Generalised Discriminative Transform via Curriculum Learning for Speaker Recognition

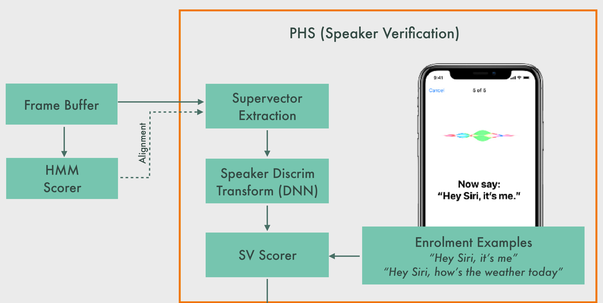
Erik Marchi, Stephen Shum, Kyuyeon Hwang, Sachin Kajarekar, Siddharth Sigtia, Hywel Richards, Rob Haynes, Yoon Kim, John Bridle
Apple Inc.

Summary

- Personalise the always-on Hey Siri detector by verifying the speaker's voice before triggering Siri.
- Leverage upon LSTMs and Curriculum Learning to reduce detection errors and improve generalisability across various conditions.
- Exploit payload data to enable extraction of speaker vectors from less-constrained text.
- Achieve a relative equal error rate (EER) reduction of 30–70% compared to the DNN baseline.

Background

Personalised Hey Siri (PHS) system



$$SV_{score}(u_a, spk) = \frac{1}{N}\sum_{i=1}^{N} \frac{f_{nn}(u_a)^\top f_{nn}(u_i^{spk})}{\|f_{nn}(u_a)\|\|f_{nn}(u_i^{spk})\|}$$

- The DNN extracts speaker-specific information based on the trigger phrase.
- The DNN system is only able to perform text-dependent speaker verification.

➡ Limited flexibility.

Premises

➡ LSTM to handle less-constrained text.

➡ Curriculum Learning to improve generalisability.

Methods & Results

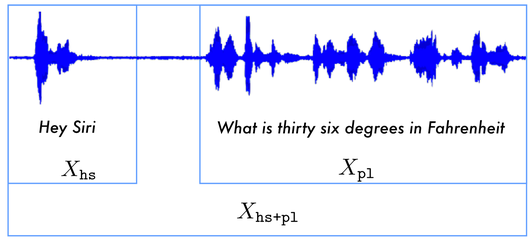
Incorporating payload speech

'*Hey Siri*' requests come in two forms:

A. Just the 'Hey Siri' trigger; or

B. The trigger followed by the payload.

Given the extra speech, we expect (B) might lead to a more reliable speaker representation.



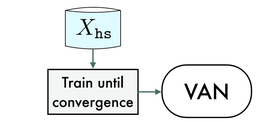
$X_{\text{all}} = \{X_{\text{hs}} \cup X_{\text{hs+pl}} \cup X_{\text{pl}}\}$ X_{all}^{sim} = simulated X_{all}

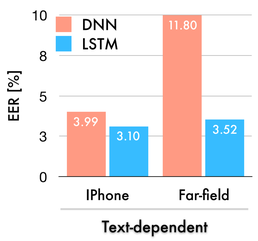
Improving generalisation via curriculum learning (CL)

The general principle of learning simpler concepts first before gradually learning more complex ones.

