Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1978–1984
Marseille, 20-25 June 2022

SNuC: The Sheffield Numbers Spoken Language Corpus

Emma Barker*, Jon Barker*, Robert Gaizauskas*, Ning Ma*, Monica Lestari Paramita†
*Department of Computer Science, †Information School
University of Sheffield
{e.barker, j.p.barker, r.gaizauskas, n.ma, m.paramita}@sheffield.ac.uk

Abstract

We present SNuC, the first published corpus of spoken alphanumeric identifiers of the sort typically used as serial and part numbers in the manufacturing sector. The dataset contains recordings and transcriptions of over 50 native British English speakers, speaking over 13,000 multi-character alphanumeric sequences and totalling almost 20 hours of recorded speech. We describe requirements taken into account in the designing the corpus and the methodology used to construct it. We present summary statistics describing the corpus contents, as well as a preliminary investigation into errors in spoken alphanumeric identifiers. We validate the corpus by showing how it can be used to adapt a deep learning neural network based ASR system, resulting in improved recognition accuracy on the task of spoken alphanumeric identifier recognition. Finally, we discuss further potential uses for the corpus and for the tools developed to construct it.

Keywords: spoken language corpora, corpus creation methodology, spoken alphanumeric identifier recognition

1. Introduction

While large vocabulary automatic speech recognition (ASR) has made tremendous progress in the last few years, there remain many challenges, for example high accuracy recognition of non-standard vocabularies used in specific applications and high accuracy recognition in noisy environments. One application that exhibits these characteristics and is of considerable interest to a range of manufacturing and equipment servicing industries, is the recognition of spoken alphanumeric identifiers, such as *serial* and *part numbers*. Speech recognition is of interest in these areas because workers who are assembling, disassembling or servicing complex machinery, such as jet engines, need their hands and eyes to get on with the job but also need to record or access information related to the components they are working on. Here, very high accuracy recognition of serial or part numbers is critical as these numbers are the key identifiers; furthermore the acoustic environments where these tasks are carried out, e.g. factories or service centres, are frequently noisy or highly reverberant, adding to the recognition challenge.

In current practice workers are often forced to interrupt their work to shift to a keyboard and screen where information is entered or retrieved; or notes are scribbled on paper to be entered later. These are time-consuming and error-prone practices. One solution to identifying components is scanning. In some applications this can work, but in many cases, either due to legacy approaches or because RFID tags are unworkable (e.g., turbine blades in a jet engine), identifiers are stamped in metal and must be read from such. In such cases, spoken language data entry and retrieval is attractive, as it exploits the strengths of the human perceptual system without overly distracting the worker from the task. Additional information, such as the condition of the component or actions taken in servicing can then also be added by voice, without the need to down tools or move to a different location.

Entering alphanumeric identifiers by voice might seem to be a trivial problem. In the simplest case there are only 36 vocabulary items (digits plus letters in the Latin alphabet). However, the problem is far from simple for reasons including the following:

- In the general case, the probability of one word (here the spoken form of a digit, such as *nine*, or of a letter, such as *B*) following another is identical for all words in the vocabulary. This will be true unless there is an organisation-specific grammar for such identifiers (e.g., "all part numbers start with a three letter alphabetic code indicating sub-assembly, followed by a 6 digit numeral, e.g. `FAN457328`"). In the absence of such a grammar, the probability of each possible alphanumeric sequence is equivalent. In the case of a 10 character sequence, say, this would be equal to $1/36^{10}$. I.e. language models play no role at all, making recognition extremely challenging.
- The set of spoken forms of English letters contains many highly confusable words *pee* vs *bee*, *em* vs *en*, *tee* vs *dee*, *see* (c) vs *zee*, which vary by a single phoneme, or even by a single phonetic feature (e.g., presence or absence of voicing within a single phoneme).
- Alphanumeric sequences are commonly spoken using a wide variety of forms. For example double or triple number or letter occurrences may be spoken as, e.g. *double nine six* for *996*, rather than *nine nine six*. *oh* is frequently substituted for 0, as in *six oh four*, for 604, rather than *six zero four*. Longer number sequences may be spoken using multi-digit rather than single digit number names,

e.g. *sixty-four oh two* for *6402* rather than *six four zero two*, or *one thousand and three* for *1003*. Additionally, in some cases phonetically identical sequences can have different meanings: *forty eight* meaning *408* vs *forty-eight* meaning *48*. These are distinguished by suprasegmental prosodic features (e.g. timing and stress patterns spanning multiple syllables) something that ASR acoustic models are notoriously bad at capturing.

