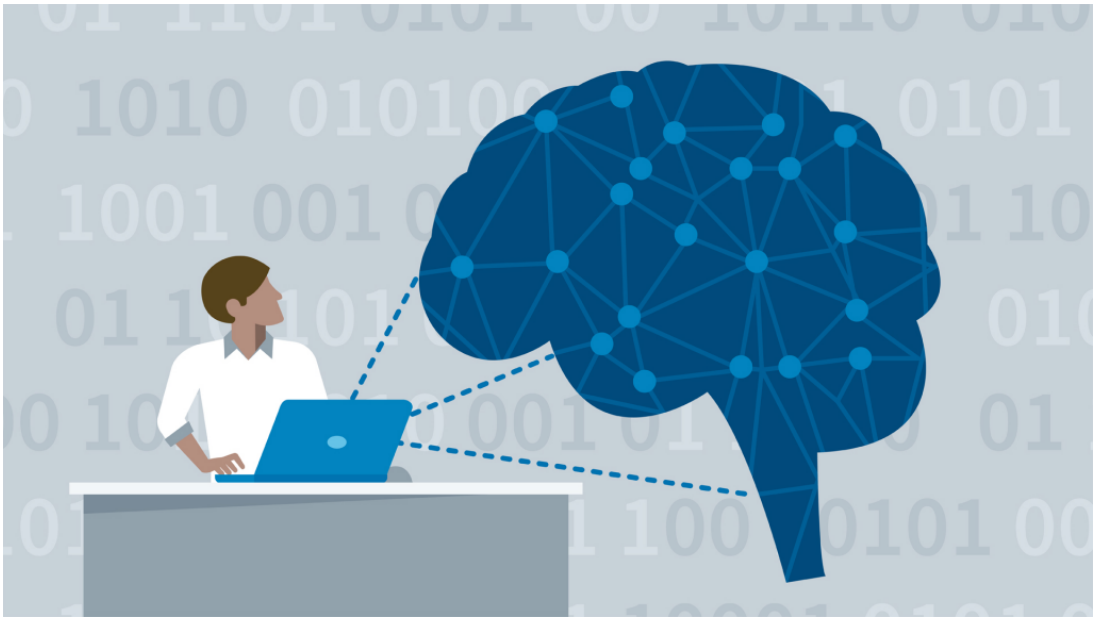
NUANCE | WHAT'S NEXT BLOG

Enterprise R&D, Innovation & Research

Delivering personalized user experiences with speaker adapted end-to-end speech recognition

Innovation at Nuance | Employee Guest Blogger

December 20, 2019



In the latest generation of Nuance's deep learning technology for speech recognition, we developed techniques for user adaptation that can reduce the error rate by up to 33 percent, exploiting regularized end-to-end learning, adaptation on text data, and minimum word error rate adaptation.

About the Author: Felix Weninger is a senior tech lead (Senior Principal Research Scientist) at Nuance Communications. His research interests include deep learning, speech recognition, speech emotion recognition, and source separation. He received his PhD degree in computer science from Technical University of Munich (TUM), Germany, in 2015. Prior to joining Nuance, he worked at Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA and in the Machine Intelligence and Signal Processing Group at TUM's Institute for Human-Machine Communication. He has published more than 100 peer-reviewed papers in books, journals, and conference proceedings.

Editor's Note: Jesús Andrés-Ferrer, Xinwei Li and Puming Zhan contributed to the research presented in this blog post along with Felix Weninger. The paper was presented at the INTERSPEECH conference in Graz, Austria, September, 2019.

A crucial capability of automatic speech recognition (ASR) systems is to cope with variability in speech,

caused by various accents, age groups, or other variations in speaking style, as well as noisy environments. Several years ago, Nuance's Dragon dictation product line pioneered the usage of deep learning technology for speaker adaptation in professional dictation systems.

Deep learning technology has rapidly transformed the way that computers perform speech recognition. Traditionally, the difficult task of building an ASR system was broken down into smaller pieces, resulting in several components for modeling the frequency and usage of words (language model), the way that words are formed of phonemes — the smallest spoken units of a language –(pronunciation model), and how phonemes are realized in the signals captured by the microphone (acoustic model).

