A Comparison of Human and Computer Recognition Accuracy for Children's Speech

S D'Arcy and M.J Russell

Department of Electronic, Electrical and Computer Engineering,
The University of Birmingham, Edgbaston,
Birmingham, B15 2TT United Kingdom

SXD130@bham.ac.uk

m.j.russell@bham.ac.uk

Abstract

Several studies have shown that automatic speech recognition error rates are greater for children's speech than for adult's speech. Investigations have demonstrated that word recognition error rates increase as age decreases, and that recognition performance for children's speech is more sensitive to bandwidth reduction, compared with adult speech. This paper presents the results of experiments to measure human recognition performance for children's speech. The paper compares human and machine recognition performance on the same children's speech data. It is shown that human recognition performance for children's speech exhibits similar effects of age and bandwidth to those observed for automatic systems. The results suggest that effects of age and bandwidth on automatic speech recognition accuracy are due to properties of children's speech rather than artifacts of the technology

1. Introduction

Automatic recognition of children's speech is a challenging task with potentially important applications in education and entertainment, [1]. Previous studies have shown that, compared with adults' speech, word recognition error rates can be 100% worse for children's speech [2], children's speech is more variable, especially for very young children, [3], and performance degradation due to bandwidth reduction is more pronounced for children's speech [4]. This paper asks whether two of these phenomena, namely the increase in word error rates for younger children and the effect of bandwidth reduction, are restricted to *computer* speech recognition, or whether they also occur in the results of *human* speech recognition experiments.

This paper reports the results of experiments which compare the effects of age and bandwidth on computer and human recognition performance for the same children's speech recognition tasks. The data used in the experiment is part of the British English children's speech data component of the EU FP5 'PF_STAR' project's children's speech corpus. This corpus also includes recordings of native children's speech in German and Swedish; non-native language children's speech in English; and spontaneous and emotional children's speech [5]. Speech data from children aged between 6 and 11, with an original bandwidth of 11 kHz, was down-sampled to 8, 6, 4 and 3 kHz bandwidths. This data was then recognised using an automatic speech recognition system trained on children's speech data, and also by 40 human listeners.

The results show that automatic speech recognition accuracy for this data is very poor. For example, for 11 year old children and a full bandwidth of 11 kHz, the word error rate is 73.39%. Human recognition accuracy is also relatively poor, with an error rate averaged over all age groups of 3.6% at 11kHz bandwidth. This suggests that the data is inherently difficult to recognise and goes some way to explaining the poor performance of the automatic system. Although human performance is far superior to that of the automatic system, the results show that age and bandwidth induce similar effects in both experiments.

The paper is organised as follows: The corpus used in the experiment is described in detail in section 2, the automatic speech recognition experiments are then described and results presented in section 3. The experiment with human listeners are presented in section 4. Finally, we present our conclusions.

2. The 'PF_STAR' British English children's speech corpus

The corpus used for these experiments is a subset of the children's speech corpus collected at the University of Birmingham as part of the EU FP5 'PF_STAR' project in 2003[1][5]. Over the course of several months over 150 children from 4 to 15 years of age were recorded, giving a total of 852 minutes of read speech. This speech was transcribed at the word level. The corpus was divided into age dependent test, training and evaluation sets which were combined for age independent experiments. For each age group, 10 speakers were chosen for the test set, and 1 male and 1 female speaker were chosen for the evaluation set. The rest of the corpus was used for training. Previous ASR experiments conducted on the corpus were reported in [1].

2.1. Down-sampling speech the data

The original PF_STAR recordings were made at a bandwidth of 11 kHz with 16 bit resolution, using high quality microphones, in varying conditions. Compared with adults' speech, children's speech contains more information at higher frequencies in the spectrum. Consequently speech recognition errors increase more rapidly for children as bandwidth is reduced [4], and this effect is more pronounced for younger children. Indeed it was suggested in [4] that this factor may partly explain the high error-rates for children's speech recognition which are reported in [6].

The goal of the current experiments is to replicate the re-

[1]This work was conducted as part of EU FP5 PF_STAR (Preparing Future Multisensorial Interaction Research)

sults reported in [4], on the PF_STAR data, in order to verify the effects of bandwidth and age on automatic speech recognition performance. In addition, human speech recognition experiments are also conducted on the same data.

The original corpus was partitioned into training, evaluation and test sets and down-sampled to 8, 6, 4 and 3 kHz bandwidths using the ESPS software.

