NUANCE | WHAT'S NEXT BLOG

Enterprise R&D, Innovation & Research

Combining the advantages of close-talk and far-talk speech recognition

Innovation at Nuance | Employee Guest Blogger

December 6, 2021

Nuance's latest end-to-end speech recognition system for transcribing doctor-patient conversations combines the advantages of both handheld devices and microphone arrays to not only improve recognition accuracy, but also preserve naturalness of the interaction.

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.

Acknowledgments: Marco Gaudesi, Ralf Leibold, Roberto Gemello, and Puming Zhan also contributed to the paper and this blog post. The paper was presented at the ASRU conference, December 2021.

Transcribing conversations between doctor and patients, as in Nuance's Dragon Ambient eXperience (DAX) solution, is a very challenging application for speech recognition. One reason is that today's speech recognition systems can reach optimal performance only for close-talk input, which requires the speaker to wear a headset microphone or use a handheld device like a mobile phone. However, in medical use cases it can be impractical and obtrusive to equip all participants (doctor, patient, nurses, etc.) with close-talk devices.

An appealing and natural way of human-machine communication is 'hands-free' speech recognition, where all speakers are captured with a far-talk device (e.g., a microphone array) that is placed at a fixed position in a room. Yet, the main challenge in far-talk speech recognition is that the accuracy is generally degraded due to the presence of room reverberation and noise (air conditioning, interfering speakers, etc.), especially with large distances (i.e., several meters) between the speakers and the microphones.

In our paper entitled *Dual-encoder architecture with encoder selection for joint close-talk and far-talk speech recognition*, we present our approach to mitigating the discrepancy between close-talk and far-talk speech recognition by combining the best of both worlds. Our paper is accepted for presentation at the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2021.

State-of-the-art speech recognition systems typically consist of an encoder-decoder architecture, where the encoder transforms the acoustic input into a high-level abstraction, and the decoder maps that abstraction to the recognition output by also considering the linguistic dependencies. Consequently, for joint close-talk and far-talk speech recognition, we propose a dual-encoder system, where one encoder is dedicated to close-talk and one to far-talk input. In contrast, only one *decoder* is present, since the linguistic content does not depend on the type of input device.

Using a dual-encoder system enables using a close-talk device for one speaker (in our case, the doctor), while simultaneously capturing the doctor along with all the other speakers with a far-talk device. Since both encoders are specific to either close-talk or far-talk input, each can achieve optimum recognition performance on the respective type of input.

For each input utterance, our speech recognition system decides which device should be used and processes the signal through the selected encoder. The decision is made by a neural network, based on the speech features of the close-talk and the far-talk device, which provide evidence about the speaker role (doctor or others) and the signal quality obtained on each device. Since only one encoder needs to be evaluated, the computational cost of the proposed system is close to the standard system with only a single encoder.

All components of the system (encoders, decoder, and encoder selection network) are part of a single neural network that is trained end-to-end; it also includes a neural beamforming component, which is used to improve the signal quality of the far-talk input. The flowchart of our dual-encoder architecture is shown below.

