B/7655BA/1400/00006/018/0
7 December 2009

SIRDCC Speech Technology WG assessment of current STT technology

Security Service have asked the SIRDCC Speech Technology Working Group to give its technical assessment of the current state of the art in Speech to Text technology, and how it is likely to develop.

Executive summary

The SIRDCC Speech Technology Working Group has evidence that current state of the art STT technology is capable of providing some business benefit in very specific circumstances. It has still to prove itself in larger-scale applications, but the potential for major benefits in productivity in the future is clear, given sufficient investment in further developing the systems for our target speech.

The Working Group believes that the most effective way to achieve these benefits is to continue to fund research and development activities. Where practical this should be supplemented with small-scale pilot deployments to explore the areas where most immediate business benefit can be got, so as to help focus the R&D investment.

The underlying technology used by all existing state-of-the-art systems is similar, and thus each is in principle capable of obtaining similar results in any given application, given sufficient effort in bespoke development and tuning. However the BBN system currently deployed at GCHQ for the last 5 years and at NSA for longer has proved itself stable, currently outperforms others on the standard measure of word error rate and is therefore recommended for operational pilots in the near term.

The decision as to when and how it is appropriate to deploy an operational pilot in any agency must depend on business decisions internal to that agency, but it is important that we share and collaborate to the fullest extent to minimise costs and maximise benefits.

Context

Security Service and GCHQ have been collaborating on research and development of capability for Speech to Text (STT), also known as Automatic Speech Recognition (ASR), for a number of years under the auspices of the SIRDCC Speech Technology Working Group. The aims are to assess the applicability of the technology to gain business benefit, and to conduct appropriate research and development to advance the technology where needed.

The other members of the Speech WG have a strong interest in the outcome as a means of informing their own future investment decisions.

UK SECRET STRAP1

DARPA evaluation programme

The DARPA evaluation programme, with significant steer from NSA, has been the main driving force behind technology improvements in the field. Unfortunately the results of the evaluations are not put in the public domain, making reference difficult.

Most of the large corpora of transcribed speech were produced under this programme for evaluation purposes: they are made of up rather artificial conversations between speakers (often college students) who are paid to take part.

Cambridge University and BBN have participated throughout the lifetime of the programme: they have joined forces for the current phase (GALE). Both have always been at the forefront. So were Dragon until their collapse and IBM until they pulled out a few years ago. IBM have subsequently re-entered with the stated objective of obtaining better than human performance, and they marginally outperformed the BBN/Cambridge entry in the most recent evaluation.

Other research labs and universities have also taken part but have never done as well as the organisations mentioned above. SAIL have never participated.

The systems used in these evaluations are research software, and not written for use by anyone other than the originating labs. A version of the BBN system is the only exception to this, having been in use at NSA for about 10 years. In this period a lot of effort has been put into giving it at least some robustness and usability, and into making it user-trainable.

Cambridge University have always taken the view that their software was for running on their own site only, though a modular toolkit HTK is publicly available.

To the best of our knowledge Security Service's purchase of Attila from IBM is the first instance of it being trained other than at its originating site, though we have reports that DSTO and CIA are also investigating its performance.

NSA programme

NSA have had the BBN speech-to-text system Byblos running at Fort Meade for at least 10 years. (Initially they also had Dragon.) During this period they have invested heavily in producing their own corpora of transcribed Sigint in both American English and an increasing range of other languages. Their application of English is to COMSEC monitoring. One of GCHQ's hopes is that NSA will give it access to the models being trained on SIGINT data, since NSA have considerable difficulty in releasing the intercept itself. This is one of the motives for GCHQ's adopting Byblos, since models trained by one system cannot be used by another.

GCHQ/Security Service approach

We have pursued our aims in this field in two main ways, evaluating systems as delivered and obtaining training data to seek to improve them. Our goals have been: (1) to evaluate the technology itself and its business applicability; (2) to perform a comparative evaluation of competing systems to decide where best to concentrate our resources.

