Comparing speech recognition for adults and children

Daniel Elenius and Mats Blomberg
Department of Speech, Music and Hearing, KTH, Stockholm

Abstract

This paper presents initial studies of the performance of a speech recogniser on children's speech when trained on children or adults. A connected-digits recogniser was used for this purpose. The individual digit accuracy among the children is correlated to some features of the child, such as age, gender, fundamental frequency and height. A strong correlation between age and accuracy was found. The accuracy was also found to be lower for child recognition than for adult recognition, even though the recognisers were trained on the correct class of speakers.

Introduction

Research on techniques for speech recognition has mostly been dealing with adult speech. However, children would certainly benefit from this modality as well - especially in the ages below which they have acquired reading and writing skills. They start using computers at early ages, in schools, at home, and at day-care and after-school centres. As voice-based applications are introduced in an increasing number, it is essential that children should be able to use these as well. It is therefore of high interest to extend the capability of speech recognition systems to be able to deal with this speaker category. This research area is one activity in the EU project PF-Star (Preparing Future Multisensorial Interaction Research), started in 2002. The runtime is two years and the partners are: ITC-irist (coordinator, Italy), the universities of Erlangen-Nurenberg, Aachen, Karlsruhe, Birmingham and KTH. Other aims of the project are synthesis and analysis of emotional expressions in speech, the face, and speech-to-speech translation.

Adapting the ASR technology

There are a number of differences between adult and children's speech. Due to a shorter vocal tract and smaller vocal folds children have higher fundamental and formant frequencies than those of adults. As they are young they may be less experienced in articulating sounds. In some cases this results in a child systematically substituting a phoneme with another, for example Swedish children may sometimes substitute /r/ by /j/. Due to their limited experience of a language they have not learned all words adults use, resulting in a smaller vocabulary. On the other hand children sometimes use their imagination and associative skills to invent their own words. Therefore the vocabulary used by children may differ from the one adults use.

Research has been conducted targeted on adapting adult speech recognition technology towards children. In particular the acoustic differences are a concrete domain to investigate. Das, Nix and Picheny (1996) have made an experiment were vocal tract length normalisation (VTLN) was applied to an adult recogniser to adapt it to children's speech. In this study a database of only five male- and six female children was used. These children read 50 commands each from a set of 400 sentences. An adult recogniser gave an average error rate of 8.33%. After speaker dependent frequency warping was applied the error rate was reduced to 2.64%.

Sjölander and Gustavsson (2001) adapted a speech recogniser were they did not have the access to the feature extraction layer, preventing VTLN from being applied. Instead of altering the feature extraction they normalised the signal, prior to feature extraction, using voice transformation techniques. In their experiment the word error rate for children before normalisation was 43% compared to 19% for adults. Using a voice transform scheme on the recorded signal the error rate for children was decreased to 31%. A second experiment was run using two age groups, three to nine years and ten to twelve years, and a larger vocabulary. The word error rate of the older children was 36% compared to 59% for younger children. Using the voice transform these error rates were decreased to 31% and 41% respectively.

Altering the signal or adjusting the feature extraction algorithms has been shown to improve recognition. However, the methods may also introduce unwanted distortion of the signal, possibly limiting the improvement otherwise obtained. Therefore it may be advantageous to introduce adaptation of the recogniser

to be familiar with the introduced distortion. Narayanan and Potamianos (2003) used a combination of VTLN and linear parameter transformation to adapt a recogniser. The parameter transform was defined as a common shift of all mean parameters in the HMM. An optimal frequency warping and shift was estimated using the maximum likelihood criterion on the observation likelihood given the HMM. In the experiment the shift of model parameters was implemented by shifting the feature-vector and keeping the model fixed. In the experiment two recognisers were used, one trained for adults and one for children. The speakers used for training the children's speech recogniser were from 10 to 17 years. These recognisers were then evaluated using 6 to 17 year old speakers. Common for all setups was that a rapid improvement of the performance was seen from an age of 7 years up to an age of 13 years. At the age of 13 the performance was close to that of adult recognition.

Table 1. Word accuracy of two recognisers, one trained for adults and one for children Narayanan and Potamianos (2003).

