Proceedings of the Institute of Acoustics

THERE'S NO DATA LIKE MORE DATA

(but when will enough be enough?)

RK Moore 20/20 Speech Ltd., Science Park, Geraldine Road, Malvern, Worcs., UK

1. INTRODUCTION

Since the introduction of hidden Markov modelling in the 1980s, there has been an increasing emphasis on data-driven approaches to automatic speech recognition (ASR). The same principles have also established themselves in other areas of speech and language technology, such as speech synthesis, language modelling, topic spotting and language translation.

The success of the data-driven approach derives from the fact that spoken language systems trained on substantial corpora readily outperform those which rely on more phonetically or linguistically-motivated priors. Similarly, the addition of extra training data almost always results in a consequent reduction in word error rate. This state of affairs led to the much quoted remark – "There's no data like more data.", and the 1990s saw the founding of the Linguistic Data Consortium (LDC) and the European Language Resources Association (ELRA) in order to service the growing international demand for speech and language data . This, in turn, fed the establishment of formal (and public) system evaluations such as those sponsored by the US Defence Advanced Research Projects Agency (DARPA) programme.

Fuelled by the relentless increase in desktop computing power, the data-driven approach has, over the past fifteen to twenty years, given rise to a substantial growth in the capabilities of automatic speech recognition, first in the research laboratory and subsequently in the commercial marketplace. The technology has reached a point where large-vocabulary speaker-independent continuous speech recognition (LVCSR) is now available for only a few tens of dollars in any high-street computer store, and where small-vocabulary voice command-and-control is becoming a familiar feature for users of telephone-based interactive voice response (IVR) systems.

However, despite this acknowledged progress, contemporary automatic speech recognition systems are not able to fulfil the requirements demanded by many potential applications, and their performance is still significantly short of the capabilities exhibited by human listeners. For these reasons, the automatic speech recognition R&D community continues to call for even greater quantities of data in order to train their system – that is, moving from 50 to 500 or even 5000 hours of speech.

This paper addresses the issue of just how much data might be required in order to bring the performance of an automatic speech recognition system up to that of a human listener and, more critically, when will enough data be enough?

2. AUTOMATIC VS. HUMAN PERFORMANCE

By far the most comprehensive comparison between automatic and human speech recognition accuracy was performed by Lippmann in 1997 [1]. Lippmann compiled results from a number of well known sources and presented comparative word error rates for a range of tasks and conditions. Figure 1 illustrates some of the key results, ranging from connected digit recognition to the transcription of spontaneous telephone speech.

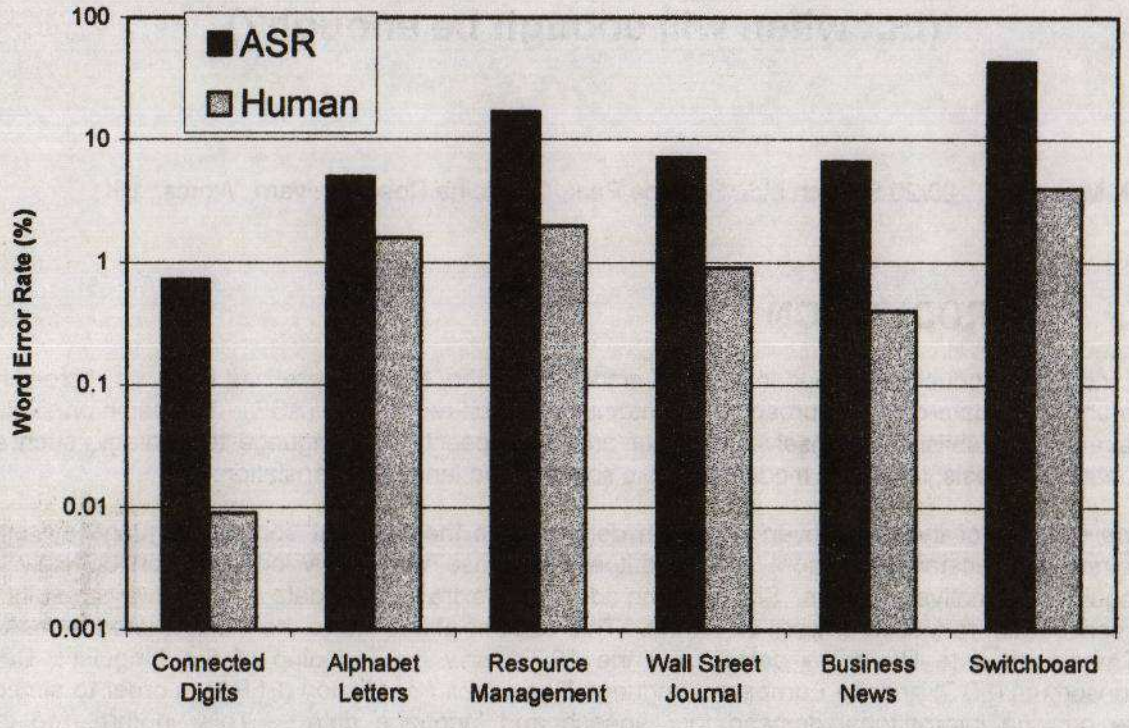

Figure 1: Comparison of human and automatic speech recognition performance.

