustic model training using collected

data, including gender dependent modeling and comparison of performance using KIDS and RT corpus. Finally we summarize our major findings and outline the future work.

2. READING TUTOR CORPUS

2.1. Project LISTEN's Reading Tutor

Project LISTEN's automated Reading Tutor (Mostow and Aist 1997, Aist and Mostow 1997) listens to children read aloud, and helps them. The Reading Tutor runs in Windows™ 95 or NT 4.0 on a Pentium™, with a noise-canceling headset microphone.

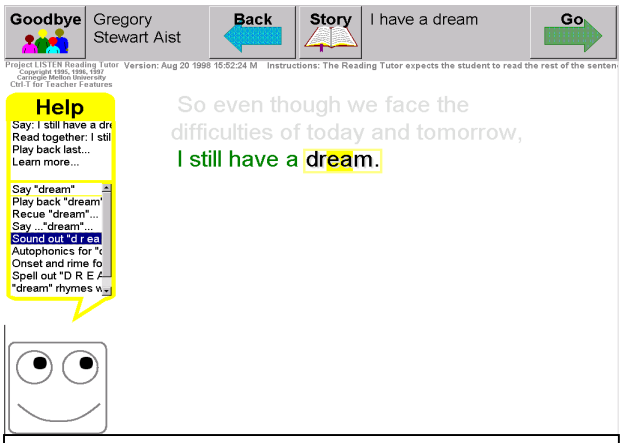

Figure 1. Reading Tutor, Fall 1998.

The Reading Tutor displays one sentence at a time and listens to the student read aloud. The Reading Tutor can interrupt the student to correct a mistake (Aist 1998). When the Reading Tutor hears the end of the sentence or a prolonged silence, it aligns the speech recognizer output against the sentence to decide which words the student read correctly. The Reading Tutor gives the student "credit" for the words it heard the student read correctly. When the student has received credit for every important word in the sentence, the Reading Tutor goes on to display the next sentence. Otherwise, the Reading Tutor responds expressively with recorded human voices. For example, the Reading Tutor may speak a word or an entire sentence. Then, the Reading Tutor lets the child reread the word or sentence. The student can also use the mouse to navigate by clicking *Back* or *Go*, or to get help on a word or sentence (Figure 1).

2.2. Data Collection and Transcription

The Reading tutor (RT) corpus was collected and automatically transcribed during children's in-school use of Project LISTEN's Reading Tutor. Challenges to corpus-quality collection of children's speech in school settings include improper microphone placement, social interactions among children, off-task speech, and environmental noise.

Collection conditions varied with Reading Tutor location. In a pilot study of eight low-reading third graders during the 1996-97 school year, a single Reading Tutor was used in a small room under individual supervision by a school aide. The pilot version of the Reading Tutor ran on a 90MHz, 64MB Pentium™ in half-duplex, causing truncation of overlapped speech, and sacrificing some accuracy for speed in the Sphinx-II speech recognizer by using the "top 1 senone" setting. The subsequent versions ran in full-duplex on 200MHz, 64MB Pentium Pro™ machines, eliminating such truncation and allowing the more accurate "top 4 senones" setting. In a July 1997 reading clinic for 62 children in grades K-6, eight Reading Tutors were used simultaneously in a school computer laboratory supervised by one graduate student. In fall 1997, approximately 200 children in grades K-4 used the Reading Tutor (one per classroom) under normal classroom conditions.

Speech data recorded from the microphone are cataloged and stored to disk as 16 bit, 16 kHz WAV files. Hypotheses and segmentation files generated by the Sphinx-II speech recognizer (Huang et al. 1993) using special-purpose language models and child-adapted acoustic models (Mostow et al. 1994, Mostow et al. 1993) are also cataloged and saved for archival by the Reading Tutor. These data are later written to recordable CDs by a human research assistant and brought back to the lab.

2.3. Data Filtering

We used three rules to extract accurate parts of the RT corpus. In the examples, the extracted part is shown in **bold**. Omitted portions are shown in *italics*.

"**Perfect**": the transcription exactly matches the sentence text. *Example* (DEC-MDLJ-1988-07-DEC17-97-09-08-45). Text: Humpty Dumpty had a great fall. Transcription: **HUMPTY DUMPTY HAD A GREAT FALL**.

"**OffByOne"**: the transcription nearly matches the sentence text, differing in only one word. The sentence text is taken to be the correct transcription, to catch cases where the student was right and the recognizer was wrong. *Example* (DEC-FKGD-1987-09-24-JUL23-97-09-57-27). Text: If the computer thinks you need help, it talks to you. Transcription: **IF THE COMPUTER THINKS YOU NEED HELP IT** *THINKS* **TO YOU**. (**TALKS** substituted for *THINKS* here.)

