SNuC: The Sheffield Numbers Spoken Language Corpus

Emma Barker*, Jon Barker*, Robert Gaizauskas*, Ning Ma*, Monica Lestari Paramita†

*Department of Computer Science, †Information School
The University of Sheffield
{e.barker, j.p.barker, r.gaizauskas, n.ma, m.paramita}@sheffield.ac.uk

Introduction

Motivation

- Spoken language (SL) data entry & retrieval attractive in task settings where hands & eyes needed
- In many such settings items must be identified via unique **part**, **serial** or other **alphanumeric identifiers**
- Thus, high accuracy recognition of spoken alphanumeric identifiers critical for SL applications to be feasible
- Our work carried out in collaboration with a major UK aerospace manufacturer with precisely this problem.

Challenges

- Vocabulary is very simple but **no language model**.
- Spoken forms of English letters contain **many highly confusable words** – *pee* vs *bee*, *em* vs *en*, *tee* vs *dee*
- **Alphanumeric sequences commonly spoken using a wide variety of forms** – *double nine six* for *996*, *oh* for 0, *sixty-four oh two* for *6402*
- Speakers often **self-correct** when reading long alphanumeric sequences
- Environments where task carried out frequently noisy/highly reverberant.

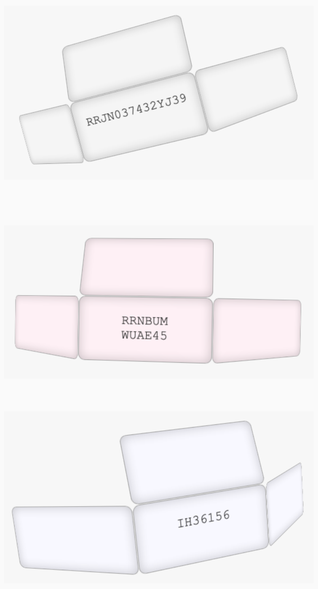
Previous Work

- ISOLET dataset: isolated letters only
- Free Spoken Digit Dataset: isolated digits only
- TIDIGITS & CLSU Numbers: digit sequences only
- CSLU Alphadigit: fixed length (6) alphanumeric sequences; telephone speech; digits spoken individually

⇒ **No existing corpus of naturally spoken,variable length alphanumeric identifiers**

Corpus Requirements

- Contain examples of the three most important alphanumeric identifier types our partner used
- Contain voices representing a wide range of British English regional accents and a mix of ages and genders
- Be recorded in a noise-free environment, with noise from real settings to be mixed in later
- Be recorded simulating the real task setting, where workers pick up physical components and read identifiers from them
- Be large enough to support training and testing of ASR systems and to be broadly representative of the British workforce
- Be collectable within the time and budgetary constraints of a 7 month short project with additional objectives



Corpus Creation Methodology

Data Capture Setup

- Participants recorded in sound-attenuating booth using Sennheiser ME 3-II headset/mic
- Numbers presented on screen, participants read the number aloud then click "Next" to progress to the next number.
- Participants asked to record for 30 minutes
- To simulate task setting, identifiers presented in randomly selected orientations/colours, perhaps split across multiple lines, with random delay in showing the next number
- For each participant, identifiers displayed + time of display are recorded and later time-aligned with the audio recordings

Participants

- Ethical approval for data collection obtained in accordance with University procedures
- 52 native English speakers recruited from University staff
- Informed consent from each participant
- Metadata collected for each participant: (1) age (18-70+) (2) gender (3) UK accent region

Post-processing and Transcription

- Audio file down-sampled from 44.1 kHz to 16 kHz
- Silence & noise at the beginning & end of each spoken number removed automatically (false starts, throat clearing, etc.) and end-points then manually checked & corrected.
- Baseline ASR system used to generate initial transcription.
- For each identifier where automatic transcription differed from the prompt shown to the speaker, human listener adjusted the transcription according to transcription guidelines.

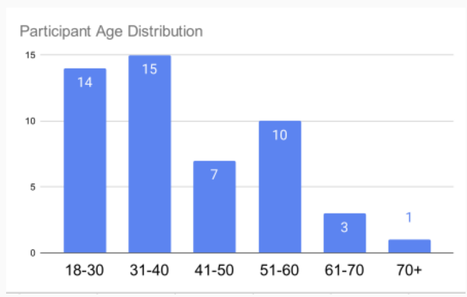
The Corpus

- Contains almost 20 hours of spoken language, spoken by 52 people (19 male/33 female), for a total of 13125 spoken alpha-numeric identifiers.
- Each speaker speaks on average ∼250 identifiers, speaking on average for around 23 minutes and taking on average 5.4 seconds per identifier.
- Errors were either **noticed** or **unnoticed**.
- Preliminary analysis shows 93.8% of numbers correctly spoken
- Of the remainder ∼1/3 were noticed errors & ∼2/3 unnoticed errors.

Summary Statistics

	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

Speaker Age Distribution



ASR Results

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%

Using the Corpus to Train an ASR System

- Trained a Kaldi-based baseline ASR system with NNet3 deep neural network setup + Librispeech + language model based on loose grammar specification of our partner's identifier types
- Trained a 2nd ASR system by (1) adapting the baseline system acoustic model using SNuC speech data + a transfer learning approach (2) refining the language model using SNuC transcriptions.
- Both systems tested on held out data recorded by speakers from our partner site.
- Results show ∼64% error reduction by SNUC-adapted system on full alphanumeric identifiers

Conclusions and Future Work

- First corpus of spoken alphanumeric identifiers, such as serial/part numbers
- > 13,000 spoken identifiers with range of British regional accents/ages/genders + transcriptions of the data.
- Demonstrated improved ASR performance of SNuC-adapted system.
- Future work includes: (1) testing performance of SNuC-adapted ASR system on other forms of alphanumeric identifiers (2) analysis of SNuC to investigate types of errors, variation in spoken forms of alphanumerics (3) comparison of spoken vs typed entry of alphanumerics

Acknowledgements

Supported by University of Sheffield Impact, Innovation and Knowledge Exchange (IIKE) fund and the Research England-funded PitchIn project: `https://pitch-in.ac.uk`.

The University Of Sheffield.

UKRI Research England

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