The results presented in Figure 1 clearly indicate that, in terms of word error rate scores, automatic speech recognition performance lags about an order-of-magnitude behind human performance.

3. ASR PERFORMANCE AS A FUNCTION OF TRAINING DATA

Data concerning the relationship between word error rate and the amount of speech training material employed is hard to find in the automatic speech recognition literature. However, a recent paper by Lamel *et al* [2] provides a very useful insight into how the performance of a contemporary state-of-the-art LVCSR system scales when trained with corpora ranging from 8 to 140 hours in duration.

In their paper, Lamel *et al* describe an investigation into what they call 'lightly supervised acoustic model training' in which labelled training data was generated from unannotated data using an automatic speech recogniser. The application was the transcription of broadcast news material, and two conditions were studied: fully automatic annotation and annotation 'filtered' using closed-captions or transcripts. The results are presented in Table 1.

Unfiltered		Filtered	
Hours	**WER**	**Hours**	**WER**
8	26.43	6	25.70
17	25.20	13	23.70
28	24.30	21	22.50
76	22.40	57	21.10
140	21.00	108	19.90

Table 1: Word error rates for increasing quantities of training data.

The tabular data shown in Table 1 are in exactly the form presented by Lamel *et al* in their paper, i.e. as a table. However, Figure 2 illustrates the same data plotted here in graphical form.

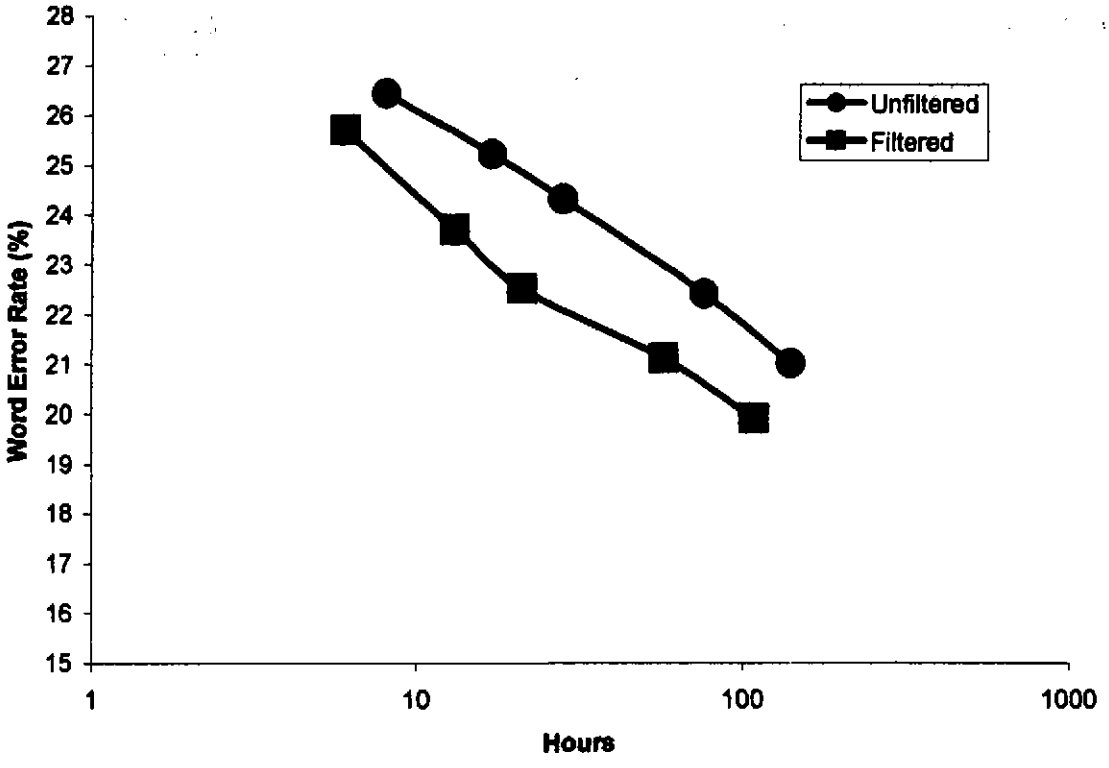

Figure 2: Table 1 plotted using a logarithmic scale for the quantitiy of training data.