Set-up	Word accuracy for 7 year olds	Word accuracy for 13 year olds
Recogniser for adults	62%	95%
Adult recogniser + Norm	76%	96%
Recogniser for children	85%	96%
Child recogniser + Norm	89%	97%

Word accuracies of the recognisers are shown in Table 1. The recogniser dedicated for children had a higher accuracy on children's speech than the adult recogniser. Both recognisers improved when a combination of VTLN and shift of feature vectors were applied. The method was more beneficial for the adult recogniser than for the child recogniser. Maybe this is because of a greater difference between the vocal tract length between adults and children than between children.

Experiment

We have previously reported on the ongoing recording activity at KTH in the PF-Star project (Elenius and Blomberg, 2003). These recordings were made with 24 bits at 32 kHz via a headset microphone and an omni-directinal device standing on a table in front of the child. Audio-prompts were read by the recording leader and the prompts were repeated by the child. This was done in a separate room to suppress noise produced by the other children at day-care and after school centres.

Our first use of this corpus is to compare the performance of speech recognition for adults and children. As we are primarily interested in the implications of the acoustical differences between children's and adult's voices, the fairly constrained vocabulary and grammar of digit-strings was used. We also believe that this choice keeps the influence that audio-prompting might have imposed on the children's natural choice of wording to a minimum.

Adult speech was gathered from the SpeeCon database (Iskra et.al 2002). The recordings were made via a head-set microphone at 16 kHz. Two sets of 60 speakers, speaking 30 digit strings each were formed. These groups were used for training and testing respectively.

Recordings of children were taken from our own database, described earlier. Each child spoke 30 digit strings. These were recorded through the same head-set as the adult speech. A training set of 60 children was formed with an equal number of children of each age group to avoid emphasising any specific age. The test-set was created with a similar sex distribution as the training-set. At the time for the experiment, the distribution of recorded children was not evenly distributed regarding age. To make the set of speakers for evaluation as large as possible, a non-uniform distribution of children over age was used. A compensation for this was made by normalising the results according to age.

The digit-string recogniser was based on HMMs. Each digit model consisted of two times as many states as the number of phonemes in the word. A standard left-to-right model was used with the extension of a transition to skip one state in order to account for variation in pronunciation clarity. The system operated on a speech signal sampled at 32 kHz, a cepstrum frame was produced each 10 ms using a mel-scaled filterbank and a Hamming window. The unusually high sampling frequency was chosen because of the higher bandwidth in children's voices. As the adult speech, stored in the SpeeCon database, was sampled at 16 kHz the sampling frequency of this data was converted to 32 kHz sampling frequency, to fit the recogniser, using linear interpolation.

Results

The recogniser that had been trained on adult speech worked well when tested on adult speech. But the performance of this recogniser was substantially lower for children (***Table 2***).

Table 2. Word accuracy of two recognisers run on child and adult speech.

Accuracy	Evaluation set	
Training set	Adults	Children
Adults	97%	51%
Children	61%	87%

The recognition for children improved when the acoustical model was retrained on children's speech. The performance was however still lower than for adults. To make the picture completed, this recogniser was tested on adult speech with a low accuracy as a result.
A correlation analysis between the accuracy and age, fundamental frequency, first formant, second formant, height of the child and gender was done. The correlation is sorted in descending order in ***Table 3***. These features are not independent. For instance the correlation coefficient between height and F0 was –0.45. The strongest correlation occurs for age. Gender exhibits quite low correlation.

Table 3. Correlation between some features and the accuracy of the recogniser

Feature	Correlation to accuracy
Age	0.48
Average F0	-0.45
Height	0.41
Average F2	-0.21
Gender	0.11
Average F1	-0.07

Recognition of some young children works well, while others were hard to recognise. As shown in Figure 1, the variance in performance is reduced for the older children compared to that of younger ones.

Discussion

The performance level of children's speech when using adults' models is much too low to be practically useful. The low accuracy is probably caused by several factors, the most important being a shorter vocal tract and deviating pronunciation. Also, some children appeared shy and spoke in a very soft, breathy, voice, possibly because they were uncomfortable duo to the recording procedure. It is likely that the resulting voice quality has lowered their recognition accuracy.

To improve the recognition rate, it will be important to determine the relative influence of the observed features and find compensation techniques for the mismatches. One way to find the important factors would be to use principal component analysis.

