The OGI kids' speech corpus and recognizers

Conference Paper · January 2000

Source: DBLP

3 authors, including:

Ronald A. Cole
Boulder Learning Inc.

255 PUBLICATIONS **5,354** CITATIONS

Some of the authors of this publication are also working on these related projects:

Project   FLORA: FLuent Oral Reading Assessment View project

Project   Enabling Automatic Language Identification Research (1992-1996) View project

THE OGI KIDS' SPEECH CORPUS AND RECOGNIZERS

Khaldoun Shobaki*, John-Paul Hosom*, and Ronald A. Cole**

* Center for Spoken Language Understanding (CSLU)
Oregon Graduate Institute of Science and Technology (OGI)
20000 N.W. Walker Rd., Portland Oregon, 97291 USA
e-mail: {shobaki,hosom}@cse.ogi.edu   www: http://cslu.cse.ogi.edu/

** Center for Speech and Language Research (CSLR)
University of Colorado, Boulder
3215 Marine Avenue, Boulder Colorado, 80303 USA
e-mail: cole@cslr.colorado.edu   www: http://cslr.colorado.edu/

ABSTRACT

We describe a corpus of children's speech, called the OGI Kids' Speech corpus, and a speaker- and vocabulary-independent recognition system trained and evaluated with these data. The corpus is composed of both prompted and spontaneous speech from 1100 children from kindergarten through grade 10. The prompted speech was presented as text appearing below an animated character (Baldi) that produced accurate visible speech synchronized with recorded prompts. The speech and text consists of isolated words, sentences, and digit strings. A phonetic recognizer was trained using an HMM/ANN framework, with training data taken from intervals of speech associated with phonetic segments in the isolated words in the corpus. Phonetic segments were derived using automatic phonetic alignment. To find out how well the recognizer is able to generalize to new words not found in the training set, we performed two test-set evaluations: one using a new set of utterances from the set of 205 words spoken in isolation (similar to the data used to train the recognizer) and one using words from the prompted sentences. Results were dramatically different (97.5% for isolated vs. 37.9% for words in sentences), and we explore methods that may be used to improve the recognizer's ability to generalize to new words.

1. INTRODUCTION

Although much effort has been invested in conducting research on male and female adult speech, efforts that target children's speech are less common. In order to address this issue, we have developed a corpus of children's speech as well as an initial speaker- and vocabulary-independent children's speech recognizer trained on this corpus. We first describe the protocol and methodology used to develop the OGI Kids' Speech corpus. We then describe development and evaluation of a recognizer trained on isolated words in the corpus. The recognizer has been ported to the CSLU Toolkit [1], and both the corpus and recognizer are freely available for research use.

2. OGI KIDS' SPEECH CORPUS

The OGI Kids' Speech corpus is a collection of spontaneous and read speech recorded at the Northwest Regional School District near Portland, Oregon. It consists of words and sentences from approximately 1100 children. A gender-balanced group of approximately 100 children per grade from Kindergarten through grade 10 participated in the collection.

Several constraints shaped the design and collection of the corpus. First, data collection was done during classroom hours, so it was not possible to take students out of their normally scheduled classes for more than 30 minutes. Second, the same protocol was to be recorded for all of the subjects. This constrained the vocabulary to a set that very young children could understand and produce. Third, it was necessary to collect a sufficient number of speaker productions of each utterance for training and evaluation. Under these constraints, the protocol was designed to provide as much coverage as possible of the most common American English biphones.

The final protocol consisted of 205 isolated words, 100 prompted sentences, and 10 numeric strings. The protocol was split into five sub-protocols to meet time constraints imposed by individual sessions.

Two computers running the Windows NT 4.0 operating system were set up at each school to perform data collection. The data collection software was written using the CSLU Toolkit. For each utterance, the text of the prompt was displayed on the screen, and a human recording of the prompt was played, in synchrony with facial animation using the animated 3D character "Baldi." The subject then repeated the prompt, which was recorded via a head-mounted microphone and digitized at 16 bits and 16 kHz using a SoundBlaster 16 PnP sound card. The recorded utterance was then played back to the subject and the data-collection supervisor. If the recording was deemed unacceptable, the prompt was repeated. After the prompted speech phase of the collection was completed, the experimenter asked the subject a series of questions intended to elicit spontaneous speech (i.e. "Tell me about your favorite movie"). The total amount of speech recorded per subject was approximately 8-10 minutes. Some additional biographical information was collected about each subject, including age, gender, languages spoken, and any physical maladies that could affect their speech.