- Speakers often make self-corrections when reading long alphanumeric sequences, i.e., cases where the speaker notices that they have made an error while reading and attempts to correct the error 'on-the-fly'. Identifying and interpreting self-corrections requires processing sophisticated prosodic cues (Shattuck-Hufnagel and Cutler, 1999; Cole et al., 2005). This remains a challenge for automatic speech recognition systems which often have weak modelling of suprasegmental features and long-span dependencies.

In order to address this challenge in the context of a collaborative project with a major UK aerospace manufacturer and in the absence of any appropriate existing resources to help us with our work, we decided to construct our own spoken language dataset consisting of multiple individuals speaking alphanumeric identifiers. Our aim was to create a resource that would support ASR system development and evaluation and would also allow us to study both how people actually spoke alphanumeric identifiers and the sorts of reading and speaking errors they made. The result of this effort is SNuC – the Sheffield Numbers (Spoken Language) Corpus. It contains recordings and transcriptions of over 50 native British English speakers, speaking over 13,000 *alphanumeric sentences* (i.e. multi-character alphanumeric sequences of the sort widely used as serial or part numbers in the manufacturing sector), totalling almost 20 hours of recorded speech. To the best of our knowledge this is the first published resource of this kind.

The rest of this paper describes in detail: our design requirements in creating the corpus (§2); the methodology we followed to create the corpus (§3); the contents of the corpus and results of some preliminary analysis of the corpus (§4); uses to which we have put the corpus and related tools, including training and testing an ASR system and comparing spoken and typed entry of alphanumeric identifiers (§5); related work (§6); and, conclusions and possible future work (§7). We are releasing version 1 of the corpus to the research community concurrently with the publication of this paper. We are also happy to release to interested parties the interface code we used to present numbers to speakers and capture their spoken responses – this can be used to extend the corpus to include further examples or other identifier formats, or to build a similar corpus for other languages.

2. Corpus Design Requirements

As noted above, our over-arching aim was to create a resource that would support ASR system training and testing and would allow us to study how people speak alphanumeric identifiers and the sorts of errors they make when speaking them. Our work was done in the context of a collaborative project with an industrial partner who had specific requirements in terms of the sorts of identifiers they wanted to capture. This partner also had experience with a number of off-the-shelf commercial ASR systems. None of them had proved up to the task of recognising these identifiers sufficiently accurately to merit deployment for use by their workforce and our partner wanted proof that an ASR system could in practice be developed to meet their standards. To that end we identified the following design requirements that our spoken language alphanumeric dataset should meet. The dataset should:

1. Contain examples of the three most important alphanumeric identifier types our partner used. There was no pre-specified, closed of list these, as new identifiers were constantly being introduced. Each of these types had a minimal grammar specification and loose length specification (see below for details), but still allowed for very considerable variability.

2. Contain voices representing a wide range of British English regional accents and a mix of ages and genders, reflecting the UK-wide workforce of our partner. While our partner did employ some non-native English-speakers, these were in a minority and drawn from such a wide pool of first languages that it was not deemed feasible to attempt to collect data to cover this space, given other requirements.

3. Be recorded in a noise-free environment, so that noise from different real settings could be mixed in later at different levels to assess its impact on recognition. In a second stage we did record some additional numbers in the intended setting of system use, which we used for held-out testing. This data may be released in a future version of the corpus, subject to approval from our partner.

4. Be recorded making at least some attempt to simulate the real task setting, i.e., one where workers pick up a series of physical components and read identifiers from them. In particular, numbers should be presented one at a time, separated by a random time interval and displayed in different sizes, colours and orientations. This is to try to prevent speakers from falling into what linguists call "list intonation", which is observed when people speak lists but is not the prosody we would expect to see in our target application, since workers would typically be reading just one part or serial number at a time.

