A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners

Roger K. Moore

20/20 Speech Ltd.
Science Park, Geraldine Road, Malvern, Worcs., WR14 3PS, UK
r.moore@2020speech.com

Abstract

Since the introduction of hidden Markov modelling there has been an increasing emphasis on data-driven approaches to automatic speech recognition. This derives from the fact that systems trained on substantial corpora readily outperform those that rely on more phonetic or linguistic priors. Similarly, extra training data almost always results in a reduction in word error rate - *"there's no data like more data"*. However, despite this progress, contemporary systems are not able to fulfill the requirements demanded by many potential applications, and performance is still significantly short of the capabilities exhibited by human listeners. For these reasons, the R&D community continues to call for even greater quantities of data in order to train their systems. 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.

1. Introduction

Since the introduction of hidden Markov modelling in the late 1970s [1][2], 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 modeling, 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 that 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. It is this state of affairs that 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) [3] and the European Language Resources Association (ELRA) [4] 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 [5].

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 Euros 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 fulfill 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 systems – moving from 50 to 500 to 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.

2. Automatic vs. human performance

By far the most comprehensive comparison between automatic and human speech recognition accuracy was performed by Lippmann in 1997 [6]. Lippmann compiled results from a number of well-known sources and presented comparative word error rates (WER) 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.

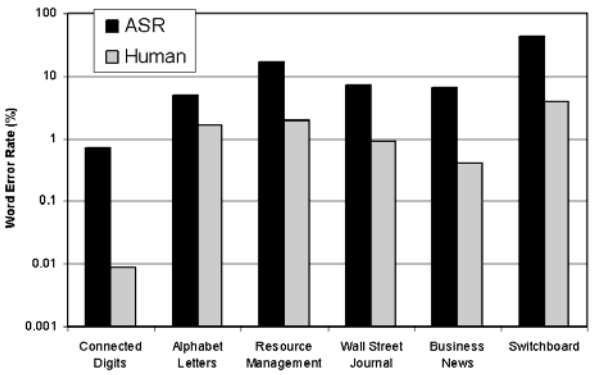

Figure 1: Comparison of human and automatic speech recognition performance (derived from Lippmann [6]).

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

3. ASR WER 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 in recent years, Lamel *et al* have provided a very useful insight into how the performance of contemporary state-of-the-art LVCSR systems scale when trained with corpora ranging from 10 minutes to 140 hours in duration [7][8].

In their 2000 paper [7], Lamel *et al* describe an investigation into what they call 'lightly supervised acoustic model training' in which labeled training data was generated from un-annotated 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.

Table 1: Word error rates for increasing quantities of training data (taken from Lamel *et al* [7]).

UNFILTERED		FILTERED	
Hours	**WER**	**Hours**	**WER**
8	26.4	6	25.7
17	25.2	13	23.7
28	24.3	21	22.5
76	22.4	57	21.1
140	21.0	108	19.9

In an earlier paper, Moore [9] discovered that Lamel *et al*'s results reported in [7] showed a clear linear relationship between word error rate and the logarithm of the quantity of training material (see Figure 2). Interestingly, the two experimental conditions have the same slope.

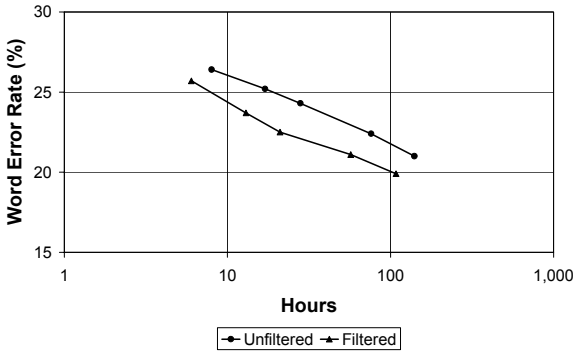

Figure 2: The results from Table 1 plotted using a logarithmic scale for the quantity of training material.

In their 2002 paper [8], Lamel *et al* extended their previous study, and again included tables of word error rate against quantities of training data for both supervised and unsupervised training regimes. In fact, two unsupervised configurations were investigated – one with a dramatically reduced quantity of language model training data (1.8M words, as opposed to >1000M) – see Table2. Figure 3 illustrates the reults in graphical form.

