Healthcare R&D, Innovation & Research

Improving automatic speech recognition from distant microphones using self-attention

Innovation at Nuance | Employee Guest Blogger

August 31, 2021

Classical approaches to multi-channel microphone array signal processing such as MVDR or fixed-weight beamformers can be used to improve far-field speech recognition when integrated into the front-end signal-processing of a standard automatic speech recognition system. This research improves upon these standard approaches via a novel self-attention mechanism, replacing fixed signal-processing with a robust, learned approach that achieves state-of-the-art performance in natural conversational settings.

About the author: Rong Gong is a senior research scientist in Nuance Communications Austria. He works on the Audio Video Processing (AVP) team, developing multichannel far-field speech recognition technologies. His main research interests are speech enhancement and far-field speech recognition.

In Dragon Ambient eXperience (DAX) research, we utilize a microphone array device to capture far-field conversational speech between doctor and patient in the form of multichannel audio. We then obtain the medical transcription from the recorded audio by using a multichannel automatic speech recognition (ASR) system.

Recent research literature demonstrates accuracy benefits from jointly optimizing a multichannel frontend and an ASR backend together when doing far-field ASR. Most of these ASR frontends are based on the beamforming paradigm. In this approach, spatial information embedded in the multichannel audio is used to enhance the speech signal consumed by the ASR backend. Two common beamforming designs are MVDR (Minimum Variance Distortionless Response)-based design [1] and non-constrained design [2]. Another ASR frontend research approach relies on deep-learning attention mechanisms instead of beamforming. Attention mechanisms can be used to either select or combine audio channels from the multichannel signal [3].

In the paper "Self-attention channel combinator frontend for end-to-end multichannel far-field speech recognition", to be presented at Interspeech 2021, we explore a self-attention mechanism-based ASR frontend – Self-Attention Channel Combinator (SACC)- to generate channel combination weights, and to give predominance to those microphone channels that can lead to a better recognition accuracy. Our experiments conducted on a multichannel simulated conversational test data show that SACC achieves a 9.3% Word Error Rate Reduction (WERR) compared to a state-of-the-art beamforming frontend [2].

The Transformer neural network architecture is a proven approach to many sequence-learning applications, including Machine Translation [4], Automatic Speech Recognition [5] and in many areas of Natural Language Understanding. Rather than relying on recurrent neural architectures it depends on the self-attention mechanism which adopts a **query-key-value** concept. The attention weight matrix in this self-attention mechanism conceptually represents the similarity between sequence elements both within or between sequences. In this work, it is used in a similar way to model the similarity between microphone channels. Channel-combinator weights which specify the predominance of certain microphone channels are then calculated by a tensor product between the **value** and the attention weights. The final single channel representation is then computed as a weighted combination of the multichannel microphone array audio using these channel-combinator weights.

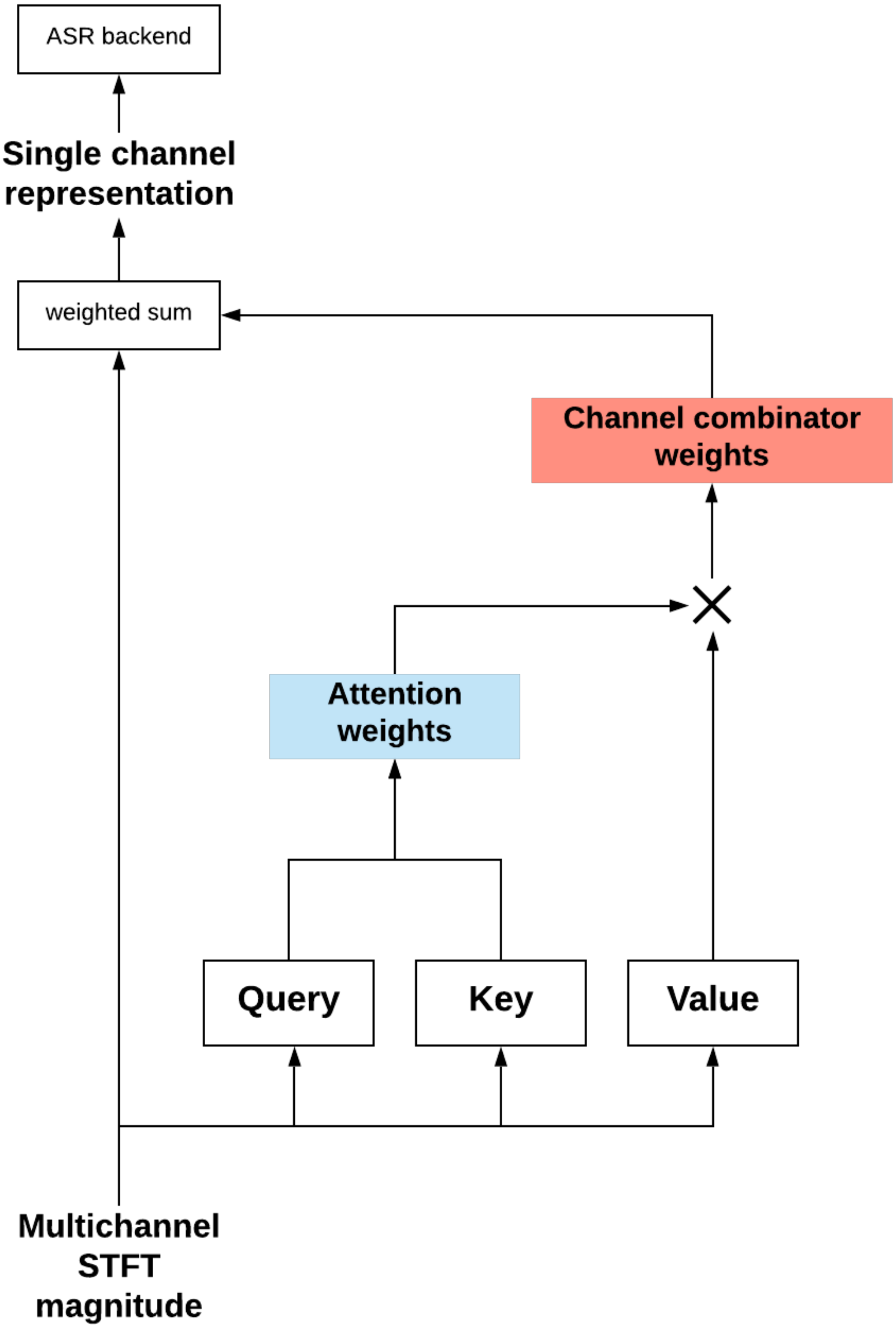
ASR backend
Single channel representation
weighted sum
Channel combinator weights
Attention weights
Query
Key
Value
Multichannel STFT magnitude