2.2. Details of the automatic speech recognition system

Bandwidth dependent sets of tied-state decision tree triphone hidden Markov models (HMMs) were generated using the HTK toolkit and the appropriately down-sampled training set from the British English component of the PF_STAR children's speech corpus. For each bandwidth the models were constructed as follows: Firstly, 46 3-state single-Gaussian-state monophone HMMs were trained on the same but appropriately down-sampled training set. These HMMs were then used to seed a full set of triphone HMMs. Tied-state triphone models were then generated using state level decision trees. Finally the model states were expanded to 4 component Gaussian mixture PDFs and re-estimated (4 component mixtures were found to be optimal in experiments on the evaluation set of the original corpus). Thus the triphone contexts were identical for all bandwidths.

2.3. Preparation of the test data

The original test data was modified so that the same data could be used for both the human and automatic speech recognition experiments. Since each utterance in the human perceptual tests would only be played once to an individual listener, the utterances had to be short enough so that the listener would be able to remember the phrase that he or she had heard for transcription. The read speech files in the corpus were made up of long lists of sentences or short phrases. These audio files were chopped up according to the known transcription files into smaller files of between 3 and 7 words. The corpus also includes words lists which were useful when recording younger children whose reading was not developed, and these lists were chopped up into single word files.

2.4. Test set used for human speech recognition experiments

The test set used in the human speech recognition experiments is a subset of the complete test set. The same original data was down-sampled to obtain data at each bandwidth. The number of files for each age group in this test 'subset' is as follows:

- 209 utterances for 6 year olds
- 234 utterances for 7 year olds
- 287 utterances for 8 year olds
- 296 utterances for 9 year olds
- 273 utterances for 10 year olds
- 246 utterances for 11 year olds

The texts on which the recordings are based were taken from a set of English phrases designed at ITC-irst in consultation with Italian teachers of English [4]. They were judged to be appropriate for reading by 10 year old Italian children learning English. Example phrases include:

- My favourite animal is the owl
- Take off your coat
- It's got three green teeth
- Five

3. Automatic speech recognition experiments

3.1. Specification of the test data

For each bandwidth two sets of automatic speech recognition experiments were conducted. The first experiment used the complete test set, down-sampled to the correct bandwidth. This experiment was then repeated using only the same test subset that was used for the human speech recognition experiments. This second experiment allows a direct comparison with the results from the the human speech recognition tests reported below.

The smaller test subset contains exactly the files that were used for the human speech recognition experiment. This means that files are repeated, as the same files will have been played to different listeners in the human speech recognition experiments. Normally when running ASR experiments repetitions of files would be ignored, however in this experiment we are essentially comparing two speech recognition systems (computer recognition versus the human speech recognition system, and we consider all the human subjects together to represent the 'average' human speech recognition system). Since the repetitions are included in the human recognition experiments to judge the 'human speech recognition system' as a whole, they are also included in the automatic speech recognition experiments so that the results are comparable. This removes any possible bias in the results due to some files being easier to recognise than others.

3.2. Data parameterisation

In the automatic speech recognition experiments the speech was represented using 12 cepstral coefficients and energy, plus the corresponding velocity (Δ) and acceleration parameters (Δ^2), resulting in 39 dimensional acoustic feature vectors. Parameterisation was done using the HTK 'HCopy' tool. An intermediate stage in the generation of this representation is 'mel scale filtering', in which the linear frequency scale log power spectrum is smoothed and converted to a mel frequency scale using a set of triangular mel-scaled filters.

In our initial experiments the same number of mel-scale triangular filters was used for all bandwidth. This results in larger filter bandwidths for the higher bandwidth frequencies. Consequently the 4 kHz system performed better than the 6 kHz and 8 kHz. It is clear, therefore, that the number of filters needs to be adjusted according to the bandwidth. To solve this problem, recognition experiments were first run on the 4 kHz data to determine the optimal number of mel-scale filters for that bandwidth. This was found to be 28. Using this as our bench mark we extrapolated to calculate the number of filters for the other bandwidths.

The bandwidth of the signal on the mel scale was calculated from the formula:

$$MEL(f) = 2595 log_{10}\left(1 + \frac{f}{700}\right)$$

The bandwidth of each filter on the mel scale is calculated by dividing the 4 kHz bandwidth by 28, giving the bandwidth of each optimal filter as 101 mels. For the bandwidth other than 4 kHz, the number of triangular mel-scale filters required was

determined by converting the bandwidth to mels and dividing by the optimum bandwidth. The speech data was then parameterised with the appropriate number of filters.

Using this method, the number of filters required for 3kHz bandwidth is 23, 6kHz is 32, 8kHz is 36 and 11kHz is 39.