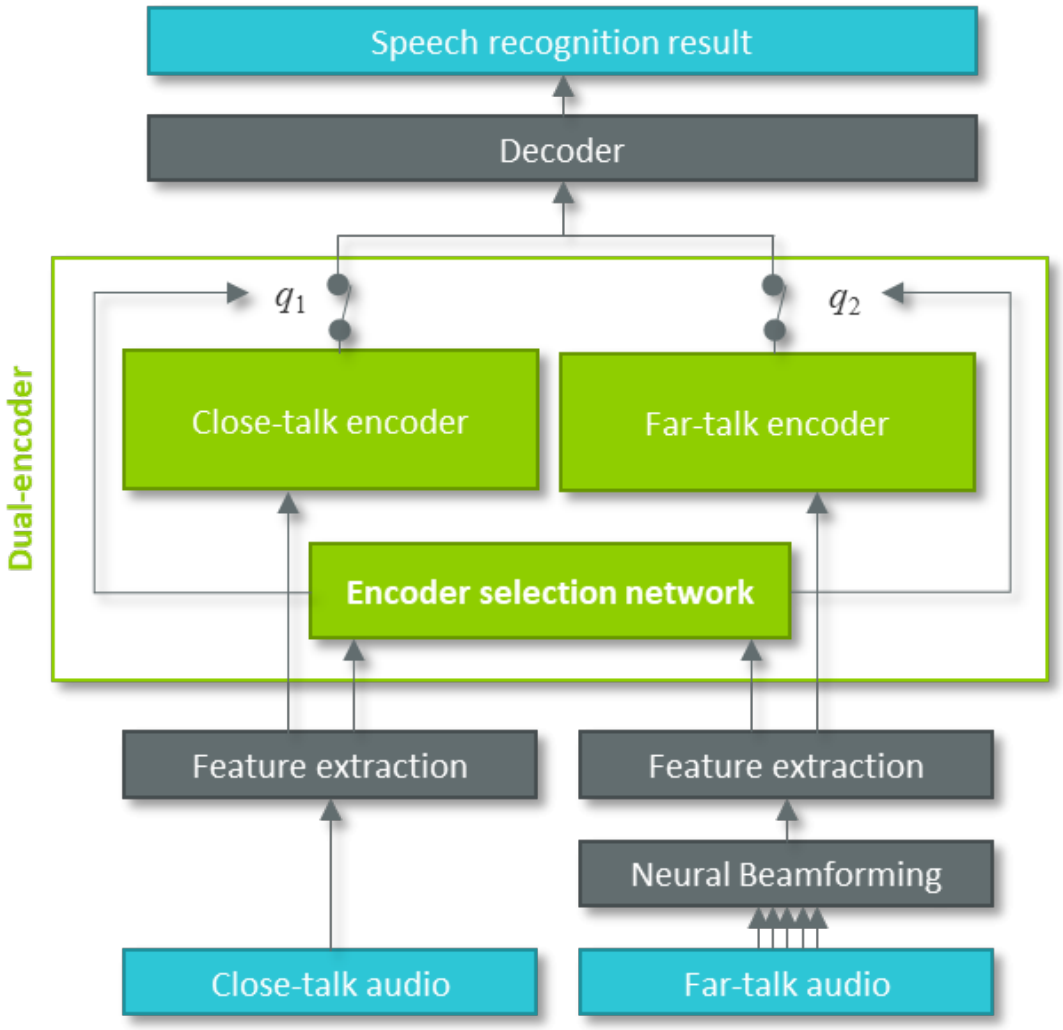

Figure 1: Dual-encoder architecture with encoder selection for joint close-talk and far-talk speech recognition.

Instead of making a 'hard' decision for one device / encoder, the proposed speech recognition system can also operate in 'soft selection' mode. Here, the outputs of both encoders are combined additively, according to the posterior probabilities q_1 and q_2 of the encoder selection network. Thus, the overall system profits from the well-known 'model combination' effect, leading to improved recognition performance: By avoiding a hard decision in case of uncertainty, we make sure to not miss valuable information from one of the inputs.

In order to make our approach work well in practice, we had to address three different issues. First, in contrast to close-talk data, which are available in large quantities, realistic far-talk data are costly to obtain. Thus, instead of training our system from scratch, we opt for a domain adaptation approach, where we initialize the parameters with the ones learnt from a large amount of close-talk data.

Second, since our system relies on combining information from different devices, it could be affected by synchronization issues between the two. To combat the sensitivity to the input synchronization, we propose training the system with random shifts between the close-talk and far-talk device signals.

Third, more than one speaker might be present in an input segment if the input is processed by automatic voice activity detection rather than manually end-pointed, and thus the optimal device can vary over time. For this case, we evaluated a frame-wise encoder selection approach and found that it performed better than the utterance-wise selection.

To validate our approach, we applied the dual-encoder paradigm to two different end-to-end speech recognition architectures: Listen-Attend-Spell and Conformer Transducer. We performed experiments with 3800 hours of close-talk speech for training a seed model and 40 hours of parallel close-talk and far-talk data for domain adaptation.

The two main takeaways from our experimental results are:

First, the dual-encoder system with combined close-talk and far-talk input outperforms the best single-

encoder close-talk and far-talk systems by up to 9% (in terms of relative word error rate reduction). Note that as expected, significant accuracy gains can be obtained by using soft selection (model combination) instead of hard selection.

Second, the accuracy of the dual-encoder system is similar to a theoretically optimal combination of the single-encoder close-talk and far-talk systems, where one queries an 'oracle' to determine the speaker role for each utterance and then chooses the close-talk system for the doctor and the far-talk system for the others.

In summary, our dual-encoder approach combines the advantages of both handheld devices and microphone arrays to improve the recognition accuracy while preserving the naturalness of the doctor-patient interaction.

Tags: Speech recognition, Research paper