- **Systems evaluation**

GCHQ has licensed the Byblos system from BBN Technologies, Boston, since 2002. This system was chosen partly because it was the best-performing system in external trials run by DARPA, but most importantly because it was already in use as a research system within NSA, who were also funding much of its development. GCHQ also funded some specific development by BBN in 2006 in order to make it more easily deployable on our systems.

Security Service (C3T) has investigated the performance of speech recognition from IBM. The initial judgement of IBM, made in 2001, was that their technology was not yet ready [1], but their comparative success in DARPA trials in 2004 led to renewed interest from Security Service who arranged for further trials on UK-accented speech by IBM. In 2009 Security Service licensed the IBM Attila system and funded IBM effort to help build and evaluate a speech recogniser specifically for Security Service product.

Security Service (A2K), with funding assistance from GCHQ, has investigated the performance of speech recognition from a European company, SAIL labs of Vienna. SAIL have licensed their system to Security Service and built a speech recogniser for evaluation.

- **Bulk transcription**

It has been recognised for several years that the main obstacle to effective STT of intercepted speech was the mismatch between the models of speech used in STT systems and the intercept. To address this using current STT technology, tens or hundreds of hours of speech must be carefully transcribed at great cost in order to provide training data. There are two deficiencies in current STT systems. Firstly their models of conversational English speech are biased strongly towards US English. Secondly, the material is gathered openly and is not representative of the speech of the majority of our targets.

GCHQ and Security Service have collaborated to acquire, transcribe and share data sets. Most of these have been UK English of various regional accents, obtained commercially, but we also have a substantial corpus of regional Arabic. A small amount (75 hours in total) has been transcribed from intercept. Of this, there is one

significant UK-regional corpus, NIRAD, which is 56 hours of mostly Northern Irish accented speech.

The very high cost of transcription for STT purposes (of the order of £1500 per hour of speech) makes it vital that we continue to collaborate and share as much as possible.

Status in December 2009

- **Systems evaluation**

The NIRAD corpus has been used to train and evaluate all three systems. The results are reported in a joint GCHQ-Security Service paper [2].

The overall figures on word error rate were: BBN 63%, IBM 82%, SAIL 101%. The figures for word accuracy were: BBN 42%, IBM 32%, SAIL 20%. Note that error rate and accuracy do not necessarily add up to 100% as the error rates are normalised with respect to the true transcript and there may be additional words incorrectly inserted by the recogniser.

The analysis shows that the BBN recogniser is better than the IBM recogniser at transcribing words by a significant margin, as measured by the number of words in each speech file that it got correct (better in 58 out of 59 files).

The analysis also shows that by this measure the IBM recogniser is better than the SAIL recogniser by a significant margin (better in 57 out of 59 files).

There is substantial variation in the recognition rates of individual words. See the Appendix for a representative sample of text as transcribed by the BBN Byblos system, and how bespoke training improves the recognition. There is also a table of the best recognised words, other than those which are recognised 100% which are mostly singletons perhaps well-recognised by accident.

For these experiments Byblos was trained by GCHQ staff with no BBN involvement. The SAIL system was trained by its developers. Attila was trained by Security Service with assistance from an IBM engineer.

Several lessons have been learnt from this evaluation. Firstly the results for Byblos are comparable with NSA's SIGINT experience (though admittedly somewhat worse), confirming that NSA's experience is applicable to our data.

Secondly this is the first time to our knowledge that the SAIL system has been objectively evaluated.

Thirdly it is the first time Attila has been trained on intercept. However there is a lot of uncertainty over the reasons for its worse performance than Byblos's. One factor,

probably, is lack of skill in its use: the IBM engineer who assisted Security Service was new to the field. Another factor is that experience from SIGINT applications has not fed into Attila in the way it has into Byblos. This was the interpretation BBN put on the result when informed of it: their lead developer commented that

> I doubt that IBM's fundamental technology is somehow irretrievably behind BBN's, but it's nice to know that the effort that you and we invest in making Byblos run "somewhat smoothly" on challenging data can pay off in this way.

Since this evaluation was completed, the IBM system has been retuned by IBM and the BBN system retuned by GCHQ (no further work has been done on the SAIL system). The current best performance is word error rate: BBN 60%, IBM 76%, SAIL 101% and word accuracy: BBN 45%, IBM 42%, SAIL 20%.

