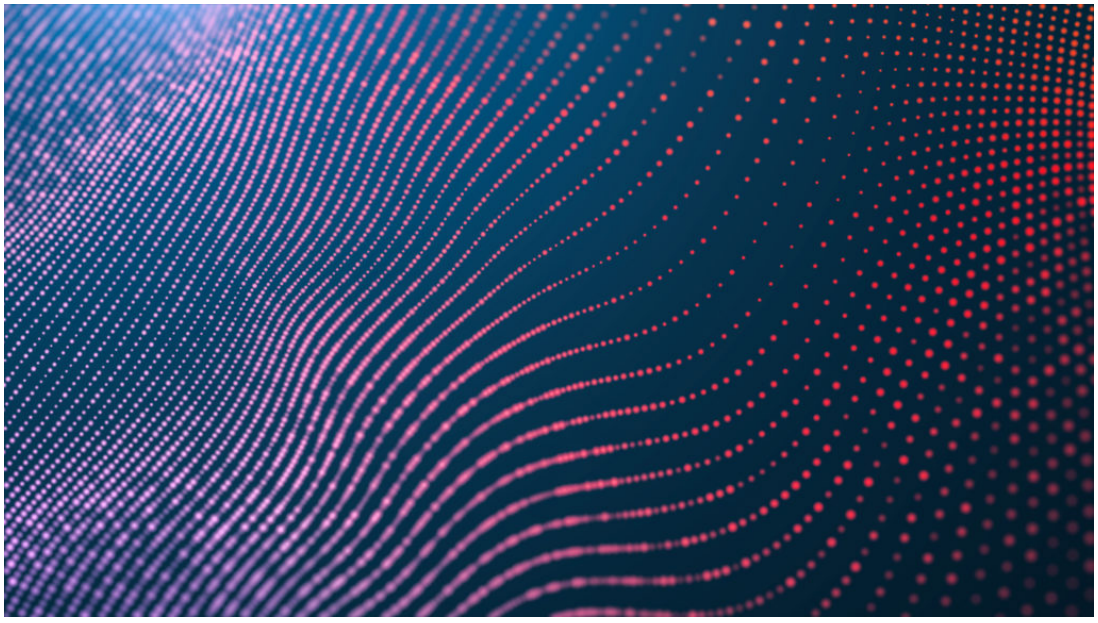
NUANCE | WHAT'S NEXT BLOG

Healthcare R&D, Innovation & Research

Improved far field speech recognition with microphone arrays

Innovation at Nuance | Employee Guest Blogger

May 2, 2022



The promise of far field speech recognition is a more natural human machine interaction. The ability for an intelligent, ambient device to be able to listen to a conversation taking place at a distance from the microphones is a challenging task due to the adverse effects of room reverberation, noise and the dynamic nature of the interaction (speaker movements). Moreover, in the ambient scenario, it is assumed that the speakers do not alter their speech articulation to accommodate a machine listener, as would be the case for voice search or dictation for example. We present a novel multi-channel front-end based on channel shortening with the Weighted Prediction Error (WPE) method followed by a fixed MVDR beamformer used in combination with a recently proposed self-attention-based channel combination (SACC) scheme, for tackling the distant ASR problem. We show that our proposed front-end used as part of an end-to-end (E2E) ASR system outperforms leading ASR systems resulting in 22% fewer errors.

About the author: Dushyant Sharma is a Principal Research Scientist, working as part of the Central Research and Development organization at Nuance on front end signal processing and acoustic data augmentation. His research interests include non-intrusive speech signal analysis, single and multi-channel speech enhancement and acoustic data augmentation and simulation. Dushyant joined Nuance in 2012 after completing his Ph.D. in speech signal processing from the Centre for Law Enforcement Audio Research (CLEAR) at Imperial College in London, UK.

The problem of Distant Automatic Speech Recognition (DASR) remains an important topic that has received much attention in the past few years. DASR represents an important component for enabling a truly natural and ambient intelligence system such as Nuance's Dragon Ambient eXperience (DAX). The current state of the art methods for DASR are based on multi-channel end-to-end (MCE2E) architectures where the front-end signal processing, acoustic modeling and language modeling are integrated into a single system that allows for a joint optimization of all sub-systems.

Current state of the art MCE2E DASR systems can be broadly grouped into three categories. The first are systems where the multi-channel front-end is based on a signal processing-based spatial filtering method with some elements of the algorithm being jointly estimated with the E2E ASR system. The second category of systems are based on neural filter and sum structures that directly optimize the ASR loss function. Finally, there are methods for speech dereverberation, which target reducing the level of reverberation in the signal. Since complete inversion of the acoustic transfer function is a challenging problem, many state of the art methods instead target channel shortening, whereby some amount of early reflections are kept intact and only the later reverberation is removed, retaining spatial localization information and thus enabling application of subsequent spatial processing algorithms.

Method

In our ICASSP 2022 paper, we propose a novel frontend based on dereverberation of microphone signals via a multi-channel Weighted Prediction Error (WPE)-based method followed by a fixed beamformer with 16 beams driving our recently proposed Self Attention Channel Combinator (SACC) for multi-channel signal combination using a self-attention mechanism.