Interestingly, the graph in Figure 2 clearly shows a strong linear relationship between the word error rate and the log of the amount of training material. It is also significant that the two graphs (filtered and unfiltered) run in parallel. These very interesting results are explored further in Section 5.

4. THE AMOUNT OF SPEECH A HUMAN HEARS

Another area of study in which there is very little published data is the amount of speech a human being is exposed to, both during their formative years and in later life. A-priori, given that after about 18 months a child is typically becoming increasingly engaged in a communicative environment through its developing ability to talk, it is reasonable to assume that the degree of exposure to speech is bound to increase dramatically around this time. Also, whilst it is known that linguistic development continues into teenage years, it would appear that speech recognition ability is certainly well established by the age of ten.

For the purposes of this paper, it is necessary to be able to determine the amount of speech exposure in the first few years of a child's life. Thereafter, it is possible to make estimates based on the proportion of time an adult spends listening to speech material.

4.1 Babies

One Dutch study [3] indicates that a very young baby receives 29 minutes of directed speech a day. This would suggest that a one year-old child would have been exposed to 176 hours of speech. Assuming an average speaking rate of 120 words/minute, this would correspond to 1,270,200 words.

4.2 Infants

A well known US study - conducted by Betty Hart and Todd Risley [4] - derived statistics of children's exposure to speech in forty-two families spanning three different social groupings. Their study took place over a period of two-and-a-half years, starting with families containing infants from six to nine months of age. Recording were made for one hour per month.

The researchers found that the children of professional parents heard, on average, 2100 words/hour, whereas children of working-class parents heard 1200 words/hour and children on welfare heard about 600 words/hour. The cumulative effect was that after one year, the children of professional parents had heard 11M words, whereas the children from working-class homes heard 6M and welfare children heard only 3M. Apparently these differences had a profound effect on each child's abilities to think conceptually by the age four.

Again assuming a speaking rate of 120 words/minute, the US study suggests that an average two/three-year old child would have been exposed to about 800 hours of speech per year.

4.3 Adults

Assuming an average of eight hours sleep per day, and that one-quarter of the waking day is spent in conversation (i.e. two hours listening), and another couple of hours is spent listening to the radio or TV, it would seem that an adult might be exposed to about 1500 hours of speech a year.

Clearly this estimate is very rough indeed (and subject to wide variance between individuals). However, it is probably accurate enough for the purposes intended here.

4.4 Summary

Based on the data outlined above, Table 2 illustrates an average human being's cumulative exposure to speech over his/her complete lifetime.

Age	Total Hours
1	176
2	976
3	1776
4	2576
5	4076
6	5576
7	7076
8	8576
9	10076
10	11576
20	26576
30	41576
40	56576
50	71576
60	86576
70	101576
80	116576

Table 2: The amount of speech a human being hears as a function of age.

This data suggests that, as a rule of thumb, a two year-old has heard 1000 hours of speech, a 10 year-old has been exposed to 10,000 hours of speech, and a 70 year-old has heard 100,000 hours of speech.

5. WHEN WILL ENOUGH BE ENOUGH?

Taking the results from Section 3 and Section 4 together, it is now possible to construct a view of the relationship between the data requirements of contemporary automatic speech recognition systems and the speech exposure of human beings. This is particularly facilitated by the fact that the data illustrated in Figure 2 is quite linear, and thus provides some confidence that it can be extrapolated to determine predicted word error rates for even larger training sets (see Figure 3).