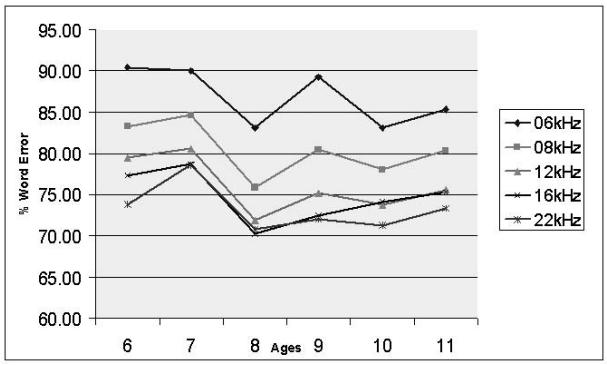

Figure 1: Automatic speech recognition results (% word errors) for different ages and bandwidths (full test set)

Automatic speech recognition results for the full test set, using this method for parameterisation, are displayed in Figure 1 above.

3.3. Note on test sets

Band Width	Full Test Set	Reduced Test Set
3kHz	90.48%	89.35%
4kHz	78.85%	81.38%
6kHz	75.72%	75.16%
8kHz	74.24%	75.25%
11kHz	73.17%	73.69%

Table 1: Percentage word errors as a function of bandwidth for ASR experiments on children's speech, using both test sets

Age	Full Test Set	Reduced Test Set
6	81.30%	80.90%
7	83.10%	83.85%
8	75.43%	76.49%
9	80.56%	77.44%
10	77.22%	76.43%
11	79.16%	81.02%

Table 2: Percentage word errors as a function of age for ASR experiments on children's speech, using both test sets

3.4. Discussion of automatic speech recognition results

The results presented in Figure 1 are broadly consistent with [4] and show that in most cases recognition accuracy decreases as bandwidth or age decreases. In all cases performance for 6kHz bandwidth is poorer than for 8kHz. As one would expect, the effect of bandwidth is greater for younger children, due to the fact that their vocal tracts are shorter and therefore more important information occurs at higher frequencies. For example, for the 6 year old children recognition accuracy drops by 18% when bandwidth is reduced from 6kHz to 4kHz, and by 54% between 4kHz and 3kHz. For 7 year old children the corresponding figures are 10% and 21%. These results confirm, for example, that restriction to telephone bandwidth will increase automatic speech recognition errors for young children. It should be noted when analysing these results that, for each bandwidth, recognition was performed using a single set of 'age-independent' HMMs trained on speech from children aged between 4 and 15, with a high proportion of the training files coming from children 10 years and younger. This may account for unexpectedly poor performance for older children.

Tables 1 and 2 summarise the results for the full test set and for the test set used in the human speech recognition experiments. Table 1 shows the average percentage word error as a function of bandwidth for both test sets and Table 2 shows the corresponding results as a function of age. The fact that the relationship between error rate and bandwidth is more consistent than that between error rate and age is to be expected. The influence of bandwidth on the speech signal is very direct, but the same is not true of age. Age is a convenient but inaccurate indicator of a range of diverse factors such as physical size, reading ability and pronunciation ability which are likely to influence variability and hence word error rate directly. Thus the relationship between word error rate and age is likely to be more sensitive to sample size.

4. Human Speech Recognition Experiments

4.1. Experimental method

The human children's speech recognition experiment involved 40 listeners aged between 14 and 50. Each listener was a native English speaker and lived in the same area as the children that were recorded[2]. The experiment was conducted using a simple program running on a computer and a set of headphones. Clicking a button on the computer screen caused the next utterance to be played, and the listener was asked to type what he or she heard into a text box. The listener was only allowed to hear each utterance once[3]

Each listener was asked to transcribe 90 utterances in this way, corresponding to six different age groups (6, 7, 8, 9, 10 and 11 year olds), five different bandwidths (3, 4, 6, 8 and 11kHz), and 3 example utterances for each combination of age-group and bandwidth. So, for example, each listener heard 3 files corresponding to speech from 8 year olds at 6kHz bandwidth, 3 files corresponding to speech from 8 year olds at 8kHz bandwidth, etc. The three files for each condition were chosen randomly, subject to the constraint that at no point did any subject hear the same utterance at different bandwidths during a session. The listeners were asked to write down what they heard to the best of their abilities. In total, the 40 candidates attempted to recognise 3,600 utterances, comprising 120 examples for each age/bandwidth combination.

At the end of each session the output of the application was

[2]In an attempt to increase the number of subjects the files were played to a group of Irish listeners. However, these results could not be used as the Irish listeners had significant problems with the accents of the children, and this resulted in much higher error rates than those scored by the English listeners

[3]When this experiment was planned it was anticipated that human speech recognition experiment for children's speech would result in very few error. Allowing the listener to hear the child's utterance only once was a mechanism to try to increase the error rate, so that effects of bandwidth and age might be observed.

the list of files played to the listener plus the listener's transcription of each file. These files were checked for correct spelling and then converted into the same format as the output from the automatic speech recognition system. The 'HResults' tool in HTK was then used to score the transcription.

