# A Low-Resource Pipeline for Text-to-Speech from Found Data with Application to Scottish Gaelic

Dan Wells,[1] Korin Richmond,[1] William Lamb[2]

[1]The Centre for Speech Technology Research, [2]Celtic and Scottish Studies
University of Edinburgh

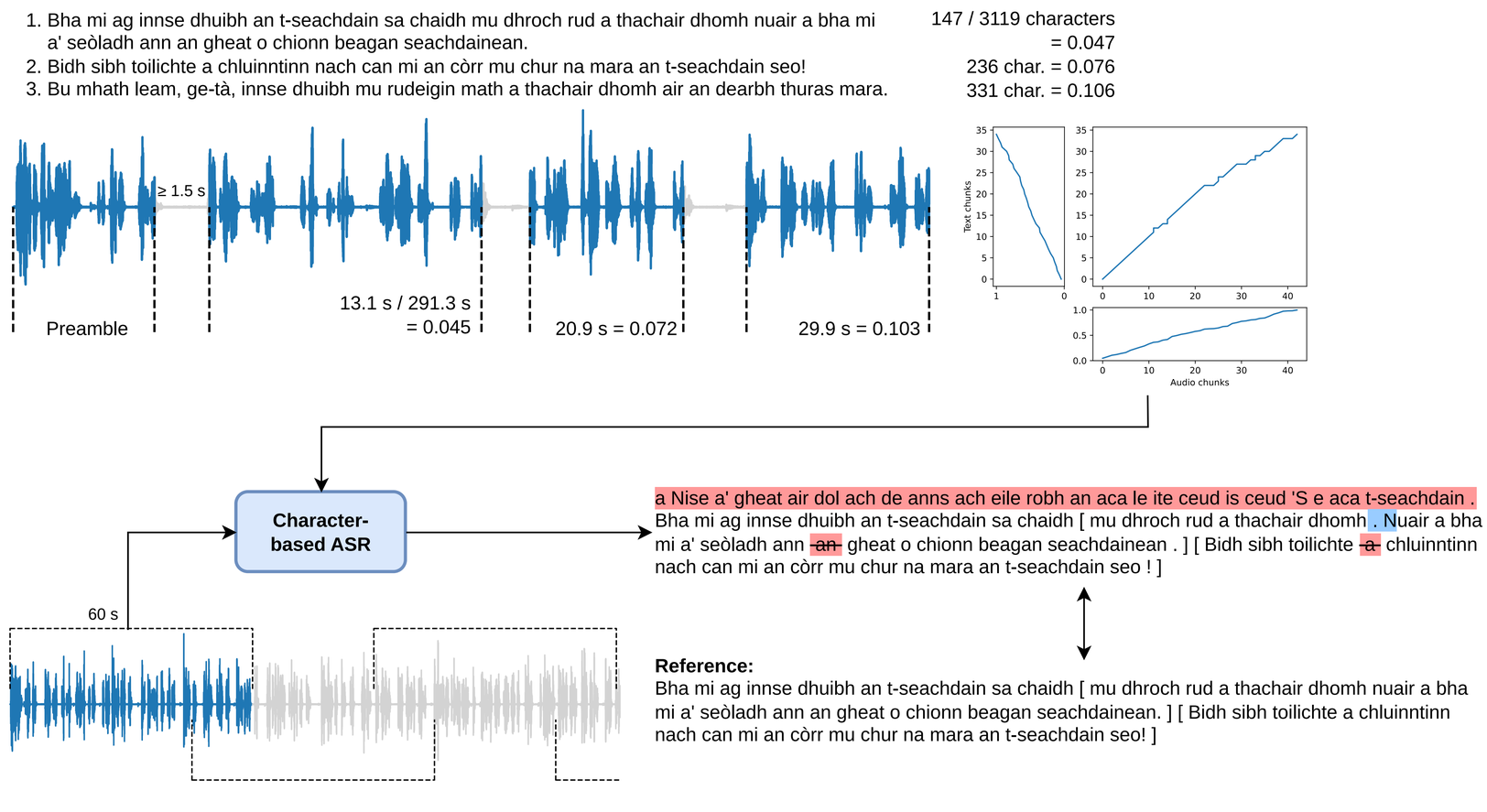
## What kind of linguistic resources do we need for TTS?

- Speech audio segmented at utterance level with text transcripts
  - Assumptions about data requirements reevaluated given more recent non-autoregressive architectures such as FastPitch
  - Increasing use of 'found' data over studio recordings, with some pre-processing
- Linguistic knowledge to process input text and represent the target language symbolically
  - Production systems rely on hand-crafted frontends and pronunciation lexicons
  - Effectiveness of character-input TTS depends on target language orthography

## What do we have for Scottish Gaelic?

- Long history of archival recordings and broadcast media, plus more recent language revitalisation efforts
- Our data comes from the *Litir do Luchd-ionnsachaidh* 'Letter to Learners' series broadcast on *BBC Radio nan Gàidheal*
  - 1,200 recordings, each around 5 minutes, all from a single speaker
  - Full text transcripts, but unsegmented
- Relatively regular orthography, plus *Am Faclair Beag* online dictionary with 35k phonemic pronunciations

## Character-based segmentation and transcript alignment




### Step 1

- Split long recordings on silences over 1.5 s
- Split long transcripts on punctuation
- Roughly align chunks based on cumulative proportions through each sequence
- Train initial acoustic model using character sequences for word pronunciations

