Bigger is Not Always Better: The Effect of Context Size on Speech Pre-Training

Sean Robertson, Ewan Dunbar
University of Toronto
sdrobert@cs.toronto.edu,ewan.dunbar@utoronto.ca

VECTOR INSTITUTE

Motivations

1. **Ecological validity**
   - Computer-Assisted Pronunciation Training (CAPT) systems often start with Automatic Speech Recognition (ASR)
   - Non-native ASR difficult
   - CAPT training commonly uses clean, read-and-record data
   - Read-and-record ≠ participating in CAPT
   - FAB data more applicable to downstream task → better training
2. **Teacher-driven labels**
   - CAPT systems rely on intrinsic criterion of *nativeness*
     - Easy for engineers: more native ≈ more probable under ASR
     - Problematic for educators: *which* native? "Good enough?"
   - Solution: assume teacher knows best and mimic her feedback
   - FAB provides both intrinsic (nativeness) and extrinsic (teacher feedback) binary label sets
3. **Absolute beginner learners**
   - Pronunciation training more impactful to beginner and intermediate learners than advanced ones (Mahdi and Al Khateeb, 2019)
   - Absolute (L2) beginners overlooked in existing corpora
   - FAB is the first corpus built from absolute beginner learners' speech

Overview and collection

Speakers	121
Segmented utterances	approx. 9000
Word segments	approx. 25000
Binary labels	approx. 17000
Word segment duration	approx. 3 hours

- Collected over two experiments at University of Toronto
- English-speaking participants knew little-to-no French
- Paired role-play tasks, consistent with task-based pedagogy (Skehan, 2003)
- Dialogue modelled by video
- Participants rehearsed dialogue with small contextual changes
- Participants took turns speaking into iPad app
- App gave feedback, including accept/reject utterance
- App actually controlled by expert in language learning
- Beginners + recall + contextual inference = very messy data!
- Data from first experiment in minimal feedback partition; second in explicit and read-and-record partitions
- Nativeness labels in minimal feedback partition; teacher feedback labels in explicit feedback partition

Minimal feedback partition

Speakers	58
Segmented utterances	approx. 4000
Word segments	approx. 12000
More/less native labels	5767
Fluency in French (1-5 asc.)	1(52), 2(8), 3(3)

ann.	κ_1	κ_2 (no.)
A	.16	.48(168)
B	.20	.52(184)
C	.17	.50(174)
D	.16	.51(154)

- Collected in first experiment (Robertson et al, 2016)
- App feedback limited to accepting/rejecting utterances
- More/less native word labels extracted by bucketed pairwise comparisons
  - Four expert annotators compared pairs, winner the "more native" of two
  - 10 instances fully ranked counting wins, choose 4th and 6th as boundaries
  - New word loses against both boundaries assigned "less native" label; if it wins both, it is assigned "more native" label
  - Inter-annotator agreement in above table; κ_1 is Cohen's Kappa including one-win-one-loss words; κ_2 only win-win or loss-loss

Explicit feedback parition

Speakers	63
Segmented utterances	approx. 5000
Word segments	approx. 13000
Correct/mispronounced labels	approx. 12000
Fluency in French (1-5 asc.)	1(59), 2(4)

- Collected in second experiment (Robertson et al, 2018)
- Word-level feedback in addition to accept/reject: insertions, deletions, mispronunciations, and *recast* recordings
- Expert feedback replaced with automated feedback in experimental condition. See *Challenges* section for results.
- Correct/mispronounced labels extracted from teacher feedback

Read-and-record partition

Speakers	19
Segmented utterances	approx. 600
Word segments	approx. 2000
Fluency in French (1-5 asc.)	1(18), 2(1)

- Collected *after* second experiment, time-permitting
- Participants read aloud and recorded as many utterances as possible
- Prompts were chosen for phonemic diversity

This research was funded by an Ontario Centres of Excellence Technical Problem Solving grant in partnership with Speax Inc., and a Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada.

Freely avaliable this year on U. of T.'s *Dataverse*: `https://dataverse.scholarsportal.info/dataverse/toronto`

Results

- Performed on nativeness labels with manual word alignments
- Pronunciation Error Detectors (PEDs) classify words as (non-)native:
  - Goodness of Pronunciation (GOP, Witt and Young, 2000): empirically-tuned threshold of

$$\log P_{\text{native}}(\text{expected words}|\text{audio}) - \max_{\text{any words}} \log P_{\text{native}}(\text{any words}|\text{audio})$$

  - Four based on GMMs trained on audio segments (Franco et al, 2014):
    - GMM1: two separately-trained GMMs per word, native vs. non-native, with empirically-tuned threshold of

$$\log P_{\text{word+native}}(\text{audio segment}) - \log P_{\text{word+non-native}}(\text{audio segment})$$

    - GMM2: same as GMM1, but GMMs MAP-adapted to Universal Background Model (UBM)
    - GMM3: UBM MAP-adapted to each word instance → extract GMM supervectors → linear SVM classifier on supervectors
    - GMM4: linear combination of GMM2 + GMM3
- Average 4-fold cross-validation with 3 partitioning strategies: split by annotator (4FA); split by participant (4FP); split to stratify database (4FS)

PED	4FA		4FP		4FS	
	Accuracy	κ	Accuracy	κ	Accuracy	κ
GOP	62%	.23	63%	.25	63%	.24
GMM1	63%	.23	63%	.25	66%	.31
GMM2	**68%**	**.34**	**66%**	**.31**	**71%**	**.41**
GMM3	65%	.27	62%	.23	67%	.33
GMM4	64%	.27	65%	.29	70%	.38

Results tuned on test set. See Robertson et al., 2016 for untuned.

Challenges

- During explicit feedback experiment, pit ASR+GMM2 against heuristic, rule-based PED
- GMM2 tuned to segments and labels from pilot phase (no task mismatch)

PED	κ
GMM2	.19
Heuristic	**.34**

- On rejected utterances, heuristic picks "hardest" word to pronounce, round-robins mispronounced word on successive rejections
- Heuristic beats GMM2 in κ → **better results when mic is turned off!**
- Possible reasons:
  - Non-native ASR difficult w/ little training data → bad alignments
  - Context effects: expert focuses on one word at a time, severity of mispronunciation, learner's motivation, etc.
- Task easier than it *should* be since teacher provides transcription
- **Need approach robust to misrecognized words, poor alignments, and dialouge effects!**