4.2. Discussion of human speech recognition results

Table 3 summarises the results of the experiment on human recognition of children's speech. The table shows percentage word error as a function of bandwidth, averaged over the different age groups, and also percentage word error as a function of age, averaged over the different bandwidths.

Bandwidth	Word error over B/W	Ages	Word error over ages
3kHz	12.4%	**6**	9.8%
4kHz	6.4%	**7**	8.3%
6kHz	4.4%	**8**	5.5%
8kHz	4.0%	**9**	5.0%
11kHz	3.6%	**10**	3.2%
		11	5.6%

Table 3: Percentage word errors as a function of bandwidth and age for human recognition experiments on children's speech

Our first observation on the results shown in table 3 is that the error rates are much larger than expected. For example, even at 11kHz bandwidth the error rate averaged over all age groups is 3.6%. This high error rate provides some explanation for the extremely high error rates reported in Figure 1 for the automatic speech recognition system. Otherwise, the trends shown in table 3 are remarkably consistent with our expectations. The second column shows small decreases in performance between 11kHz and 6kHz bandwidth, a larger increase in error rate of 31% as the bandwidth drops from 6kHz to 4kHz and an even larger increase in error rate of 94% between 4kHz and 3kHz. This follows a similar broad trend to that observed in the automatic speech recognition results from Figure 1.

With the exception of the result for speech from 11 year old children, the final column of table 3 shows human word recognition error rate increasing as the age group of the child decreases, as observed in automatic speech recognition experiments.

Table 4 presents the results of the human speech recognition experiments for each combination of age group and sampling rate.

	3kHz	4kHz	6kHz	8kHz	11kHz
6ry	12.4%	13.9%	13.8%	5.3%	5.2%
7yr	19.0%	7.7 %	1.4%	8.5 %	4.0%
8yr	8.7%	7.6%	3.2%	4.2%	3.7%
9yr	12.0%	4.6%	5.5%	2.0%	2.7%
10yr	6.0%	1.9%	2.9%	2.5%	3.3%
11yr	12.9%	5.5%	5.6%	1.9%	2.2%

Table 4: Detailed breakdown of results (%Word Error Rate) of experiments in human recognition of children's speech

5. Conclusions

There were two motivations for this study. The first was to replicate results reported elsewhere on the effects of age and bandwidth on automatic recognition of children's speech, using data from the British English component of the new 'PF_STAR' children's speech corpus. The second motivation was to discover whether human recognition of children's speech is subject to similar effects.

Tables 1 and 2 demonstrate that automatic speech recognition experiments on the new corpus exhibit similar effects due to age and bandwidth to those reported elsewhere. In particular, degradation of performance as a function of bandwidth is already evident between 6kHz and 4kHz bandwidths, especially for younger children. In other words, restriction to telephone bandwidth is likely to result in increased error rates for children's speech. As mentioned previously, it should be noted that the results were obtained using an 'age-independent' children's speech recognition system, and this may increase the word error rate for older children (though it is interesting to note that the experiments with human listeners also show an increase in word errors for 11 year olds).

Overall, the error rates obtained in the experiments involving human recognition of children's speech are greater than expected. However, the results show similar trends to the automatic speech recognition results, though the actual level of performance is, of course, much higher. Table 3 shows a monotonic increase in errors as bandwidth is decreased, with a 31% increase between 6kHz and 4kHz bandwidth and an even larger increase of 94% between 4kHz and 3kHz. The results also show a similar effect of age on word error rate. With the exception of the average result for 11 year old children, there is a consistent increase in error rate as the age of children in the test set is decreased. The result for 11 year olds is difficult to explain, other than as an effect of the relatively small sample set size. Finally, these results suggest that effects of age and bandwidth on automatic speech recognition accuracy are due to properties of children's speech rather than artifacts of automatic speech recognition technology.

6. References

[1] Mostow, J., Roth, S.F., Hauptmann, A.G. & Kane, M. "A prototype reading coach that listens", Proc. 12th National Conference on Artificial Intelligence (AAAI'94), Seattle, WA, pp 785-792, 1994.

[2] Wilpon Jay G., Jacobsen Claus N., "A Study of Speech Recognition for Children and the Elderly", Proc. ICASSP'96, Vol. 1, pp. 349-352, 1996.

[3] Lee, S., Potamianos, A. & Narayanan, S., "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," Journal of the Acoustical Society of America, pp. 1455-1468, Mar. 1999.

[4] Li, Q. & Russell, M., "An Analysis of the Causes of Increased Error Rates in Children's Speech Recognition", Proc. ICSLP 2002.

[5] Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., & Wong, M., "The PF_STAR Children's Speech Corpus", submitted to Interspeech 2005.

[6] Russell, M.J., D'Arcy, S. and Wong, M. "Recognition of Read and Spontaneous Children's speech using two new corpora",
ICSLP, jeju, S.Korea, 2004.