Speech and Language Processing Technical Committee Newsletter
Winter 2014 edition

Personalizing Cortana: a close-up view of the speech technology powering Cortana

Fil Alleva and Michael Tjalve

SLTC Newsletter, November 2014

Introduction

Cortana is the most personal assistant who learns about her user to provide a truly personalized experience. Behind the scenes, there's a wide range of speech and language technology components which make the natural interaction with Cortana possible. Personalization of the speech experience happens seamlessly while the user simply experiences better speech recognition accuracy and a more natural interface. Here are a few examples of recent research from our labs.

Acoustic Model Adaptation

Speaker independent models are designed to work well across the span of the user demographic but each interaction represents an individual use case with characteristics in accent, articulation, context, devices and acoustic environment. This means that there is significant potential gain in recognition accuracy from personalized models. However, with tens of millions of parameters in the DNN models, storing personalized models in a large-scale deployment becomes a challenge of practical limitations and scalability. Xue et al. [1] describes the method of singular value decomposition (SVD) bottleneck adaptation where a matrix containing the speaker information is inserted between low-rank matrices in each layer. With this approach, adaptation can happen by updating only a few small matrices for each speaker. With only a small number of speaker-specific parameters to maintain, the storage requirement for the personalized model is less than 1% of the original model thus dramatically reducing the cost of personalized models while maintaining accuracy gains.

Language Model Adaptation

One of the key sources for improving language models is data from real live usage. Automated unsupervised language model (LM) adaptation based on anonymized live service data can provide

significant recognition accuracy improvements without the cost of manual transcriptions if you can suppress the error attractors and retain the most valuable data. In [2], a framework for discriminatively filtering adaptation training data is presented. An initial updated LM is generated by incorporating live user data. The framework then learns recognition regression patterns between the recognition results from the baseline LM and the initial adapted LM focusing on differences in LM scores. Error attractor n-grams with their frequencies are computed from the regression pairs and applied to the data filtering process and a new adapted LM is trained on the filtered adaptation data leading to substantial recognition error reduction.

Accented Speech Recognition

With the wider adoption of our devices and services, Cortana is exposed to more variation of accents from a diverse set of users. Accented speech is challenging because of the large number of possible pronunciation variants and the mismatch between the reference model and the incoming speech. [3] presents a multi-accent deep neural network acoustic model targeted at improving speech recognition accuracy for accented speakers. Kullback-Leibler divergence (KLD) regularized model adaptation is applied to train an accent-specific top layer while avoiding overfitting. This top layer models the accent specific information whereas the bottom hidden layers are shared across accents. This allows for maximum data sharing of speech across accents. Building on top of this work, we have now taught Cortana how to better understand non-native accented speech.

References

1. "Singular Value Decomposition Based Low-footprint Speaker Adaptation and Personalization for Deep Neural Network" by Jian Xue, Jinyu Li, Dong Yu, Mike Seltzer, and Yifan Gong, in Proceedings of ICASSP 2014.
2. "Improving Unsupervised Language Model Adaptation with Discriminative Data Filtering" by Shuangyu Chang, Michael Levit, Partha Parthasarathy, Benoit Dumoulin, in Proceedings of Interspeech 2013.
3. "Multi-Accent Deep Neural Network Acoustic Model with Accent-Specific Top Layer Using the KLD-Regularized Model Adaptation" by Yan Huang, Dong Yu, Chaojun Liu, and Yifan Gong, in Proceedings of Interspeech 2014.

Fil Alleva and Michael Tjalve are with Microsoft's speech technology group