TC-Cap February 2018 Status

Narrative Report

Prepared by Jim Kyle, John Martin,

Lorenzo Benoni, Ed Toms

81 Macrae Rd, Ham Green, Bristol BS20 0DD

www. bbitg.co.uk

supported by

Vassall Centre, Gill Avenue, Bristol BS16 2QQ

A project funded by Innovate UK

|in the programme of the Technology Strategy Board:

Emerging and Enabling Technologies Round 1

For further information contact jim.kyle@bbitg.co.uk

Contents

1.0 TC-Cap Project Summary –the starting point- May 2017

Of 1 in 7 in UK with significant hearing loss, 500,000 speak well but can't use a voice phone, negatively impacting: public services, health provision, employment and access in emergency. Advanced (costly) publicly-funded telecoms relay in USA & Australia contrast with time-consuming UK Text relay services - funded by levy on a BT monopoly which passes on charges (possibly £1m pa) to 100 communication providers CPs. BT text reaches only 3% of potential users.

TC-Cap works to develop a cost-effective innovation providing a real-time, visual display of respondent's speech in a call. TC-Cap aimed to deliver faster, more reliable communication through a distributed network of re-speakers and automated smart-voting system with identified speaker profiles and attributed context. *However, as can be seen from the text below, we have moved significantly beyond this point to a fully automated approach.*

TC-Cap integrates respondent's voice-text automatically slashing costs a priority in Australia's public consultation (March 2016) and for the FCC in USA. TC-Cap-Video augments lip-reading with simultaneous respondent text & video. Australian data projected to UK implies over 2 million calls per annum with societal benefits over £20m (Ofcom, 2012 figures). By selling to businesses worldwide, TC-Cap can create economies of scale. TC-Cap is the socially-inclusive solution for employees, service providers, end-users.

Scope of proposed development

TC-Cap enables hard of hearing end-users to re-connect to telephony. TC-Cap is a user-driven, demand-sensitive enabling technology designed to reduce social isolation, improve employability and personal security. TC-Cap positively impacts social care, health delivery, family networks as well as ensuring access to work and increasing productivity. TC-Cap offers significantly reduced costs to 100 UK communication providers and impacts the BT monopoly of the use of typists and retro text connections for an OfCom approved service. More advanced text relay services in the USA and Australia while demonstrating an increased demand in captioned telephony when operators are not declared at the start of the call, also seek to reduce costs.

TC-Cap combines different technologies (voice and video telecoms, speech recognition (SR), parallel processing of voice input, a probabilistic mathematical model (voting protocol) for profile matching and context identification), within an advanced infrastructure and service model which tracks user characteristics and provides a billing framework. *TC-Cap has now focused its development on a machine learning framework which will allow it to grow with use and to become more accurate and effective.* TC-Cap is operating-system agnostic, and can be implemented on smartphones, desk-based and set-top boxes. There is no learning process for end-users - the smart network recognises their hearing status and inserts TC-Cap; similarly, incoming calls are identified and TC-Cap is invoked. Users simply talk and respondent's speech appears onscreen. Hard-of-hearing users can choose to have voice carry-over augmented by text display, or full video voice and text, allowing enhanced lip-reading and visual empathy. *The visual display also helps to reduce cross-talk where one party anticipates and elaborates before receiving the partner's response*.

Krull & Humes (2016) in Ear & Hearing, confirm the positive impact of augmenting lip-reading and telephone usage in this way. TC-Cap builds upon our existing telecoms software (TC-Phone) and our network infrastructure (TC-Network) with full back-office monitoring, billing and security systems. TC-Cap is telecoms standards compliant and is also embedded in user consultation in the service design process, through engagement of the Deaf Studies Trust.

TC-Cap sits on the digital economy timeline exploiting mobile networks and smartphones, creating inclusion and innovation in a way which requires no new learning by the end-user and virtually no change in behaviour by the respondent in the call.

2.0 Executive Summary of Project to February 2018

Use of the telephone is a central feature of our society and it creates the capacity to interact with fellow citizens and organisations. It satisfies the need for time-critical exchange of information and mutual sharing of emotion. Now expanded in terms of video calling, it offers a new dimension in reducing the social and personal distance between people. However, it remains speech centric and requires hearing capability. For those people with a hearing loss, these latter requirements may not be met and for a significant but un-documented (and marginalised) number, telephony in its current form is not accessible without further human intervention to mediate.

TC-Cap has approached the development of an automated speech-to-text system for telephony in a number of parallel actions:

1. Recording and establishment of a telephone call audio database incorporating scripted calls (for direct comparability) and unscripted calls (drawing on UK based calls). This is currently over nearly 1000 minutes of calls.
2. Initial processing of this database by transcription and timecoding of utterances to create source-reference files (Source-Ref) (approximately half of these have been completed; there is a ratio for this process of around 1:20 i.e. to complete the transcription and timecoding of the existing database is over 350 hours of work.)
3. Training of a selection of speakers as (a) re-speakers (Re-Spk) and (b) as acoustic and language (SR) profiles to function in Nuance's Dragon Dictate v15. Re-speakers listen to the audio in a call and 'dictate' the words into the speech recognition system, thereby, in theory, overcoming the problem of source dialect and diction.
4. Establishment of a secondary database of speech dialogue and transcribed speech as a training environment for the TC-Cap post-processing framework. This transcript collection constitutes several thousand hours of material.
5. Most recent additions to the audio phone database, include the use of a pangram sentence at the start of the call, which may aid the construction of a speaker–specific SR profile for incoming calls.
6. Managing and collating the database in the form of audio, and timecoded text output based on actions 1-3 above.
7. Developing two major programs which analyse the source texts in conjunction with the (Re-speaker) Re-Spk texts and Automatic Speech recognition (ASR) profiles. The first compares text output for every original source (Source-Ref) file with the corresponding Re-Spk and ASR profiles by generating a metric for the extent of transposition required at character level, to create a perfect match. The closer the performance match, the more accurate the performance of that ASR profile to the known Source-Ref.
8. The second program examines the syntax of the various files by using automated part-of-speech tagging (POS-Tags). Similarly, closeness of match to Source-Ref is calculated for each utterance.
9. In both cases, calculations are made by utterance (linked to the timeline), by overall accuracy at the end of the audio call and by cumulative measurement as the call develops. These calculations are then validated against the precise verbatim measurement (word for word) taking the Source-Ref as the standard.
10. Graphical representation and cluster analysis is supported by regression calculations in order to identify the most effective speech to text performance.

11. From these 2 programs, concordance between the various ASR outputs can be established leading to confidence that the Re-Spk output can be discarded and the most consistent text output presented to the partner in the call.
12. It has further become apparent that although the text display envisaged (in the TC-Cap proposal) was meant for the hard of hearing viewer alone, it is clear that the dialogue itself has a significant role to play. As a result, speech from both parties in the call are being passed through ASR and the transcript of the hard of hearing caller can be used to cue the potential responses of the hearing speaker. In order to bring this into play, we have adopted principles from the field of Discourse Analysis.
13. In addition, we have now applied "post-post-processing" involving an artificial intelligence model (which will continue to learn as the database of calls increases), and can demonstrate how probabilistic calculations on the co-occurrence of words and component phrase can trap errors (even when presented from the concordance measures) and lead to more accurate output of text.
14. This component is still in development but engages with developments in Deep Learning and Natural Language Processing (NLP) in order to assign probabilities to each of the ASR outputs. While in theory, Dragon claims to be using Deep Learning, their application focuses on monologue dictation and different ASR profiles applied to new speech dialogues, produces errors. Using our embryonic AI post-post process, we can trap these errors.
15. Work remains to be done to tie this advanced analysis (which is a post-processing and post-post processing functionality wholly owned by the project) to continuous time-based decision-making within a call.
16. Using our existing TC-Phone video telephony application and also an open source video and instant messaging telephony app (Linphone), we have been able to demonstrate live STT performance in telephony calls.
17. User feedback on these supports the view that video calls with captions, may be most effective, not necessarily to allow lip-reading as such but rather to be able more effectively to monitor expected output and thereby to avoid cross-talk.
18. The development of the dialogue approach as indicated above, also opens up the reality that the participants naturally repair mis-spellings in display and that in circumstances, where STT is problematic, the third voice of the system can intervene and request clarification.
19. Market analysis and exploitation plan has been developed in the context of deaf needs for telephony, the move to ubiquitous real-time text (RTT) applications on all mobile devices (by the FCC in the USA) as well as the UK's prioritisation of AI developments.

20. Future Developments:

Phase 2 of the project involves

(a) additions to the audio telephony database with a wider range of non-scripted calls
(b) use of pangram and speaker-specific SR profiling at the start of the call to enable creation of new ASR profiles
(c) maximising the range of ASR by operationalising the existing database recordings as new profiles to be used – offering up to 80 additional ASR profiles.
(b) fine tuning and introduction of time constraints on the algorithms for best performing SR profile through concordance – taking into account obvious variables of

age, gender, accent and diction
(c) introduction of more advanced machine learning models to produce new algorithms for speech dialogue
(d) More extended lab trials with TC-Phone and mobile phone apps; then field trials of the system as a whole.

Phase 3 involves
(a) refinement of all points above
(b) production of the software app
(b) creation of a pilot service in the UK
(c) service development to scale, offering Total Conversation – video, voice and text according to user choice

3.0 Speech to text: Mythology and Reality

HAL[1] has a great deal to answer for in raising general expectations of the extent of real time processing of natural speech. Nevertheless, in the 50 years since the film's appearance, there has been significant progress in automatically recognising speech.

We encounter this in call centre automation where at its simplest level a computer recognises numbers spoken in sequences and repeats those back in voice but also in a more advance form when the callers are asked to briefly state the purpose of the call, in order to be transferred to the appropriate queue. We see it widely advertised now in-home systems: Amazon's Alexa for example, which are usually command based and are driven by "what can you say to Alexa" structures.

Considerably more extensive and advance systems have taken existing large text databases and used these to create speech identification for specific purposes – e.g. doctor dictating patient notes into a pre-determined script. Systems such as IBM's Watson can be packaged and sold to companies wishing to reduce human staff involvement at their customer-facing interfaces. The largest companies Apple, Microsoft, Google all offer online speech dictation systems for their software. Nuance offers various standalone products (i.e. which function without an Internet connection) and the latest version Dragon Dictate v15 for Windows, has been used as the processing engine for TC-Cap.

Various claims for these systems are made – usually in comparison to verbatim accuracy. Such claims can be disputed as there is acknowledgement that they do not function effectively for all speakers and for lower sound quality input and speech variations. Nevertheless, in dictation system which are speaker dependent (i.e. the speaker has "trained" on the system), accuracy rates in the high 90 percentages are common.

However, to date, despite research efforts (such as work on the British National Corpus), there has been no significant breakthrough in determination of meaning (in ASR) nor in the generation of text from speech interactions i.e. more than one person involved. Captioned telephony where it exists (e.g. USA) remains a domain driven by speaker dependent speech

[1] HAL was the automated speech recognition and speech production computer in the film 2001: a Space Odyssey, released in 1968.

recognition – i.e. the call is intercepted and directed to a remote listener-speaker, who shadows the speech and generates text through a program like Dragon Dictate.

At the present time, there is no automatic speech recognition service available for telephony.

TC-Cap tackles this issue head-on. BBITG have an existing video telephony infrastructure and 20 years of experience in providing video calling, scaled to large numbers of users. BBITG also developed and implemented a T140 integrated text delivery system which sits inside the TC-Phone application. This T140 telecoms standard creates synchronous dialogue (unlike instant messaging apps which are asynchronous and sequential) so that text from both participants is displayed simultaneously, and in timely fashion.

What is now required is to find a way to bridge the communication gap between hard of hearing users and hearing speakers in text-based telephony as well as video telephony. This is the goal of TC-Cap.

3.1 Audio Database Creation and Management

Although there are databases of telephone calls available in the USA, we were unable to discover a database of phone calls for the UK. A first task of TC-Cap was to create such an audio database.

A simple protocol for simultaneously but separately recording both parties in a call was developed and implemented. Ethical guidelines were followed in regard to explanation and consent, storage and use of the data and all participants were offered the possibility to erase their call behaviour if it was in some way problematic.

Scripts for calls were created to allow the possibility of fixed comparability but call recordings also include different scenarios without script. Participants recruited by Deaf Studies Trust were required to take part in four phone calls – 2 scripted and 2 unscripted. Almost 1000 minutes was recorded by 68 speakers.

The audio calls had then to be included in a reference database which identified the participant by age and gender and the call by its content and directionality (e.g.. to- or from-the doctor).

3.2 Speech to Text Engines and Applicability

At the beginning of the project it was considered that there might be a range of STT engines which could be applied. The following systems were tested:

Dragon 6 on Mac
Dragon 15 on Windows 10
Google Speech online and in Soundwriter on Google Docs
Cortana (in MS Dictate)
Kaldi
CMU Sphinx

And for sound management:

Audacity, Soundflower, Sound Forge

Best results were obtained and greatest control could be exercised using Nuance's Dragon Dictate v15 and the main results reported below come from the application of this stand-alone software.

Although the Dragon package comes with training texts for various specialist areas such as medical and legal, these are not specifically used and in fact, the expectation is that the software is to be used out of the box, to begin speech recognition. Typically such software is designed to work with high quality recording devices, whereas TC-Cap (at this feasibility stage) used a simple sound card USB device to input the audio; this resulted in a noticeable reduction in volume and clarity. We can predict higher quality STT output when we implement a more powerful system in phase 2.

However, in dictation mode, it is expected that speakers will record and correct their STT sessions. When sufficient numbers of these corrections are created then the option exists to fine-tune the acoustic model and the language model for that speaker. The ASR profile is then in theory, strengthened and should become more accurate for the domains in which that speaker has been shown to be operating.

Six ASR profiles were "trained" in this way.

3.3 Re-Speaker Performance: Characteristics and Output

Re-speaking requires an intermediary to listen to live or recorded speech and to shadow that speech with his/her own voice. This voice output is then fed into the STT system which generates a text output from this. In our case, when this text stream appears onscreen the end point of the text output (i.e. each utterance) was timecoded using software (InqScribe).

Differences in performance were noticeable in the case of male and female profiles. In early testing, it appeared that texts which were re-spoken were more accurate than the ASR profiles.

However, what also became apparent was that the process of re-speaking was different from the direct ASR process. In effect, the re-speaker was inserting an intervening stage of psycholinguistic analysis to create what is the natural language processing of human interaction. While ASR using Dragon, went from the acoustic and language model to generate the highest probability match to English and then produced an output, the re-speaker was (a) waiting for enough text to be received in order to make sense and (b) filtering out repetitions and diversions and sounds with no meaning (stutters, "eh", " I mean.." and so on). A major part of this difference occurs because ASR has to make sense of the sequence of sounds (albeit intelligently by matching to previous utterances) while the re-speaker is able to process larger chunks semantically and pragmatically. This leaves us with a problem (in real operation of STT in telephone calls) as most attempts at measuring accuracy rely on mapping words or characters to an original source.

A further issue is that a re-speaker and a listener, engage in processes which involve prediction of what someone is about to say, and not just on what they have heard. Interaction and dialogue proceeds smoothly because the interactants are individually predicting what the other is about to say. The inbuilt redundancy in English text/speech helps this process.

While an "obvious" solution may be to use the re-speaker as the reference text (instead of the Source-Ref) we have still to balance the user's perceived need to have verbatim text which becomes significant when the user has some hearing or where the call is in video and the lip-patterns of the source may be detectable.

It is also the case, that the re-speaker's production is more subject to omissions when the text stream for that utterance becomes elongated. In effect, the re-speaker may not be able to

keep up. In a live conversation, one solution would be for the re-speaker to ask for repetition when he/she detects these omissions or errors.

The issues raised here in regard to the role of re-speaker is still under discussion, but as later work showed, we may be able to completely remove the re-speaker from this process.

4.0 Application of ASR to Audio Database:

4.1 Issues of Concordance of text output at character level

The critical phase of the work is reached when we apply the metrics to the text output from the STT engine. Various formatting and arrangements are required.

The first step is to arrange text output according to when it appeared on screen, in each timed second of the call (Figure 1). This creates a problem as the Source-ref is time coded according to the start of the utterance (which makes sense logically and linguistically), whereas

Figure 1: text output arranged vertically as a time line

the ASR output has to be timecoded to the end of the displayed utterance, again logically as the speed of the read-out (display) means that the viewer has to wait until the whole utterance is present on screen.

The net result is that it becomes clear that an ASR enabled interaction (using a screen display) is functionally different from a one-to-one interaction in speech. Users will need to be aware of this and the way in which this functionality affects their ability to use the redundancy of English to predict the end of the utterance and to prepare the response.

Text files were grouped into folders for analysis.

Figure 2: Timecoded text outputs grouped and analysed by folder (single speech source)

The principles for this analysis are described in Appendix 2 and require a mapping file which sets out the iterations required for the multiple comparisons. The comparisons made are at character level and these are compared with the detailed analysis carried out painstakingly word by word (see below).

The output of this analysis appears complex.

The creation of the necessary algorithms is anything but trivial. While the usual application of accuracy measurement is applied to static texts, in our case we have not only varied text to deal with but text appearing at different times along the timeline of the call. The program has to sample at every second to determine which ASR text now matches to an earlier Source-Ref text. When the texts are likely to be different it may be difficult to determine commonality. Using a process of iterative *chunking*[2], text matches were forced and analysed.

This process is further complicated by the fact that discrete utterances in the Source-Ref are often concatenated in the ASR Profile output, forcing decisions about which elements to match and which to enforce consistently, when working across different profiles.

The primary metric used for comparison was the Levenshtein Distance. This calculates the number of letters which have to be moved/transformed in order to match the original (Figure 3).

[2] The chunk is often shorter than the utterance and the procedure has to determine comparability across a set of different texts appearing at different points along the timeline. Comparable chunks have to be detected automatically.

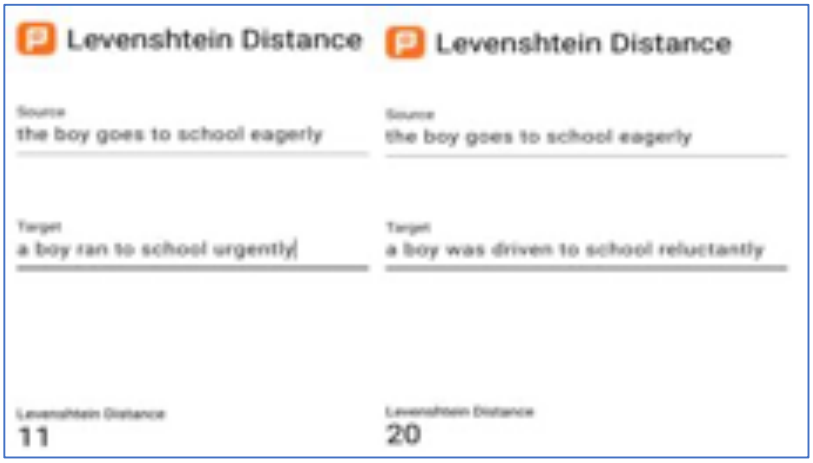


Figure 3: Character level calculation of accuracy between 2 texts

As the text diverges from the original, the number of transforms needed increases. The fewer the transforms, the more accurate the text is deemed to be. However, the measure is not perfect in matching as can be seen in Figure 4 where reduction in text output can be deemed almost as accurate as a closer match with more words. The value 13 is returned (compared to value 11) when two whole words are omitted.