Figure 1: Simplified Self-Attention Channel Combinator (SACC) flowchart

In Figure 2 we show the channel combinator weights learned from jointly optimizing the SACC frontend and an ASR backend. The lower and upper figures represent the weights of two test examples of the same utterance, recorded by speakers located in two different positions. We can clearly observe that the learned weights vary depending on the position. In the upper figure, the 2nd and 3rd channels are weighted higher, and the 7th and 8th channels are weighted lower than the other channels. In the lower figure the 1st channel is weighted much lower than the other channels.

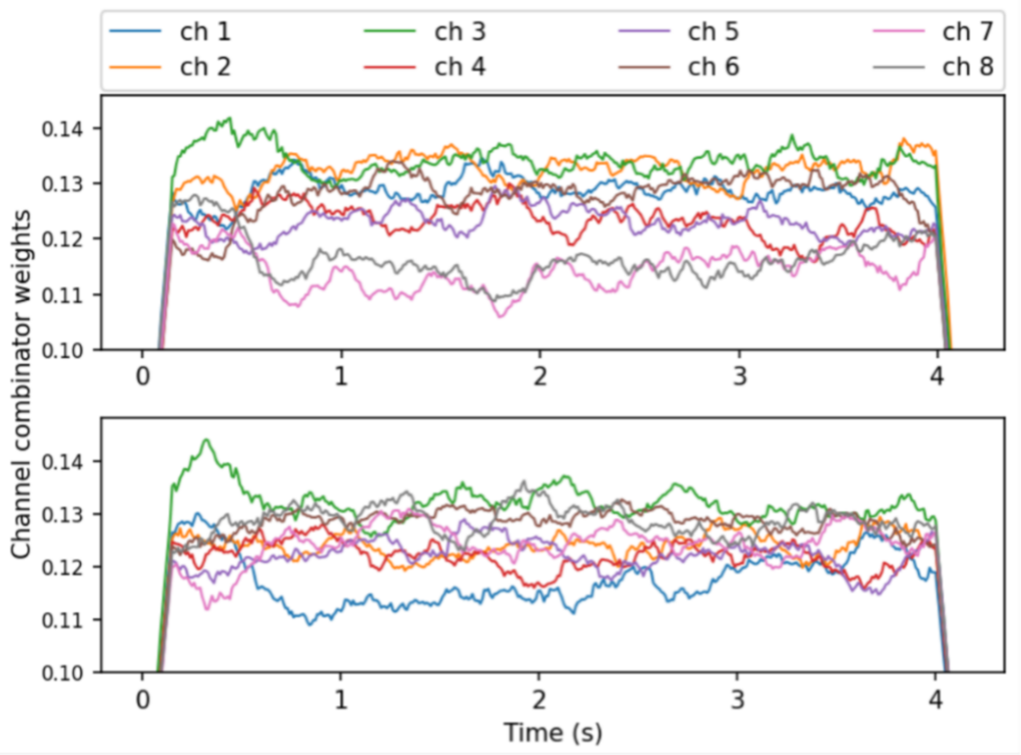

Figure 2: Channel combinator weights per channel.