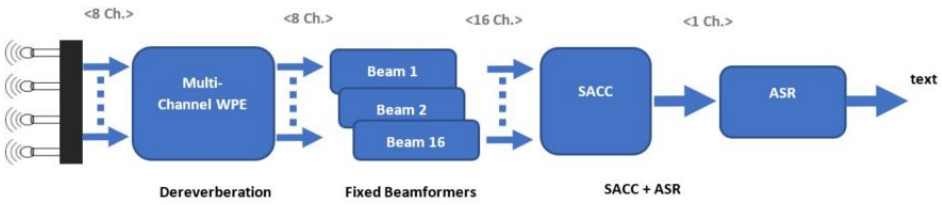


Figure 1 Outline of the proposed system.

The above figure shows the overall architecture of our proposed system. Audio form an 8-channel microphone array are first processed by an 8 channel WPE-based dereverberation algorithm to reduce the level of late reverberation in the signals. These are then processed by 16 fixed beamformers resulting in a 16 channel beamformed signal. The SACC component produces a single channel signal in the short time Fourier transform domain by taking an element-wise product and sum over the channel dimension of a weight matrix (computed via a scaled dot-product self-attention mechanism) and the normalized logarithmic power of multi-channel input. This is then used as input to an E2E ASR system and the SACC weights are learnt jointly with the ASR.

To train the ASR systems, a 460 hour multi-channel training dataset was created by convolving a clean subset of single channel speech from the LibriSpeech and Mozilla Common Voice datasets with simulated room impulse responses (RIRs) and additive ambient noise followed by a level augmentation. For test data, a set of clean speech utterances from the Librispeech test-clean partition were simultaneously played back through artificial mouth loudspeakers and recorded by a wall mounted 8 microphone uniform linear array. The playback and recording was performed from four positions in a typical office room (3m X 3.7m), simulating two pairs of conversation positions.

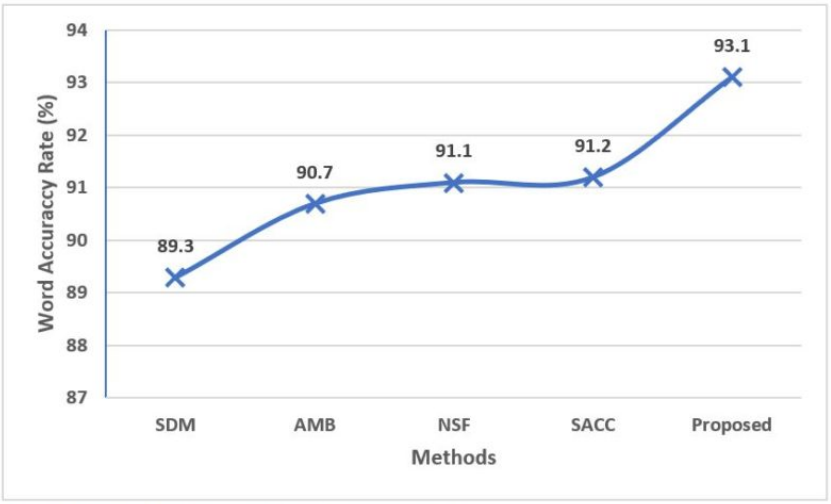


Figure 2 Results of the proposed and baseline methods.

The ASR test results for the different systems are presented in Figure 2, where it can be seen that our proposed system provides the highest word accuracy rate of 93.1%, nearly 36% fewer errors than the single distant microphone (SDM) baseline. Furthermore, as shown in Figure 3, the average SACC weights for a test utterance correlate with the source position. This is an added advantage of our proposed architecture, in that using the SACC weights, source location information may be inferred and used for other tasks such as speaker diarization.

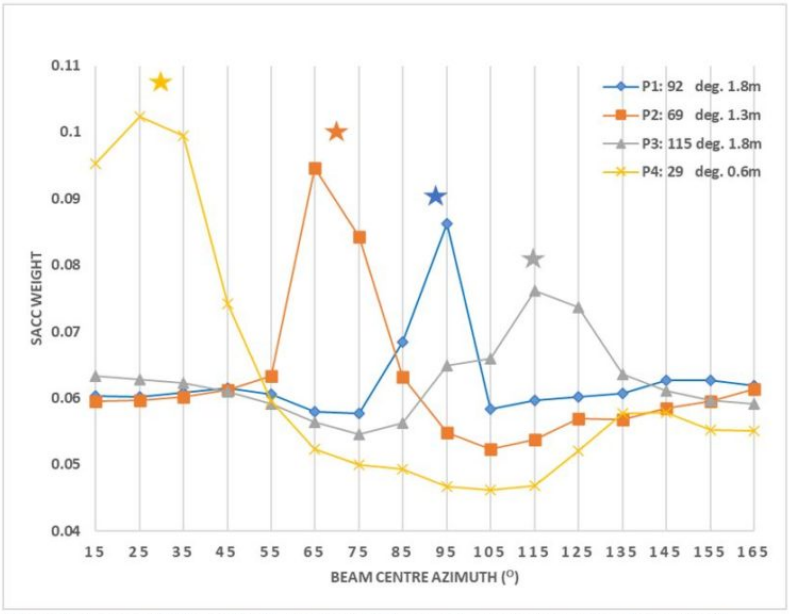


Figure 3. Average SACC weights for the test set from each playback position. Ground truth positions are marked by stars. Note that the x axis shows the 16 beam center azimuths.

In conclusion

we have shown how a novel multi-channel frontend, comprising a multi-channel dereverberation followed by a set of fixed beamformers, can be used in conjunction with a self-attention channel combinator-based E2E system to deliver state of the art far field ASR.

Tags: Speech recognition, Dragon Ambient eXperience, Research paper