Adaptation techniques for children's speech recognition

Daniel Elenius @ KTH/TMH 2004

1 Introduction

Speech recognisers have mostly been trained and evaluated on adult speech. Applications using this technology might work for adults but what about children? Does the technology work for them and if not, is it possible to adapt the technology in order to improve recognition of children's speech?

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

2 Adaptation techniques

A speech recogniser might be divided into four parts: the digitising, feature extraction, acoustic modelling, and language modelling part. When the signal is digitised an appropriate sampling frequency needs to be chosen to catch the high-pitched voices of children. Feature extraction often involves approximating the spectrum of the signal. During this phase it is possible to use signal processing to alter the spectrum prior to recognition. This may be used to decrease the sensitivity against background noise or to normalize some spectral characteristics of the recorded speech. Another method to target the current spectral characteristics may be to adapt the acoustic models to better match the speech. Adjustment of the language model may be used to take the children's vocabulary and pronunciation into account. As was seen adaptation may be performed on many levels. Adaptation has become linked with altering the acoustical model in the speech recogniser. In this paper however the term will be used somewhat more freely incorporating other techniques to adapt the speech technology.

2.1 Vocal tract length normalization

Speakers have different lengths of their vocal tract. Speakers with a short vocal tract tend to have a high-pitched voice and vice versa. This difference in speakers' voices results in a mismatch between training and usage of a speech recogniser and hence performance may suffer from the acoustical mismatch. During feature extraction it may be possible to compensate for this difference. This is sometimes referred to as vocal tract length normalization (VTLN). The basis of this method is that extending a tube by a factor alpha scales the frequency of the spectrum by the same factor. For instance a resonance at 100 Hz may be changed to 200 Hz by doubling the length of the tube. In theory rescaling the frequency axis may therefore compensate for a shorter vocal tract.

Narayanan and Potamianos (2002) have investigated the performance of a speech recogniser as a function of age and have tried VTLN as a means of improving the recogniser. They trained two recognisers, one for children and one for adults. 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.

Set-up	Word accuracy at an age of 7 years	Word accuracy at an age of 13 years
Recogniser for adults	62%	95%
Adult recogniser + VTLN	76%	96%
Recogniser for children	85%	96%
Child recogniser + VTLN	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 by applying VTLN. 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.

2.2 Voice transformation

Sometimes one may want to adapt a speech recogniser to which one does not have the access to the feature extraction layer. This was the case for Sjölander and Gustavsson (2001). Instead of altering the feature extraction they normalized the signal using voice transformation techniques. In their experiment the word error rate for children 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%. These children were also divided into two groups. One group consisted of children of three to nine years, and the other contained children of 10 to 12 years. The word error rate of older children was 36% compared to 59% for younger children. Using the voice transform these error rates were decreased to 31% and 41% respectively.

As was seen in the previous studies, Sjölander and Gustavsson (2001), and Narayanan and Potamianos (2002), a high word error rate may result from using an adult speech recogniser that has not been adapted for children. In particular young children seem trickier to recognise than older children. It was also seen that the word error rates of the recognisers was decreased when the technology was adapted for children.

2.3 Model adaptation

In section 2.1 it was seen that the feature extractor in the speech recogniser might be adjusted in order to improve the recognition of a child's speech. It was also seen that an alternative to this adaptation technique was to alter the recorded signal using voice transformation techniques. In this section the focus will lie on techniques for adapting the acoustical model of a speech recogniser. Such a model may be adapted by using a transform of the model parameters based on suitable speech material. Two properties that limit the precision of this kind of adaptation is the size of adaptation material and how the adaptation is performed.

Adaptation using a speech material alters the acoustical models to better suit the current speech material. Since all acoustic features of this material may possibly affect the adaptation of the model, the material needs to be chosen carefully. An adaptation towards the room acoustics might be achieved by recording a large number of speakers in a given room. In this case it is important that a large number of different speakers is recorded to avoid adapting towards any particular speaker as well as the properties of the room. If no regard to any particular room or any individual speaker is to be taken in a straightforward adaptation, a large number of speakers recorded in a number of rooms are needed. Normalisation using this technique quickly increases the demands on the adaptation set as the number of factors grows. Other methods to reduce the influence of some factors are therefore needed. For instance, to reduce the effects of the room acoustics a close talking microphone might be used. This reduces the effects of the room

acoustics, as the intensity of the speaker's voice is then much higher than the echoes produced by the room.

A phoneme recogniser may use one HMM per phoneme. When the current acoustics differ from the training conditions, a number of probability density functions of each such HMM need to be adapted. Transforming each phoneme separately may require a large set of adaptation data, especially as some phonemes may be quite rare in spoken utterances. Some phonemes may need a similar transform, which makes it possible to use clustering techniques to reduce the demand on the size of the adaptation material.

If the adaptation set is of medium size, coarse transforms might first be estimated separately and then grouped based on similarity to form more reliable transforms. The problem, in this case, is that the poor initial estimates of the transforms might lead to hasty conclusions regarding which transforms are similar. Another problem is that this technique does not allow adapting phonemes that are not present in the adaptation set. In this case some other method is needed to cluster transforms.

A more indirect clustering technique may be based on the assumption that some common characteristic of phonemes would indicate that they ought to be transformed similarly. For instance if phonemes have similar probability density functions it might be argued that these phonemes ought to have similar density functions also after the transform. One method of meeting this criterion is to use the same transform for these phonemes. This clustering strategy is useful in practice since the data for clustering is stored in the HMM and hence the adaptation data is not used to choose which phones to group together. Thereby it is possible to deduce transforms for phones not present in the adaptation material.