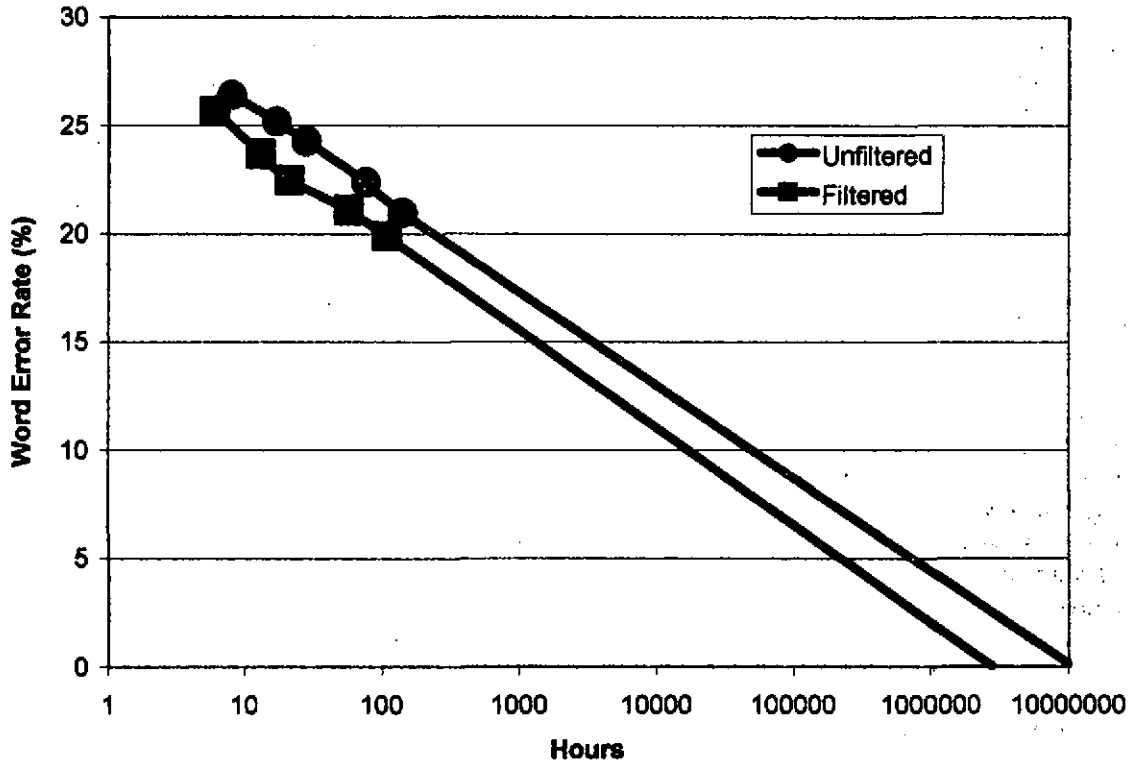

Figure 3: Extrapolated word error rates for increasing quantities of training data.

Whilst current systems would appear to be trained on an order of magnitude less material than a two year-old infant, increasing the amount of data to that received by a ten year-old would still only reduce the word error rate of the automatic system to 12%. In fact the extrapolated results illustrated in Figure 3 indicate that word error rates approaching 0% would require up to 10,000,000 hours of speech training data, and comparison with Figure 2 reveals this to be equivalent to 100 human lifetimes exposure to speech!

6. CONCLUSION

This paper has presented a comparison of the amount of speech a state-of-the-art automatic speech recognition system uses for training, and the amount that a human listener hears over the course of their lifetime. Clearly the training of the recognition capabilities of a human being is conducted in an unsupervised manner - the speech that a child hears is coupled with a multitude of other events in the acoustic-visual world and enbedded in a set of complex connections and relations which themselves have to be learnt. This is presumably a considerably harder task than the supervised training of a conventional automatic speech recognition system.

However, this paper has compared the human data with both supervised (filtered) and unsupervised (unfiltered) training of an automatic speech recognition system. In both cases the results are much the same; a fantastic amount of speech would seem to be needed to bring the performance of an automatic speech recognition system up to that exhibited by a human listener.

Also, it is interesting that the two extrapolated graphs shown in Figure 3 remain parallel, even over a data scale spanning four orders of magnitude. This would seem to suggest some stability in the results, that gives credibility to the overall conclusions.

Finally, the main conclusion from this study would seem to be that simply demanding more and more training data is not going to provide a satisfactory soution to approaching human levels of speech recognition performance. What is needed is a change in approach that would alter the slope of the data presented in Figure 3. In other words, true progress is not only dependent on the availability of more and more data, but on the development of more structured models which are able to better exploit the information available in existing data. Indeed, these results suggest the automatic speech recognition research community may be squandering its data resources, and thereby missing out on understanding - and thus exploiting - the underlying mechanisms which enable a human being to develop his/her astonishing listening skills in the course of only a few years.

REFERENCES

[1] R. Lippmann. Speech recognition by machines and humans. *J. Speech Communication*, vol.22, pp 1-15, 1997.

[2] L. Lamel, J.-L. Gauvain and G. Adda. Lightly supervised acoustic model training. *Proc. ISCA Workshop on Automatic Speech Recognition*, pp 150-154, 2000.

[3] A. Cutler. *Personal Communication.* 1996.

[4] B. Hart and T.R. Risley. Meaningful differences in the everyday experiences of young American children. *Baltimore: Paul. H. Brookes Publishing Company*, ISBN 1-55766-197-9, 1995.