"**RefInHyp**": the sentence text is strictly contained in the transcription – the reference string is a proper substring of the hypothesis string. The corresponding speech is extracted. *Example* (DEC-FSDL-1988-12-04-OCT14-97-10-59-12). Text: She showed them how a bee gets its honey from flowers. Transcription: *SHE SHOWED* **SHE SHOWED THEM HOW A BEE GETS ITS HONEY FROM FLOWERS**.

For each of these heuristics, we measured the *purity* -- the percentage of words in the relevant portion of the transcript that are actually correct. The purity of the "Perfect" heuristic was 94%; OffByOne, 88%; and RefInHyp, 97%. These heuristics extracted 9977 utterances from the Reading Tutor (RT) corpus. Only those portions of the utterance which matched the extracted "good" part of the transcription were intended for use in acoustic training.

3. ACOUSTIC MODELING

3.1. Experimental Setup

We used Microsoft's WHISPER speech recognition system (Huang et al. 1995, Alleva et al. 1996) in our experiments. Briefly, WHISPER processes 16kHz PCM data using a MEL-scale cepstrum along with its dynamics into a multi-dimensional feature vector. WHISPER can use either semi-continuous or continuous density HMMs. In section 3.2, we used a compact set of semi-continuous HMMs for experiments with command-and-control type task constrained by context-free grammar. In section 3.3, we used a set of HMMs with continuous-density output probabilities consisting of 6000 senones (Hwang et al. 1993) and statistical bigram. A mixture of 20 Gaussian densities with diagonal covariances was used for each senone. The phonetic modeling in the system consists of position and context dependent within-word and crossword triphones.

Training acoustic model with RT data, we performed some pre-processing to make waveform data matching the transcription text for category "OffByOne" and "RefInHyp". We want to make sure that acoustic models are not contaminated.

3.2. Gender-Dependent Modeling

There is an assumption that gender makes little or no difference for children's speech recognition since both boys and girls tend to have about the same pitch range (Lee et al. 1997). Is this assumption really true? We conducted an experiment to explore whether gender-dependent modeling would result in any difference.

We used a compact set of acoustic model with semi-continuous HMMs. The task was a simulated command-and-control task constrained by a context-free grammar. We created a grammar specifically for each test utterance. For each word, we randomly selected a list of alternative words from the dictionary. We imposed a constraint on alternative words selection for perplexity purpose: the selected words must have the same number of phones as the original word. One advantage of this task is that you can adjust the perplexity of the task by controlling the length of the alternative word list.

We trained acoustic models with boys' speech only (MODEL-BOY), girls' speech only (MODEL-GIRL) and both boys' and girls' speech (MODEL-ALL). We tested them on boys' and girls' test data respectively. What we found is that the girls' model (MODEL-GIRL) performed best on the girls' test set. However, for the boys' test set, the mixed model (MODEL-ALL) was the best choice (slightly better than MODEL-BOY). Of course, MODEL-BOY was still better than MODEL-GIRL on the boys' test set.

3.3. Modeling With KIDS and RT Corpus

How good is automatically collected and transcribed data? We performed an experiment using both KIDS and RT corpus for acoustic model training.

We took the large-vocabulary continuous speech recognition system with speaker-independent adult female acoustic model as baseline. We used the 5,000-word vocabulary and language model (LM) from closed-5K Wall Street Journal (WSJ) task. We adapted the model by doing additional training using either KIDS corpus or RT corpus. We also adapted the language model by interpolating the WSJ text corpus with KIDS and RT training materials to reduce the perplexity. Here we report the results with both WSJ and interpolated language model.

The perplexity of the WSJ LM on children's speech is about 435. After interpolating, perplexity was 115, close to the perplexity of the WSJ LM on standard Nov92 test set (128). The OOV rate is only about 1.8%; that is quite normal.

USING WSJ LM	KIDS	RT
Adult Female Baseline Model	65.1%	107.2%
Adapted with KIDS	18.6%	80.7%
Adapted with RT	33.3%	49.2%
Adapted with KIDS+RT	16.8%	51.6%

Table 1. Recognition error rate with various models on KIDS & RT test data with mismatched WSJ language model

Using Interpolated LM	KIDS	RT
Adult Female Baseline Model	26.3%	93.3%
Adapted with KIDS	7.0%	62.4%
Adapted with RT	8.8%	27.6%
Adapted with KIDS+RT	6.3%	25.6%

Table 2. Recognition error rate with various models on KIDS & RT test data with interpolated language model

From Table 1 and Table 2 we find:

Difficulty of RT data. RT data are much more difficult for recognition than KIDS data. With various acoustic models, the error rate on RT data was more than the error rate on KIDS data by a factor of 3~9. For the baseline, RT data has an error rate around 100%!

Best performance achieved. For "clean" KIDS data, we achieved 6.3% error rate, which is comparable to the performance on adult speech with a similar task. However, the best error rate on RT data remains quite high (25.6%).

Effect of language model interpolation. As expected, the interpolated language model performs significantly better than WSJ language model with various acoustic models.