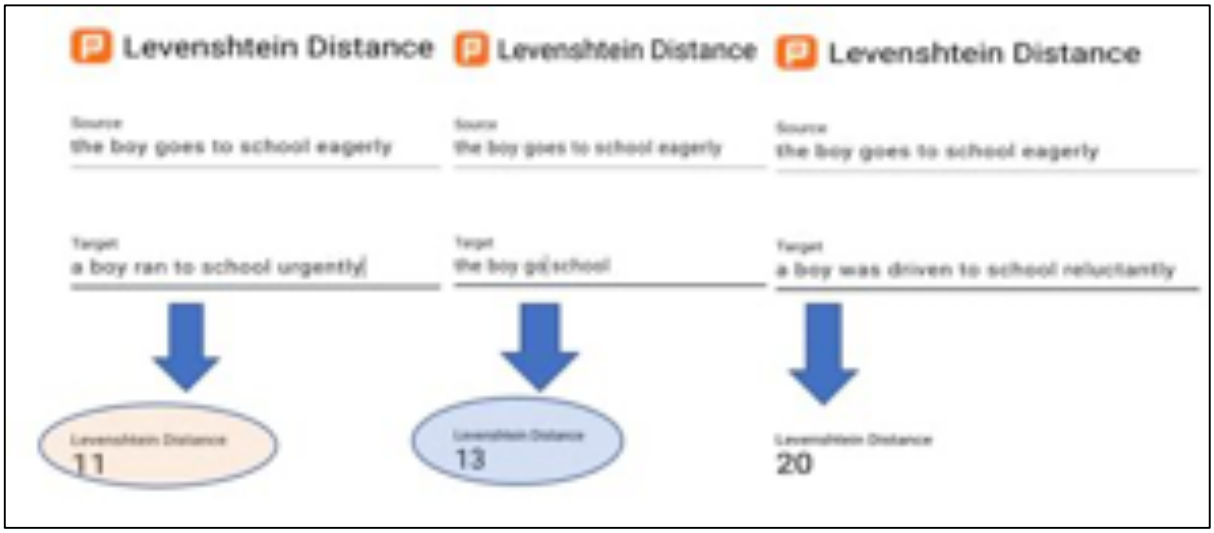


Figure 4: Levenshtein Distance does not always behave quite as desired

Nevertheless, there is a positive correlation of the measure with the closeness to the verbatim Source-Ref. The metric can be applied across all texts, looking for the comparable chunks in output. Figure 5 shows the calculations of distance across the outputs and across the timeline.

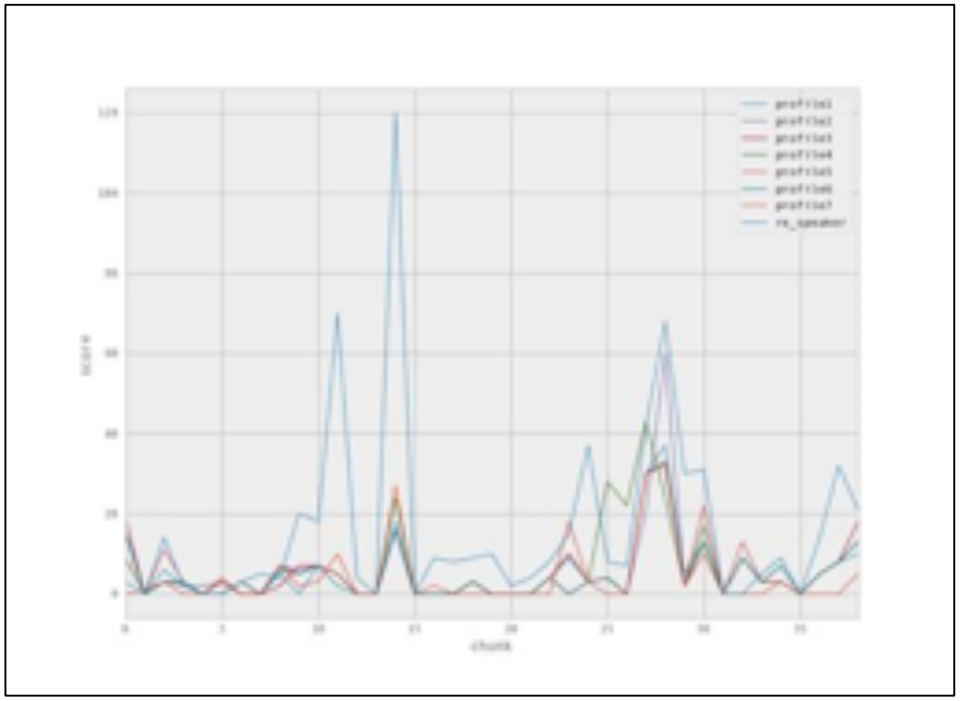


Figure 5: Analysis of folder (various SR profiles against the Source-Ref) arranged by matching chunks of text across the timeline

This character level algorithm generates a metric with the lower scores indicating a closer match. On the assumption that the relative performance of the ASR changes and settles over the duration of the call, we can examine the figures as a cumulative accuracy score (Figure 6).

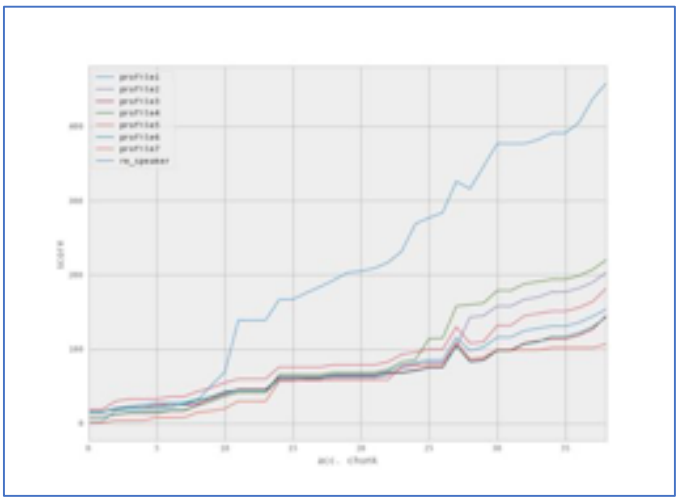


Figure 6: Cumulative accuracy measurement showing increasing deviation over the duration of the call.

In this case it begins to be apparent that the Re-Spk$_1$ increasingly deviates from the Source-Ref as the extent of psycholinguistic processing being carried out interacts with the volume (text) of the utterance i.e. the re-speaker omits section of text or paraphrases.

Instead of using character level analysis, we can examine the outputs exactly by matching each word to the Source-Ref. This is hugely time consuming and impractical in an actual call. However, the character level analysis turns out to be (overall) consistent with the word for word analysis carried out on the same profiles (Figure 7).

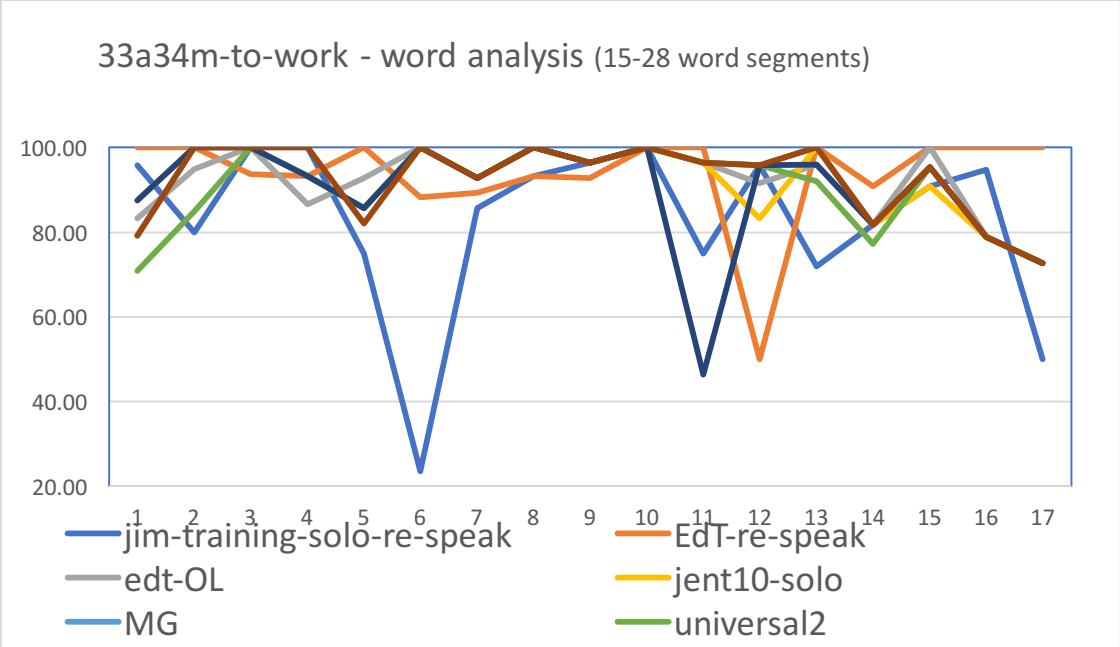


Figure 7: word by word comparison of verbatim accuracy

Word for word analysis has to be carried out manually, using spreadsheets to match words. One set of files in a folder takes around 3 hours to analyse word for word. In Figure 7, in an utterance by utterance comparison of verbatim accuracy, the higher scores indicate closer match. When utterances are shorter, the likelihood of perfect accuracy tends to be higher or even maximal. This figure also shows the drop in accuracy of the Re-Spk$_1$ at utterance 6 where there is a significant omission. The omissions and substitutions of the Re-Spk arise from the internal processing of the message as mentioned earlier.

In the next step for clarity, we can extract the three profiles which provide the best match (Figure 8).

In this case we group the utterances to try to create comparable utterance word lengths. Again, we find a second re-speaker omitting text.

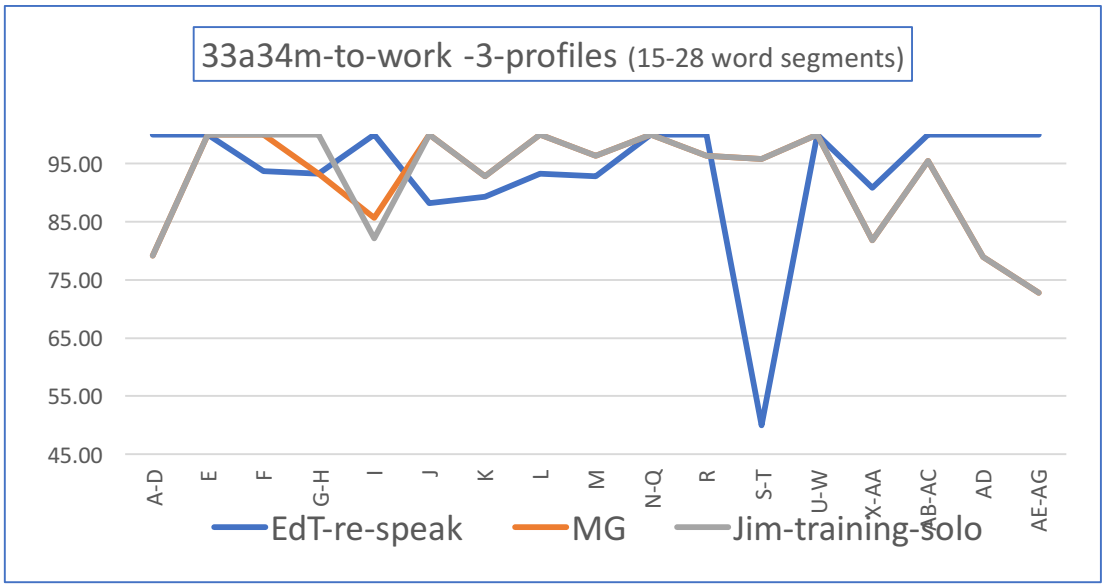


Figure 8: Word by word comparison of accuracy with utterances grouped for comparable lengths (% accuracy)

As in the case of the character level analysis we can compute cumulative word by word verbatim scores.

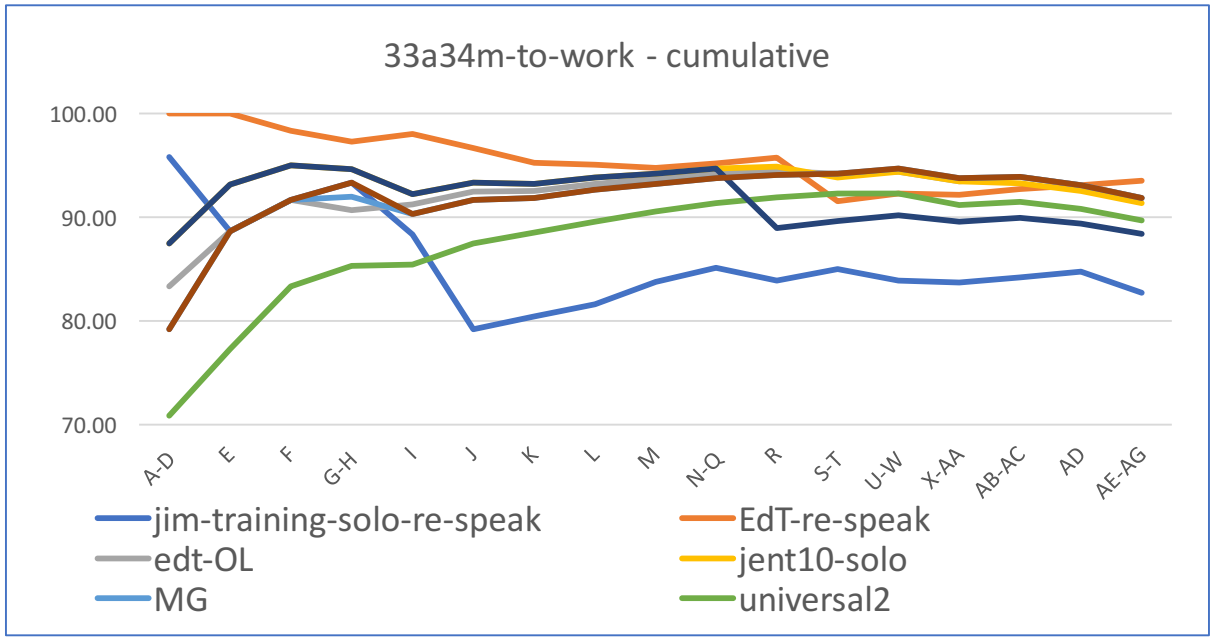


Figure 9: Cumulative word by word comparisons

Again, we see the poorer performance of Re-Spk$_1$ while Re-Spk$_2$ performs better. This can be seen more clearly in choosing the 3 best preforming profiles. (Figure 10).

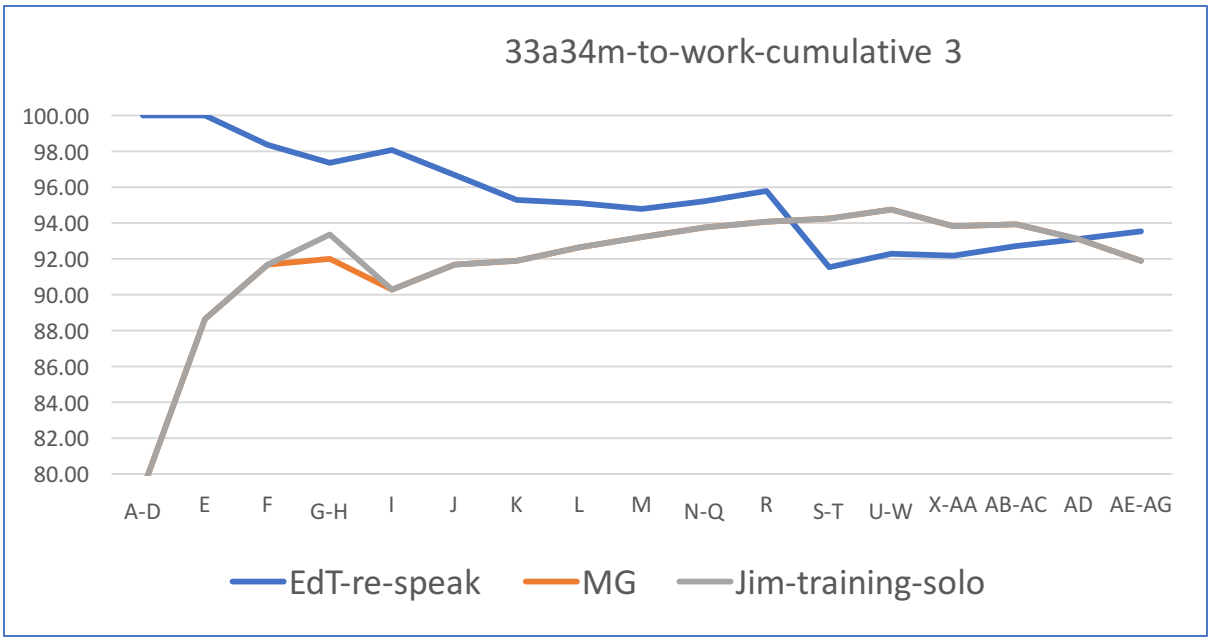


Figure 10: Word for word verbatim accuracy for 3 best performing profiles

We can see now that the performance of Re-Spk$_2$ is better overall although despite the omissions in the first two utterances, the SR profiles perform well.

In summary, we find the word level analysis produces a similar result to the character level. With the male re-speaker poorer than the ASR female profiles. Accuracy level is around 90% overall.

Ideally, we could carry out a similar analysis at semantic and syntactic level; where we might expect the re-speaker to do better.

However, it was also possible to create an original-speaker profile on-the-fly – that is by using the data we have collected on the person's speech in the other telephone calls he/she carried out.

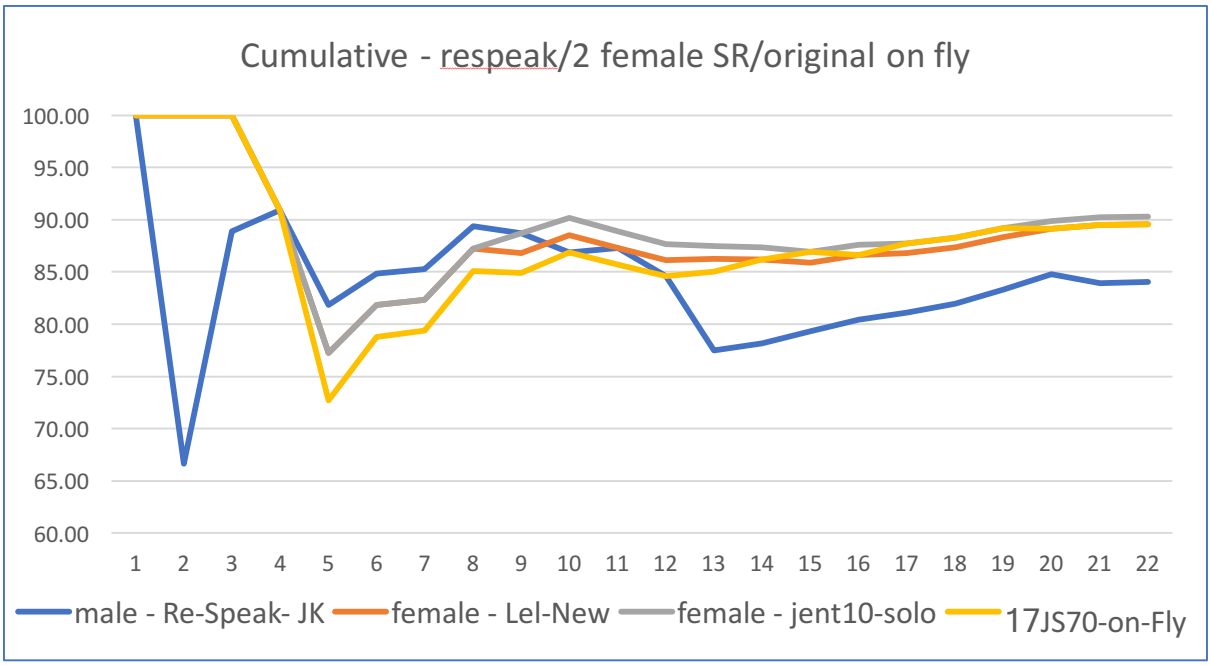


Figure 11: Cumulative word for word verbatim accuracy
but including a profile for the source speaker based on performance on other calls

Here we find that the speaker-specific profile generated, on-the-fly, produces a cumulative performance very close to the trained profiles. That is, using the person's own voice from other separate calls, can produce comparable accuracy. Possibilities here are of interest, since it might overcome some of the issues of diction and accent.

