IMPROVEMENTS IN CHILDREN'S SPEECH RECOGNITION PERFORMANCE

Subrata Das, Don Nix, Michael Picheny

IBM Research Division
Thomas J. Watson Research Center
P. O. Box 218
Yorktown Heights, NY 10598

ABSTRACT

There are several reasons why conventional speech recognition systems modeled on adult data fail to perform satisfactorily on children's speech input. For instance, children's vocal characteristics differ significantly from those of adults. In addition, their choices of vocabulary and sentence construction modalities usually do not conform to adult patterns. We describe comparative studies demonstrating the performance gain realized by adopting to children's acoustic and language model data to construct a children's speech recognition system.

1. INTRODUCTION

Speech recognition systems are usually modeled on adult data. The acoustic data employed for designing its acoustic model component consists of a speech corpus collected from a number of adult male and female speakers. An example of this is the IBM speaker-independent continuous speech recognition system [1]. This system is developed by utilizing a speech database of several hundred adult speakers from both genders. Similarly, the design of the language model component of a speech recognition system is normally based on a collection of textual data generated by adults. The language model data used in the IBM system, for instance, includes regular office correspondences taken from an archival database.

It is well known that vocal characteristics of young children differ markedly from those of adults. For instance, a recent study [2] documents the progression of important speech parameters such as formant frequencies, as the age group under study changes from young children to mature adults. As a consequence, conventional speech recognition systems tend to perform poorly on children's speech. This type of degradation is verified in [3] for digit and phrase recognition tasks. The authors experiment with signal processing and acoustic modeling procedures to improve their performance. For instance, speaker normalization by frequency warping brings down their error rate by 45 percent in one case.

In this paper, we discuss some aspects of our work with a children's speech recognition system. We try frequency warping technique with moderate success. We collect both acoustic and language model data from children. The recognition system constructed from these data perform with a word recognition error rate of 10.46 percent.

Remainder of this paper is organized as follows. In the next section, we describe the databases used for our experiments. Section 3 is concerned with frequency warping studies. Section 4 describes how we construct a recognition system for children and the results of our experiments with this system. We conclude this paper with a summary and some observations in Section 5.

2. CHILDREN'S DATABASES

The first acoustic database called KidCC consists of utterances of five male and six female children reading desktop command and control sentences, such as, "Go to Lotus organizer", in a normal continuous manner. They range in age from 8 to 13. Each child reads 50 such sentences from a total list of 400 sentences. This database is used both conventionally and frequency warped to test the performance of a command and control speech recognition system.

In the next part of our work, we collect a second acoustic database, called KidAM, for the purpose of experimenting with the design of an acoustic model for children. We assume that the vocabulary size for young children could be considerably smaller than the tens of thousands of words, typically used for building a regular (adult) dictation system. In our case, the vocabulary, taken from a popular children's story, consists of 376 words. We record data from 395 children between the ages of 5 and 10, each of them reading about 70 isolated words from the story. We record the data at a relatively high sampling rate of 22 kHz. Children's voice tends to reach up to a frequency higher than that of adults. We want to study if we can exploit this higher bandwidth to improve the performance of our recognizer.

We collect a textual database, KidLM, as well for building a children's language model. It consists of 240,000 words of text composed by children. The subject matter relates to diverse categories ranging from essays, fictions and book reports to movie reviews, travelogues and poems. We edit this database minimally by correcting misspellings, but do not attempt to fix the ungrammatical or incomplete sentences.

Finally, the database, KidTEST, used for testing the system performance consists of speech taken from 6 male and 6 female children, who are asked to speak in an isolated-speech mode. Each of these children is 6 years of age. KidTEST data are recorded at a 22 kHz sampling rate, for the reasons given previously in connection with the description of the KidAM database.

3. FREQUENCY WARPING

Frequency warping is studied by a number of groups [3-5]. The goal is to reduce the mismatch between the training and test data. An optimal warp factor could be picked for each sentence, or, even for each phone. However, due to practical considerations, we decide to investigate two types of warping: speaker-dependent and speaker-independent.

As a first step, we decode the KidCC database sentences without any frequrncy warping, using a conventional (adult model) command-and-control speech recognition system. This is the baseline result. Word recognition error rates are listed in Table 1 under the column labeled "Unwarped".

Table 1. Error rates for unwarped and warped data

Spkr	Unwarped	SD warp	SI warp
F1	10.3	0.0	2.05
F2	5.1	0.0	0.0
F3	0.73	0.36	1.09
F4	19.08	1.14	1.91
F5	1.91	1.14	1.53
F6	29.37	8.33	8.73
M1	2.19	1.82	2.92
M2	4.73	4.73	4.73
M3	7.36	4.26	4.26
M4	3.57	1.59	6.35
M5	7.29	5.67	8.91
Average	8.33	2.64	3.86