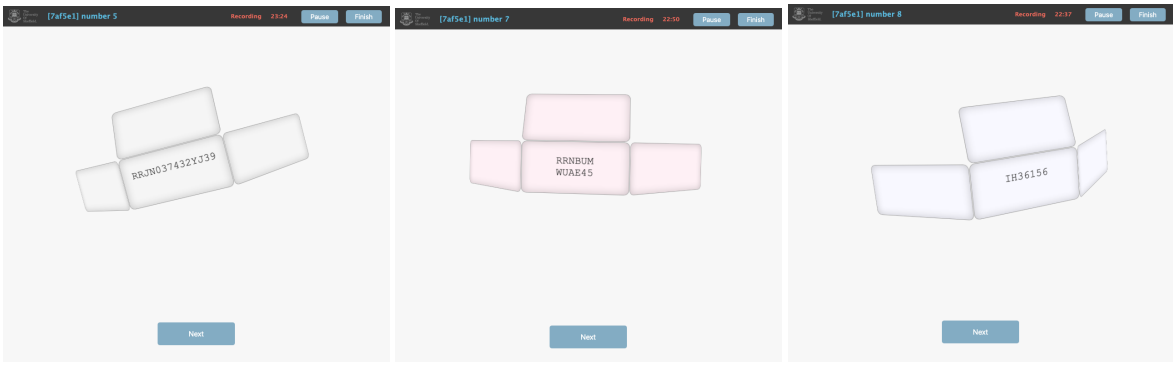

Figure 1: Examples of Numbers Presented for Reading.

5. Be large enough in terms of number of speakers and number of identifiers spoken per speaker to provide sufficient samples for training and testing and to be as representative of the adult British workforce as feasible.

6. Be collectable within the time and budgetary constraints of a 7 month short project, of which corpus collection was merely one part.

The next section details how we created a corpus to meet these requirements.

3. Corpus Creation Methodology

Building the corpus meant carrying out a number of activities. Chief amongst these were: (1) building a data capture setup that would display numbers to participant speakers, subject to the requirements mentioned in Section 2, and record their spoken output; (2) recruiting a sizeable and representative set of native speakers willing to take part in speaking numbers; (3) post-processing the recordings to tidy them up and to provide transcriptions (in general one cannot assume that a speaker has correctly read and said the number that was presented, so a transcription of what was actually said is needed).

3.1. Data Capture Setup

Participants in the number reading exercise were seated in a sound-attenuating booth in our Speech and Hearing Research Group lab, especially designed for recording human speakers for purposes of speech research. Speakers wore a Sennheiser ME 3-II headset microphone and their speech data was recorded via a Zoom H1n handheld digital audio recorder connected to a PC. Recording gain was adjusted carefully to avoid clipping. Participants were positioned in front of a screen and keyboard (earlier stages of preparing the participants for the exercise are described in the next section). They first completed a metadata page which asked for: (1) their age range in one of 6 roughly ten year age range bands from 18 to over 70; (2) their gender; (3) a self description of their accent region. When they were ready to begin recording they clicked on a "Start" button and were then taken to a page on which a number would appear (see Figure 1 for examples). At this point they simply needed to read the number and then either click on the "Next" button or press right arrow or the space bar in order to progress to the next number. Recording begins when the "Start" button is clicked and a new .wav file is written each time the participant moves to start a new number as signalled by clicking the "Next" button or pressing right arrow or the space bar. No visual feedback was provided; i.e., participants were not shown ASR results as they were speaking, nor even a graphical display of audio input level.

Participants were asked to record for 30 minutes and a timer was displayed in the top right corner of the screen, counting down to 0. At any time participants could pause the process to take a break, by clicking a "Pause" button or complete the process by clicking the "Finish" button. A message suggesting they take a break was shown every five minutes, which participants could either heed or ignore.

As shown in Figure 1, identifiers of varying length are presented, in different, randomly selected orientations and colours and perhaps split across multiple lines (addressing requirement 4 in the previous section). There was a random delay in showing the next number when requested, ranging from 0.5 – 3 seconds.

The numbers the participants were shown were one of the three types mentioned above:

1. ID Type 1: two letters followed by 5 digits, e.g., "AX73123" or "XH95372". In some cases, these numbers may also contain an additional suffix "P01", e.g., "EE29425P01".

2. ID Type 2: this number type contains a fixed four-character prefix ("CAT2"), followed by a space and another 9 digits, e.g., "CAT2 614422266". Some numbers may contain an additional suffix "-A" at the end, e.g., "CAT2 346472867-A".