4.2 Implications of character-level analysis

These analyses have been repeated with four folders (bear in mind that this analysis is very time intensive – requiring close to 2 person-days to complete per folder). The results are similar and indicate the relative performance of the different profiles, this strengthens the argument for post-processing of the STT output and the parallel examination of performance by utterance and consequently, by time.

4.3 Issues of Concordance of text output at word/part of speech level

The search for a metric which could be applied in a post processing stage continued. In this post-processing stage it is envisaged that a large number of ASR sessions would occur in parallel and that the metric of accuracy would be traded against time lag. Different settings, even different users would set the priorities.

We required a better automated language analysis system which would offer greater sensitivity to the text output in terms of syntax and potentially in terms of semantics. Part-of-speech tagging is a well-established approach to text analysis and is a fundamental of the British National Corpus development. (see http://www.natcorp.ox.ac.uk/ and http://ucrel.lancs.ac.uk/claws/). The possibility to carry out the analysis at word level using such a system appeared to be a considerable step forward. The limitation was that it would not address the timeline variable.

Details on part of speech (Pos-Tags) is provided in Appendix 2. In brief, the purpose is to examine word-syntax and function in the text. Our argument is that this is a more intelligent means of text analysis than simple character level or even word level. It takes us closer to the workings of natural language. Each word or word inflection is allocated a tag. This process is automated and very fast. Accuracy is claimed in the high 90 percent. The classification used in the analysis is the CLAWS 5 tags (http://ucrel.lancs.ac.uk/claws/).

The online tagging tool provides an immediate print out:

The_AT0 online_AJ0 tagging_AJ0 tool_NN1 provides VVZ an_AT0 immediate_AJ0 print_NN1 out_SENT

Each component of the sentence is assigned a part-of-speech (POS) tag. We chose then to assign a weighted value to each component based on the perceived significance of that component in an ASR output of a phone call. We also grouped various components – e.g. we gave all verb components an equal value and all nouns an equal value. The program which was created then calculated the weighted score for each utterance and compared this to the Source-Ref weighted score. If an ASR profile produced a value the same as or very close to the Source-Ref, then we are predicting better performance in ASR. The task will then be to

choose that ASR profile in a timely fashion to provide the user with the most accurate STT output.

In order to understand this process better, we have been able to express the relationships mathematically.

The first step is to map the Pos-Tag value of the Source-Ref against the Re-Spk values and against the ASR Profile values. Removing the timeline consideration we simply use all values for each utterance. These plots of concordance can then be examined as linear regressions. If an SR Profile Value and the Source-Ref value are matched perfectly (verbatim output), then the regression line will offer a slope of 45 degrees.

The following series of figures illustrate the outcome of the application of the metric described generated from the program we have developed.

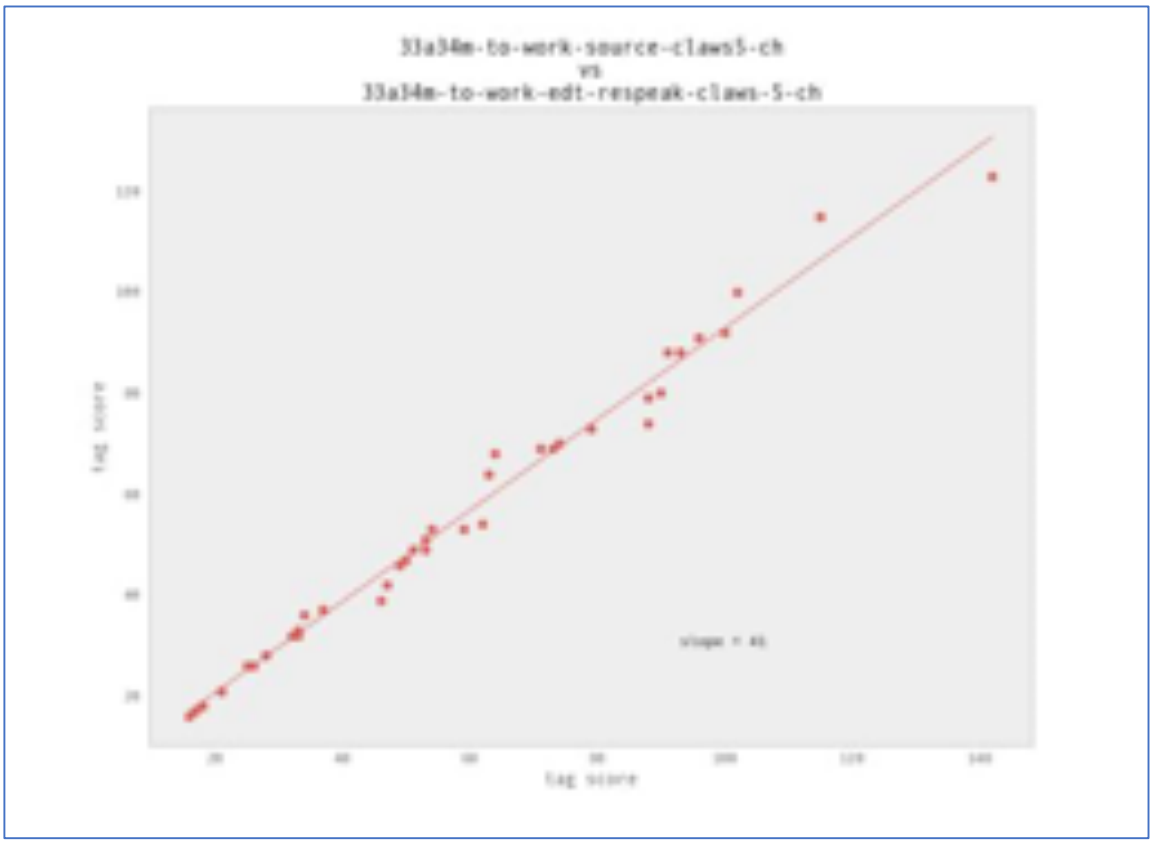


Figure 12: Plot of Pos-Tags scores of Source-Ref against the re-speaker (Re-Spk$_2$)

Figure 12 illustrates the closeness of the STT output from the re-speaker to that of the original source. In effect, the metric illustrates close concordance of the re-speaker text with the original i.e. better performance.

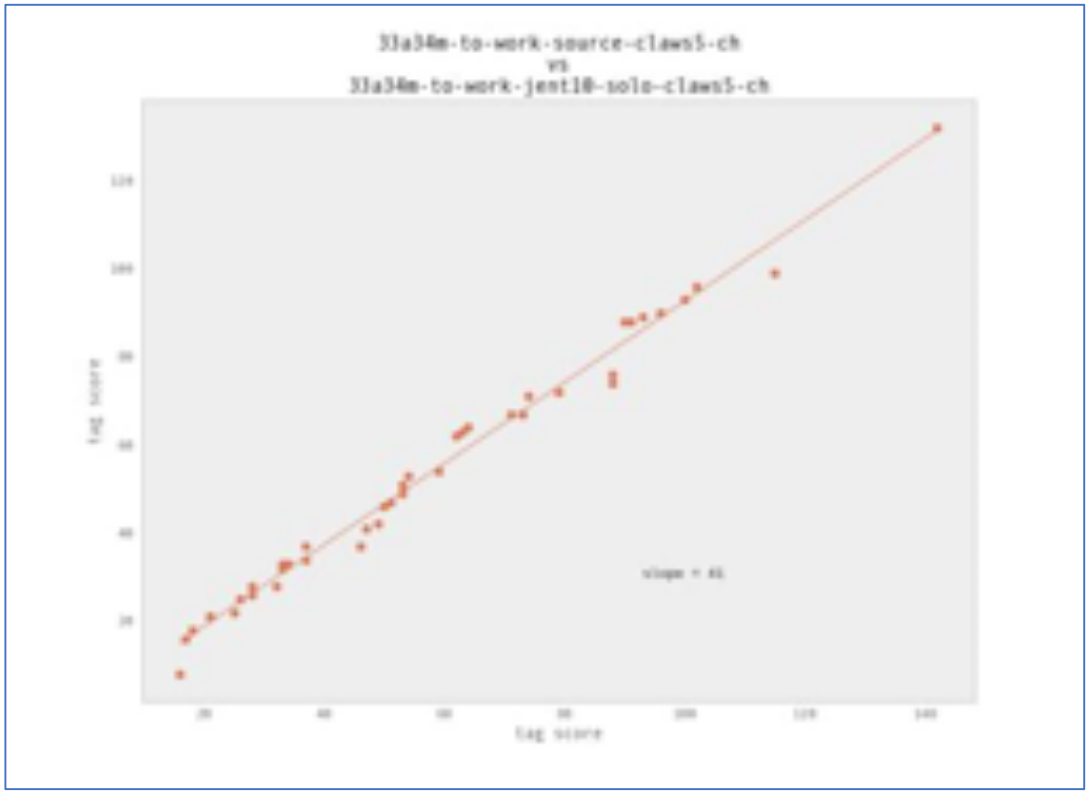


Figure 13: Plot of Pos-Tags scores of Source-Ref against the ASR Profile, Pr_2)

A similar relation can be found in Figure 13, which is more significant as this represents the relation of ASR and the source text, implying the working of a process which does not need human intervention.

In contrast, in Figure 14, a different male SR profile compared to the source is much less accurate. In particular we see a particular damaging outlier in the relationship (point 2,500, 2,400). This is likely to entail considerable omission of the original text.

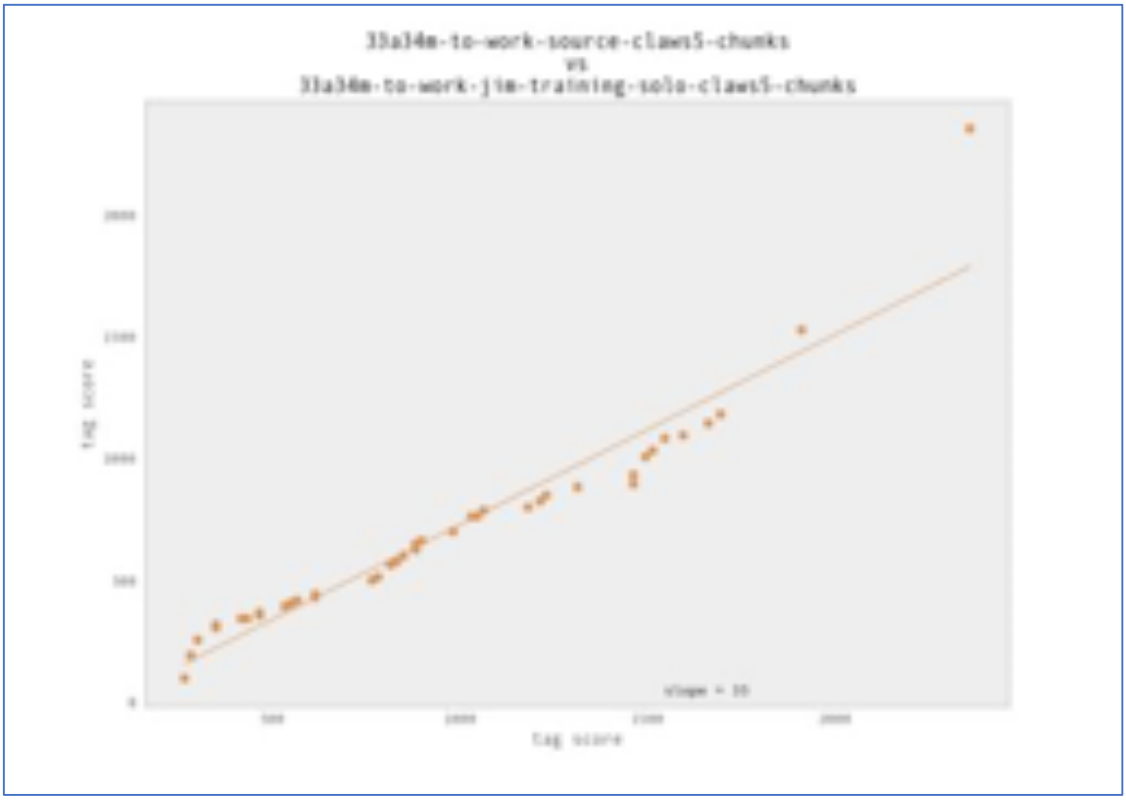


Figure 14: Plot of Pos-Tags scores of Source-Ref against a male ASR profile (Pr_5)

Figure 15 shows the extent of calculation of the relationships within the plots of Pos-Tags, illustrating the post-processing which we aim for in determining the best ASR profile to use in TC-Cap.

We can express this more succinctly by considering the correlation matrix of these values (Table 1) and this in turn opens up the possibility of meaningful cluster analysis of these inter-relations.

Correlations marked in yellow indicate high levels of accuracy in regard to the source. Correlations marked in blue suggest the relation to the re-speaker text. Correlations marked in green show the cluster of agreement of the ASR profiles … i.e. when human intervention is removed.

For the first time, this allows us to envisage post-processing of the ASR-Profiles which produce a meaningful accurate rendition without knowledge of the original source (as in a typical phone call) and without knowledge of a re-speaker operator.

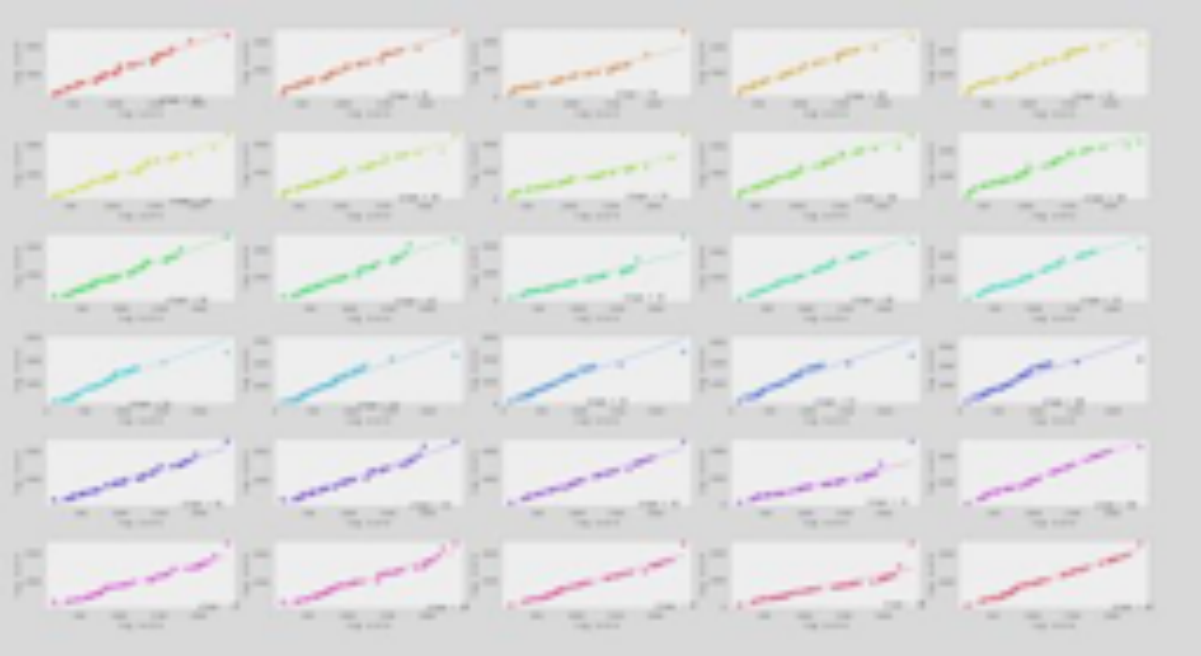


Figure 15: Plots of the combinations of SR Profiles and Source-Ref

	source	**edt-respeak-**	**jent10-solo**	**jim-training-solo**	**MG**	**universal2**
source	1.00	0.97	0.95	0.58	0.95	0.94
edt-respeak-	0.97	1.00	0.95	0.60	0.96	0.95
jent10-solo	0.95	0.95	1.00	0.66	0.98	0.97
jim-training-solo	0.58	0.60	0.66	1.00	0.64	0.63
MG	0.95	0.96	0.98	0.64	1.00	0.99
universal2	0.94	0.95	0.97	0.63	0.99	1.00

Table 1: Correlation Matrix of Pos-Tag scores for each profile

We are still in the process of refining and automating this set of calculations but if offers a new means of post-processing of STT which can be envisaged as concordance generated from a wide range of profiles.

5.0 Discovering Dialogue: Potential Role for Bilateral STT

While the project initially envisaged working only with the text output to the hard of hearing person, it became apparent that dialogue works dynamically between the two participants. In telephone calls without visual context, the two participants not only speak and receive spoken information but also jointly construct the meaning and create a transaction. There is a more detailed explanation of this aspect in Appendix 3.

Conversation proceeds from an intention on the part of one person to interact, usually with a purpose which needs to be designed and shaped by the interacting utterances. Intention is constructed as meaning which finds a vehicle in its lexicon and syntax, which then invokes its process of articulation. This is fairly obvious.

However, the words as spoken aloud are transitory – they do not exist physically once they are spoken and received. As the words are uttered, the words disappear:

Jack..
and..
Jill..
went..
up..
the..
hill..

The magic of speech is that the words are not only received but perceived – identified as meaningful as each word appears. Psychologists talk about a short-term memory to hold short sentences (7 or so items) but in reality speech perception does not require attention to each single word as it appears because …. the listener has already predicted the word sequence and output (to the extent, sometimes, that they seem to mis-hear). In terms of the internal lexicon, the listener is cueing an area of meaning and a likely range of words, prior to the reception of the incoming speech.

This is helped enormously by the syntax and thereby the redundancy of English. We know well in advance of receiving the last word that it will be “hill”.

In a telephone dialogue, the listener is already actively involved in understanding prior to the response of the other participant.

Telephone or spoken language dialogue allows participants to jointly create meaning but not sequentially. That is, the processing of the incoming speech precedes the actual output of those words from the other partner. The dialogue proceeds as a matching process of what a person is saying to what is expected.

This creates a fairly obvious difference between spoken dialogue and text dialogue. While the speaker-listener dialogue proceeds more or less in parallel, the speaker-reader dialogue has to wait for the visual reception of the words, which in the case of STT appear as whole utterances and NOT as a sequence of individual words. Meaning construction for the reader is reliant on a different set of perceptual skills and cannot begin effectively until the last word in the utterance is "seen".

It is unlikely that TC-Cap can create the dynamic fluency of spoken dialogue because reading responses in text does not allow the same processes to be implemented. Given that analysis, there remains much to learn from the nature of dialogue. While psychologists have worked on speech perception, linguists have created another field: discourse analysis.

5.1 The Relevance of Discourse Analysis

There is a considerable amount of research on the way in which language is learned through interaction. In fact, most linguists would disagree with a notion that language is taught and would rather explain that language (and meaning) is constructed by negotiation with a partner in speech. As a child, this is an interaction with someone whose language development is more advanced (but who is able to adjust the nature of his/her speech to fit the current level of language competence of the child).

Discourse analysis divides speech dialogue into: interaction, transaction, moves and acts. Details on this approach are set out more extensively in Appendix 3. Figure 16, illustrates a fragment of a telephone conversation where A's first speech act creates a move to "elicit" information; B's act "informs" and A's response "acknowledges".

These building blocks are fundamental to the intended interactions.

What it implies is that in a telephone call, both parties are responsible for the creation of meaning and establishing the outcome.

Instead of TC-Cap providing a representation of the other's speech to a hard of hearing person, the task is to allow both parties to engage in meaning construction.