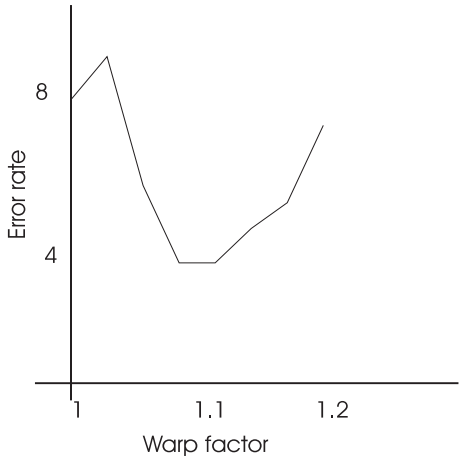

Figure 1. Effect of warp factor on error rate

We see error rates vary over a substantial range, from a low of 0.73 percent for the female speaker F3 to a high value of 29.37 percent for the female speaker F6. These results are compared with the improvements obtained after frequency warping, as listed under columns headed "SD warp" and "SI warp". "SD warp" refers to speaker-dependent warp where the warp factor is separately optimized for each individual speaker. In the case of speaker-independent "SI warp", a common warp factor optimized over all 11 speakers is used. On the average, "SD warp" works the best as expected, reducing the word error rate from 8.33 percent to 2.64 percent. Average error rate for "SI warp" is 3.86 percent over all test speakers.

Note that error rates go up for some speakers in this case. For instance, it increases from 3.57 percent to 6.35 percent for speaker M4. He is a 13 year old male with an adult voice. Consequently, speaker-independent warp hurts rather than helps his performance. The graph in Figure 1 shows how average error rate changes as a function of the SI warping factor. We see that the optimal warping factor in this case is 1.12.

4. A CHILDREN'S SPEECH RECOGNITION SYSTEM

Next, we turn our attention to developing a children's speech recognition system. We use the KidAM database of children's speech to design two acoustic models for children. The first one is made from data which are signal processed at a 22 kHz rate. This is called KidAM22. We downsample the data to an 11 kHz rate and construct the second acoustic model from this downsampled data. This is referred by the name KidAM11.

In addition, we experiment with two language models. One is our regular (adult) trigram language model, called AdultLM. The second one, KidLM, is developed by utilizing the KidLM database. This KidLM, constructed by utilizing our usual software package, is a trigram language model as well.

We conduct a number of experiments using these acoustic and language models. The configurations studied and the corresponding word error rates are listed in Table 2. The KidTEST database is used to carry out the tests in all cases. When we test a configuration with the KidAM22 component, KidTEST data are signal processed at a 22 kHz rate for compatibility. When KidAM11 is used as the acoustic model, KidTEST data are downsampled to an 11 kHz rate during signal processing.

Comparing the results of experiments 1 and 2, we see that 22 kHz processing helps to improve the recognition performance. But the greatest improvement is due to switching to the KidLM, as seen from the results of experiment 3. The final error rate is 10.46 percent. Table 3 lists the error rates observed for each child in this case. Note that these children are different from the ones listed in Table 1. Scrutinizing the results, we observe that some of the errors are due to pronunciation problems. For example, "gives milk" which is a portion of a sentence uttered by F9 is decoded as "is not". First part of the word "gives" is barely audible, causing one decoder error. In addition, "is not" is preferred by the language model over "gives milk". We see that we need to pay more attention to features pertinent to a children's speech recognition system.

Table 2. Error rates for different configurations

Expt.	Sampling Rate(kHz)	Acoustic Model	Language Model	Word error rate (%)
1	11	KidAM11	AdultLM	23.67
2	22	KidAM22	AdultLM	20.73
3	22	KidAM22	KidLM	10.46

Table 3. Error rate for each child

Spkr	Error rate
F7	6.9
F8	14.0
F9	18.6
F10	23.3
F11	6.5
F12	0.0
M6	14.6
M7	16.7
M8	2.0
M9	8.0
M10	6.0
M11	10.0
Average	10.46

5. CONCLUSIONS

We conduct a number of experiments with children's data with the goal of developing a children's speech recognition system. Using some command and control data, we observe the benefit of frequency warping. Next, we collect both acoustic and language model data for children. We construct acoustic and language models for children using these data and study word recognition error

rates under different configurations. We conclude that use of a children's language model substantially improves the recognition performance. Future work is directed towards expanding the vocabulary as well as improving the performance of the recognizer.

6. REFERENCES

1. L. Bahl, S. Balakrishnan-Aiyer, J. Bellegarda, M. Franz, P. Gopalakrishnan, D. Nahamoo, M. Novak, M. Padmanabhan, M. Picheny and S. Roukos, "Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task," vol. 1, pp. 41 - 44, Proc. ICASSP-95, 1995.

2. S. Lee, A. Potamianos and S. Narayanan, "Analysis of children's speech: duration, pitch and formants," Proc. Eurospeech'97, vol. 1, pp. 473 - 476, September 1997.

3. A. Potamianos, S. Narayanan and S. Lee, "Automatic speech recognition for children," Proc. Eurospeech'97, vol. 5, pp. 2371 - 2374, September 1997.

4. S. Wegmann, D. McAllaster, J. Orloff and B. Peskin, "Speaker normalization on conversational telephone speech," Proc. ICASSP-96, pp. 339 - 341, May 1996.

5. L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," Proc. ICASSP-96, pp. 353 - 356, May 1996.