### Step 2

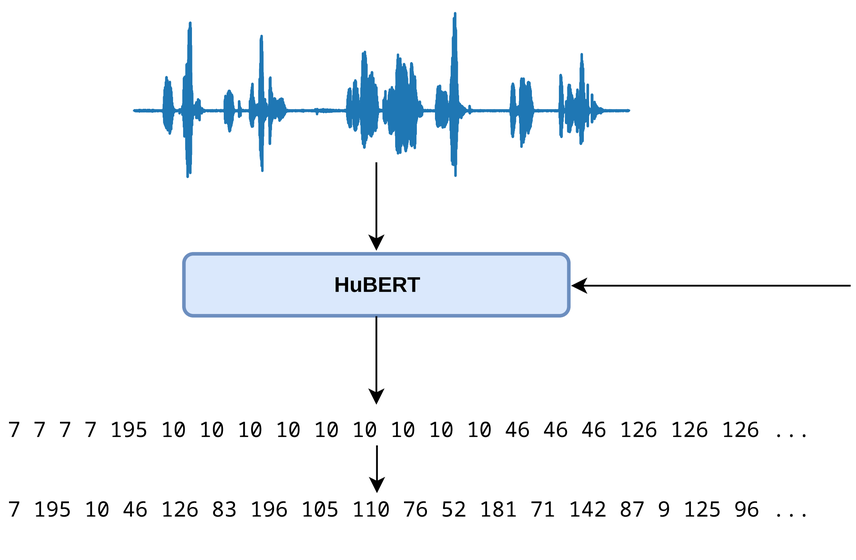
- Decode 60 s chunks of audio with language model trained on full text transcripts
- Smith-Waterman alignment between ASR hypotheses and reference text
- Successfully ignores audio portions not in transcripts, e.g. recording preambles

| Unsegmented | Yield | # Utts |
|---|---|---|
| 100 hours | 86.7 hours | 31,174 |
| 8 hours | 5.5 hours | 1,203 |
| 2 hours | 55 mins | 253 |

## Acoustically-driven data set selection

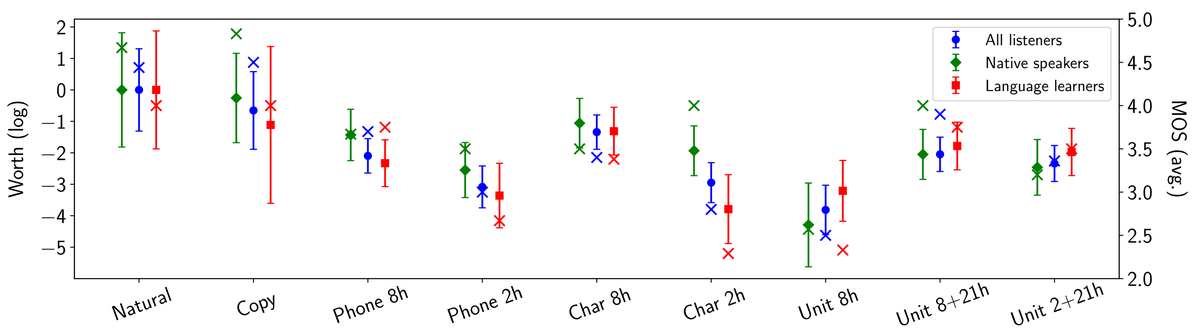
- Aim to maximise phonetic coverage in smaller TTS corpora
- Substitute discrete acoustic units from pre-trained HuBERT (English-only) + Gaelic k-means (200 clusters)
- Greedily select utterances based on unit trigram coverage
  - Achieves 80% of reference triphone coverage for given corpus size

| Data set | # Utts | Triphone coverage | |
|---|---|---|---|
| 2h clean | 475 | 12,376 | 30.2% |
| 8h clean | 2,207 | 22,995 | 56.1% |
| 21h noisy | 7,335 | 24,510 | 59.8% |
| Validation | 380 | 7,412 | 18.1% |
| Test | 418 | 7,112 | 12.4% |




## TTS evaluation

- FastPitch models with different inputs
  - Characters
  - Phones (G2P from *Am Faclair Beag*)
  - Acoustic units (text-to-unit supplemented with noisier recordings)
- Hard to recruit listeners (57k speakers)
  - 3 native speakers
  - 3 'confident/intermediate' language learners
- Best-worst scaling design to compensate for small number of participants




Subjective rankings from best-worst comparisons; crosses indicate average MOS

## Conclusions

- Best-worst scaling and MOS results generally consistent
- Character-based system works well for Gaelic, avoiding effort of building lexical resources and possible G2P errors
- Acoustic unit-based systems seem more consistent when reducing corpus size – perhaps simpler resynthesis task
- Judgements of characters vs. acoustic units different for native speakers and language learners – perhaps reflecting linguistic knowledge vs. audio quality preference

**Acknowledgements:** We would like to thank Ruairidh MacIlleathain for allowing us to use his voice for this work. We also thank BBC ALBA and MG ALBA for granting permission as copyright holder and funder of the *Letter to Learners* series of recordings, respectively. This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

**Get the code!**

dan.wells@ed.ac.uk
https://wellsd.net

THE UNIVERSITY of EDINBURGH informatics

THE UNIVERSITY of EDINBURGH UKRI Centre for Doctoral Training in Natural Language Processing

UK Research and Innovation