3. ID Type 3: this type always starts with a two-letter prefix ("RR"), followed by a combi-

	Total	Mean	Median	Min	Max
Participants	52				
Male	19				
Female	33				
Ids Spoken (/speaker)	13125	252.4	251.5	193	499
Duration (/speaker)	19.8 (h)	1373.2	1438.2	779.6	2230.3
Duration (/identifier)	19.8 (h)	5.4	5.4	3.5	9.0

Table 1: Summary statistics for the SNuC Corpus. All times are in seconds unless marked (h) (hours).

nation of letters and digits of various lengths (10–12 characters), e.g. "RRISJFGR8850SY", "RRPT7704359267". These numbers may be displayed in a single line or broken into two lines.

The identifiers were randomly generated, subject to the constraints listed above and uniformly distributed across these three types.

The numbers displayed to a participant are recorded and output in a single JSON file which contains a participant id, the metadata captured at the start of the session and for each identifier presented, the identifier itself, the display form used (e.g. whether or not a line break was included and where), which of the three types it was, and the time at which it was displayed. This time is subsequently used to align the number displayed with the audio recording of the speaker reading the number.

3.2. Participants

Ethical approval to use human participants to gather spoken data using the setup described above was sought and obtained via the University of Sheffield's Research Ethics Review procedures (application 031449). An email was sent to all staff, both academic and non-academic, at the University inviting them to participate (around 8,500 persons). Students were not invited as we did not want the age balance of the corpus to be skewed towards younger voices, which would have been the case if they were included. University staff are quite representative of the age ranges found in a typical UK workforce. We accepted all those who agreed to participate, without further action to try balance the corpus by gender or regional accent, as time constraints on the project did not afford us this luxury.

When participants arrived to be recorded, our project and the data capture setup were explained to them and they were shown the sound booth and how to use the display software and microphone headset. They were told to speak the identifiers naturally and neither to try to avoid any particular spoken forms (e.g. *double three* for *33* or multi-digit number names, such as *seventy two* for *72*), nor to strain to include them. They were given the opportunity to ask questions and a chance to have a practice session with the data capture setup. Informed consent was obtained from all participants, whereby they agreed to participate in the exercise and for their data to be held anonymously and redistributed for research purposes.

3.3. Post-processing and Transcription

Each audio file (one file per spoken identifier) was post-processed in four stages:

1. *Signal Processing* The audio file was down-sampled from 44.1 kHz to 16 kHz, which is the standard used for ASR and makes the files smaller and easier to handle.

2. *End Pointing* Silence and noise at the beginning and end of each spoken number was removed automatically. This included false starts, throat clearing, etc. Speech regions were detected by identifying boundaries where the energy averaged across the spectrum was above an ad-hoc threshold. The end-points were then manually checked and corrected.

3. *Transcription generation* A baseline ASR system (see section 5 below) was used to automatically generate an initial transcription.

4. *Transcription correction* A human listener listened to each number where the automatic transcription differed from the prompt presented to the speaker, and adjusted the automatic transcription generated in the previous step as necessary. A set of transcription guidelines were developed and used by the set of five transcribers who did the work. These included prescriptions on how to transcribe spoken digits, non-alpha-digit words and common filler words as well as how to annotate restarts, clipped segments, etc. The full guidelines will be released with the corpus.

4. Description and Analysis of the Corpus Contents

Here we report summary statistics about the corpus and some preliminary investigations we have carried out into the frequency and types of spoken errors.

4.1. Summary Statistics

Summary statistics for the corpus are shown in Table 1. As is evident from the table, the corpus contains almost 20 hours of spoken language, spoken by 52 people (19 male and 33 female), for a total of 13125

ASR System	Baseline Test: SNuC	Baseline Test: Partner Data	SNuC-Adapted Test: Partner Data
Word-Level Accuracy	98.4%	96.6%	98.5%
Sentence-Level Accuracy	81.7%	77%	91.7%

Table 2: Using SNuC to Adapt to an ASR System for Spoken Identifier Recognition

spoken alpha-numeric identifiers. Each speaker spoke on average ∼250 identifiers, speaking on average for around 23 minutes and taking on average 5.4 seconds per identifier (after end-pointing).