Effect of acoustic adaptation. RT adapted model did well on clean KIDS test data with relatively small error increase (7.0% to 8.8% with INT-LM) compared to KIDS adapted model. However, this does not hold true vice versa. Namely, KIDS adapted model did much worse on RT test data than RT adapted

model (27.6% to 62.4% with INT-LM). This may indicate that it is very important to get real-scenario data in order to improve the performance for a particular application.

Effect of automatically transcribed speech. Most importantly, the model adapted with both KIDS+RT data gives the best performance on both KIDS and RT test set with INT-LM. Therefore, automatically collected and transcribed data are good for children's speech recognition, *even across tasks*.

4. SUMMARY

In this paper we have described an inexpensive way to collect children's speech using Project LISTEN's Reading Tutor system. We also discussed how to purify the automatically collected and transcribed data with heuristics to ensure good quality. In addition, we found that gender dependent modeling is worthwhile considering for children's speech. Finally, experiments show that the automatically collected data can dramatically improve the recognition performance on children's speech even across tasks.

For future work, we need to more closely examine the effects of data filtering on recognition performance. We could also explore combinations of age-dependent (Potamianos et al. 1997) and gender-dependent children's acoustic modeling.

5. ACKNOWLDEGEMENTS

The development of the Reading Tutor and the Reading Tutor corpus (Section 2) was supported in part by the National Science Foundation under Grants No. IRI-9505156 and CDA-9616546 and by the first author's National Science Foundation Graduate Fellowship and Harvey Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the official policies, either expressed or implied, of the sponsors or of the United States Government. We thank Weekly Reader for permission to use materials in the Reading Tutor.

6. REFERENCES

1. Aist, G. S. Expanding a time-sensitive conversational architecture for turn-taking to handle content-driven interruption. To appear in ICSLP 1998, Sydney, Australia, 1998.

2. Aist, G. S., and Mostow, J., "Adapting Human Tutorial Interventions for a Reading Tutor that Listens: Using Continuous Speech Recognition in Interactive Educational Multimedia", *In Proceedings of CALL 97: Theory and Practice of Multimedia in Computer Assisted Language Learning*, Exeter, UK, 1997

3. Alleva F., Huang X. and Hwang M., "Improvements on the Pronunciation Prefix Tree Search Organization", *IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta, GA, May, 1996

4. Das S., Nix D. and Picheny M., "Improvements in Children's Speech Recognition Performance", *International Conference on Acoustics, Speech and Signal Processing*, Seattle, WA, May, 1998

5. Eskenazi, M., "KIDS: A Database of Children's Speech", *Journal Acoustic Society of America*, Vol 100:4, Part 2, December, 1996

6. Huang X., Acero A., Alleve F., Hwang M.Y., Jiang L. and Mahajan M., "Microsoft Windows Highly Intelligent Speech Recognizer: Whisper", *IEEE International Conference on Acoustics, Speech and Signal Processing*, Detroit, MI, May, 1995

7. Huang, X. D., Alleva, F., Hon, H. W., Hwang, M. Y., Lee, K. F., and Rosenfeld, R. 1993. The Sphinx-II Speech Recognition System: An Overview. Computer Speech and Language 7(2):137-148.

8. Hwang, M.Y. Huang X. and Alleva F., "Predicting Unseen Triphone with Senones", *IEEE International Conference on Acoustics, Speech and Signal Processing*, Minneapolis, MN, April, 1993

9. Lee S., Potamianos A. and Narayanan S., "Analysis of Children's Speech: Duration, Pitch and Formants", *European Conference on Speech Communication and Technology*, Rhodes, Greece, September, 1997

10. Mostow, J., and Aist, G. S., "The Sounds of Silence: Towards Automatic Evaluation of Student Learning in a Reading Tutor that Listens", *In Proceedings of the 1997 National Conference on Artificial Intelligence (AAAI 97)*, pages 355-361, 1997

11. Mostow, J., Roth, S. F., Hauptmann, A. G., and Kane, M. 1994. A Prototype Reading Coach that Listens. In Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle WA. Selected as the AAAI-94 Outstanding Paper.

12. Mostow, J., Hauptmann, A. G., Chase, L. L., and Roth. S. 1993. Towards a Reading Coach that Listens: Automatic Detection of Oral Reading Errors. In Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93), 392-397. Washington DC: American Association for Artificial Intelligence.

13. Potamianos A., Narayanan S. and Lee S., "Automatic Speech Recognition for Children", *European Conference on Speech Communication and Technology*, Rhodes, Greece, September, 1997

14. Wilpon J. and Jacobsen C. N., "A Study of Speech Recognition for Children and The Elderly", *International Conference on Acoustics, Speech and Signal Processing*, Atlanta, GA, May, 1996

15. Zavaliagkos G. and Colthurst T., "Utilizing Untranscribed Training Data to Improve Performance", *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, Feb, 1998