Table 2. Word error rates for increasing quantities of training data (taken from Lamel *et al* [8]).

SUPERVISED		UNSUPERVISED		UNSUPERVISED (reduced LM training)	
Hours	**WER**	**Hours**	**WER**	**Hours**	**WER**
0.2	53.1	4	37.3	0.2	65.3
1	33.3	12	31.7	4	54.1
33	20.7	27	27.9	12	47.7
67	19.1	53	26.0	27	43.7
123	18.0	135	23.4	53	41.4
				103	39.2
				135	37.4

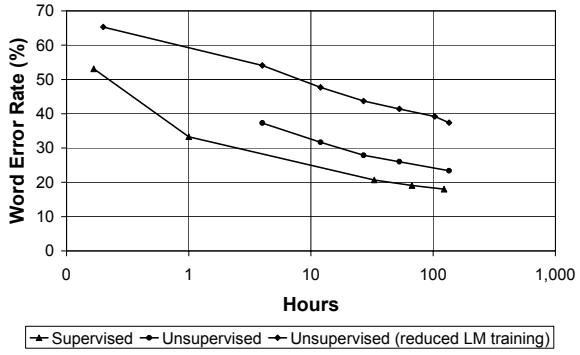

Figure 3. The results from Table 2 plotted using a logarithmic scale for the quantity of training material.

As can be seen from Figure 3, Lamel *et al*'s more recent results again show a near linear relationship between word error rate and the logarithm of the amount of acoustic training data. All these very interesting trends shown in Figures 2 and 3 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. However, 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.

4.1. Babies

A study by van de Weijer [10] indicates that a very young baby receives about 20 minutes of directed speech a day.

This would suggest that a one year-old child would have been exposed to around 120 hours of speech. Assuming a speaking rate of between 60 and 120 words per minute, this would correspond to between 500K and 1M words.

4.2. Infants

A US study conducted by Hart and Risley [11] 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. Recordings were made for one hour per month.

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

Assuming an average speaking rate of 120 words per minute, the US study suggests that a two/three year-old child would have been exposed to about 800 hours of speech (~6M words) per year.

4.3. Adults

There appears to be no published data on adult exposure to speech. However, it is possible to make some fairly crude estimates based on the following: (i) assuming an average of 8 hours sleep per day, and that one-quarter of the waking day is spent in conversation, an adult might be exposed to two hours listening; and (ii) another 3.7 hours is spent listening to the radio or TV [12]. Based on this, it would seem that an adult might be exposed to about 2,000 hours of speech (14M words) 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 all the estimates outlined above, Figure 4 illustrates an average human being's cumulative exposure to speech over his or her lifetime.

These figures suggest the following 'rules of thumb':

- a two year-old has heard ~1000 hours of speech;
- a 10 year-old has heard ~10,000 hours;
- a 50 year-old has heard ~100,000 hours;
- an 80 year-old has heard ~150,000 hours of speech.

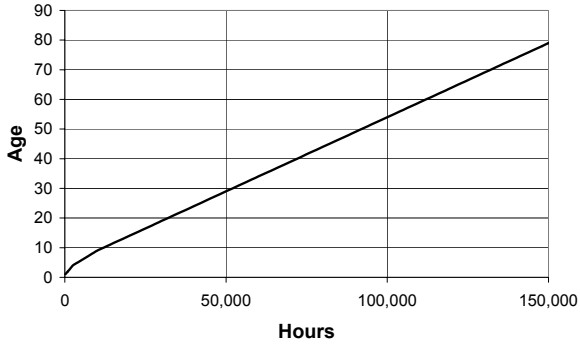

Figure 4: Estimated amount of speech a human being hears as a function of age.

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 Figures 2 and 3 are reasonably linear, and thus can be extrapolated to determine predicted word error rates for even larger training sets.

Figure 5 illustrates the extrapolated word error rates for the data presented in Figure 2, and Figure 6 illustrates the extrapolated word error rates for the data presented in Figure 3.