The age distribution of the participants in shown in Figure 2.

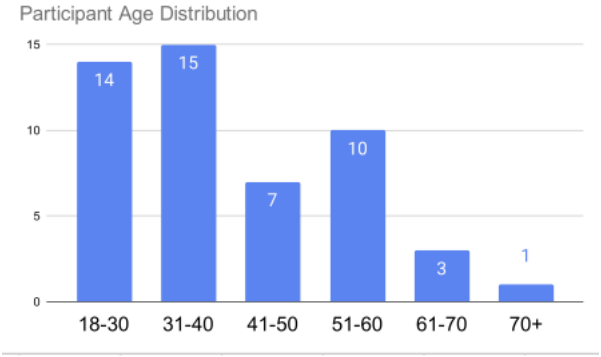

Figure 2: Age distribution of speakers in SNuC

4.2. Errors and Self-Correction in Spoken Alphanumeric Identifiers

SNuC includes examples where participants repaired their speech, in some cases starting again. There are also examples where the spoken utterance did not match the prompt. In a preliminary analysis of the data we distinguished between:

1. *Noticed errors*: i.e. cases where an explicit cue (e.g. cough, pause, words like *sorry* or *oops*, intonational shift) is followed by a repeat, correct reading of part or all of the prompt sequence, e.g. to the prompt `LP58644P02` the speaker says: *L P FIFTY EIGHT SIX FOUR FOUR NINE OH SORRY [noise] [restart] L P FIFTY EIGHT SIX FOUR FOUR P ZERO TWO*; and

2. *Unnoticed errors*: i.e. those cases where what was spoken did not match the prompt and there was either no or a faulty attempt at repair. These consisted of a mix of missed characters (deletions), added characters (insertions), one character substituted for another (substitutions – most commonly "1" for "I" or vice versa) and permutations, a special case of substitution, where a sequence of characters was reversed or reordered, e.g. "KF" for "FK" or "010" replacing "101".

Preliminary automatic analysis of the data has shown that 93.8% of the numbers were correctly read and spoken by participants without mishap. Of the remaining 6.2%, 31.8%, or roughly 1/3, were errors noticed by the speaker and corrected in the course of speaking. The remaining 68.2%, or roughly 2/3, of the errors were unnoticed by the speaker. A full quantitative analysis of the distribution of these error types and their correlation with, e.g., individual speakers, identifier types and identifier lengths remains to be carried out.

5. Example Use of the Corpus

The primary use we have made of SNuC to date is for adapting and testing an ASR system.

We trained a baseline ASR system based on Kaldi (Povey et al., 2011) with the NNet3 deep neural network setup using Librispeech (Panayotov et al., 2015). The baseline ASR system also adopted a language model trained using randomly generated numbers following the loose grammar specification used by the SNuC corpus. This system was used as the baseline ASR system for generating initial transcription, as discussed in Section 3.3 above. We then used the SNuC corpus to develop a second ASR system by adapting the baseline system. In the SNuC-adapted system, the baseline acoustic model was adapted using the SNuC speech data based on the transfer learning approach (Wang and Zheng, 2015). The language model was refined using transcriptions from the SNuC corpus.

We evaluated the resulting system using the held-out data recorded at our partner's site using the same data capture protocol as was used for creating SNuC (section 3.1), but with speakers recruited from our partner's workforce and the recording taking place in the real task setting. This dataset contains around 2500 spoken identifiers, recorded from 16 speakers. Table 2 shows the results of the evaluation. In the table, word-level accuracy refers to the accuracy of recognising individual spoken letters or digits; sentence-level accuracy refers to the accuracy of recognising a sequence of such "words" that comprises a complete identifier (these range in length from about 6 to 14 "words").

The results show that the baseline system trained on LibriSpeech can achieve reasonable word-level accuracy on both noise-free data (SNuC) (98.4%) and on real-world data (96.6%). However, accuracy at the sentence-level is considerably lower both on noise-free data (81.7%) and particularly on real-world data (77%).

This last figure amounts to almost 1 in 4 spoken identifiers containing a recognition error, which is below the level of acceptability for use in a real application (which minimally should be less than 1 in 10).