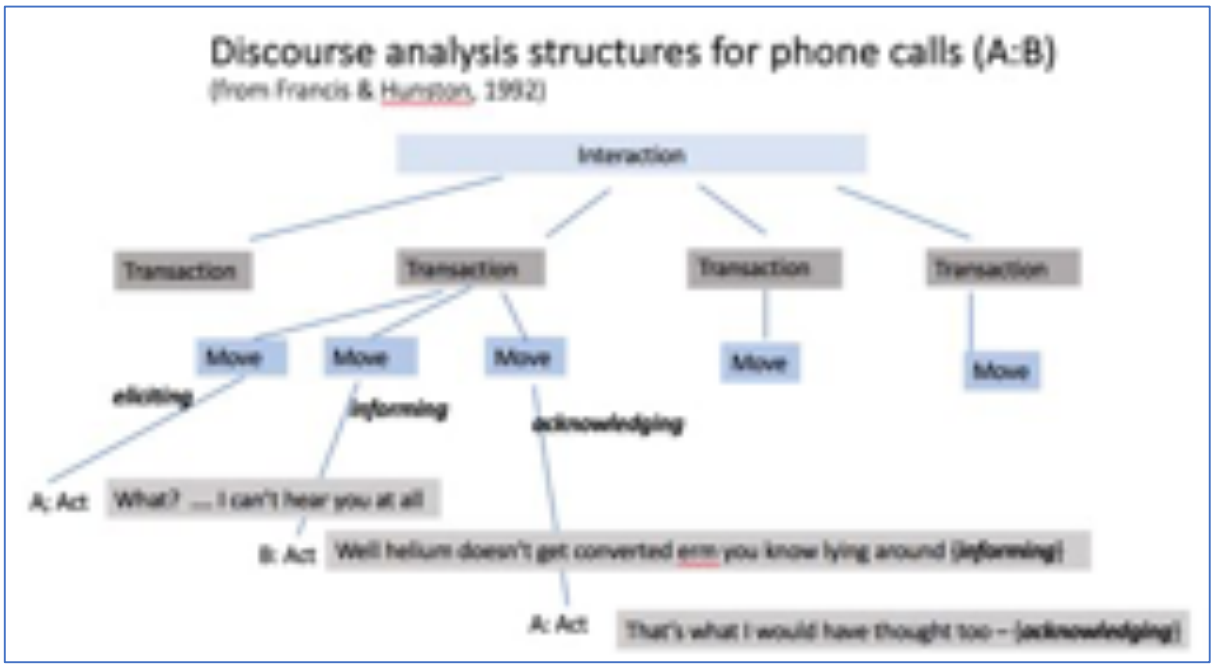


Figure 16: Acts and Moves as fundamental to transactions in the interaction

5.2 Implications of Including Dialogue

This pushes TC-Cap towards a different framework for development in several aspects:

1. Our STT system needs to be applied to *both parts of the dialogue* – the hard of hearing person's speech as well as the hearing person's responses. This does not imply displaying the hard of hearing person's speech as text but rather extracting from it, cues to expected meaning. In effect, we need to be able to classify the hard of hearing person's utterances in a way which will make more accurate calls to the STT engine for relevant output.
2. The dialogue which is created through STT now can be seen as imbalanced. In hearing-to-hearing interaction both parties produce speech and receive speech (usually assuming that the spoken words are the same ones as perceived by the listener). However, in hard-of-hearing to hearing dialogue through STT, the hearing person may be unaware of the words as perceived by the hard of hearing person (unless we use software which returns the text back to the hearing person's handset).

In hearing to hearing conversation errors of mis-articulation/mis-perception are trapped and corrected by the participants (usually jointly). This may be more difficult in the hard-of-hearing-to-hearing-speaker dialogue.

Training of the participants to understand this aspect, is one possibility but perhaps the more effective solution is to provide a means for the hearing person to monitor their output.

3. Since both participants in the dialogue can be seen to be important, then this has implications for the way in which we set up our database. Instead of flattening it as Person A's utterances and separately Person B's utterances, we may need to interleave the acts, moves and transactions.

   This creates a different issue as now the speech characteristics of each participant are different and the ASR profile which is functioning best for one participant may not be effective for the other. The simplest example is male-female interaction. The implication is that the multi-ASR profile approach has to be applied to each part of the conversation separately, even though the *text* database which has been created does not distinguish the speakers.
4. This will require the separation of both streams of the conversation – which is where a bespoke application is needed. Each speaker can then be processed through a favourable ASR profile.
5. These insights begin to indicate that simple dictation STT (Acoustic model then superficial language model) are insufficient for this dialogue. We need to incorporate a more effective Natural Language Processing (NLP) model and ultimately a machine-learning AI model (Appendix 5 and below).

6.0 Initial Lab Trials

As well as working on the accuracy of the STT output and developing the multi-ASR-profile approach, we have also tested the concepts of ASR in telephony with potential users. This has been done using a previously developed Total Conversation application (laptop based APS-50) which connects users in video, voice and text in an Internet call. The software underpinning this was developed in the REACH112 project where we set up a UK national pilot for Total Conversation with over 2,000 users. Unique telephone numbers are assigned

and users can call each other directly or can engage with a hearing voicephone user, automatically connected through the relay service. The pilot programme ran from 2010 until 2014 but was not funded after this due to disputes between central government and the telecoms operators as to who might be liable for financial support. In the period of austerity, the government departments were unable to provide support for what needed to be a human intermediary 24-hour service. (In contrast, TC-Cap is proposing an automatic service without the need for human attendance).

In the TC-Cap trials, users were connected to the hearing-speaker in the same network but in different rooms. Figure 17 illustrates the screen display for the hard of hearing user.

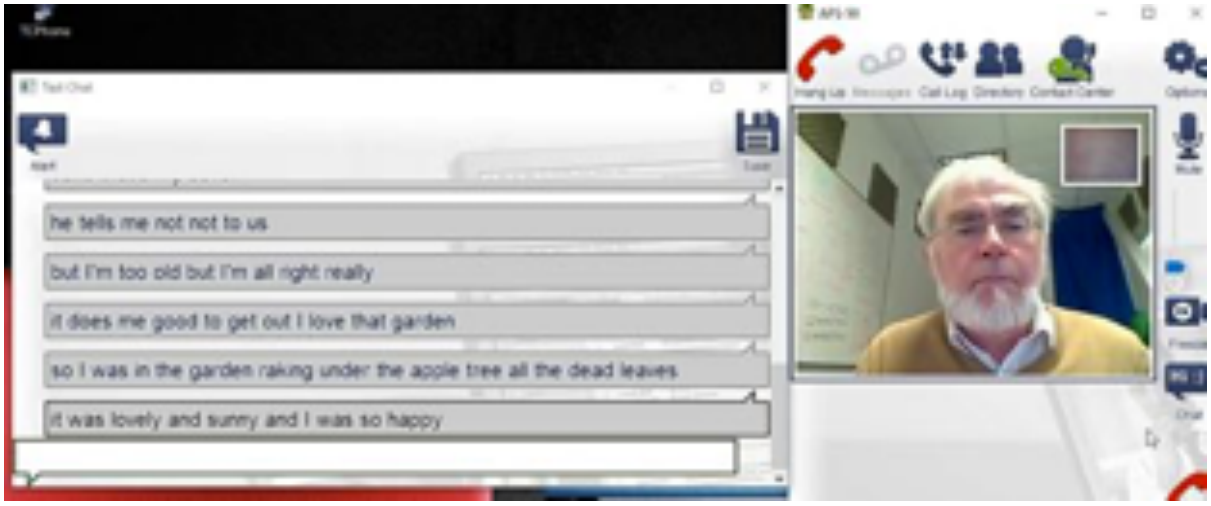


Figure 17: APS-50 application
user sees the hearing-speaker and the STT output appears in the box on the left.

The text box on the left, displays the words spoken by the hearing person. After each utterance, the text box moves upwards and a new line is displayed. All text in the interaction can be saved by the user for later reference.

The user had several conditions:

(a) Could see the hearing caller onscreen (but not hear his/her voice) and had the possibility to interact visually, although this tended not to occur as the user was focused on the text output.
(b) Could not see the hearing caller and had to use only the text output, as it appeared onscreen

In both these conditions, there was no audio available (of the hearing caller).

(c) Could not see the hearing caller but could just about hear parts of what he/she was saying, due to low volume and background noise.

In a fourth condition (d) some deaf users typed their comments/questions/responses to the hearing person, who responded by speaking and having that speech displayed as text.

The eight lab trials taken as a whole, were meant to be a proof of concept of STT in a telephone interaction. All interactions appeared to be successful (that is STT accuracy was in the high 90% and close to 100% in the most recent when the database alterations were made) and the qualitative data collected was highly informative.

Users felt they had to concentrate primarily on the text screen and there were pauses in the interaction as STT process was underway and further delay as the text was displayed on screen and had to be read. There was a tendency for cross-talk, where the user tried to fill gaps in the interaction and some STT output arrived after the user had commented further. Users considered that having a visual display of the hearing person reduced the cross-talk issue, as they could tell when the speaker was active and when he/she had stopped.

An interesting issue from the perspective of the hearing speaker was that the ability to see his/her own STT output was helpful in determining exactly what was being received by the user. In this case, obvious errors could be corrected, although in most cases the user was able to retrieve the meaning. In common, user-to-hearing interaction, envisaged in the future TC-Cap application, hearing speakers may be using only a ‘normal’ voice phone and so will be unable to see the text which is being sent.

These lab trials were also partially replicated with the Linphone mobile phone app (Figure 18).

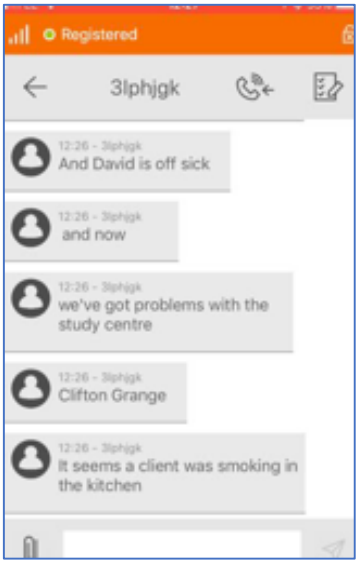


Figure 18: Linphone mobile phone app user sees a text display which scrolls upwards

In this case, the user has a text display only of the hearing person's speech.

Based on these trials we consider that the incorporation of TC-Cap software and STT software into the video telephony application is perfectly possible and desirable.

7.0 Going beyond: The role of Artificial Intelligence

Despite the optimism in the previous section concerning feasibility, it was still the case that use of multiple ASR profiles, produced different errors and there were clear gender differences; that is, if a male profile was applied to an incoming male voice, then differences were noted if a female profile had been applied. These differences were the reason for the planned use of many profiles and the post processing which was envisaged.

Surprisingly, perhaps even the post processing had to work with output from Dragon Dictate which was ungrammatical and often nonsensical.

The insights in Section 5 led us to the view that both parts of the dialogue needed to be analysed and at some point, a means of cueing semantic, syntactical, lexical and pragmatic areas of the knowledge base should be implemented based on the hard of hearing person/'s queries and comments.

It also became clear that a much larger database of voice-call transcripts were needed.

7.1 Utilising the whole of the TC-Cap recorded database

However, as a starting point, we added all of the Source-Ref texts from the 1,000 minutes of calls to the 'training' function of Dragon. As far as we can tell for each STT profile which we designate, this process is designed to identify new vocabulary and to adjust 'style'. It is primarily, designed for single speaker profiles for dictation but we utilised this for five different profiles.

7.1.1 Accuracy and output

The ASR process was then run with audio which had already been transcribed and had been used for training. The result was much improved performance and accuracy. This would seem reasonable if Dragon was 'learning'.

These profiles were then applied to voice call recordings which had not been encoded (and were not included in the 'learning' process. As before, the performance is enhanced (see Table in Appendix 8).

Nevertheless, there continued to be some errors and the problem of different profiles producing slightly different errors remained.

7.2 Augmenting the speech dialogue database

Given that much of the development work on speech recognition has been carried out using very large-scale databases, it seemed appropriate to search for additional resources for TC-Cap to use. As indicated already, there is very little public domain audio transcripts of telephone calls, but given our analysis of dialogue, it was decided to broaden the scope of the search for more data. Appendix 7 sets out the details of this work and indicates the identified database additions.

A subset of the Hansard transcripts (in 2017, dealing mainly with disability) and the transcripts of the TV/radio programmes In Touch, Andrew Marr, Moral Maze and You & Yours were added to Dragon Dictate for learning.

A new audio recording of a phone call was then processed using five different ASR profiles. The results (Appendix 8) were remarkably accurate but with one particular area of disagreement where four different text versions of the comment:

"What a mess!"

were produced:

"one mess", "1 domestic", "one May's", "one May's" and "What a mess".

Since the incorrect text of "one May's"[3] would have been selected by a simple voting system, we needed a different approach to trap this error.

The solution comes in the field of machine learning.

7.3 Machine learning tools to determine statistical probability of accuracy

It was essential that a statistical probability could be generated for each of the outputs, in order to reverse the errors created by the simple STT system and the likely voting solutions as shown in Section 4.

Appendix 5 & 6 set out the approach taken. This involved the capture of a much larger database of speech coupled with the telephone database which TC-Cap had already collected. Our aim was then to determine the likelihood of co-occurrence of particular sequences of words, based on their presence in the much larger database.

7.4 Implementation of AI in trapping errors in multiple ASR

The approach taken involved a technique called word embeddings within the concept of Deep Learning. (See Appendix 5 and Appendix 6). The tool used was Word2Vec which creates a multidimensional space of the co-occurrences of particular words. This mirrors a supposed conceptual or semantic space in psycholinguistics. Put very simply it can create a "map" of words which co-occur and predict which words are likely to precede or to follow particular sequences.

[3] Interestingly, the output "one May's" was produced by both of the female SR profiles, but not by the male profiles.

This Deep Learning process is resource hungry and the extent of processing can be ever-changing as new data is added. Given the current phase of the project, we choose to focus on 300 words from the overall extent of 148,563 in the database but allowed dimensionality of 300 (extent of interconnectivity). In later phases of the development considerably larger databases will be imported and dialogue features will be explicitly examined.

We were then able to apply an algorithm to test the statistical likelihood of the four STT outputs as indicated above. By adding rules for the presence or absence of certain components we were able to generate a log likelihood value for each of the four outputs.

As detailed in Appendix 6, this rejected the nonsense outputs and chose the most likely and correct output.

While the Deep Learning phase is complex and time-consuming, the testing of specific outputs is fast and feasible within a phone call timeline.

7.5 Potential for this process

We consider this development a significant one. While Deep Learning is applied within Dragon Dictate, it does not yet deal with dialogue nor does it account for application where the `ASR profile is not the profile of the speaker, who is of course, unknown in the call.

We have moved significantly beyond the need for a re-speaker, and beyond even our earlier character level and Pos-Tag analyses to having a function which can assess likelihood of the phrases produced by different ASR profiles.

8.0 A product with a third-party system voice

There are two absolutely crucial insights which we have gained from the detailed examination of STT in telephone calls. The first is that text dialogue displayed on screen is not the same as an audio interaction - the display modalities are different, the fact of having to read a text is quite different from having to hear the speech and the pace of the interaction is slowed.

The second is that spoken dialogue is not a linear question and answer process. In fact, the interaction of the two speakers creates a *shared meaning*. We can consider this as a negotiation but it is reliant to a large extent on the two participants sharing the same conceptual domains. The interaction proceeds through a series of transactions, constructed through the speech moves of the participants and these moves are underpinned by the relevant speech acts. *Meaning is created from already known information* and is agreed upon through the medium of conventional language. Both speakers have a vested interest in constructing this meaning and both engage in a highly dynamic predictive interaction. Where confusion occurs or lack of reference (knowledge) in one participant occurs, then the transaction has to be repaired. This is done through requests for re-sending of the message or paraphrasing. It is the responsibility of the non-understanding participant to intervene (even in the middle of an utterance) to request such a clarification.

While perfect clarity in a telephone call is a supposed goal, many factors come into play which make this goal unattainable. Words are misheard, connections are poor, background noise occurs in one or other of the participant's environments. In those cases, the users have to repair the conversation …. and crucially, *will expect to repair the conversation*.

Our goal in TC-Cap does not have to be perfect verbatim representations of speech, but rather the challenge is to supply the necessary tools such that the participants can jointly negotiate the meaning. Where inaccurate text occurs in an ASR session, it remains the responsibility of users themselves to negotiate the meaning.

However, rather than simply leaving it to the users to sort out incorrect ASR displays, it is feasible for an intelligent system to trap errors and to communicate with one or both of the participants. There is some indication that if a third-party system voice appears in the hearing person's speaking/hearing, that this will create an acceptability problem for the hearing person. This may discourage a hearing person from using the system.

However, if the system can send text information to the hard of hearing person, then this is likely to be more acceptable.

Where the post processing of the ASR output is inconclusive, the system may display to the user:

{unsure} … "the room is not ready for them" ….. "their own is not ready for ten" *{ask to clarify}*

or where the post-post-processing, indicates that none of the options for output are considered to be statistically likely, the TC-Cap voice displays only:

{I am not understanding this speech … please ask for clarification or repetition}

The intention here is to ensure that the control of the interaction is `with the hard of hearing user.

Ideally, we wish not to indicate to the hearing person that he/she is communicating with a deaf or hard of hearing person and the aim will be to make the interaction seamless. The deaf or hard of hearing user will have a unique telephone number which looks identical to that of any other number, but which automatically routes incoming or outgoing calls to the TC-Cap system.

Users will also have the option to switch off or on, the TC-Cap ASR when the hard of hearing person feels able to cope by listening. At the same time, TC-Cap can record and later transcribe the call for the reference of either or both parties. There will be instruction to the user of the form of the explanation that the call may need to be recorded e.g.

"Just to let you know that I am hard of hearing, and I would like to record the call so that I can review it later. This is not to be used by anyone other than myself, without your permission."

Further insights were gained from the participation of deaf sign language users (who do not speak but rather type their responses). Although the typing slowed down the interaction (a well-known aspect of text telephony), nevertheless, the conversation flowed well from their perspective (as they are controlling the text input and the responses from STT appear quickly). However, they also queried the situation that if the hearing person did not have the TC-Cap app, then he/she would not be able to read the text. We know that this situation will change with the directive from the FCC in the USA that all smartphones should be enabled with T140 text functionality i.e. that inbuilt software will allow the display of text. However, a far more elegant solution which would then also include standard telephone lines, is for their text output to be reversed through the process of text-to-speech and then played to the hearing person as speech, i.e. through their standard ear piece.

Taking account of the above, we are imbuing the TC-Cap systems with *an intelligent voice* in the interaction,

9.0 Exploitation

We are continuing to analyse the development and implementation route for the TC-Cap product and service. This section is of necessity, preliminary. More detail is supplied in a separate paper on the development plans.

9.1 Marketplace and Need

The need for this service remains as pressing as envisaged at the start of the project. Despite progress in dictation models of STT, there are no applications which begin from the need of people to interact in live conversation. This lack is acutely felt by deaf and hard of hearing people whose existing services of text relay are primitive given the technology advances in communication. The numbers of people affected cannot be determined precisely as many, if not most, of the 10% of the UK population which do not hear well, simply avoid using the telephone and rely on others in cases of dire need. As well as affecting quality of life and efficiency/employability in the workplace, the lack destroys independence and reduces personal security.

Providing the TC-Cap solution, would empower and enable, would give control over remote communication and bring back into mainstream society, a significant number of people. The fact that hearing people are mostly unwilling to change their mode of interaction (bear in mind, our finding that hearing people rarely initiate calls through the text relay system), then the TC-Cap which rests the control of the interaction with the hard of hearing person and allows the hearing person to conduct the call as he or she would normally, is a major step forward.

9.2 Relevance and innovation in AI

The development is timely also in a technological sense as we become more aware of the potential of Deep Learning and as we develop the necessary algorithms to make appropriate conclusions on the output from improving speech to text applications. The general awareness and potential funding opportunities for innovation in this area point to a major opportunity. Appendix 5 provides a more detailed review of this field, as does the development plan paper.