- **Bulk transcription**

The need for additional bulk transcription can be seen from the data presented in the Figure at the end of this report. It shows data points derived from NSA experiments on a variety of languages, as well as data points drawn from NIST evaluations sponsored by DARPA. Each point shows the measured word error rate (or character error rate for Korean and Mandarin) for a given number of hours of transcribed training data. All points are got using the Byblos system, and all except those labelled "DARPA English" correspond to experiments conducted on transcribed SIGINT data.

There are three lines drawn on the figure. The bottom one labelled "DARPA English" shows the performance of models built on public data, assessed on such data. There is a clear trend of improved performance associated with the use of more training data, but note that the improvement is only logarithmic.

The top one, labelled "Unclass. system on IA English" shows the performance of these same models on an Information Assurance application, where the speech to be transcribed is US English. The trend is the same, but there is a significant performance gap - of the order of 20 percentage points.

The middle line, labelled "IA English" shows the improvement that can be got by training a bespoke model for the task. There is still a substantial residual gap of around 7 percentage points between the DARPA line and the IA English line. The reason for this gap is not known, but it is clear that there has been a substantial improvement of performance – of the order of 13 percentage points – by using bespoke training.

The remaining points for other languages have much more variation, but overall are compatible with the existence of a similar trend of better performance associated with using more data. We have no information for these other languages on how much worse the performance would have been if public data had instead been used in the system training, these points are all drawn from models built using intercept.

The point for NIRAD English is high in comparison with the broad trend for all the non-IA English languages – one would have expected a word error rate of closer to 50% rather than the 62.5% measured. This may be due to the nature of the data, as it has been recorded with both sides of the conversation merged which is known to have an adverse effect on the performance of speech processing algorithms.

We cannot explain the substantial gap between the performance on IA English and that on all other languages; it may be attributable to an inbuilt bias in current speech recognition systems towards features of US English caused by decades of intensive research driven by US funding using US speech data.

GCHQ operational experience

GCHQ has been making operational use of Byblos since around 2004. The transcripts it produces unaided have not been of sufficient accuracy to have any value, but the technique of language-model biasing has enabled GCHQ to tailor Byblos to specific keywords or strings of interest. (The possibility of sharing techniques of this sort is a further reason to aim for compatibility between agencies.)

The first application was to strings of digits spoken by Caribbean drugs runners. GCHQ was able to detect spoken telephone numbers with high reliability using an out-of-the-box recogniser whose error rate was greater than 100% under the standard metric. Since then several instances of number detection have been deployed. In one recent case the digits are recognised with sufficient accuracy for it to be worth reporting their values to analysts, rather than just reporting their detection.

GCHQ has one deployed example of keyword detection other than spoken digits, but has had difficulty in persuading analysts to propose suitable search strings. GCHQ expects to be able to extend the range of deployments over the next couple of years, owing both to the wider range of languages available and to improved accuracy as Sigint corpora get transcribed. The operational benefit in the short term is likely to remain small compared with other technologies such as diarisation, gender and speaker ID.

Conclusion

The current state of technology is that systems are capable of automatic transcription with word error rates of between 30% and 40%, given amounts of training data of the order of hundreds of hours. The cost of transcribing this amount of training data is substantial – of the order of £0.5M for 300-400 hours of material.

The accuracy required of a system in order for it to provide business benefit will depend on the business application, and we do not yet have a good understanding of this. GCHQ have successfully deployed several STT applications to locate the

existence of spoken numbers such as telephone numbers in speech. They have also deployed a STT application which locates the existence of specific keywords.

In each of these applications, success has been achieved using an extremely poor core STT model (the default unclassified one supplied by BBN), with the performance enhanced by tailoring the language model. As the performance of STT systems improves, either by providing more training data or by technical advances in the algorithms used, so the range of applications for which they can provide business benefit will expand.

In the long term it is difficult to predict how the technology will evolve. Our judgement is that the recent improvement in performance driven by large-scale US investment is likely to plateau as the performance of STT on transcription of cooperative or public speech attains levels approaching 90% accuracy. US investment is now moving towards follow-on applications such as machine translation of the recognised speech.