Text-dependent → text-independent

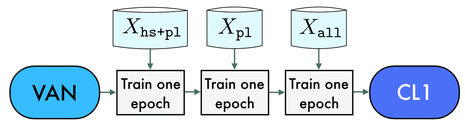
1. Learn fixed phrase

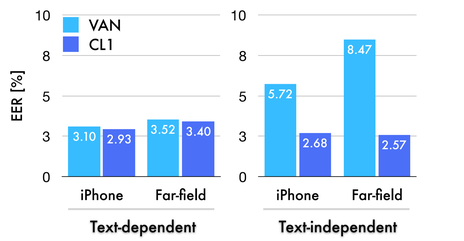




▸ DNN → LSTM:
↑ Accuracy
↑ Flexibility

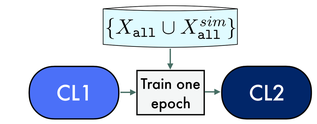
2. Learn less-constrained text content

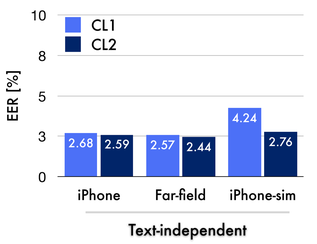




▸ Fixed text → less-constrained text:
↑ Generalisability towards text scenarios
↑ Using payload further reduces EER

3. Learn acoustic conditions





▸ Clean → augmented:
↑ Robustness under various acoustic conditions

DNN: 442 supervector → 4x256 sigmoidal → 1x128 linear (→ 1x18k softmax)
LSTM: 20 MFCCs → 1x512 LSTM → 1x128 linear (→ 1x18k softmax)

VAN: Vanilla model only trained on X_{hs}
CL1, CL2: Curriculum Learning models

Datasets

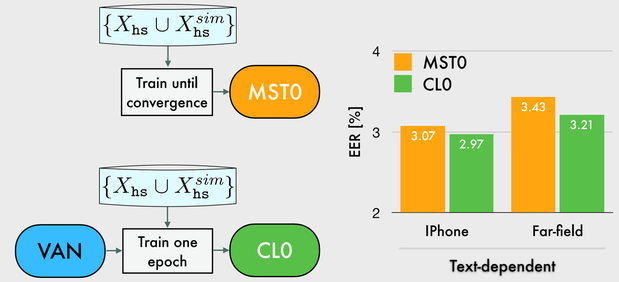
- iPhone data: 'Hey Siri' requests sent to our servers.
- Far-field data: 'Hey Siri' requests from various distances (6-15ft) recorded in various rooms of ten different houses.
- Simulated iPhone (iPhone-sim): a subset of iPhone data convolved with various RIRs and/or corrupted with car noise.

		#utts	#spks	#utt/spk
Train ($X_.$)		2.5M	18k	>20
Train ($X_.^{sim}$)		1.5M	18k	>20
iPhone	enrol	2.5k	500	>4
	test	53k	500	>40
Far-field	enrol	490	98	>4
	test	11k	102	>20
iPhone-sim	enrol	490	98	>4
	test	11k	102	>20

Discussion

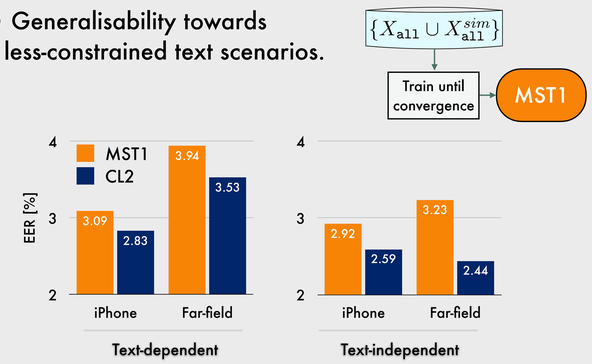
Is CL better than multi-style training (MST)?

- Acoustic robustness in text-dependent tasks.



↑ CL better than MST on all acoustic conditions

- Generalisability towards less-constrained text scenarios.



↑ CL better than MST on less-constrained text scenario

↓ MST1 takes longer to converge

Conclusions

- The proposed curriculum learning procedure improves:
  - ➡ The robustness against various acoustic conditions.
  - ➡ The generalisability towards less constrained-text scenarios.
- A single generalised discriminative transform that performs speaker verification on both text-dependent and text-independent tasks.
- Using payload speech further reduces the EER.

E. Marchi, S. Shum, K. Hwang, S. Kajarekar, S. Sigtia, H. Richards, R. Haynes, Y. Kim, J. Bridle, Generalised Discriminative Transform via Curriculum Learning for Speaker Recognition, ICASSP 2018. Contact: emarchi@apple.com