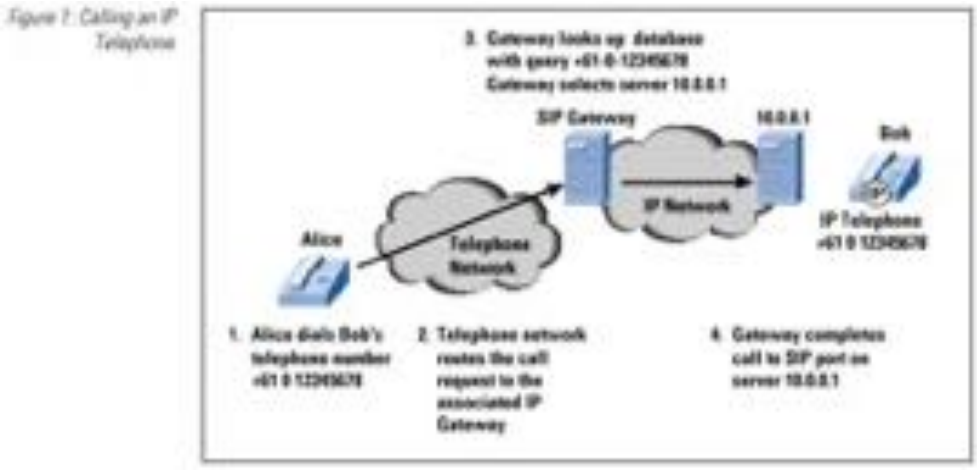
9.3 User interface

Our focus on the TC-Cap product development has drawn on our experiences with Total Conversation and relay services. It seems clear that attempts to alter the pattern of communication of the majority of hearing people, is unlikely to work in the short to medium term. However, given the view that telephone interaction is a joint negotiation of meaning, then we can ensure that this functionality can be respected by giving control of the interaction to the deaf or hard of hearing user. In effect, we do not envisage any alteration in the hearing person's telephone behaviour since he/she will hear voice and respond in his/her own speech. No new app is required by the hearing person.

The TC-Cap application will be registered to the deaf or hard of hearing person on laptop or on smartphone and all calls in or out of that supplied ENUM, will be connected seamlessly.

Huston (2002)[4] explains the principle:

"Consider an example. When Alice, on a normal telephone, wants to call Bob, on an Internet phone, all Alice needs to do is simply dial Bob's telephone number, or his E.164 address (Figure 1). Of course, because Bob's phone is connected to the Internet and can't directly receive Alice's call request, a gateway is necessary. The telephone system should be able to map Alice's call request to the Internet telephony gateway that is configured to act as Bob's gateway agent. The gateway then needs to translate Bob's E.164 phone number into an IP address. Then the gateway has to map the telephone network signals associated with Alice's call request to corresponding signals within an Internet session initiation protocol, and then send these IP packets to Bob's Internet phone. If Bob answers the call, the phone uses the same protocol to inform the gateway, which then sends a corresponding telephone call code across the telephone network to Alice."



[4] Huston G (2002) ENUM—Mapping the E.164 Number Space into the DNS - *The Internet Protocol Journal - Volume 5,* Number 2

This means of connection allows the deaf and hard of hearing user access to TC-Cap in their telephone call and provides the STT and TTS functionality.

9.4 Business case

As well as the simple application driven by the needs of deaf and hard of hearing people, there are many aspects, there are many other circumstances where interaction by telephone may be enhanced.

9.4.1 Emergency services

Most pressing is the need to enable the 999 service with a speech to text function, where not only hard of hearing people are able to read what the operator is saying, but in the ambulance service, where instructions on caring for the person in the incident are given over a periods of several minutes, then having a text display which can be read and re-read may be a particular advantage to hearing people under stress.

9.4.2 Call Centres

It is perhaps rather obvious that those agents who work in call centres may deal with a wide range of callers whose grasp of English may not be perfect and for whom, the possibility to read what is being said and then to save that dialogue for later inspection, could be very important. A particular advantage to the call centre is that no alteration to their working is required in terms of special equipment.

A particular addition which would help the TC-Cap performance is the use of the agent's own STT profile which could be trained relatively easily by supplying existing recorded calls for processing i.e. no new actions would be required by the call centre agent.

9.4.3 Calls to services by non-native English speakers

It is widely recognised in the field of second language learning that conducting a telephone call is one of the hardest tasks in the early stages of learning the language. The lack of contextual clues and the differences in accent of the speaker, make it often, stressful to the language learner. Provision of a STT display could significantly improve the non-native English speaker's performance in the call.

10.0 Conclusions

There has been considerable progress in this short feasibility study. We began with a wide range of expertise in Language, Deaf issues, Video Telecommunications, Relay Services and capacity to develop to scale of public telecom services and have taken a concept of automatic speech recognition into that domain. Despite considerable technological progress in AI, Deep Learning, speech dictation (eg Dragon Dictate), various chat bots, command driven recognition (eg Amazon Alexa) and so on, there has not yet been the prospect of an automated telephony application for these technologies.

Having taken the components of this concept apart and analysed each systematically, we have

- demonstrated the feasibility of offering speech to text to deaf and hard of hearing people, in video calls, text calls and through smartphones
- shown that we can dispense with the intermediary, re-speaker, and then process incoming speech by using multiple SR profiles
- improved the STT performance by adding domain specific (ie spoken dialogue) texts to the recognition process
- shown the possibilities of character level and then syntactic post-processing, where regression analysis can determine appropriateness of output against the source reference text
- now introduced the notion of discourse analysis and the concept of negotiated, shared meaning, where partners in the call have the responsibility to repair any breakdowns
- consequently, understood that the TC-Cap system does not have to "know" the original source text as its verbatim representation but needs to recognise the dialogue characteristics of the two parties in the call and the consequent cueing and predicting of speech acts.
- applied AI techniques in a Deep Learning framework to an augmented database of spoken language texts
- In order to apply this additional post-processing, produced a new TC-Cap algorithm which generates probabilities of particular competing output texts (from the multiple ASRs)

- additionally, set a specification of the user interface and the third-party system voice which will allow the deaf and hard of hearing user to control the telephone call and the dialogue which occurs.
- produced a route map for development of the products and services, the technological back office, the marketing and offering of TC-CAP to scale

We consider that major steps have been taken towards the goal of access to telephony for all those who have difficulty in hearing another's voice remotely.

Appendix 1: What is Automatic Speech Recognition?

Ed Toms, BBITG

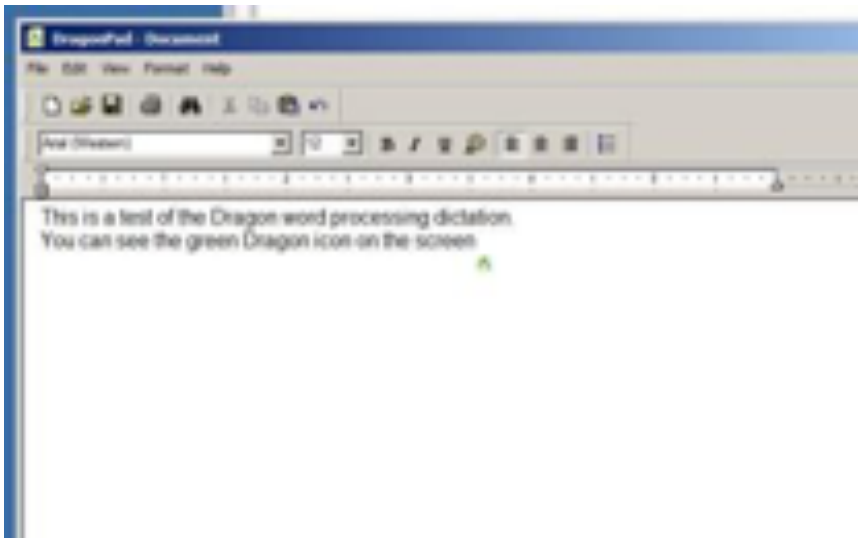
Automatic Speech Recognition (ASR) is the process of transforming the analogue waveforms of speech into the corresponding text using computer systems. The field has been in development since the early 1950s, with the early technology being limited to single-speaker systems with limited vocabularies of around ten words. Modern systems support multiple speakers, with vocabularies of thousands of words and in some cases claimed accuracy rates of up to 99%.

The current systems are made effective by building up a large database of trained examples with which to compare the current waveform being analysed. ASR systems do not necessarily need to understand the meaning of a text all of the time, they just need to apply a set of rules learned from previous examples, which could be 'mistaken' for understanding due to the accuracy of the predictions.

A good example of this is choosing between homophones. In many cases when distinguishing between the appropriate spelling of a word such as *two* and *too*, the system may not 'understand' why one form is chosen over the other, but it will still be able to select the correct form by looking at previous correct examples. ASR tools have different ways of building their database depending on the environment in which it will be used. *Nuance's Dragon Dictate* originally had a training process which a speaker had to go through in order to build a user-specific speech profile. Errors had to be corrected in the same session in order to train the system, or the text would be deemed correct and used for future decision-making. The latest version of Dragon Dictate (v15) does not require the speaker to train but once a significant corrected set of texts exists, a new acoustic model and new language model can be generated. Other tools with a very large user base such as Apple's *Siri* build individual user profiles, but are also able to collect vast amounts of data to refine the systems due to the

number of users. This builds a *language model*, which is a set of expectations about what someone is likely to say.

Jill Duffy: PCMag - 12 August 2014



Phonemes are the units of sound that distinguish one word from another in a particular language, for example the opening sound in the words *kill* and *kiss*. English has is said to have 44 phonemes. At its base level, ASR operates by analysing and comparing sequences of phonemes and selecting the most likely sound based on the preceding and subsequent phonemes. The important breakthrough which led to the rise in the accuracy of ASR systems was the ability to remember the current sentence rather than just the current word, allowing context to be used to make better predictions. However, as sentences became longer, the number of grammatically possible but nonsensical options increases exponentially, so very long sentences can be more difficult for ASR systems to transcribe accurately. Another difficulty comes from the reduction in clarity of conversation when compared with dictation. When dictating clearly, phonemes are distinct and easier to understand by the system. In conversation, speakers tend to run phonemes together, making it difficult to distinguish when one ends and another starts. As well as this, speakers are less likely to use full sentences that make grammatical sense, which can confuse systems looking to follow a set of standard grammar rules. The ambiguity of phonemes falls under one of the four main sources of variability in ASR systems, which are *Task Domain*, *Speaker Characteristics, Speaking Style* and *Recognition Environment*.

Task Domain - This refers to the type of vocabulary and whether it will be restricted in some way. For example, an ASR system used in a medical environment would need medical terminology, whereas an automated phone system may only need to recognise the numbers 1-10.

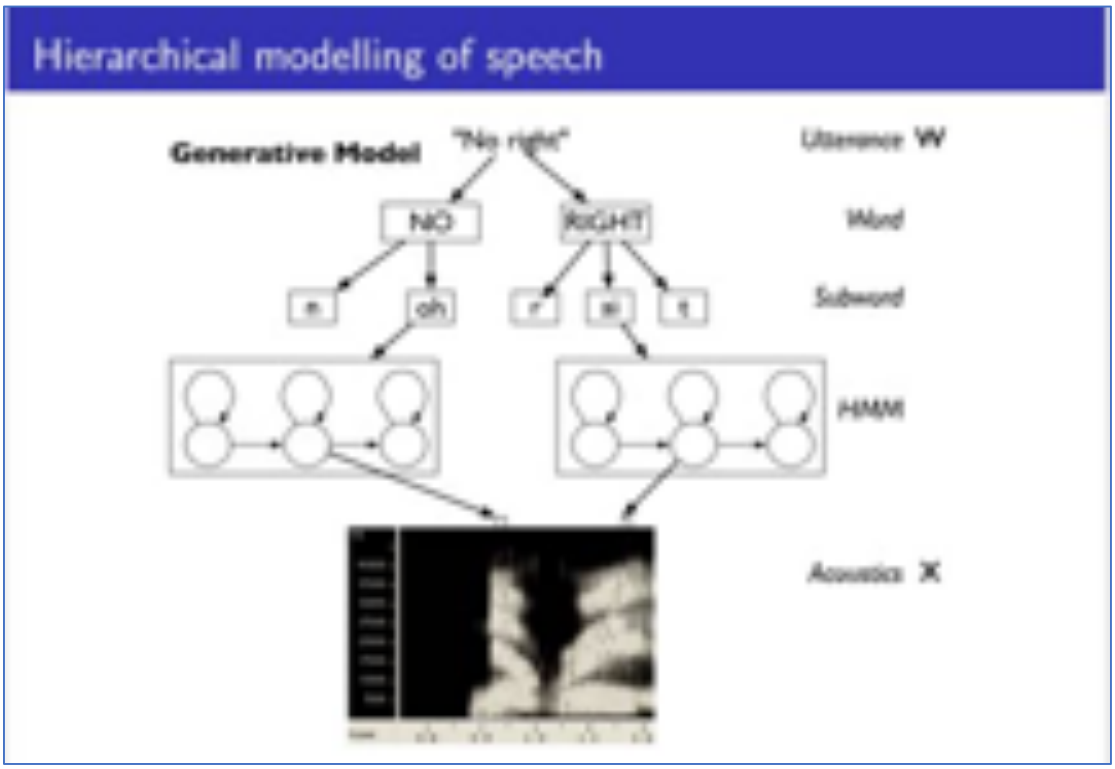
Speaker Characteristics - Speakers may pronounce the same phoneme in different ways, due to age, regional accents or other factors.

Speaking Style - The dictation style may vary, for example a casual speaker is more likely to run words together, whereas someone clearly dictating is more likely to take pauses between words; or dictation in a conference speech may be different from the same speaker in a social telephone call.

Recognition Environment - The acoustic environment in which speech is recorded, as well as transmission channel and microphone quality can have a large impact on the accuracy of an ASR system. The largest errors occur in situations with high additive noise, multiple acoustic sources or environments with high reverberation.

In order to carry out these analyses, the acoustic signal must be broken down, and then passed through a number of computer systems. The most common way of representing speech is using Mel Frequency Cepstral Coefficients (MFCCs). This allows the features for identifying linguistic content (the acoustic features) to be extracted, and discards other unnecessary information carrying the information such as background noise. Calculations and predictions are then carried out at this level. Analysing spoken words by identifying phonemes is often referred to as the *beads-on-a-string model*. These 'beads' are then modelled using a Hidden Markov Model (HMM). This represents the 'beads' the computer thinks are sitting on the 'string', based on the information it has obtained from the sound spectrum input.

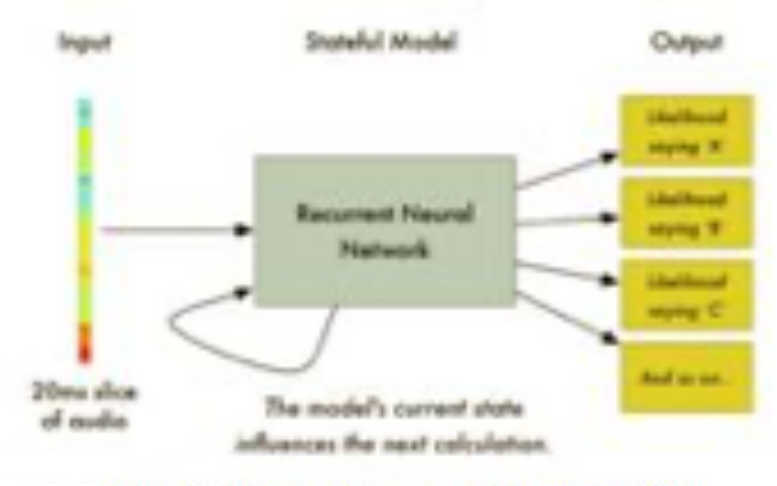
The basis of the model is a sequence of states (in this case the 44 phonemes) which change between each other with given probabilities. These probabilities are calculated based on the training data the system has, and it is a 'hidden' model as the current state is not known (as that is what is being calculated), so only the effects of it may be observed. The probability of a phoneme following the current phone is different for each of the 44 phonemes, so new probabilities are calculated with every change. These are also used in conjunction with a Gaussian Mixture Model to form a HMM/GMM.



Steve Renals & Hiroshi Shimodaira:
Automatic Speech Recognition - ASR Lecture 1 - 16 January 2017

These models may then be used alongside neural networks. Neural networks can find patterns in spectrograms much more easily than in a raw sound waves, so a spectrogram of the data is built and passed through a recurrent neural network, meaning a neural network that has a memory that will influence future predictions. Each audio chunk is analysed by the network, with the most likely letter (or letter combination) for each phone being calculated. However, due to the nature of the network predicting one character at a time rather than sounds or words, the output can be inaccurate, often reading like a phonetic translation of a word (this aspect, we later came to call the *Bruce Zuma problem* – where instead of "presumably" the ASR output was "Bruce Zuma"). In order to get around this, the pronunciation based predictions are compared against likelihood scores based on a large database of written text.

Recurrent neural networks (RNNs)[5] have taken over many aspects of speech recognition, mainly through the introduction of *Long short-term memory* (LSTM). This network can learn "Very Deep Learning" tasks that require memory of events that happened thousands of time steps ago (i.e. in long term memory). Conventional RNNs struggle with learning long-term dependencies, which in terms of language modelling means the systems struggle to remember data from several sentences ago and use it as context for future predictions. RNNs can generally predict short sentences, but references to information from several sentences ago is generally hard for the system to "understand". LSTMs are explained in more detail in Appendix 7.



Adam Geitgey: medium.com - 24 December 2016

Due to context being required to make correct word predictions, this is very important for speech. The application of Deep Learning led to a claimed 30% decrease in word error rate, and in 2015 Google's speech recognition reportedly experienced a claimed performance increase of 49% through the use of an LSTM.

[5] See Appendix 5 for more detail on Recurrent Neural Networks and Deep Learning

Appendix 2: Measurement of Accuracy in STT output

Lorenzo Benoni, BBITG

A2.1 The task

In the TC-Cap Service pipeline a group of ASR (Automatic Speech Recognition) engines generate text from the same audio feed, the phone call. Separate profiles can be used for specific speakers and it is especially important to 'train" such profiles in order to make them more efficient. See Appendix 1 for an explanation of this ASR process.

In regard to the timeline of the original/source/reference audio, these texts appear at different points in time as different profiles appear to be sensitive to different aspect of the utterances even though the same STT program is being used. While we cannot see inside the box of Nuance's Dragon Dictate engine, we find text generated only when the confidence level reaches a certain threshold of comparative accuracy or when the process itself, times out.

Figure 1: Three time-annotated sequences showing how similar texts are produced at different points in time

A2.2 The Strategy Adopted for Improvement

There is the need for an **automated and immediate comparison** of all the text produced by ASR profiles (from the same input audio source) to determine which profile generates the **most accurate text** in comparison to the original speech. Attempts to calculate this by inspection and human examination of outputs, is tedious and impractical in the context of live phone calls.

However, the concept of "accuracy" in this application is very difficult to define. One can insist on verbatim performance but after initial examination of initial speaker source input, re-

speaker output and ASR output, this strict definition is impractical as this is just not the way in which dialogue progresses in a phone call.

We sought quantitative approaches to the measurement of accuracy among the text outputs which would allow comparison relative to the source/reference text but which might include absolute similarity score, depending only on and strongly related to the structure of the utterance.

Also, the fact that even with the same engine different stored speaker profiles produce representations of the source text at different times introduces a **synchronization issue** (for the ASR comparison) and there is a need to identify, at any given point on the timeline of the call, if the ASRs all produced some text and, more important, if the texts are similar enough to be compared.

Using a subset of the audio database, two broad approaches were taken and the appropriate algorithms invoked in the software application which was written.

A2.3 The post-processing algorithms

A **Python program** has been developed to run such a test: as an input it was given a folder with the timecoded transcription of the call for the reference text (created by hand) (Source-Ref), the re-speaker output (Re-Spk$_1$, Respk$_2$ …) and the ASR profiles (Pr$_1$, Pr$_2$, Pr$_3$ …..). Two versions of the accuracy metric were developed:

(a) **Levenshtein distance** [http://www.nltk.org/api/nltk.metrics.html]: The program goes through the whole audio recording (of one side of the conversation) second by second, buffering all the utterances and testing the similarities, as determined by a character-by character comparison. Whenever the similarity test gives a result within a certain threshold, a score ("accuracy") function is computed. The accuracy function compares the generated text (Pr$_1$ ….) and the reference text (Source-Ref) according to the number of transformations (eliminations, substitutions, insertions) required at character-level to completely match Source-Ref to Pr$_1$…..

Figure 2: Python code that checks the texts and compute the scores

A simple calculator and example for Levenshtein distance can be found here

http://levenshtein.net/

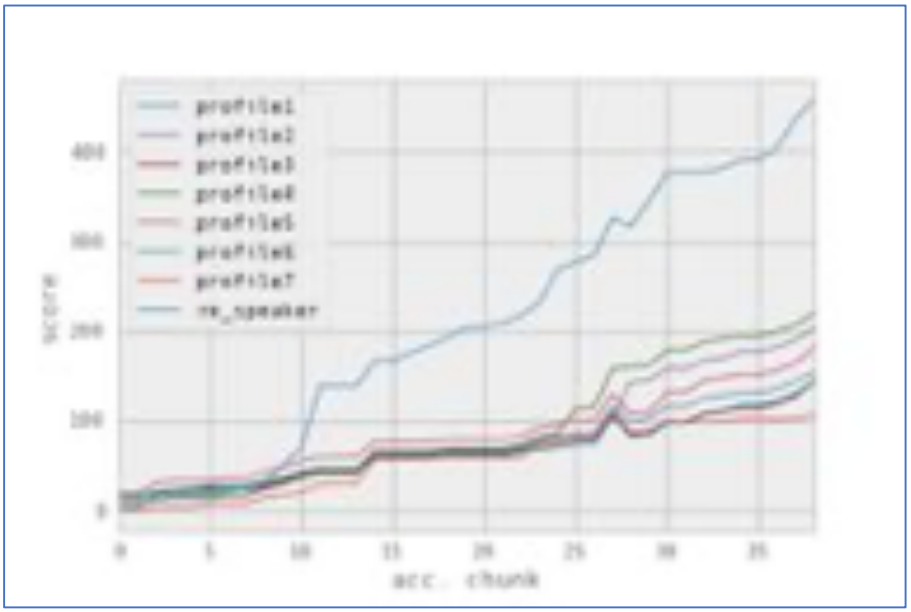


Figure 3: Graph generated from the output score vector for cumulative text (lower lines the greater the accuracy)

These results can then be displayed as 'time-sequence' matches in graph form, the intention being to determine when Re-Spk$_n$ or Pr$_n$ reaches a probability of accuracy which would allow the other profile outputs to be rejected. This implies a calculation of end-point accuracy as well as the intermediate time-sequenced calculations.

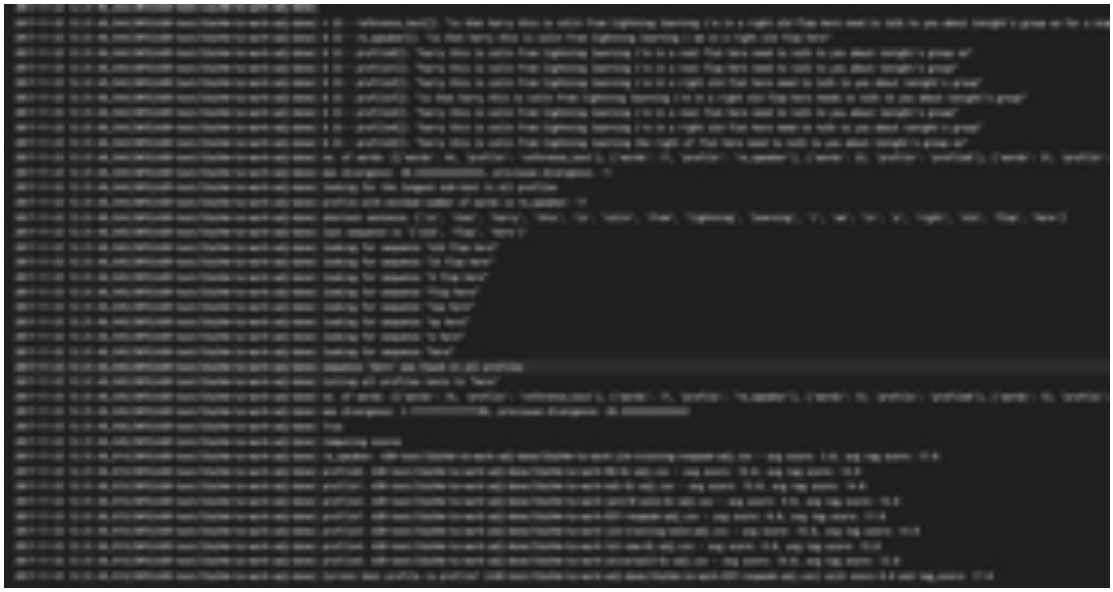
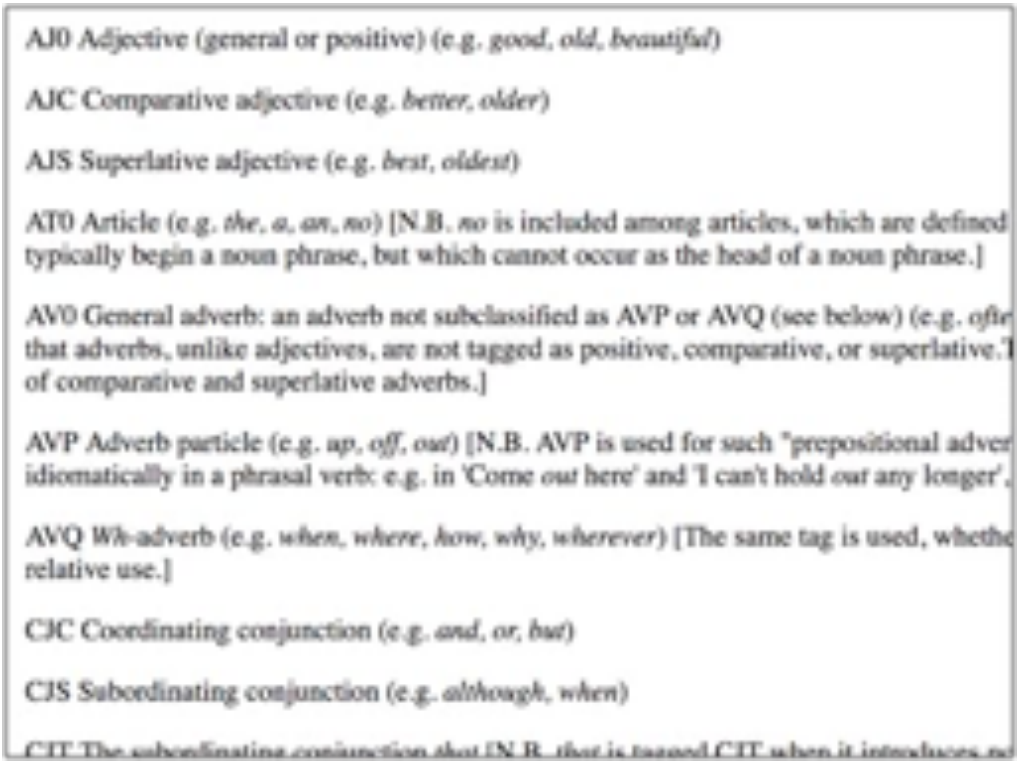


Figure 3: showing how the text similarities and the scores are computed as well as the choice of the best profile

As it is possible to see in Figure 3 there are three visual areas: in the first one the buffered texts for all profiles are shown, in the second (middle) area the texts are checked to find a common sub-text based on the actual words and in the last section the scores are computed and the current best profile is chosen.

A2.4 Weighted Part of Speech (POS-tag) analysis

An approach to evaluate text similarities and correlation different from this word-analysis is to analyse the structure of the phrase through Part Of Speech (POS) tagging. The POS-tagged resulting phrase will then show structures and patterns that can be considered highly correlated to their actual meaning under the strong assumption that although the texts are generated by different ASR engines, they have all started from the same source and therefore have a shared context.


AJ0 Adjective (general or positive) (e.g. *good, old, beautiful*)

AJC Comparative adjective (e.g. *better, older*)

AJS Superlative adjective (e.g. *best, oldest*)

AT0 Article (e.g. *the, a, an, no*) [N.B. *no* is included among articles, which are defined typically begin a noun phrase, but which cannot occur as the head of a noun phrase.]

AV0 General adverb: an adverb not subclassified as AVP or AVQ (see below) (e.g. *ofte* that adverbs, unlike adjectives, are not tagged as positive, comparative, or superlative.] of comparative and superlative adverbs.]

AVP Adverb particle (e.g. *up, off, out*) [N.B. AVP is used for such "prepositional adver idiomatically in a phrasal verb: e.g. in 'Come *out* here' and 'I can't hold *out* any longer',

AVQ *Wh*-adverb (e.g. *when, where, how, why, wherever*) [The same tag is used, whethe relative use.]

CJC Coordinating conjunction (e.g. *and, or, but*)

CJS Subordinating conjunction (e.g. *although, when*)

CJT The subordinating conjunction *that* [N.B. *that* is tagged CJT when it introduces no

Figure 4: an example of Part Of Speech Tags taken from the BNC (British National Corpus) [http://www.natcorp.ox.ac.uk/docs/c5spec.html]

As an example, we can consider the two sentences

The boy goes to school

The boys go to their school

resulting in the following two POS_tagged sequences

*The_**AT0** boy_**NN1** goes_**VVZ** to_SENT school_**NN1***

*The_**AT0** boys_**NN2** go_**VVB** to_PRP their_SENT school_**NN1***

And by the analysis of the POS Tag sequence we can decide that they share a **Common Longest Sequence** of type

AT0_NNx_VVx_NN1

By assigning proper weights to the POS-Tags, it is possible to evaluate the similarity of two sentences based on their own structure; that is, counting less for the less significant parts (like prepositions and pronouns) and more for the predicted important elements (nouns, verbs)

This score was computed both at utterance level and for the cumulative text.

In this way, the program is able to identify in real time the profile that is currently generating the most accurate text.

A second and ongoing type of similarity analysis was then designed to better measure the text similarities by means of the computation of the correlation coefficients for all the possible pairs in the set [reference_text, profile_1, …, profile_n] (where a profile might be a re-speaker).

The program currently runs on a single folder chosen from those used to feed the first program and compares the Weighted-POS tagging score at utterance level as well as the cumulative text and generates for both cases a correlation matrix and a graph for each compared pair …. along with the linear regression line to highlight the level of linear correlation.

As a next step in the data analysis the resulting score on the cumulative text can also be plotted against the number of generated words and clustering techniques (k-Means) can be used to group similar text together isolating potentially wrong texts as outliers.

Appendix 3: Discourse Analysis

Having worked through Character Level analysis and Pos-Tagging, and examined the outputs, it was realised that a potential enhancement could be found by considering both parts of the conversation. This had not been done at that point, as the participants' speech had been recorded on separate files and the task set had been to produce STT output for one side of the conversation. The realisation that responses in speech might be contingent on the other person's utterance led to a shift in emphasis from monologue to dialogue.

A3.1 Introducing Discourse Analysis

In a basic model of discourse analysis (eg Coulthard and Mongomery, 1981) there are three basic elements/moves:

Element		Move
Initiation		eliciting
Response		informing
Follow-up		acknowledging

There are some issues in regard to whether one utterance/response links to the previous exchange and this may be determined by intonation or the complexity of the information to be supplied. As a result, an exchange may continue over repeated eliciting and informing.

Francis and Hunston (1992) set out to analyse a phone call using this form of discourse analysis. They identify four levels: Interactions, Transactions, Moves and Acts. For our purposes the classification at the level of moves and acts is most relevant (in terms of building a model).

Acts combine to form moves.

Moves	**Acts**
Framing	*framer (marking boundaries)*
Opening	*metastatement. conclusion, greeting, summons*
Answering	*acquiesce, reply-greeting, reply-summons, summons*
Eliciting	*inquire, neutral proposal, marked proposal, return, loop, prompt*
Informing	*observation, informative, concur, confirm, qualify, reject*

Acknowledgement	*terminate, receive, react, endorse, protest*
Directing	*directive*
Behaving	*behave*

We do not need to go into the detail of each of these acts here; suffice it to say that they represent the classification of what the participants say during each transaction and then as part of the whole interaction. The analysis here is descriptive but there is a good deal to learn in regard to the predictive nature of the moves in the dialogue.

An example of a section of Francis and Hunston's transcription is shown below in Figure 1.

A:	(la) What do you mean good exercise it	ret	h	eliciting	I^b	Clarify
B:	(la) I mean walking round looking for the fair was exercise	i	h	informing	R	
A:	Yeah my *feet* hurt	end	h	acknowl	F	
	Looking for the what?	ret	h	eliciting	I^b	Clarify
B:	Looking for the fair	i	h	informing	R	
A:	Mm	(eng)				
B:	the trade fair or whatever it was					
A:	Mm	(eng)				
B:	autumn something fair					
A:	Yeah (mid key)	rec	h	acknowl	F	
	We we don't walk enough my feet really hurt (2)	obs	h	informing	I	Inform
B:	Mm (low key)	ter	h	acknowl	R	
	Yeah bit of a let-down	obs	h	informing	I	Inform
A:	Mm (mid key)	rec	h	acknowl	R	
	Still Ben had a nice time	obs	h	informing	I	Inform (incompl
B:	Especially when the Chinese opera turned out to be (#) er a group of Chinese madrigal singers or something	obs	h	informing	I	Inform
A:	If it was	ref	h	acknowl	R	
B:	(la) Wh-whatever	end	h	acknowl	F	

Figure 1: Extract of discourse analysis of telephone call (Francis and Hunston, 1992)

Single lines indicate the boundary of an exchange; with a broken line suggesting an overlap or bound–elicit.

When we apply the analysis we can begin to see the potential for a model of the interaction/transaction (Figure 2).

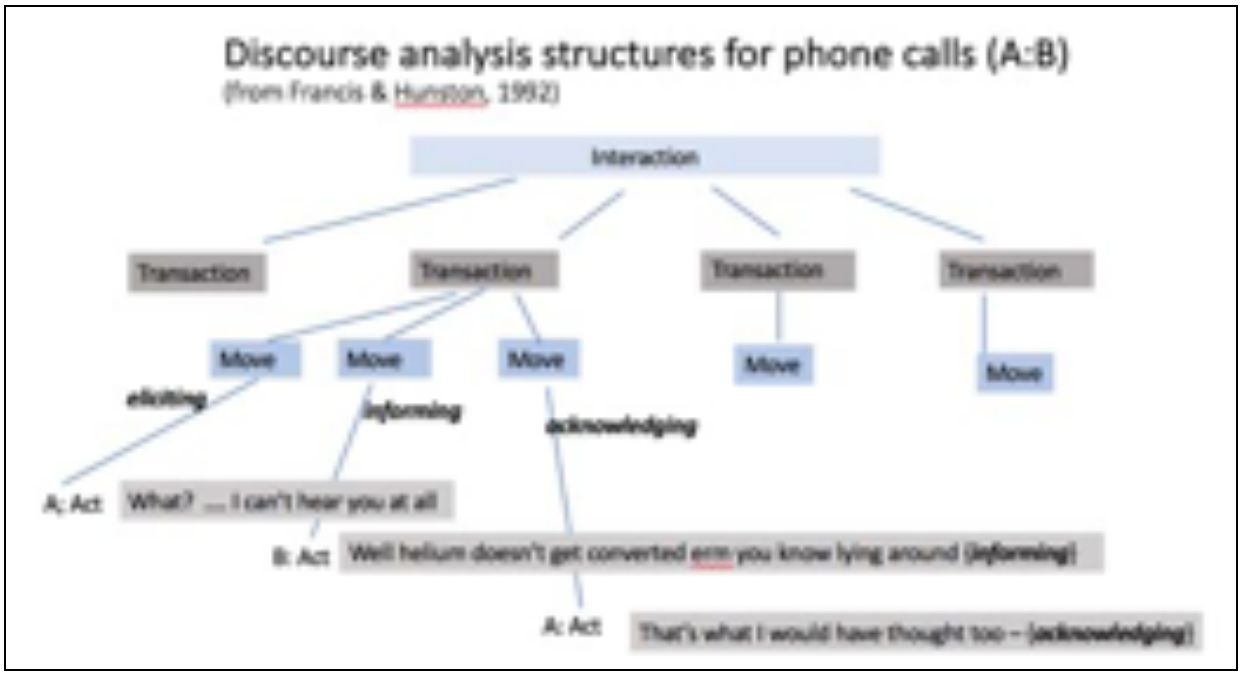


Figure 2: Discourse analysis as a model of phone interaction

Such an approach may well prove to be significant in the next phase of the project development. We can begin by framing our process model taking dialogue (and thereby the linked discourse) into account (Figure 2).

A3.2 Implications for TC-Cap

There is much work to do to fit discourse theory into a programmable approach to our STT issues. However, we can begin to see the psycholinguistic challenges and can consider the issues of timeliness in the interaction.

If we start by a conceptual view of the hearing-to-hearing (or to-deaf) dialogue in a phone call (Figure 3),

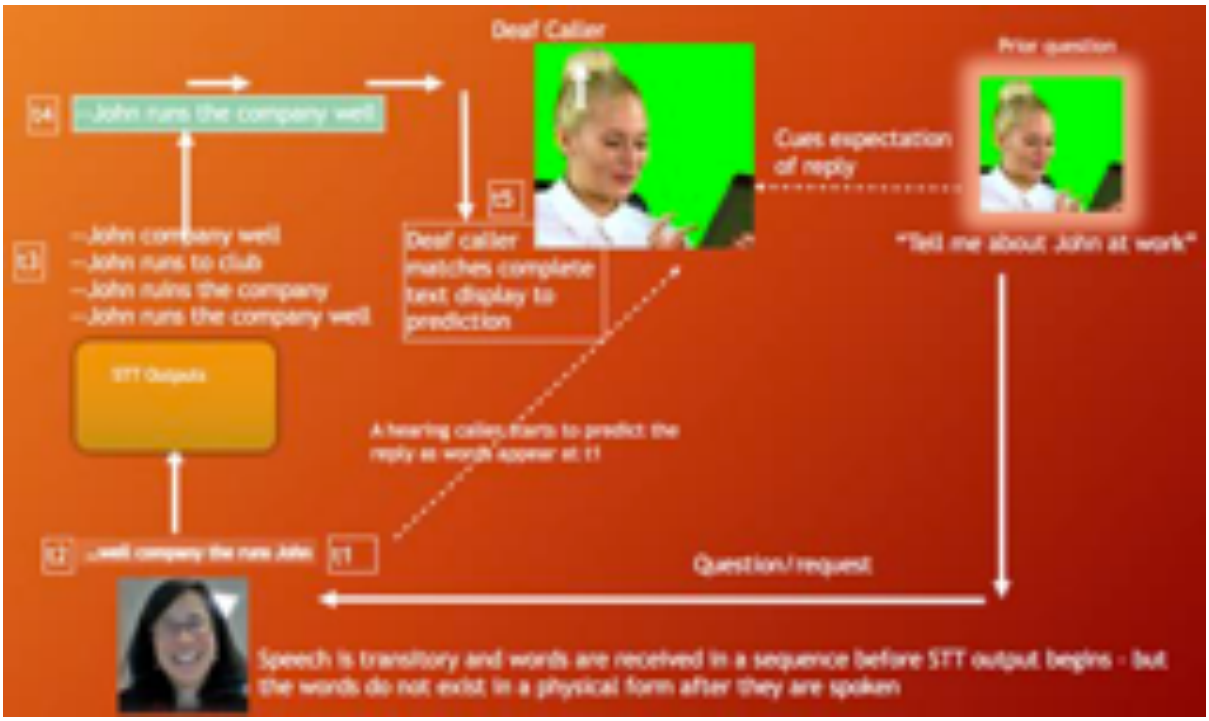


Figure 3: Dialogue: the process starts in the prior question

Several issues become apparent here and distinguish a deaf-hearing call from a hearing-hearing call. As a general principle, the dialogue starts from the prior question/request (top right in Figure 1). This is transmitted to the partner in the conversation and prompts a response. However, the question/request sets up a cue for particular expectations and highlights domains of knowledge in the first person … in preparation for the answer.