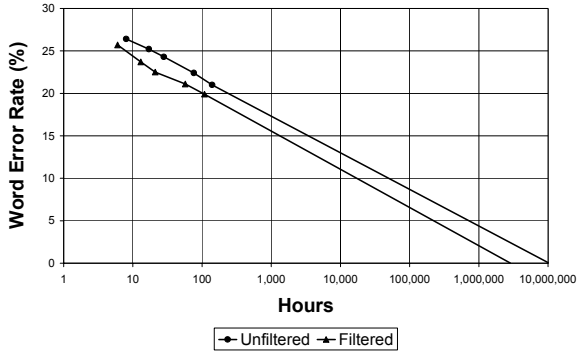

Figure 5: Extrapolated word error rates for increasing quantities of training data (based on Figure 2).

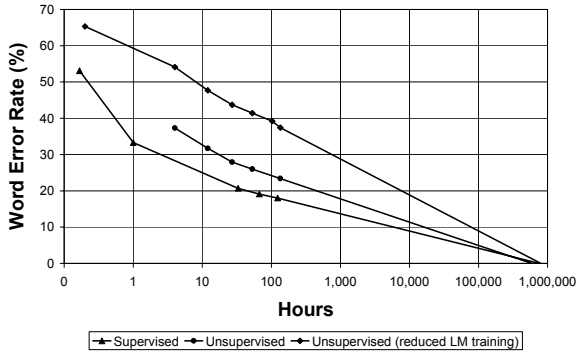

Figure 6: Extrapolated word error rates for increasing quantities of training data (based on Figure 3).

The extrapolated results presented in Figures 5 and 6 make for some very interesting comparisons.

First, 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 child (i.e. the amount that would seem to be adequate to train a human listener) would still only reduce the word error rate of an automatic system to between 10% and 20%.

Second, the extrapolated results derived from [7] (and illustrated in Figure 5) indicate that word error rates approaching 0% would require from 3,000,000 to 10,000,000 hours of acoustic training data, and the extrapolated results derived from [8] (illustrated in Figure 6) indicate that word error rates approaching 0% would require from 600,000 to 800,000 hours of acoustic training data.

Comparison of both these results with the data illustrated in Figure 4 reveals this to be equivalent to between 4 and 70 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 audio-visual world and embedded 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 and unsupervised training of an automatic speech recognition system. In both cases the results indicate that 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. In fact it is estimated that current techniques would require two to three orders of magnitude more data than a human being.

Therefore, 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 solution 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 Figures 5 and 6. In other words, true progress is not only dependent on the availability of more and more data, but on the development of methods that are able to better exploit the information available in existing data.

7. References

[1] Baker, J. K., "The DRAGON system – an overview", *IEEE Trans. Acoustics, Speech and Signal Processing*, 23, 24-29, 1975.

[2] Jelinek, F. "Continuous Speech Recognition by Statistical Methods", *Proc. IEEE*, 64, 532-556, 1976.

[3] Linguistic Data Consortium, http://www.ldc.upenn.edu

[4] European Language Resource/Data Association, http://www.elda.fr

[5] Benchmark Tests, US National Institute of Standards and Technology, http://www.nist.gov/speech/tests/index.htm

[6] Lippmann, R., "Speech Recognition by Machines and Humans", *J. Speech Communication*, 22, 1-15, Elsevier, 1997.

[7] Lamel, L., Gauvain, J.-L. and Adda, G., "Lightly Supervised Acoustic Model Training", *Proc. ISCA Workshop on Automatic Speech Recognition*, 150-154, 2000.

[8] Lamel, L., Gauvain, J-L. and Adda, G., "Unsupervised Acoustic Model Training", *Proc. IEEE Int. Conf. On Acoustics, Speech and Signal Processing*, I, 877-880, 2002.

[9] Moore, R. K., "There's No Data Like More Data: But When Will Enough Be Enough?", *Proc. Workshop on Innovations in Speech Processing*, UK Institute of Acoustics, 2001.

[10] Weijer, J. van de, "Language Input for Word Discovery", PhD Thesis, University of Nijmegen, 1998.

[11] Hart, B. and Risley, T. R., "Meaningful Differences in the Everyday Experiences of Young American Children", Baltimore: Paul. H. Brookes Publishing Company, 1995.

[12] Ninomiya, N. "TV Evolution Continues in the Broadband Age", *NHK Monthly Report on Broadcast Research*, The View from Atagoyama, 17, January 2002.