There remains a significant gap between the performance measured on public data and the performance measured on intercept data, which may limit the potential for transcription of intercept data to accuracies of the order of 80% using current technology. However, to achieve such levels of accuracy will need substantial investment in bespoke training, and we should not wait for them to be achieved before seeking applications.

It is premature to choose between the IBM and BBN systems in terms of performance on classified material, as we only have one experiment to guide us. However the fact of the long experience of BBN in developing systems for use on SIGINT material makes it the preferred system for operational deployment in the short term.

State of the art speech recognisers are not shrink-wrapped products and require substantial training in order to understand how to use them and exploit them. There is no standard for STT models, and so models built for one recogniser are not portable to another. STT models are not cheap to build, requiring of the order of a year of CPU time (depending on the amount of data). These factors mean that there is considerable benefit to be had in UK agencies agreeing to use a common system in the long term, which would allow pooling of expertise and sharing of built models.

[redacted]

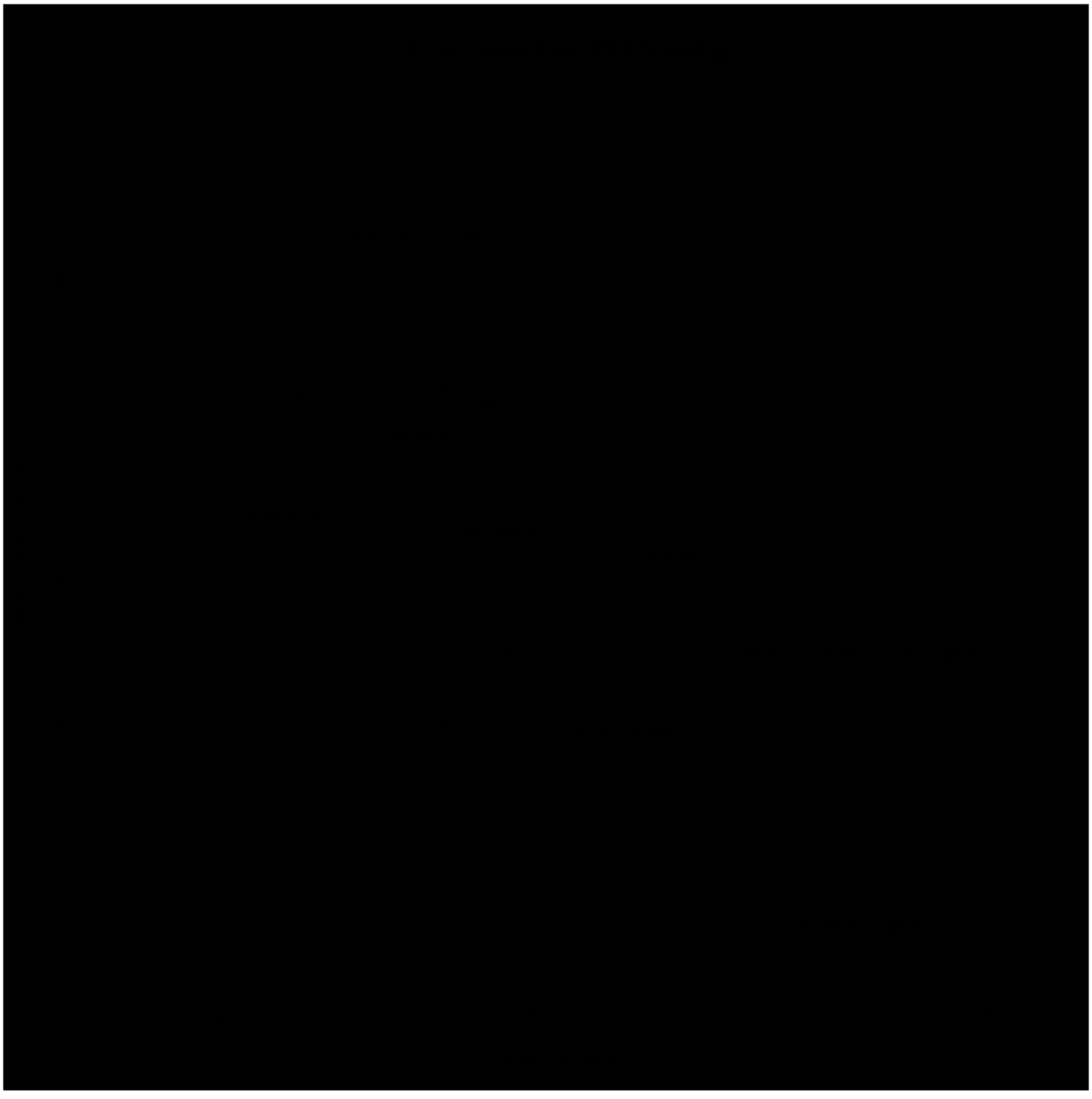
Chair, SIRDCC Speech Technology Working Group