As was seen, performance is dependent of the age of the speaker. We are currently running experiments were MAP and MLLR are used to create age-dependent models. Preliminary results indicate a reduction of the error rate by 30%.

In a study by Das, Nix and Picheny (1996), vocal tract length normalization (VTLN) improved recognition of children by 6%-units. An improvement is not surprising since a recognizer for children is confronted by a diverse set of vocal tract sizes in a fairly large range on the relative scale. We also saw in our study that a correlation between accuracy and the height of the child and F0 existed, indicating sensitivity against a difference in the size of the vocal tract. Normalizing for this attribute reduces the variability and thereby improves performance. Narayan at.al (2002) also combined VTLN with a simple model transform resulting in an improvement of up to 14%-units. As VTLN has proven useful for other languages, we are going to try it also for our recordings of Swedish children.

As can be seen in Figure 1, even in the youngest age group, there are individual children that reach perfect recognition. Evidently, this is possible in spite of the expected smaller physical size of these subjects. It is probable that these children have a pronunciation that conforms to that of the older children and that this overrides the size differences. A possible method to improve the recognition accuracy of the poorer children is to include deviating pronunciation in the digit models.

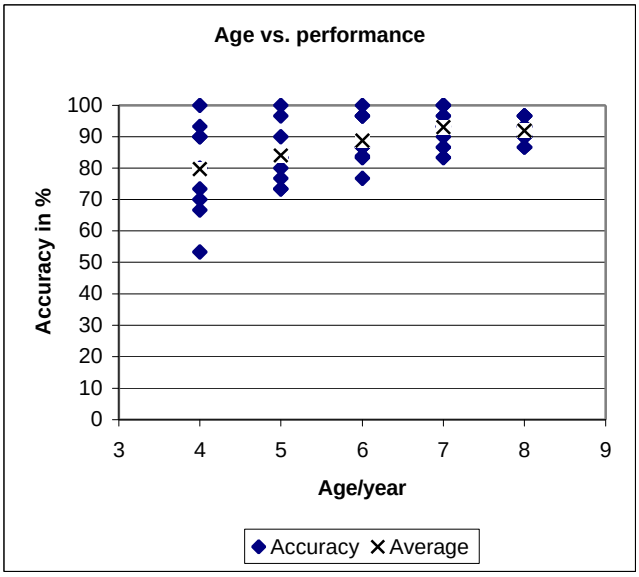

Figure 1. Digit accuracy of a connected-digit recogniser as a function of age. Each data point indicates the individual accuracy for one child.

Conclusion

A digit-string recogniser trained on adult speech performed much worse for children. After training the acoustical model on utterances spoken by children a substantial improvement was found. It was also observed that the variability in performance for young children was larger than for older children. Further improvement should be possible by using age-dependent models, age-group adaptation, vocal tract length normalisation and modelling of deviating pronunciation of young children.

Acknowledgement

This work was performed within the PF-Star project, funded by the European Commission (IST-2001 37599).

References

Das S., Nix D., Picheny M. (1996): “Improvements in Children's Speech Recognition Performance”. Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume: 1 , 12-15 May 1998. pp. 433 – 436.

Elenius K. (2000). Experiences from collecting two Swedish telephone speech databases. Int. Journal of Speech Technology 3:119-127.

Elenius, D., Blomberg, M. (2003). Collection and recognition of childrens speech in the PF-Star project. Proc. of Fonetik 2003 Umeå University, Department of Philosophy and Linguistics PHONUM 9, 81-84

Gustafson, J and Sjölander, K (2002): ”Voice Transformations For Improving Children's Speech Recognition In A Publicly Available Dialogue System”. In the Proceedings of the International Conference on Spoken Language Processing 2002, pp 297 - 300.

Iskra, D., Großkopf, B., Marasek, K., v. d. Heuvel, H., Diehl, F., Kiessling, A. (2002). SpeeCon - speech data for consumer devices: Database specification and validation. Second International Conference on Language Resources and Evaluation 2002.

Narayanan, S and Potamianos, A (2002): “Creating conversational interfaces for children”. IEEE Transactions on Speech and Audio Processing, Volume: 10, Issue: 2, Feb. 2002, pp 65 – 78.