Although the supervisor verified each utterance during data collection, two individuals subsequently verified each utterance independently. Each utterance was rated on a three-point scale consisting of "good" utterances (the word is clearly intelligible

with no significant background noise or extraneous speech), "questionable" utterances (intelligible but accompanied by other sounds) or "bad" utterances (unintelligible or wrong word spoken). Good utterances were selected for training and testing via this process. Additionally, the spontaneous speech components of the corpus were orthographically transcribed according to published CSLU transcription conventions [2, 3]. Phonetic transcription of the corpus is currently underway, and will be described in [4]. A summary of the number of subjects and recordings is shown, per grade, in Table 1.

Grade	Male Subjects/Recordings	Female Subjects/Recordings
K	39 / 1142	49 / 1915
1	58 / 3921	31 / 2032
2	53 / 3584	61 / 2032
3	63 / 4194	52 / 3516
4	47 / 3178	45 / 2976
5	49 / 3361	49 / 3362
6	57 / 3912	55 / 3774
7	46 / 3136	51 / 3499
8	49 / 3362	50 / 3431
9	69 / 4606	40 / 2677
10	75 / 5084	29 / 1989

**Table 1.** Number of subjects and recordings, per grade, in the OGI Kids' Speech corpus.

3. RECOGNITION SYSTEM OVERVIEW

A children's speech recognizer has been developed using the OGI Kids' Speech corpus and the CSLU Toolkit's hybrid HMM/ANN framework [5], as illustrated in Figures 1 and 2. The system uses a 10-msec frame rate and MFCC features with a window size of 16 msec. Delta values of the 13 MFCC features are also used, for a total of 26 features per frame. Cepstral-mean subtraction (CMS) is performed in order to reduce convolutional noise. A neural network is used for classification of features into context-dependent sub-phonetic categories. The neural network is input a set of features for the frame to be classified, as well as a set of features from adjacent frames: –60, –30, 30, and 60 msec relative to the frame of interest, for a total of 130 inputs to the network. This "context window" of features is used to provide the network with information about the dynamics of the speech signal, in addition to the information provided by the delta values. As the speech in the Kids' corpus had not yet been labeled with time-aligned phonemes, the training data were phonetically aligned using a standard forced-alignment procedure. The pronunciations for each word were obtained from the CMU dictionary; pronunciations not in this dictionary were obtained from the Festival speech synthesis system's letter-to-sound rules.

The system was trained to recognize context-dependent units. The phonetic contexts were grouped into clusters of phonemes using a classification and regression tree (CART) in order to reduce the number of output categories from the network. Each target phoneme was split into one, two, or three sub-phonetic parts. The initial part is dependent on the context of the preceding cluster, the center part (if any) is context independent, and the final part is dependent on the following cluster. Phonemes that remain as a one-part phoneme can either be context-independent (for example, /.pau/) or dependent on the following phoneme (for example, /th/).

At each frame, the neural network classifies the features in the context window into phonetic-based categories, estimating the
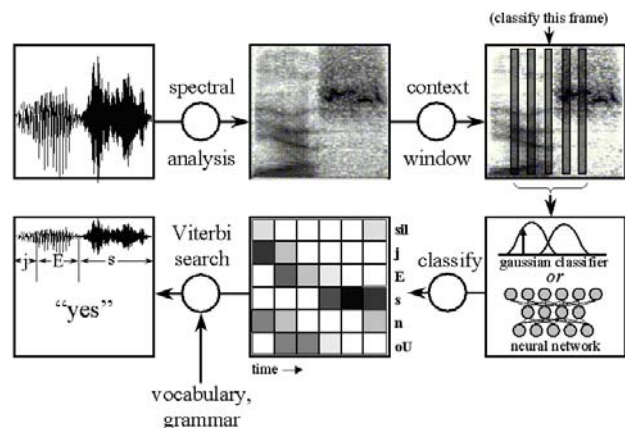
Figure 1. Illustration of the recognition process using the
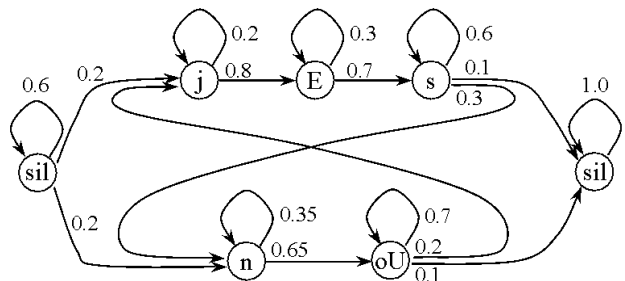
Figure 2. Illustration of a state sequence in an HMM or HMM/ANN hybrid system.

probabilities of each category being represented by that set of features. Fully connected feed-forward networks were trained for 30 iterations using standard back-propagation. A Viterbi search is used to find the best path through these probability scores, constrained to the state sequences that are allowed by the vocabulary and grammar. This is illustrated in Figure 2, where each state represents a phonetic-based category, and there are certain probabilities of transitioning from one state to another.