We compare the SACC with four baselines: (1) single channel distant microphone (SDM), (2) random channel distant microphone (RDM), with a channel chosen at random per training utterance, (3) MVDR beamformer, (4) neural beamforming (NBF) [2], with fixed beamformer weights learned for 8 beam directions. The NBF and our SACC frontends are jointly optimized with the ASR backend. The ASR backend is an attention-based encoder-decoder (AED) system. The encoder is a variant of the ContextNet architecture [6], and the decoder is a single layer LSTM network.

The training data is based on a 460-hour subset of speech from Librispeech and the English partition of Mozilla Common Voice. The subset was selected based on various signal characteristics extracted using the NISA [7] algorithm. This base material was convolved with 8 channels room impulse responses (RIRs) generated using the Image method, followed by addition of ambient, babble and fan noise. The testing data is a set of speech material simultaneously played back and recorded in a meeting room. The room was setup with an 8-channel analog Uniform Linear Array with 33 mm inter-microphone spacing.

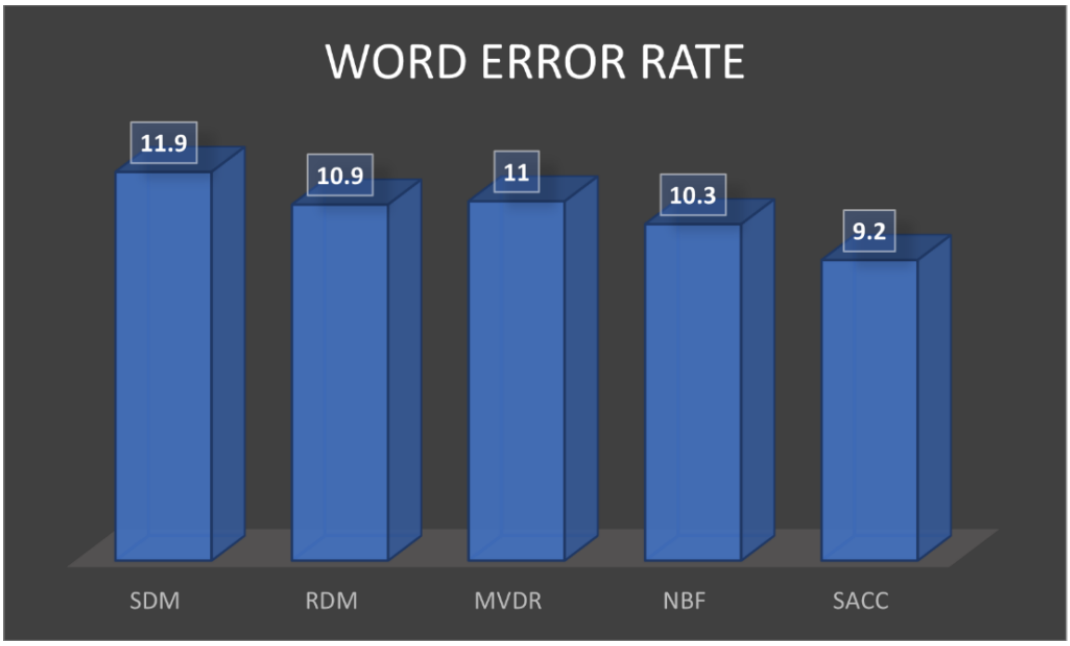

Figure 3: The Word Error Rate for the baselines and the SACC.

In conclusion, the SACC frontend achieves a 9.3% word error rate reduction (WERR) over NBF and represents a promising path towards building state-of-the-art multichannel ASR systems eschewing the beamforming paradigm.

References

[1] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. Xiao, "Unified Architecture for Multichannel End-to-End Speech Recognition With Neural Beamforming," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, Dec. 2017.

[2] W. Minhua, K. Kumatani, S. Sundaram, N. Strom, and B. Hoffmeister, "Frequency Domain Multi-Channel Acoustic Modeling for Distant Speech Recognition," in ICASSP 2019, May 2019.

[3] S. Braun, D. Neil, J. Anumula, E. Ceolini, and S.-C. Liu, "Multichannel Attention for End-to-End Speech Recognition," in Interspeech 2018, Sep. 2018.

[4] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[5] Dong, S. Xu, and B. Xu, "Speech-Transformer: A No Recurrence Sequence-to-Sequence Model for Speech Recognition," in ICASSP 2018, Apr. 2018.

[6] Han, Z. Zhang, Y. Zhang, J. Yu, C.-C. Chiu, J. Qin, A. Gulati, R. Pang, and Y. Wu, "ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context," in Interspeech 2020, Oct. 2020.

[7] D. Sharma, L. Berger, C. Quillen, and P. A. Naylor, "Non-intrusive estimation of speech signal parameters using a frame based machine learning approach," in Proc. of European Signal Processing Conference, Amsterdam, The Netherlands, 2020.

Tags: Dragon Ambient eXperience