Using SNuC to adapt the baseline system, while showing only minor improvement in word-level accuracy, shows a very significant increase in sentence-level accuracy on the real-world data. This validates our hypothesis that adapting an ASR system trained on a general corpus by using data specific to the challenge of spoken identifier recognition can provide substantial performance improvements – in this case making the difference between a technology that reaches the level required for deployment in a real application and one that does not.

There is a notable difference in the pattern of errors when comparing the baseline tested on SNuC (column 1) and when testing on the partner data using SNuC adaptation (column 3). The word-level accuracy is the same for both, but in the latter case the sentence-level accuracy is much higher (91.7% vs 81.7%). The lower sentence level accuracy matches that which would be expected in the conditions where word errors are occurring independently within every sentence at the word-level error rate. The higher sentence level accuracy in the partner data is explained by word-errors clustering in certain utterances. This could be due to the greater variability of the partner data with some speakers being better matched to the SNuC adaptation data than others (e.g., either due to accent or to noise background), an hypothesis that merits further investigation.

6. Related Work

To the best of our knowledge SNuC is the first published corpus of variable length spoken alphanumeric identifiers of the sort typically used in part and serial numbers in manufacturing (and in many other areas such as train, airplane, cinema or theatre booking references).

The ISOLET dataset (Cole, R.A. et al., 2008b) contains recordings of 150 subjects of mixed ages, balanced by gender, speaking the name of each letter of the alphabet twice. This is not sufficient for training systems for alphanumeric identifier recognition, both because the data consists only of letters and because the letters are read as isolated samples, not as part of a spoken sequence.

The Free Spoken Digit Dataset (FSDD, 2018) is a growing resource, which at this point contains 3,000 recordings of 6 speakers speaking each digit 50 times, using English pronunciations. As with the ISOLET dataset, this dataset is of limited value because it consists only of digits and because the digits are read as isolated samples, not as part of a spoken sequence of mixed letters and digits.

TIDIGITS (Leonard, R.G. and Doddington, G., 1993) is a corpus of spoken connected digit sequences. It contains recordings of over 300 speakers of mixed gender and ages each speaking 77 digit sequences. This addresses the problem of isolated digits; however, it is still a resource consisting solely of digit sequences and not of mixed alphanumerics.

CLSU Numbers (Cole, R.A et al., 2009) is "a collection of naturally produced numbers taken from utterances in various CSLU telephone speech data collections". It contains a mix of isolated digits, continuous digit strings, and ordinal/cardinal numbers and is drawn from sources including phone numbers, numbers from street addresses and zip codes. It was gathered from $\sim$13K speakers speaking $\sim$ 24K utterances, in most cases in response to a request for a caller's phone number, birthdate or zip code. This corpus has the advantage of being naturalistic speech, but again contains only spoken numbers and not mixed sequences of alphanumerics.

The only corpus we are aware of that contains mixed alphanumerics is CSLU Alphadigit (Cole, R.A. et al., 2008a). Alphadigit "is a collection of 78,044 utterances from 3,025 speakers saying six-digit strings of letters and digits over the telephone for a total of approximately 82 hours of speech." Like SNuC this corpus does contain a mixture of letters and digits. However, it differs in several important respects. First it is speech recorded over a telephone line sampled at 8khz, while SNuC is recorded using a headset microphone in a sound attenuating booth at 44kHz, down-sampled to 16kHz, making it more suited to developing applications where noise from real work environments needs to be taken into account, since this can be mixed in as as required. Second, the Alphadigits numbers are all exactly 6 tokens, while SNuC identifiers are of three sorts, each of different forms and of variable lengths, making them much more realistic exemplars of serial or part numbers as found in real world applications. Third, the digits in Alphadigits are spoken strictly as digits, so the 6 token prompts are already read as 6 spoken words from a 10+26 word vocabulary. In contrast, SNuC allows talkers to read the sequences in whatever way feels most natural, e.g., "44" can be read as "forty four" or "double four", etc. This significantly increases the complexity of the task.

7. Conclusions and Future Work

We have presented SNuC, the first published corpus of spoken alphanumeric identifiers of the sort typically used as serial and part numbers in manufacturing. We believe this corpus will be of use to those working to develop voice-based applications in manufacturing, particularly in areas such as maintenance, repair and overhaul, where scrupulous attention to detail in recording which parts have been replaced or serviced and when, is often critical, and where voice-based applications have real potential, as they can support manual workers who need to keep their hands and eyes free for their work while also needing to record or access data keyed by identifier.

