Improving Human-Computer Interaction in Low-Resource Settings with Text-to-Phonetic Data Augmentation

Adam Stiff, Prashant Serai, and Eric Fosler-Lussier

Motivation

Off-the-shelf speech recognizers are error-prone in specialized domains; we aim to mitigate the impact of these errors for downstream classification tasks without in-domain speech training data. In this work, we study how to mitigate the effects of the lack of speech training data when converting a typed chatbot to a spoken language interface.

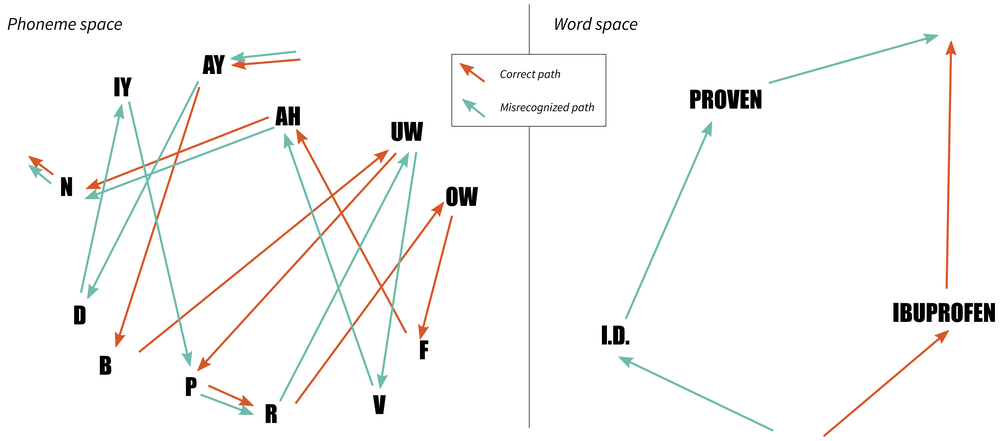
Summary

- We ensemble text CNNs trained on word representations, and inferred acoustic representations of available in-domain text data.
- We generate likely recognition errors for our text examples and sample from them during training.
- We experiment with three different methods of representing of word boundaries

Our application domain is a virtual patient. We treat the multi-turn conversation as an iterated question classification task.

Data

- Training data come from typed conversations with a virtual patient: 4,330 examples of 359 classes in a long-tailed distribution [1].
- Phonetic representations of the input are looked up or inferred using CMUdict and Phonetisaurus.
- A small test set was collected by capturing recognition results from volunteers reading additional typed conversations; only 756 examples.

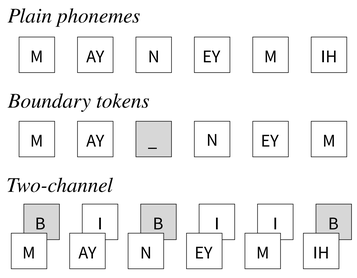


Motivating illustration: In the word space, misrecognition paths can be very divergent from the correct paths; in the acoustic space, they are much more similar.

Error generation and sampling

- We generate up to a hundred likely misrecognitions for each text input sentence, by sampling errorful phones from a distribution derived from confusions observed in a general purpose ASR system, and decoding [2].
- At training time, we train on alternative examples at varying rates, and sample these errorful alternatives according to their likelihood of being generated.
- Sampling is intended to augment training data without making development sets too easy.

Word boundaries



Architecture

Results

	Sampling	Phonemes	Words	Combo
Baseline (typescript)	N/A	System trained as combination only		69.9
Baseline (speakscript)	N/A	System trained as combination only		65.7
All alternatives	N/A	64.95	65.48	65.74
Plain phonemes	0%	67.15	66.27	67.55
	5%	66.89	66.76	67.68
	10%	66.75	66.40	67.73
	20%	66.75	66.00	**68.30**
	50%	66.36	66.09	67.50
Boundary tokens	0%	66.45	66.09	67.64
	5%	66.67	66.05	67.86
	10%	66.58	66.88	67.90
	20%	65.88	66.76	67.77
	50%	65.96	66.31	67.99
2-channel bounds	0%	**67.37**	**66.89**	67.37
	5%	66.67	66.58	67.59
	10%	66.48	66.40	67.42
	20%	66.62	**66.89**	68.12
	50%	67.11	66.36	67.95

All Combo results are significant improvments over the speakscript baseline using a chi-squared test and Benjamini-Hochberg multiple tests correction with a false discovery rate of 10%.

Conclusions and future work

- The best result recovers 62% of the volume of errors induced by naively feeding recognition results to a model trained on text.
- The benefit of word boundary information, and different ways of representing it, are unclear.
- Sampling generated errors seems to generally provide a benefit.
- The phoneme and word sub-ensembles seem to be learning complementary information.
- Ongoing work has collected much more spontaneous spoken data, so future work can give a better statistical footing.

References

[1] Lifeng Jin, Michael White, Evan Jaffe, Laura Zimmerman, and Douglas Danforth, "Combining cnns and pattern matching for question interpretation in a virtual patient dialogue system," in Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, 2017, pp. 11–21.

[2] Prashant Serai, Peidong Wang, and Eric Fosler-Lussier, "Improving speech recognition error prediction for modern and off-the-shelf speech recognizers," IEEE International Conference on Acoustics, Speech, and Signal Processing. (ICASSP)., 2019.

The Ohio State University