Sufficiency Quantification for Seamless Text-Independent Speaker Enrollment

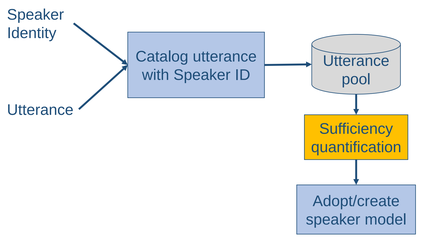
Gokcen Cilingir, Jonathan Huang, Mandar S Joshi, Narayan Biswal

Intel Corporation

Motivation

- **NOW**
  - ✓ **Voice-enabled platforms are taking off**
  - ✓ **Text-dependent speaker recognition is a technology that is already in the market in smart speakers**
- **FUTURE**
  - ✓ **Speaker recognition (SR) on natural speech (Text-independent speaker recognition)**
- **PROBLEM**
  - ✓ **Enrolling for SR with natural speech is NOT user-friendly**
- **SOLUTION**
  - ✓**Seamless enrollment/adaptation**

Seamless enrollment/adaptation flow

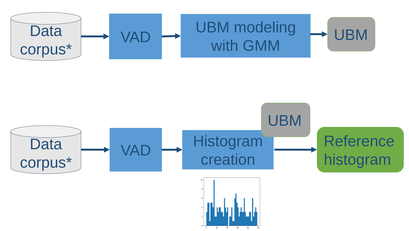


Sufficiency quantification

What criteria will you use to decide that you have enough data in the utterance pool to create a phrase-independent model of the speaker?

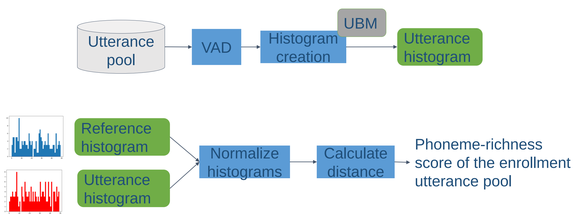
- **Current enrollment model:** Ask user to read a phoneme-rich passage, or a couple of such sentences
- **Naïve approach:** Use speech duration
- **Our approach:** Define a metric to quantify phoneme-richness

Phoneme-richness quantification – offline process



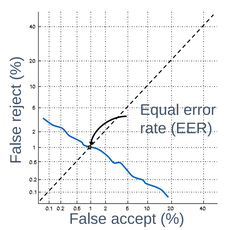
*We used TIMIT. TIMIT contains 10 phoneme-rich sentence utterances from >650 speakers

Phoneme-richness quantification over enrollment utterance pool

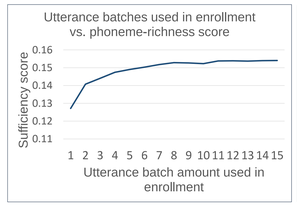


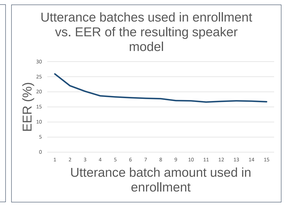
Data set

- A proprietary dataset of 40 speakers.
- Each speaker uttered 180 short commands, 10 repeats of 20 trigger words and a short phoneme-rich passage.
- 180 short commands are split into 25 batches, where each batch contained ~3 secs of speech content.
- 10 batches left out for testing. Equal error rate (EER) is as accuracy metric.



Experiment Results





Sufficiency metrics	Success criteria (Pearson correlation with EER)
Phoneme-richness score	-0.99
Speech duration	-0.79

Experiment Results

Enrollment scenarios	Speech duration (seconds)	Phoneme-richness score	EER (%)
10 "hello computer" repeats	3.02	0.092	37.44
10 repeats of 4 trigger words	13.32	0.120	30.12
Short phoneme-rich passage	12.33	0.152	23.96

Sufficiency metrics	Success criteria (Pearson correlation with EER)
Phoneme-richness score	-0.96
Speech duration	-0.68

Sufficiency quantification usages beyond seamless enrollment

- Improve UX for traditional text-independent SR enrollment
- Confidence modeling during detection/test.
- For multi-session enrollment and model fusion

Future Direction

- Proving metric utility over low quality data, spontaneous and noisy speech
- Proving metric utility over confidence modeling task