SNuC contains recordings of over 13,000 spoken identifiers, with voices from a range of British regional accents and ages and both genders. Also included are manual transcriptions of the data. We have validated the utility of the corpus by using it to adapt a deep neural network based ASR system and showing that the adapted system demonstrates significantly better performance than a system trained on a general spoken language corpus at the task of correctly recognising full spoken identifiers in real world data. We anticipate that a SNuC-adapted ASR system will also demonstrate improved recognition of alphanumeric identifiers of types other than those in the corpus, such as post codes, telephone numbers or ISBN numbers. However, further testing is needed to establish this definitively, something we intend to carry out as future work.

Aside from its utility in improving ASR performance for the task of alphanumeric identifier recognition, SNuC and the tools we have created to construct it make possible a variety of further studies and related work. In particular they facilitate:

1. Studies of the types and frequencies of errors person make in reading and speaking alphanumeric identifiers. We have begun to look at these (section 4.2), but much more could be done here, including studying the distribution of error types and their correlation with individual speakers, identifier types and identifier lengths. Such studies could help inform the design of alphanumeric identifier schemes as well as the design of voice or multimodal interfaces to support user identification and correction of errors.

2. Studies of variation in spoken forms in alphanumeric identifiers. What are the types and extent of variation in spoken forms of alphanumeric identifiers, such as those mentioned in section 1? (e.g. use of *double* or *triple*, replacement of *zero* by *oh*, use of multi-digit number names, and so on). Understanding this can inform the design of ASR systems (e.g. in language modelling) for the task of alphanumeric identifier recognition.

3. Studies of typing versus spoken data entry of alphanumeric identifiers. We have done preliminary work on using an adapted version of the data capture interface to allow direct comparison between typing a set of randomly presented alphanumeric identifiers versus speaking these identifiers into an ASR system. This setup allows us to test hypotheses about whether humans can enter alphanumeric identifiers faster or more accurately using a keyboard or by voice, giving insight into the strengths and weaknesses of these two modalities for this type of data entry and informing future system design.

SNuC is freely available at `https://doi.org/10.15131/shef.data.19673772` under the CC BY-NC 4.0 licence.

8. Acknowledgements

The authors would like to acknowledge support from the University of Sheffield Impact, Innovation and Knowledge Exchange (IIKE) fund and the Research England-funded PitchIn project [1].

9. Bibliographical References

Cole, J. S., Hasegawa-Johnson, M. A., Shih, C., Kim, H., Lee, E.-K., Lu, H. Y. D., Mo, Y., and Yoon, T. (2005). Prosodic parallelism as a cue to repetition and error correction disfluency. In *Proc. Disfluency in Spontaneous Speech (DiSS 2005)*, pages 53–58.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December. IEEE Catalog No.: CFP11SRW-USB.

Shattuck-Hufnagel, S. and Cutler, A. (1999). The prosody of speech error corrections revisited. In *Proceedings of the Fourteenth International Congress of Phonetic Sciences: Vol. 2*, pages 1483–1486, December.

Wang, D. and Zheng, T. F. (2015). Transfer learning for speech and language processing. In *Proc. of APSIPA*.

10. Language Resource References

Cole, R.A. et al. (2008a). *CSLU: Alphadigit Version 1.3*. distributed via LDC: LDC Catalog No.: LDC2008S06, ISLRN 569-415-930-320-1.

Cole, R.A. et al. (2008b). *CSLU: ISOLET Spoken Letter Database Version 1.3*. distributed via LDC: LDC Catalog No.: LDC2008S07, ISLRN 707-184-716-094-7.

Cole, R.A et al. (2009). *CSLU: Numbers Version 1.3*. distributed via LDC: LDC Catalog No.: LDC2009S01, ISLRN 144-817-035-468-1.

FSDD. (2018). *Free Spoken Digit Dataset*. available at: `https://zenodo.org/badge/latestdoi/61622039`.

Leonard, R.G. and Doddington, G. (1993). *TIDIGITS*. distributed via LDC: LDC Catalog No.: LDC93S10, ISLRN 177-353-807-744-3.

[1] `https://pitch-in.ac.uk`.