References

[1] Minutes of SIRDCC Working Group Meeting on Speech Technology, 2001-12-03

[2] Comparative evaluation of three commercial speech recognisers, TRMCA/Inf/623, 2009-09-04 (revised 2009-12-07)

Figure: Error rates from training Byblos recogniser on different amounts of data



B/7655BA/1400/00006/018/0
7 December 2009

Appendix: Illustrative text and 100 well-recognised words

BBN Byblos transcription – correct words are marked in red

As delivered 2007

Truth: great o. k. that that's that's perfect o. k. well listen [talking] to derry give me i'll expect you there i will expect a call maybe some time thursday morning

Byblos: critical credit book books post post purple it was miles to go before you on the show communal experts will but the coma mission and mourn

Bespoke trained 2009

Truth: great o.k. *** that ** that's that's perfect o.k. well listen

Byblos: right o.k. but that is that's that's perfect o.k. **** what

Truth: [talking] to derry and [talking] give me i will expect you there i i will expect a call maybe some **** time thursday morning

Byblos: ********* on the fariones should give me * **** ******* ****** all to go to the hospital call maybe some cunt was a morning

The best-recognised words (other than 100%) with their frequency counts

94%		78%		73%		69%		66%	
CRAIC	17	SOMEBODY	18	LAST	26	NO	261	FELT	3
FUCKING	204	WEEK	22	PROBLEM	26	KNOW	390	FIFTY	15
SCALLY	9	FRIDAY	26	BELFAST	11	TOMORROW	45	FIND	12
MORNING	30	TWELVE	13	SIX	33	TOLD	35	HOPEFULLY	3
DIFFERENT	7	SEVEN	42	GIVE	76	NUMBER	57	JOB	9
MUMMY	14	AGAIN	29	RIGHT	284	[BREATH]	136	JOKING	3
NINETY	7	AIRPORT	8	TALKING	18	PHONE	47	LEAST	3
YEAH	339	ALREADY	4	REALLY	25	SAYS	135	MARATHON	3
WEEKEND	12	CHECKED	4	CHANCE	7	HALF	28	MOVING	3
BACK	103	DEAD	8	DRIVING	7	HUNDRED	86	MUCH	33
CLEAR	5	DUBLIN	4	ELEVEN	28	BEDROOM	3	NIGHTMARE	3
COUPLE	15	EACH	4	MOBILE	7	BLAME	3	OPPOSITE	3
DRINK	5	EXACTLY	8	PEOPLE	21	BRILLIANT	12	PASSPORT	3
KEPT	5	HOURS	8	NEXT	24	CHRISTMAS	6	PRESSURE	3
HELLO	100	KNOWS	4	BIG	17	CLEAN	6	PUB	3
COMING	19	LIVERPOOL	8	HOUSE	40	DATE	3	QUID	6
MINUTE	19	PARK	4	MONDAY	10	DERRY	3	SEAN	3

B/7655BA/1400/00006/018/0

7 December 2009

O'CLOCK	19	PICTURES	4	SOMEWHERE	10	DRINKING	3	SECONDS	3
DOUBLE	9	THIRTEEN	4	ANYWAY	23	DRUNK	3	SIXTY	9
REMEMBER	9	GRAND	15	TWENTY	36	DURING	6	SLOWLY	3

UK SECRET STRAP1