One major difference between this framework and standard HMM systems is that the phonetic likelihoods are estimated using a neural network instead of a Gaussian mixture model (GMM). Using a neural network to do this estimation has the

advantage of not requiring assumptions about the distribution or independence of the input data, and neural networks easily perform discriminative training [6]. Also, neural networks can be used to perform recognition much faster than standard HMMs. A second difference is in the type of context-dependent units; whereas standard HMMs train on the context of the preceding *and* following phonemes, our system splits each phoneme into states that are dependent on the left or right context, or are context independent.

Three speaker-independent partitions were created from the corpus: 3/5 for training, 1/5 for development testing, and 1/5 for final testing. The total number of files used for training was 20448, with isolated-word utterances from grades 2 through 10. In addition to training on data from the OGI Kids' Speech corpus, 50 examples of each category were obtained from the TIMIT corpus in order to "pad" those categories that had insufficient training data. Development and test-set evaluation of the recognizers was done by recognition of 205 isolated-word utterances in the corpus, on a total of 6826 files. In addition, test-set accuracy on 100 words not seen in training was evaluated using 900 files (with each word extracted from continuous speech sentences in the corpus).

4. TRAINING METHODOLOGY

Because hand-labeled data were not available at the time that the recognizer was trained, phonetic labels were created using automatic alignment. However, as the vocal-tract properties of children change with age, it was felt that the most accurate alignments could be obtained by training one grade at a time, from higher grades to lower grades. The recognizer for each grade was trained on the data for that grade and all higher grades. Data from kindergarten and first grade were not included in training or evaluation.

An adult-speech recognizer currently in the CSLU Toolkit was used to automatically align the data for the 10th-grade speech. This 10th-grade recognizer was used to automatically align the 9th grade data, and the 9th-grade recognizer was then trained on the data from *both* the 9th and 10th grades, as well as the 50 samples per category of TIMIT data. Up to 8000 samples per category from each grade were used in training. The 9th-grade recognizer was used to automatically align the 10th, 9th, and 8th grade data, and the 8th-grade recognizer was trained on the data from all three grades (with up to 3300 samples per category from each grade) as well as the same TIMIT data. This process continued until a single recognizer had been trained on data from grades 2 through 10, inclusive. For grades 7 through 2, the requested number of training samples per category *from each grade* was 2500, 2000, 1666, 1143, 1000, and 888, respectively. Thus, the total number of requested training samples per category was about 8000 for all recognizers.

The method of training each new recognizer on the data from the current grade and all higher grades carried the risk of making the data harder to learn (because the different voice qualities of the children in each grade implies a larger variance of the data within each category), but had the advantage of yielding a single recognizer that is not specific to any one grade.
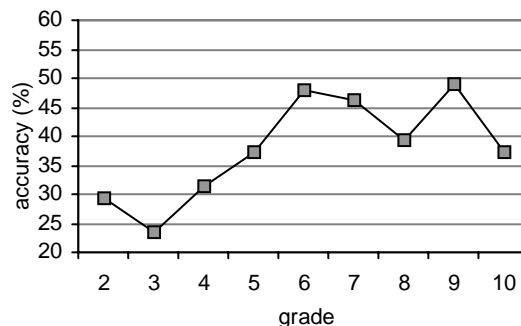
5. RESULTS

In order to evaluate the performance of isolated-word recognition on the words in the training corpus, the recognizer vocabulary was set to the list of 205 isolated words used in training (although different utterances of these words were used in training and evaluation). Test-set evaluation of the recognizer was done separately for each grade level. The results for each grade are plotted in Figure 3; it can be seen that accuracy ranged from 93.1% to 99.5%, with an average of 97.5%.
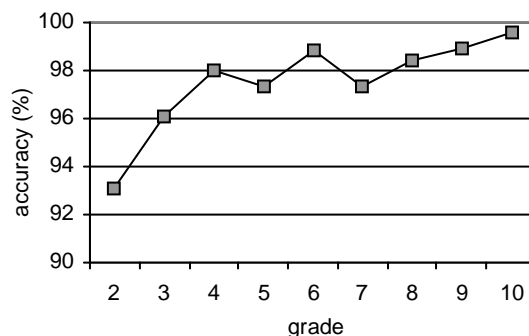
Figure 3. Test-set results, evaluated at each grade on words in the training corpus. The grade is on the *X* axis, and accuracy is on the *Y* axis.