Another criteria for clustering of phones that would provide transforms of unseen phones, in the adaptation material, is to group phones that are articulated similarly. As the vocal tract has a similar shape it might be argued that these phonemes would be transformed in a similar fashion as the speaker grows up. This line of argument is in some sense similar to the spectral argument.

A method to estimate a transform of a model is described by Leggetter and Woodland (1995). The method is called maximum likelihood linear regression (MLLR), which may be adapted to different sizes of adaptation material by creating regression classes by clustering techniques. A transform may then be estimated for each regression class using MLLR. This method adapts Gaussian distributions by transforming the mean vector, while keeping the covariance matrix fixed. An assumption is therefore made that the variance is independent of the absolute value of the mean values. This reduces the demands on the size of the adaptation material compared to a case in which the variances would be transformed.

Gauvain and Lee (1994) describe a method called maximum a posteriori estimation (MAP), which may be used for adapting Gaussian density distributions. This technique requires a larger adaptation set than MLLR, which was described in the previous section. The reason for this is that the parameters of the model are re-estimated rather than transformed. Re-estimation involves estimating more parameters than a transform does, and hence a larger adaptation set is needed.

3 Experimental set up

Four experiments were set up to measure the performance of a digit string recogniser for a set of children. A combination of training and testing on children and adult speech was used. Adult speech was gathered from a database, SpeeCon, previously recorded at KTH/TMH. The recordings of the children were made in a separate room, at after school and day care centres. For this purpose a computer was prepared with a program that prompted for utterances, using text

prompts, and which stored the recordings on disc. The children repeated the prompts that the adult read from the screen.

The experiment was carried out with 116 four to eight year old children, who spoke ten three-digit sequences each. A training set was formed with speech from 60 children. These children were evenly distributed according to age, in order to avoid emphasizing any specific age. The test-set was created with a similar sex distribution as the training-set. But the distribution of recorded children over age demanded for either a smaller set of speakers, for evaluation, or different sized subpopulations. In this experiment the number of speakers was not the same for each age. Each age group contained ten children except the seven year olds, who where 16 speakers.

For the experiments a digit-string recogniser using one HMM per digit was used. The number of states was chosen to be twice the number of phones in the digit-word. As each digit may be pronounced more or less carefully, transitions were inserted to make it possible to skip one state as shown in Figure 1.

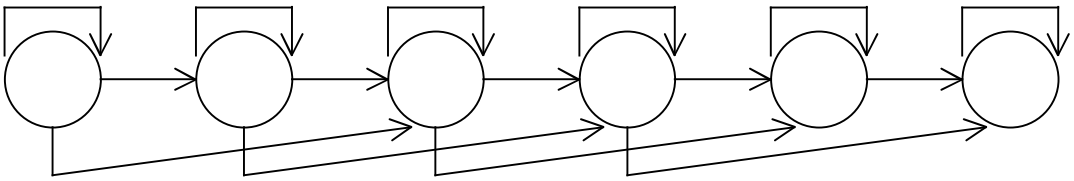

Figure 1. A left-to-right HMM with a skip transition to reduce the constraints of the pronunciation.

The speech samples were read at 32kHz, a cepstum frame was produced each 100 ms using a mel-scaled filterbank and a hamming window. The unusual high sampling frequency was chosen because of childrens higher-pitched voices. As the SpeeCon data of the adult speech was sampled at 16kHz the sampling frequency of this data was converted to 32kHz sampling frequency using linear interpolation.

4 Result

The digit-string recogniser was run and the results are shown in Table 2. Recognition of adult speech worked quite well. Recognizing children speech with the adult model however, worked quite poorly. The acoustical model of the recogniser was then retrained using children's speech with a large improvement as the result. This model was then tested on adult speech with similar results as recognizing children's speech with the adult model.

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

	Evaluation set	
Training set	Adults	Children
Adults	97%	51%
Children	64%	87%

Performance was measured as a function of age, see Table 3. Performance was higher for older children than for younger children.

Table 3. Word accuracy as a function of the age of the speaker.

	Age of the speaker				
Training set	4	5	6	7	8
Adults	41%	41%	48%	55%	68%
Children	78%	83%	88%	93%	92%

5 Conclusions

The performance of a recogniser targeted for adult speech was reduced if run on a child's speech. When the acoustical model was retrained using children's speech the performance was improved dramatically. Recognising children's speech after adapting a recogniser therefore seems possible.

When the adaptation material is large retraining the acoustical model increased the performance of the recogniser. In the cases were the adaptation material is more limited some other methods such as: vocal tract length normalisation (VTLN), voice transformation, maximum a posteriori estimation (MAP), and maximum likely hood linear regression (MLLR) might be useful.

6 Future work

Recognition was poorer for younger children than for older ones. As the younger children probably have a smaller vocabulary than the older children, individualization of the language model might improve performance of younger children. There is also more to be done on the acoustic level. As full retraining demands for a large training set, the technique used here is not possible to use in an age dependent recogniser. However, further adaptation might be performed using MLLR.

7 References

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.

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.

Leggetter, C. J. and Woodland, P. C. (1995): "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models ". Computer Speech and Language (1995) 9, 171–185.

Gauvain, J.-L.; Chin-Hui Lee (1994): "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains". IEEE Transactions on Speech and Audio Processing, Volume: 2, Issue: 2 , April 1994, pp 291 – 298