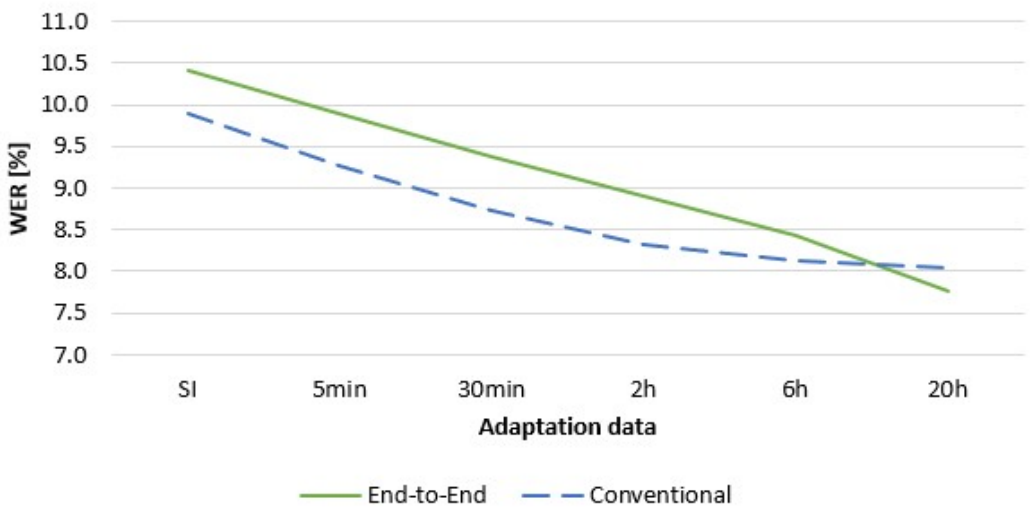
Given the advances in deep learning, these days, it is possible to subsume the acoustic, pronunciation and language models into a single deep neural network (DNN), which is also known as "end-to-end" ASR. While end-to-end learning is certainly advantageous in terms of simplicity and accuracy, it does not solve the problem of variability in speech; thus, the need arises for effective adaptation schemes for end-to-end ASR.

In the paper titled "Listen, Attend, Spell and Adapt: Speaker Adapted Sequence-to-Sequence ASR", we explored the adaptation of state-of-the-art end-to-end ASR, demonstrating up to 33 percent reduction in word error rate (WER) on a dictation task. Such large reductions are possible because our end-to-end learning methods adapt to the user's specific vocabulary, sentence structure, accent, microphone, etc., all at the same time, in contrast to conventional systems which adapt the acoustic model and language model separately. Our paper is accepted as an oral presentation at INTERSPEECH 2019, the world's largest conference on spoken language understanding.

One challenge that we faced in implementing end-to-end ASR adaptation is that DNNs comprising the acoustic, pronunciation and language model need to be huge in order to be effective: In fact, the model used in the paper contains more than 100 million parameters. Adapting these parameters on speech data from a single user can easily result in an optimization problem known in the literature as "overfitting", which yields undesirable side effects. For example, a user's dictations could always start with the words "This is Jane Smith dictating a letter ...". If the model is trained on too many examples of similar kind, it will lose its ability to output general sentences.

We solved this problem by changing the way the optimization is done, by employing a strategy known in the literature as "regularization". Rather than only minimizing the error rate on the user-specific sentences, we also discourage the output of the model from deviating too much from the outputs of our off-the-shelf (speaker-independent, SI) speech recognizer. Furthermore, we showed that we do not need to adapt all the parameters in the model: carefully selecting a subset can yield very good performance as well.

In the result, we could get significant error rate reductions even with a few minutes of speech, and up to 25 percent with 20 hours of speech:



As can be seen from the graph above, end-to-end ASR adaptation already outperforms the conventional acoustic model adaptation given enough data. However, we were able to obtain even larger gains by also exploiting pure text data sources from the same users (for example, written documents), rather than only their recorded speech data. This was not trivial to achieve since there is no separation into acoustic and language models in end-to-end systems, and text data can only be used to train a language model. Our strategy is to use the text data to fine-tune an external language model integrated into the end-to-end system. Similarly to the full end-to-end system, we regularize it using techniques we had previously developed in the paper "Efficient language model adaptation with noise contrastive estimation and Kullback-Leibler regularization" (INTERSPEECH 2018).

Another contribution of our paper is to change the optimization criterion in adaptation to directly reflect the WER metric (minimum WER adaptation). Overall, it turned out that the gains from regularized end-to-end learning, adaptation on text data, and minimum WER adaptation added up nicely, so that we were able to obtain up to 33 percent WER reduction from adaptation.

In conclusion, the advanced end-to-end deep learning technologies we employed do not only make ASR systems simpler and more powerful – they also deliver a much more personalized user experience, through seamless adaptation of the way computers perceive both acoustic and linguistic aspects of speech.

Tags: