A semi-automatic pipeline for transcribing and segmenting child speech

[1]**Polychronia Christodoulidou,** [2]**James Tanner,** [2]**Jane Stuart-Smith,** [3]**Michael McAuliffe**
[4]**Mridhula Murali,** [4]**Amy Smith,**[4]**Lauren Taylor,** [4]**Joanne Cleland,** [4]**Anja Kuschmann**
[1]Aristotle University of Thessaloniki, Greece; [2]University of Glasgow, UK
[3]McGill University; [4]University of Strathclyde, UK

VariCS
Variability in Child Speech

The problem

Getting reliable acoustic phonetic measures from (lots of) child speech is **challenging**[1,2,3,4]

- child speech is highly variable[1,5,6]
- speech style and spoken variety[2,3]
- extremely time consuming to prepare child speech recordings for analysis[3,7]

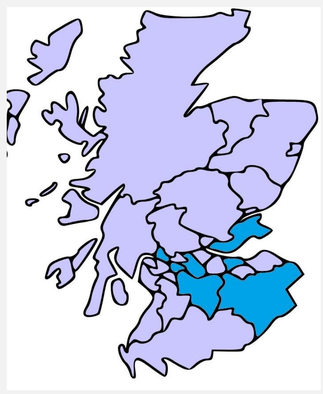
automated transcription ⇨ manual adjustment ⇨ forced alignment ⇨ manual check

- **strong performance from WhisperX**
- **adjusting WhisperX output improves quality**
- **adapting MFA model improves quality**
- **far less manual input (= time!) to achieve robust acoustic measures**

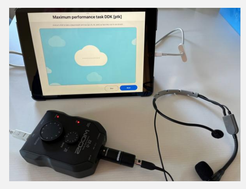
VARICS: large child speech reference corpus

- 275 children
- 227 typically-developing children
- age: 5;0 – 11; 9
- gender: 124 male, 150 female, 1NB
- 23 schools, 7 councils in Scotland, UK

May – November 2023 | November – June 2024 | May – November 2024 | November – June 2025



Data collection



- DEAP screening task
- 2 non-speech tasks
- 3 connected speech tasks
- **single word naming task** (3x)
  CVC /i e ɪ ɛ a ʌ ɔ o ʉ/
- bespoke iPad app
- Shure SM35 mic/Zoom-U22 interface

Method

- 783 recordings / ~13 hours / 273 children
- WhisperX[13] model `small`, language `English`
- team of phonetically-trained student assistants
- Python conversion script *Whisper2textgrid.py*
- Montreal Force Aligner[17] *english_us_arpa*

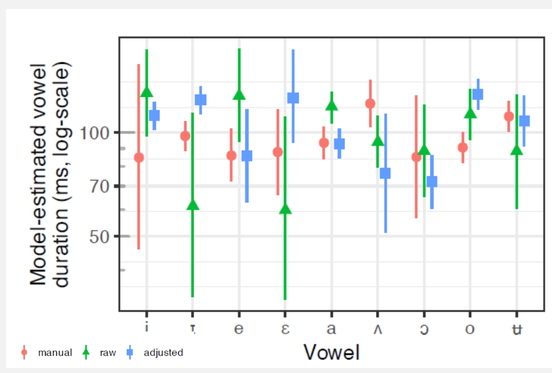
➢ acoustic vowel measures (F1, F2, duration) in *Praat* from MFA / manual boundaries

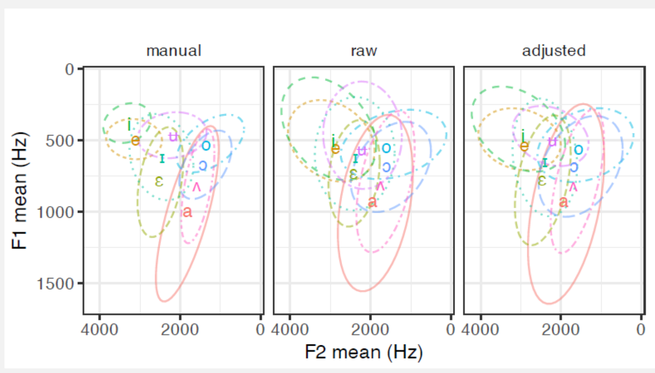
Stage 1: adjusting WhisperX output

- ~ 150 hours manual adjustment of WhisperX output
- 73% recordings all speech transcribed
- adjustment needed: 70% 'a little' / 'no'; 24% 'some' ; 5% 'a lot' / 1% 'all'

evaluation against manual segmentation

- 64 recordings / 3695 sec / 22 children
- Linear Mixed Modelling on log. durations / F1, F2
- **manual** vs '**raw**' WhisperX vs **adjusted** WhisperX





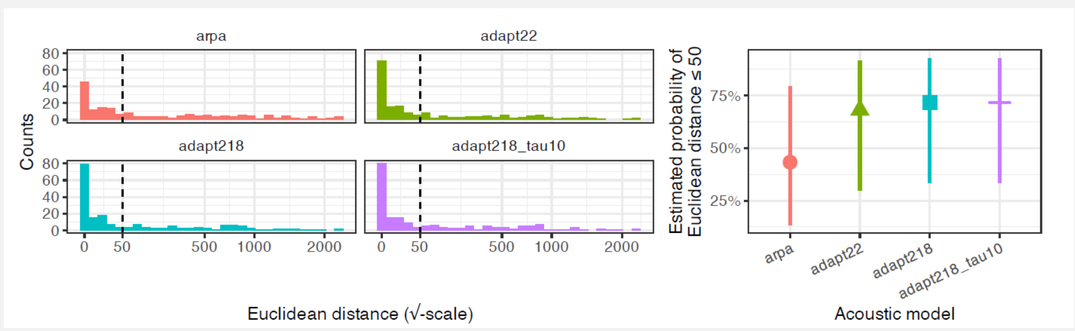
➢ durations: raw < manual = adjusted
➢ formants: raw / adjusted show more variable vowel space; vowel specific differences (especially /i/)

Stage 2: adapting MFA acoustic models

- used *mfa_adapt*[26] to make three new acoustic models from MFA *english_us_arpa* (**arpa**)
  adapt22: 64 recordings / 3695 sec / 22 children
  adapt218: 654 recordings / ~ 11 hours / 218 children
  adapt218_tau10: 218 recordings / ~ 11 hours / 218 children, with `–mapping-tau=10`

evaluation against manual segmentation

- 9 recordings / 418 sec / 9 children
- log. regr. on Eucl. Dist. from manual (≤ / > 50 Hz)
- **arpa** vs **adapt22** vs **adapt218** vs **adapt218_tau10**



➢ all three adapted models perform better than *english us_arpa*
➢ no improvement with greater speaker sample

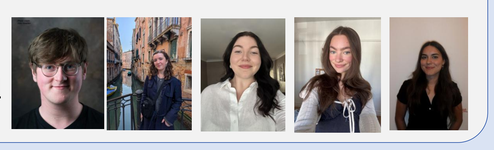
Reflections

- both WhisperX and MFA disrupted by
  - extraneous noise (field recordings)
  - consistent child speech hyperarticulation
- adaptation of MFA model only needs small speaker sample (McAuliffe/Gunter LSA 2025)

references in paper

more about VARICS adapted models here

Thanks to the **VARICS Correction Crew = Zara Johnson, Roxy Patton, Abbie Halls, Thea Foster, Sofia Moscato, Jack Louw, Ollie Gotts & Zack Barr ;** and to all **schools, children & parents**



University of Strathclyde Glasgow
University of Glasgow
GULP
UKRI Economic and Social Research Council
ES/W003244/1