In order to evaluate the ability of the recognizer to generalize to

Figure 4. Test-set results, evaluated at each grade on words not in the training corpus. The grade is on the *X* axis, and accuracy is on the *Y* axis.

words that had not been encountered in the training data, content words from the prompted sentences in the test-set partition of the corpus were extracted and used for evaluation. In this case, a list of 100 words was used (one word from each sentence), and each sentence was randomly selected from all grades, for a total of 900 utterances in the test set. This task is much harder, not only because the contexts of the phonemes are different, but also because the words were recognized from continuous speech and therefore have a higher degree of coarticulation than isolated words.

In this second evaluation, with results plotted in Figure 4, accuracy ranged from 23.5% to 49.0% correct, with an average of 37.9%.

6. CONCLUSION

The OGI Kids' Speech corpus provides a large database for research on children's speech, including data from grades K through 10. These data have been used to train a speaker- and vocabulary-independent recognizer of children's speech. In subsequent articles, we will report results of detailed analyses of speech produced by children of different ages, following the work of Lee, Potamianos, and Narayanan [7].

The method of iterative training that includes data from all higher grades proved to yield high accuracy when evaluated on the same words found in the training set. When different words were used in evaluation, however, the performance of the recognizer changed dramatically. We believe that the large variance of the data from all grades was not a factor in the poor performance, and that failure to generalize to new words can be ascribed to differences between the training and test sets. When the recognizer was trained and tested on different occurrences of the same set of 205 words, it was highly tuned to the acoustic feature patterns of phonemes in these specific words. For example, features at the boundaries of the context window (+60 msec; -60 msec) often measured periods of relative silence before and after the word. When presented with words in fluent speech, the features differed in two significant ways from words in isolation. First, phonemes within the fluent speech test words were more variable due to coarticulation with preceding and subsequent words. Second, rather than periods of silence before and after each word, the network is now presented with acoustic features of preceding and following words.

We are now phonetically labeling the fluent speech in this corpus, conducting research to understand differences between prompted and extemporaneous speech, and will be training recognizers on both types of data.

To develop recognizers that will generalize to new vocabularies, including prompted and extemporaneous speech, we plan the following:

(a) to use the read sentences and spontaneous continuous-speech data in training, thereby providing a richer set of phonetic contexts,

(b) to train recognizers using phonetic hand-labels when available,

(c) to train recognizers using a new method of automatic alignment [8] when hand labels are not available, to better focus the training on specific phonetic properties,

(d) to evaluate different clusterings of the phonetic contexts, and

(e) to add a model for breath noise to the recognizers.

We expect that the first four of these items will improve recognition performance of words not found in the training corpus, and that training a model for breath noise will reduce errors caused by this common phenomenon

7. ACKNOWLEDGEMENTS

8. REFERENCES

1. Sutton, S., Cole, R., de Villiers, J., et al., "Universal Speech Tools: the CSLU Toolkit," ICSLP '98, Sydney, Australia, Nov. 1998, pp. 3221-3224.

2. Lander, T., "The CSLU Labeling Guide," technical report CSLU-014--96, Center for Spoken Language Understanding, Oregon Graduate Institute, Jun., 1996.

3. Details about the corpus can be found on-line at CSLU's web site: http://cslu.cse.ogi.edu/corpora/kids

4. Kawai, G., Shobaki, K., Lander, T., Durham, T., Corson, L., Krech, H., and Cole, R., "A corpus of American English spoken by children," *in these proceedings*.

5. Hosom, J.P., Cosi, P., and Cole, R.A., "Evaluation and Integration of Neural-Network Training Techniques for Continuous Digit Recognition," ICSLP '98, Sydney, Australia, Nov. 1998, pp. 731-734.

6. Bourlard, H., "Towards Increasing Speech Recognition Error Rates," Eurospeech '95, vol. 2, Madrid, Spain, Sep. 1995, pp. 883-894.

7. Lee, S., Potamianos, A., and Narayanan, S., "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," JASA, vol. 105, no. 3, Mar. 1999, pp. 1455-1468

8. Hosom, J.P., "Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information," Ph.D. thesis, Oregon Graduate Institute, May 2000.