At time "t1", the second (hearing) person begins a sentence in response and these words are received in a transitory sequence where no one word exists physically at the time a second word appears. The first (hearing) person has therefore to find a means to identify and store the words in the sequence; at the same, matching these with predicted responses. Clearly, the first (hearing) person is already "understanding the response" by time "t2"; and will be ready to assimilate and respond further.

However, if the first person is deaf, then the message is not perceived until time "t5" after it has gone through the STT process and is chosen as the output and presented as a whole, *to be read*.

The dialogue is now materially different for deaf-hearing interaction from hearing-hearing interaction.

Instead of focusing on the monologue, as had been planned, it becomes apparent that TC-Cap needs to understand the cueing process in dialogue.

A3,3 New Basic principles for the model (for telephone calls)

In the first analysis, we consider the Hearing speaker (HS1) and the Hearing listener (HL1). Participants (HS1 & HL1) construct, perceive[6], deconstruct, create meaning, interact and reverse this process. The message is created in a series of steps (which might occur almost simultaneously) ….

1. HS1 has an intention to communicate
2. the meaning of that intention is built into a coherent set of concepts by HS1
3. an internal (and personalised, idiosyncratic) language model is applied to turn the idea/concept into a systematic sequential output by HS1 – ie a sentence (although often phone calls are not bound in sentences) - better to refer to 'utterance'
4. (personalised, idiosyncratic) articulation occurs … in speech this is always sequential BUT with the significant difference to writing say, that each word disappears as it is spoken and the link to the previous and next "word" are not physically present as the word is being spoken (nor as the word is perceived)
5. The HL1 at no point in time has the whole utterance within his/her auditory reception. Immediate memory holds the sequence of words as long as it is no longer than 7 items; but a separate predictive process is implemented in the receiver ie he/she begins to anticipate and predict what is coming next in the sentence.
6. The formulation of the response begins long before the utterance is wholly processed by HL1, leading to the view that the HL1/HS1 earlier dialogue sets a context and into this, specific concepts are introduced by both of them, but to which HL1 responds.
7. The interaction can often be highly redundant….. ie HL1 can easily predict the ends of utterances:

" Hello Mike, how …..? <are you> "can't …… <complain>

8. And it is often framed to focus attention …

[6] We refer to perception here as the words are not simply 'heard' acoustically but rather heard, matched to knowledge, understood and interpreted in context … ie perceived.

"I am phoning about xxxxx" {ie pay attention to xxxx – possible request for response}
" Did you hear what happened to zzzz" {ie reference to shared knowledge … prompting limited responses as "yes, no, maybe, not sure …"}

9. A further issue is the capacity of HS1 and HS2, jointly and individually, to separate out speech sounds from non-speech sounds and from partial articulations, repetitions, hesitations and inadvertent vocalisations (laughs, coughs, giggles, erm, y'know).
10. Additionally, we recognise that speech has no punctuation symbols. Pauses, slowing of speech, and intonation carry the additional meaning; creating some issues for STT output and automated displays.
11. Our task is now to better understand this dialogue process in order to better model the language which can be used to generate the STT output.
12. This then calls forth better representation of the STT output.

In this way, the rules of dialogue in theory, can be invoked to improve the STT output … or at least to act upon it after it has been displayed.

The major task is how to describe this call to conceptual knowledge. … and then significantly how to build it into an ASR system.

We can distinguish between a hearing English user (HEU), a hard of hearing English user (HoHEU) and a deaf reader (DR) {a person with different, possibly limited, but unknown experience of English}.

We need also to differentiate from the HS1-HL1 calls from partially heard calls and video calls (where non-speech signals of emotion, intonation and non-speech actions (use of props, looking away), may also allow parts of our 12 point model to be implemented.

At the present time, there is no complete statistical/mathematical model of spoken language discourse. As a result, we need to create a natural language model for spoken language interaction – potentially drawing on what we know of discourse analysis. Instead of applying STT profiles to the second person speech, we need also to apply the same to the first person and to use this as a means to narrow down the possible output from the second person.

This sits as a *post-post process*, after (a) the process of STT from multiple profiles and (b) after our post-process of concordance analysis at Character stage and Pos-tag stage which generates a probability calculation (list) for that utterance.

Appendix 4: Telephone dialogue and the TC-Cap system voice

In workpackage 5, lab trials of ASR and STT output were conducted with hard of hearing and deaf users. A software application (APS-50) developed for the REACH112 project (2010 – 2014) was used as the main test application, and an open source smartphone app (Linphone) was also tested. This work is set as a proof of concept and qualitative data was collected as the priority. The trials were conducted in November 2017 and then again in January-February 2018, when the database had been incorporated into the Dragon ASR profiles. This improved accuracy significantly. The earlier trials used scripted calls; the later trials used free interaction on spontaneously generated topics.

Eight hearing, hard of hearing and deaf (sign language users) took part as users while the trained STT profiles were used to generate the text in the other part of the call.

The figure above shows the screen of the deaf user. The sentences spoken are displayed on the left. In this example, the words are spoken by the person in the video screen but this can be done just as easily by feeding the speech from the voice call directly to the TC-Cap system.

The trials were designed to test feasibility. The following results were obtained.

1. Users found it easy to interact with the hearing user

2. Accuracy of the text display was higher in the January-February tests that is, after the database contexts had been imported into the Dragon software.
3. Users spent most time watching the text display, rather than the video.
4. However, they also indicated that a video display would help to know when speech was being delivered.
5. Deaf non-speakers typed their responses to the hearing person and this tended to slow down the interaction as mis-typing occurred and had to be corrected. As the text is presented as Real Time Text, the characters entered and the mistyping, regressions etc were visible to the hearing person.
6. Deaf users pointed out that this could only occur if the hearing person also had the app installed. However, assurances were given that a text-to-speech functionality would be available which would mean that the hearing person would hear what the deaf person had typed.
7. Users preferred that the hearing person did not know that they were talking to a deaf person. The system third person voice should not address the hearing person.
8. With the use of ENUM, deaf user's incoming and outgoing calls would go directly to the TC-Cap system, without any intermediate message to announce that "you are talking to a deaf person".
9. Where the TC-Cap system had not resolved the most accurate output text, the deaf user was to be informed and could then choose to request clarification or repetition.

These points were made/received with satisfaction.

The TC-Cap system is a feasible support to deaf and hard of hearing interaction with hearing callers.

Appendix 5: AI: Natural Language Processing (NLP)

Speech processing has advanced a great deal in recent times as part of huge machine learning financial stimulus, more powerful computers and greater demand for text versions of audio data. Hall and Pesenti (2017) in their report to UK Government set out recommendations for the development of Artificial Intelligence (AI) or machine intelligence applications in the UK. They propose that this has to be high priority for funding, intellectual excellence and sharing of data. This last one being one of the most important aspect – the setting up of Data Trusts and sharing of resources. We will explore this work in greater depth in the TC-Cap exploitation plan.

In the USA, Zilis (in 2014) offered this diagram as a guide for investment and development.

But this has been superseded by a more extensive map of the players in the field (see Harvard Business Review, November 2016, Zilis and Cham)

Notably, those originally involved tended to be large companies but now there is clearly a shift to much smaller enterprises and to the investment opportunities therein.

The stimulus and growth in this area is enormous but requires a different way of looking at the problems.

> "Most IT groups think in terms of applications and data. New machine intelligence IT groups will think about applications, data, and models. Think of software as the combination of code, data, and a model. "Model" here means business rules, like rules for approving loans or adjusting power consumption in data centers. In traditional software, programmers created these rules by hand. Today machine intelligence can use data and new algorithms to generate a model too complex for any human programmer to write.
>
> With traditional software, the model changes only when programmers explicitly rewrite it. With machine intelligence, companies can create models that evolve much more regularly, allowing you to build a lasting advantage that strengthens over time as the model "learns." (Harvard Business Review, November 2016, Zilis and Cham)

TC-Cap in working towards a niche application (deaf telephone calls) can now also draw on this AI mainstream.

A5.1 AI and Language

One of the significant challenges for Artificial Intelligence is to "understand" human speech. The nature of "understanding" here continues to be the subject of much theorising. However, an even greater problem is natural language production, where utterances have to be produced as visual text or as synthesised speech, coherently and cohesively with the prior reception of (natural language) speech/text. At the present time, there is a belief that systems such as Siri, Alexa and so on, are close to mastering this process. However, as even a short period of use will attest, this is not the case and only subsets of the whole human language processing task are being demonstrated.

There are several components to the dialogues humans engage in, which can be described in simple terms:

Speaker: Intention to communicate – coherent meaning constructed internally – preparation of language model – expression using the internal lexicon, presented in standardised morphological and pragmatic form – articulation in sequential form (grammatical output)

Listener: pre-processing by cueing specific areas of knowledge/expectation - perception of first elements of speaker's utterance ie identifying "words" as being present in the lexicon – matching of following elements to the redundancy of standard English – cueing of expected areas of knowledge/experience – decision: interrupt or wait to end of incoming utterance? – listener becomes speaker and reverses the process – adds self-monitoring of own output.

Much of the research in NLP has been devoted to monologue and to the task of dictation or else to the analysis of fixed written texts. Software such as Dragon Dictate v15 with practice can produce high levels of accuracy in dictation, using an acoustic model (to extract the features of sound which correspond to specific words) and a "language model" which, in theory, attempts to learn and then represent the language knowledge of the person dictating. However, Dragon frequently produces utterances which are syntactically incorrect and also semantically incorrect. This can be seen often when different speech profiles are used to recognise speech from a range of different speakers – ie not those on whom the speech profile was trained.

A5.2 Moving from monologue/dictation to dialogue

Much of "language" takes place in interactions between people. As a result, any natural language processing should take into account the nature of dialogue. A key approach is to see the dialogue as an overall transaction. We may then consider dialogues in terms of

An opening transaction: ie finding the right person and introducing participants and the topic;

A: Hello, Rail Enquiries B: Is that Gloucester Riverside Station? A: Yes, Michael, here, how can I help you ?

A **question-answer transaction** – {pre-condition: A has information}

B: when does the train for Swindon leave? A: 4:45pm, platform 5 B: thanks

But in some cases, there is a misconception which needs to be corrected (in this case, a direct answer would be misleading):

B: When does the train for Corsham leave? A: There is no train station in Corsham. Take the 4:30 to Bath and then catch a bus to Corsham. B: thanks

In both cases, "train" and "when" should cue a particular domain of knowledge, with "Swindon" and "Corsham" identifying a specific local area of knowledge. There is also a contextual aspect which implies that what is required is "the next train" information rather than a list of train times.

A **tutorial dialogue transaction** – might occur in a restaurant:

B: what do you recommend today? A: are you looking for a meat course or vegetarian? B: what is good in terms of meat? A: well there is some tender lamb with greengage sauce or maybe the t-bone steak in pepper sauce. B: greengage sounds interesting, are these locally sourced? A: all the dishes are from our farm. B: OK, I will try that, then.

In this dialogue, A does have information but has to work to determine which information is to be supplied appropriately. The conversation has to be guided by providing options to person B. At the same time, "these locally sourced" is ambiguous in referring to greengages or both meat dishes. Finally, A then provides a global response but with the substitution of "dish" for "food".

An alternative dialogue, which works at a higher concept level, involves inferring a **plan of action** on the part of the questioner. Here A makes a leap to infer that B has a plan for lunch and responds to B's additional comments.

B: can you tell me where the Asian Star restaurant is? A: on the High Street, but it isn't open at lunchtime B: is that Indian restaurant open, I forget its name? A: Sri Lankan, I think you mean. It should be open. B: yes, but I am not sure where that is it exactly? A: Martin's Close just off the High Street, 5 minutes walk.

In this case, the inference made is "going to eat now" and "needing directions to". These utterances prompt the use of A's internal scripts about restaurants, local geography and Asian food.

However, this type of analysis comes after the fact; that is, we have already transcribed the two sides of the dialogue into an approved script. It is more challenging to program a computer to manage the interaction – to produce the appropriate responses of person A.

Nevertheless, these type of dialogue "partners" are highly sought after and are beginning to be used in call centres to reduce the need for human interventions. At their simplest level they are governed by decision trees, asking the caller to "choose option x" but more recently there has developed a market in more intelligent systems which can work with the higher level planning as above.

IBM Watson is a well-established system for turning speech into text and for identifying key words which can be used in business decision making and marketing. The system is commercially designed for non-real-time applications, although it claims to work with telephone calls.

A5.3 The need for massive databases

One of the major requirement for the development of NLP is the need for extensive databases. Typically, researchers may use the whole of the Wikipedia text as a database to learn sequences (sentences/utterances) of the English language. This has been used mostly for 'smart-search' operations where a specific query is posed.

A5.4 Deep Learning[7]

This relatively recent approach makes use of very large databases of images or text in order to apply algorithms which 'learn' at different layers of the data. The aim is to make sense of a large arrangement of data (in our case, words in sentences/phrases) by examining the co-occurrences in that data. Not surprisingly, when the database stretches to several million occurrences of words, then the interconnections of these words and the mapping of them turns out to be extremely complex and when it is set as a learning task, it requires a great deal of computer resource. However, even in our case where the data base is (currently) under 150,000 words, the results can be particularly useful in determining what are common and useful sequences of words.

The implication of this mapping of words is that to some extent it reflects the way in which semantic knowledge is stored in the human system. We know that semantically similar words are stored close together (experiments on reaction times to closely co-occurring words and unrelated words, shows much faster response when they are semantically similar). It is also true that this applies in the area of concepts with connections between elements of a concept being much closer together. The principle is then that applying Deep Learning techniques gives us partial insight into the semantics of knowledge. This is a significant step beyond our character level analysis and the part of speech (syntactic) analysis.

A5.5 Word Embeddings and Word2Vec

Natural language is rule-bound and as a result it makes some sense to determine the co-occurrence of words and phrases in the database of all natural use of that language. When there are large numbers of examples of these co-occurrences it is possible to generate probabilities of any one word in the same context being preceded by or followed by a specific other. This might lead to a pairwise or even pairs-of-phrases wise matching/prediction procedure.

---

[7] Lecun Y, Bengio Y and Hinton G (2015) Deep learning, *Nature volume 521,* 436–444 (28 May)

While this gives us sequences of words, it also implies semantic/syntactic relationships between each word assuming the database, is sufficiently extensive.

Feature specification for concepts is an old idea and was at one point thought to be the way in which children built up concepts

Exemplar A: it moves, makes noise (barks), has four legs, has fur, wags tail = DOG and can be distinguished from Exemplar B: it moves, makes noise (miaow), has four legs, has fur, does not wag tail = CAT.

Children's early development of language is full of over-extensions where all four legged animals are "DOG". However, children typically build these distinctions from supervised learning experiences in specific visual and tactile contexts (as well as linguistic) and begin to differentiate through various forms of 'reinforcement' or reward. At that stage of learning, the syntax of utterances is much less significant and adults tend to use both child-centred speech and 'babytalk'. Semantics is built from naming, co-occurrence and reward (interaction). The child "learns how to mean".

The concepts in machine intelligence of word embeddings and word vectors draw upon these notions albeit by starting from the wrong end, that is, by assuming all word knowledge should be present from the start (ie a massive database) instead of imitating the gradual growth from a small set of meanings/words towards self-generated standardised sequences of speech, rewarded by interaction with the caregiver.

However, the latter remains problematic and researchers have chosen to identify frequencies and co-occurrences of words in the existing database and then attempt to represent these as a set of probabilities of features.

So the DOG might be 0.99 (moving), 0.99 makes noise (0.99 barks) 0.80 (four legs), 0.99 (fur) and (0.75) wags tail while CAT is 0.99 (moving), 0.99 makes noise (0.0 barks; 0.9 miaows) 0.80 (four legs), 0.99 (fur) and (0.1) wags tail

The learning has to build word vectors for elements (words) in the database, although it may choose to focus on only a large subset of the whole. The intention is than to be able to generate/match/test for output words as in

> "The TC-Cap project has been ---------- by the Innovate UK programme."
>
> The options then being "funded"; "supported" "criticised" "rejected" and so on …

The objective in training is to maximize the conditional probability of obtaining the actual output word (the focus (ideal) word – "funded") given the context words which precede (Continuous Bag of Words model) or to determine the words which follow (Skip Gram model), by using the weights (or probabilities implicit in the vector).

Word2Vec (https://code.google.com/archive/p/word2vec/) is a tool for extracting/constructing word vectors from a database and Word2phrase can offer a similar analysis for phrases, which provides a stronger simulation of semantic NLP.

Examples of this approach in action are shown in Appendix 6.

Appendix 6: TC Cap Application of word embedding and word vectors for Natural Language Processing (NLP)

Lorenzo Benoni, BBITG

A new audio source (ie an audio phone call which had not already been added to the database) was processed by several ASRs operating different Dragon v15 speech profiles and generated texts which were over 90% accurate but had several differences which might confuse the reader.

The challenge was to automatically choose the output text which is closest to, or exactly what was spoken.

This is a very difficult challenge that can be tackled from different angles and using different techniques: we can use POS Tagging (Appendix 2) to discriminate ill-formed sentences from a grammatical standpoint, we can use concordance and co-correlation methods or clustering to implement a democratic approach and pick the text that the greatest number of ASR outputs agree upon but this approach is not totally resistant to errors as in mispresented words (ie homophones) or mistakenly understood lexical choices.

The AI-based development (see Appendix 5) which we applied, required that in a post-STT processing stage, all the generated text was presented to a bespoke “discriminator” built around language-bound techniques to reduce the number of candidate texts and from there funnel the result into a model that computes a "sentence likelihood" based on

- a cross-check among the candidates (ie the most likely sentence in that very set of sentences)
- the single words present in each sentence to account for words that just do not fit or do not make sense

The first discriminator is based on co-correlation and a grammar likelihood score based on POS-Tagging and the results of that procedure were documented in the main report and in Appendix 1.

The second stage in the flow is statistical in nature and has to depend on the context it is modelled on.

Different approaches were investigated: Recurrent Neural Networks (RNNs) and other statistical models. We chose to implement this using Word Embeddings and the word vectors generated by Word2vec (See Mikolov at al, 2013) applied to our database of transcribed calls.

Word Embeddings represent a major breakthrough in Deep Learning applied to NLP and are now widely used and claim to show an improvement in the field of Machine Translation, Sentiment Analysis and other fields where probabilistic analysis of text is needed in order to create correct output.

We trained our model on a large corpus made of telephone calls, conversations and political debates (taken from Hansard, 148,363 unique words). A total of 300 word vectors were calculated (this being an optimum number for power-efficiency – chosen by a word2vec algorithm). The vectors have been generated using 300 dimensions but modelled around the entire dataset. Needless to say, we might increase the number dimensions but at a cost of a higher computational effort. We could use ALL words even if they appear only once in the database (the current frequency constraint is that a word is kept into account if it appears to be used at least three times).

We fed the resulting candidate sentences into the model to predict the likelihood of the co-occurrence of the word sequences (generated by ASR). Below is shown an example of how the system is able to pick the correct sentence out of a set of sentences generated by the Dragon ASRs of which at least two of them were meaningless. Note that the output is a comment by one speaker, and is not a response to a query, nor is it contingent on the content of the other speaker's utterance. As a result, contingency with the previous utterance is low and effectively unpredictable. Therefore, the ASR output has to be analysed according to its own internal coherence.

Output from 5 speech profiles from dragon dictate v 15:
correct output should be "What a mess!"

s1 = "1 domestic"
s2 = "one mess"
s3 = "one May's" (2 ASR exemplars)
s4 = "what a mess"

```
Determine potential 'sentence' outputs from ASR array;
first determine presence of elements in the database.
sentences = [s1, s2, s3, s4]
print_most_likely_sentence_in (sentences)
s1 did not find a word vector for '1', factor is now 10.0
s1 did not find a word vector for 'domestic', factor is now 100.0
s3 did not find a word vector for "May's", factor is now 10.0
```

If word is not found in the database, in that context, penalise it by a factor of 10

```
sentences: [['1', 'domestic'], ['one', 'mess'], ['one', "May's"], ['what', 'a', 'mess']]
original scores: [  0.      -30.812973   -6.3469734  -6.971523 ]
adjusted scores: [0.0, -30.812973022460938, -63.46973419189453, -6.971522808074951]
adjusted scores: [-1000000000000000, -30.812973022460938, -63.46973419189453, -
6.971522808074951]
```

choose sentence with highest value (expressed as a log likelihood)

from sentences: ['1 domestic', 'one mess', "one May's", 'what a mess']
most likely sentence is: **what a mess**

A second example is designed to differentiate output text where the candidates are possible meaningful utterances, but one is more probable. Four output texts were produced:

0: 'probably my concern is about um bill and Bob',
1: 'oh my concern is about um bill and Bob',
2: 'oh my concern is about um bill and Bob',
3: 'oh for public my concern is about um bill and Bob'

The algorithm applied to our database then produced:

0 did not find a word vector for um factor is now 10.0
0 did not find a word vector for Bob factor is now 100.0
1 did not find a word vector for oh factor is now 10.0
1 did not find a word vector for um factor is now 100.0
1 did not find a word vector for Bob factor is now 1000.0
2 did not find a word vector for Oh factor is now 10.0
2 did not find a word vector for um factor is now 100.0
2 did not find a word vector for Bob factor is now 1000.0
3 did not find a word vector for oh factor is now 10.0
3 did not find a word vector for um factor is now 100.0
3 did not find a word vector for Bob factor is now 1000.0

[['probably', 'my', 'concern', 'is', 'about', 'um', 'bill', 'and', 'Bob'], ['oh', 'my', 'concern', 'is', 'about', 'um', 'bill', 'and', 'Bob'], ['Oh', 'my', 'concern', 'is', 'about', 'um', 'bill', 'and', 'Bob'], ['oh', 'for', 'public', 'my', 'concern', 'is', 'about', 'um', 'bill', 'and', 'Bob']]

[-37.848774 -26.486488 -26.486488 -40.112125]
[-3784.877395629883, -26486.488342285156, -26486.488342285156, -40112.125396728516]
[-3784.877395629883, -26486.488342285156, -26486.488342285156, -40112.125396728516]

['probably my concern is about um bill and Bob', 'oh my concern is about um bill and Bob', 'Oh my concern is about um bill and Bob', 'oh for public my concern is about um bill and Bob']

most likely sentence is: **probably my concern is about um bill and Bob**

By applying the above process, to an increasing database of telephone and speech dialogues, we can 'intercept' errors produced by Dragon ASR output. Since the proposed systems will have many more parallel ASR outputs than the 4 or 5 on which this has been tested, it is likely that at least one will be accepted by the system as effective. However, in the circumstance where none of the outputs can be assessed appropriately by the process, then intervention from the system to both parties to indicate that the previous utterance was not understood. (see *third-party system voice* in section 8.0 above).

A6.1 Next steps

This process is to be refined.

- database is to be expanded
- the pre-training of the database in in terms of word embeddings using Word2Vec and Word2Phrase is to be applied [this process is completed separately and outside of the timeframe of telephone calls]
- the 'penalties' are to be applied more sparingly as the database lexicon is significantly expanded and unidentified words are reduced in number
- word vectors and co-occurrences are to be expressed in terms of the probabilities of the component features (see Appendix 5)
- the output from the ASR application is to be held in a matrix to which the above decision process is to be applied.

Appendix 7: Database management and encoding

Ed Toms, BBITG

A7.1 Timecoding of the database of audio calls

A total of 156 phone calls have been transcribed and timecoded by utterance. Since the recording of each participant was done separately, the number of source time codings is double the number of calls, as there is one for each speaker per call. Average duration of calls was 3.2 minutes, creating a database of nearly 1000 minutes of audio. These are arranged in folders with up to 7 STT output texts from each audio participant in the call.

The initial time codings were adjusted several times to format the text more appropriately for different analyses techniques. This necessitated a number of stages – cleaning up the text to remove non-words, typos and stutters, removing timecodes and generating text–only files.

The resulting database has been stored in folders in a central server for further analysis.

A7.2 Bigger data

Most commercially available language models (i.e. STT software) are very well trained, meaning they have been through an extensive process of analysing written text and learning the patterns of the language, in order to produce accurate text transcriptions. However, this is generally only effective when the speaker is dictating or speaking using a restricted lexicon. This is due to the training process used to create the model; the training texts used are mostly texts by famous authors, transcriptions of film/television scripts or articles from the internet. All three of these (except perhaps films scripts, although still more accurate than conversation) are for the most part perfect English, which is very far from the structure and style of conversational English. Informal English usually contains ‘shortcuts’ in speech, repetitions of words or nonsensical changes in topic.

A7.3 Neural Networks

Artificial neural networks are learning systems consisting of layers of interconnected groups of nodes (artificial neurones) which carry out calculations and pass the outputs to each node in the next layer. They contain an input, output, and one or more ‘hidden’ layers. The

networks are said to be based on biological brains and carry out calculations in the same way that scientists believe animal neurones do. Recurrent neural networks follow this architecture, except the connections between units form a directed cycle, meaning a node can be reached by following the path coming from itself. This allows the outputs to be passed back in the network at all points, constantly re-training and altering the network.

Long short-term memory (LSTM) networks constitute powerful recurrent neural networks, and increasingly can be seen in speech recognition systems. This was the logical choice for training a language model.

Pre-built neural network packages such as TensorFlow and Keras in Python provide a starting point. Four trainable and pre-trained language models available on GitHub were used. Experiments with different network structures, altering the number of nodes, hidden layers and epochs, as well as training on written text and our own corpus, were attempted. However, results were inconclusive, often producing nonsensical sentences or words. On its own, this did not prove to be an effective direction to take.

A7.4 New Databases

Although the project has collected a significant number of telephone calls with a range of speakers and topics, this database is still small in relation to the tasks required of it, and for certain learning tasks there is a bias due to repeats in scripted calls.

It was necessary to locate and to determine the usability of several other database sources. This was simplified to some extent when we removed the need for the audio to consist of phone calls. We were now able to look for natural spoken audio, preferably in dialogue form. The obvious examples of this were TV talk shows and radio shows, as they contain multiple speakers in dialogue, and although the theme is scripted, the actual conversation is not. This means we can develop a language model which is better at generating conversational English thanks to the variety of themes and speakers, who are engaging with one another rather than dictating a script. The changeable nature of talk/debate shows is similar in style to phone calls, so is applicable to phone call modelling as the text is conversational.

Eight sources with usable transcripts were located online. None of them came as a compiled corpus, so each file had to be individually downloaded and then combined into one complete

file. However, having individual files per episode allows learning tasks (such as Dragon training) to be carried out on smaller scales.

Nearly all transcripts from broadcast shows had to be manually downloaded, although in some cases, web scraping could be applied to speed up the process. The following database sources were investigated:

Source	Content	Extent	Comments on use
BBC Andrew Marr Show	Transcripts of The Andrew Marr show, generally consists of two speakers in dialogue, each speaker is labelled when they speak.	149,793 words, 796,691 characters, 15,309 lines. Contains a range of themes (although generally political) and speaker styles.	Contains longer periods of speaking so is harder to use for training a dialogue model, but speech is still conversational so follows the same patterns and flows.
Fight CPS	Transcripts of the Fight CPS radio show. Contains callers speaking to the host about American Child Protection Services, so limited to the legal theme.	180,619 words, 962,293 characters, 2,979 lines. Various speaking styles but topic is very restricted to be legal terminology relating to CPS.	Unused as only contains American text. Could be usable in future for expansion of the project.
BBC Radio 4 In Touch	Radio 4's In Touch transcripts. A show containing news, views and information for blind or partially sighted people.	279,626 words, 1,467,984 characters, 11,865 lines. Wide range of speakers and themes, although all have some relation to blind/partially sighted people.	Speech is in much shorter bursts so is more conversational making it more useful for training. However, requires more processing (removing speaker labels etc.) before use.
James Corden Interviews	A collection of interviews with James Corden. Discusses a variety of themes with some variance in style.	Very limited: 19,300 words, 98,500 characters, 686 lines.	Most likely too short to be of any use, but could be combined with others to make a larger combined corpus.
BBC Radio 4 The Moral Maze	Debating show discussing moral issues behind one of the week's news stories. Contains a variety of different speakers.	66,731 words, 365,498 characters, 1,619 lines. Different theme and speakers per episode creating a diverse corpus.	Useful as there are many question-answer dialogues, speech is in shorter bursts and is unscripted. Requires some pre-processing before use but could still be used, however corpus is relatively small.

Warwick University Lectures	Lectures in four subject areas. Main speaker is the same for each lecture per subject (four main speakers overall). Less conversational but there is still some question-answer dialogue with the audience.	Extensive: 160 lectures containing 1,590,659 words, 8,335,609 characters, 116,165 lines.	The largest corpus of one style, it is of a size usable for training a language model or neural network, although is less conversational. Parts could be combined with another corpus in order to create a varied and extensive training corpus. Could also be used to train language models in specific subject areas.
BBC Radio 4 You & Yours	Radio 4's consumer affairs programme, containing news and discussion. Contains a different presenter each episode as well as different guests	1,731 words, 441,287 characters, 3,125 lines.	Useful for creating a broader, non-specific language model due to the variety of speaker styles and topics. Requires some pre-processing before it can be used.
Hansard – transcripts of all parliamentary business in the UK	Very extensive and up to date speech transcripts	vast	mostly longer turns in the speech and mostly "cleaned up" – removing hesitations, repetitions and non-speech.

A7.5 Further developments for the approach

Upon carrying out more research on the inner workings of ASR systems, several new possibilities emerged.

An n-gram is a contiguous sequence of n items for a given sample of text. In the context of the aims of the project an n-gram approach may allow us to estimate the probability of a word given the last n words in a sequence.

As a starting point, trigrams were used (using the previous two words to estimate the third in the sequence). This is calculated by applying a Markov Assumption to the normal chain rule of probability, the result of which is the probability of the sequence. One may think that the chain rule could just be applied without needing the Markov Assumption, but in practice

there are far too many possible sequences, meaning the calculation gets out of hand very quickly, becoming almost impossible to compute.

It is likely that results would be more accurate without the adjustment, but for now it is simply impossible to be done in that way.

$$\prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

The Chain Rule

$$\prod_i P(w_i \mid w_{i-k} \ldots w_{i-1})$$

Applying the Markov Assumption to the chain rule, adjusting the number of words to be within a
certain amount of one-another (depending on which n-gram is being used).

The application of n-grams could be extremely useful for the project as a system for validating text outputs as a form of post-processing, rather than generating our own outputs which was the initial idea when using LSTMs. When the Dragon profiles produce their text, this can be fed into an n-gram model, and the probability of each sentence occurring will be calculated. The output with the highest probability is then chosen and will be presented to the TC-CAP user, minimising the likelihood of nonsense text generation. Having a larger database is extremely useful for any machine learning tasks, as more training data helps immensely in producing accurate predictions. With n-grams in particular, a large dataset means predictions will not be limited as the system is more likely to have seen a particular sequence of words before. A pre-processing system could be implemented in conjunction with the variety of corpora. By training different models on the themed corpora, we can create a range of specialised language models which can be activated when a specialised topic is detected. For example, if law terminology is detected, a specialised ‘law’ model could be invoked which is trained on law-related speech, for example transcripts from courtrooms. Using pre-processing to narrow down the domain of speech will allow for quicker generation (due to fewer speech profiles being used) and a lower chance of words unknown to the model being said, and a different word being output due to it being the

closest ‘known’ word. Using these two systems on top of the current systems could reduce error and generation times, allowing the conversation to flow closer to real time, and avoiding cross-talk.

Appendix 8: ASR performance after adding dialogue texts to Dragon v15

(timeline in seconds in left column) Green implies missing component; yellow implies section of interest, red implies incorrect text)

AC56F : source	RV55m-J-solo-train	RV55M - MG	RV55M-Jent	RV55M-Lel	RV55M - Universal
0,					
1,					
21,					
22,					
23,	Hello	Hello	Hello	Hello	
24,I is that Harry?					
25,	Yes	Yes	Yes	Yes	Yes
26,Uh this is Karen from lightning learning					
27,					
28,					
29,I'm in a right old flap here need to talk to you about tonight's group					
30,					
31,					
32,					
33,					
34,Well so for a start we've got no Peter					
35,	OK	OK	OK	OK	OK
36,					
37,He's on annual leave and his team manager forgot to tell us					

38,					
39,	Okay good start	Okay good start	Okay good start	Okay good start	Okay good start
40,					
41,					
42,And David is off sick and now we've got a problem with the study centre Clifton Grange					
43,					
44,					
45,	oh what's wrong with Clifton Grange	oh what's wrong with Clifton Grange	oh what's wrong with Clifton Grange	oh what's wrong with Clifton Grange	oh what's wrong
46,					
47,					
48,It seems a client was smoking in in the kitchen					
49,					
50,					
51,Well near the door maybe anyway					
52,	in the kitchen but they all know about not smoking indoors my God it's been years now	in the kitchen but they all know about not smoking indoors my God it's been years now	in the kitchen but they all know about not smoking indoors my God it's been years now	in the kitchen but they all know about not smoking indoors my God it's been years now	in the kitchen but they all know about not smoking indoors my God it's been years now
53,					
54,					
55,					
56,					

57,					
58,It set off the fire alarm and the fire service were automatically called out to the building so the building's closed right now while they check it over					
59,					
60,	what a drama	what a drama	what a drama	what a drama	what a drama
61,					
62,					
63,					
64,					
65,Everyone's been evacuated and all the art stuff has just been left in the group room					
66,					
67,					
68,	so all the stuff is still in the room then	so all the stuff is still in the room then	so all the stuff is still in the room then	so all the stuff is still in the room then	so all the stuff is still in the room then
69,					
70,					
71,					
72,Yes I'm not convinced that when we they all get back in there there's going to be enough time for them to clear the room					
73,					

74,					
75,					
76,					
77,					
78,To tidy away the paints easels and water they usually need half an hour					
79,	1 domestic	one mess	one May's	one May's	what a mess
80,					
81,					
82,Basically we need to either find another venue or cancel tonight's group					
83,					
84,					
85,					
86,And we need another moderator to jump in now at such short notice to be the leader					
87,					
88,					
89,					
90,	you could try Robert	you could try Robert	you could try Robert	you could try Robert	you could try Robert
91,					
92,					
93,Robert Mason?					
94,Tried him on holiday					
95,					

96,					
97,	and Max he's on the harm minimisation conference in Manchester	and Max he's on the harm minimisation conference in Manchester	and Max he's on the harm minimisation conference in Manchester	and Max he's on the harm minimisation conference in Manchester	and Max he's on the harm minimisation conference in Manchester
98,					
99,					
100,Yes he's still out of town					
101,					
102,					
103,	well just thinking	well just thinking	well just thinking	well just thinking	well just thinking
104,					
105,					
106,					
107,					
108,	as long as we've got a paid member of staff in the building for insurance purposes I could always facilitate the group myself with the volunteers	as long as we've got a paid member of staff in the building for insurance purposes I could always facilitate the group myself with the volunteers	as long as we've got a paid member of staff in the building for insurance purposes I could always facilitate the group myself with the volunteers	as long as we've got a paid member of staff in the building for insurance purposes I could always facilitate the group myself with the volunteers	Rosie got paid member of staff in the building for insurance purposes I could always facilitate the group myself with the volunteers
109,					
110,					
111,	I'm fine with that I've done it before	I'm fine with that I've done it before	I'm fine with that I've done it before	I'm fine with that I've done it before	I'm fine before
112,					

113,Yes yes I suppose so yes um					
114,					
115,					
116,I will I will check with James who else will be in the building tonight					
117,					
118,					
119,					
120,But we still need a sign on the door saying group has been moved					
121,					
122,	I can do that	I can do that	I can do that	I can do that	I can do that
123,					
124,					
125,	plus I'll send a text out to everyone and so that participants know	plus I'll send a text out to everyone and so that participants know	plus I'll send a text out to everyone and so that participants know	plus I'll send a text out to everyone and so that participants know	plus I'll send a text out to everyone and so that participants know
126,					
127,					
128,For next week we'll need another room anyway as there are some visitors					
129,					
130,					
131,I'm just checking the Google diary bear with me					

132,					
133,					
134,					
135,Tuesday Wednesday four o'clock looks like the training room at Brunswick Court is free until six					
136,					
137,					
138,	if you need me for next week as well then sure I can finish up by 545 instead of 6	if you need me for next week as well then sure I can finish up by 545 instead of 7	if you need me for next week as well then sure I can finish up by 545 instead of 8	if you need me for next week as well then sure I can finish up by 545 instead of 9	if you electric as well then sure I can finish up by 545 instead of six
139,					
140,					
141,Can you be out by six for both weeks?					
142,					
143,					
144,	and then we can all be cleared up and out just before 6 is that right	and then we can all be cleared up and out just before 6 is that right	and then we can all be cleared up and out just before 6 is that right	and then we can all be cleared up and out just before 6 is that right	and then we can all be cleared up and out just before 6 is that right
145,					
146,					
147,					
148,					

149,					
150,					
151,					
152,					
153,Phew					
154,Okay what a day I'll book this training room now					
155,					
156,	okay great so Brunswick court for tonight and next week too	okay great so Brunswick court for tonight and next week too	okay great so Brunswick court for tonight and next week too	okay great so Brunswick court for tonight and next week too	okay great so Brunswick court for tonight and next week too
157,					
158,					
159,					
160,					
161,Yes that's it have you got the door code?					
162,					
163,	oh no can you send it to me	oh no can you send it to me	oh no can you send it to me	oh no can you send it to me	oh no can definitively
164,					
165,Yes I I'll just tell you now Harry can you put it on into your phone or somewhere safe don't write it down					
166,					
167,					

168,	okay I got my phone	okay I got my phone	okay I got my phone	okay I got my phone	okay got my phone
169,					
170,					
171,					
172,	we'll put it under BC Yep go ahead	we'll put it under BC Yep go ahead	we'll put it under BC Yep go ahead	we'll put it under BC Yep go ahead	we'll put it under BC Yep go ahead
173,					
174,					
175,					
176,Okay so it's two zero zero six enter to go in and two zero zero six exit to leave					
177,					
178,					
179,					
180,					
181,					
182,					
183,					
184,					
185,You won't need the alarm codes as people should still be in the building					
186,					
187,					
188,	okay yes just check with James	okay yes just check with James	okay yes just check with James	okay yes just check with James	okay yes just check with James

189,					
190,					
191,					
192,Yeah yeah will do					
193,	but I'll assume it's all okay for now	but I'll assume it's all okay for now	but I'll assume it's all okay for now	but I'll assume it's all okay for now	but I'll assume it's all okay for now
194,					
195,Okay thanks Harry cheers bye					
196,					
197,	okay thanks	okay thanks	okay thanks	okay thanks	okay thanks
198,	Bye	Bye	Bye	Bye	Bye