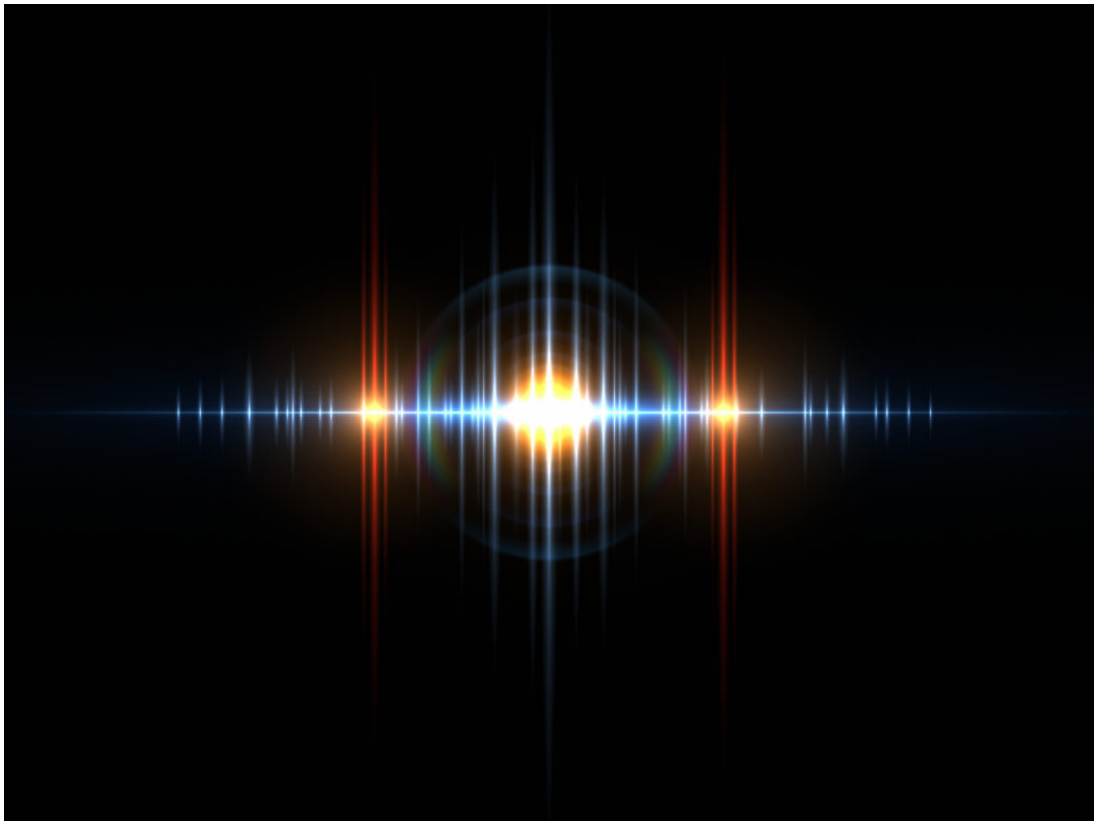
NUANCE | WHAT'S NEXT BLOG

Enterprise R&D, Innovation & Research

Data augmentation for real-world multi-channel speech recognition

Innovation at Nuance | Employee Guest Blogger

December 8, 2021



Data augmentation techniques are widely applied for training end-to-end models to maximize speech recognition accuracy. However, only a few studies so far have applied data augmentation to the multi-channel scenario. Our work proposes a new technique for making multi-channel models more robust to mismatched input audio conditions.

About the author: Marco Gaudesi is a Senior Research Scientist at Nuance Communications. He received his PhD degree in computer and control engineering from Politecnico di Torino, Italy, in 2015. His research interests include deep learning, speech recognition, and artificial intelligence. Felix Weninger, Dushyant Sharma, and Puming Zhan also contributed to the paper and this blog post. The paper was presented at the ASRU 2021 conference, December 2021.

Nowadays, the end-to-end modelling technique is widespread in the field of automatic speech recognition (ASR), with its main characteristic of being able to bring together all the parts in a single neural network: acoustic, pronunciation, and language modelling. To improve the audio quality in the input to the network,

multi-channel audio capture devices (microphone arrays) are used. The end-to-end multi-channel ASR system signal processing part often includes a beamformer to map the multi-channel signal to an enhanced single-channel speech signal. This can be realized through a neural beamforming technique jointly trained with the rest of the network.

Since neural networks require a large amount of data during the training process, data augmentation techniques are applied to generate a virtually infinite amount of data, by artificially perturbing the training examples. The most popular approach for data augmentation in end-to-end systems is SpecAugment, which randomly drops frequency bands and/or time frames from the time-frequency features (usually the Mel spectrogram).

In our paper entitled ChannelAugment: Improving generalization of multi-channel ASR by training with input channel randomization, we present a new data augmentation technique designed for multi-channel ASR: it randomly varies the set of input channels during training to improve the generalization to different microphone array configurations. Furthermore, in some neural beamforming configurations, such as the neural Minimum Variance Distortionless Response (MVDR), this technique can also reduce training time, computational cost and memory consumption required for the end-to-end ASR system. Our paper is accepted for presentation at the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2021.

End-to-end multi-channel ASR systems achieve state-of-the-art performance by joint training of a multi-channel front-end together with the ASR model. One of the limitations of such systems is that they are usually trained with audio data captured with a fixed array geometry; the performance of those systems is optimal when they operate in matched condition (i.e., by using input data captured from a device similar to the one used to capture training data), but it can suffer from a degradation in accuracy when input data are taken from a different device configuration. From that, the need to train different models arises, which leads to an increase in training cost and deployment. To address this limitation, we introduce the ChannelAugment technique for on-the-fly data augmentation in end-to-end multi-channel ASR training; by randomly masking some channels in the input, this technique allows the trained model to be more robust when operating in mismatched conditions in real-world applications and increases the training efficiency.

The flowchart is shown below:

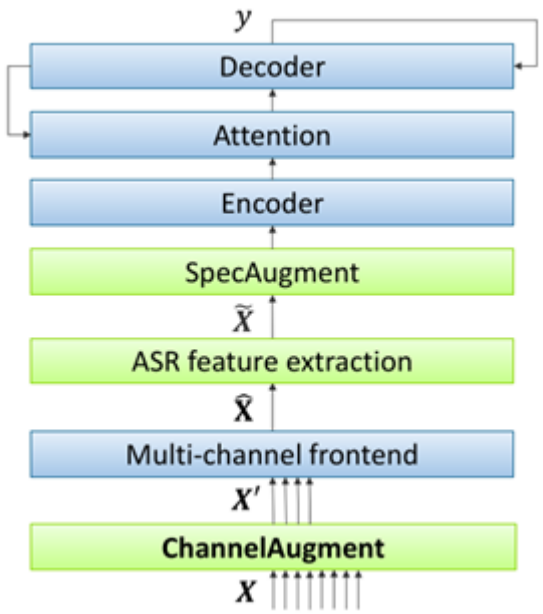


Figure 1 – Flowchart of the proposed ChannelAugment technique, applied during training to an ASR end-to-end model, together with SpecAugment data augmentation.

Two different channel masking approaches were developed for the proposed technique, which works in the complex spectrum domain:

- Frequency independent: randomly selects and masks some input channels entirely in an input utterance; it is parametrized by a minimum and a maximum number of channels to keep (see Figure 2.a).
- Frequency dependent: masks a different set of channels for each frequency of the input utterance; it is parametrized by a single value indicating the probability of keeping a channel for each frequency (see Figure 2.b).

In both the approaches, only the selected channels are passed to the multi-channel frontend (that contains the neural beamformer).

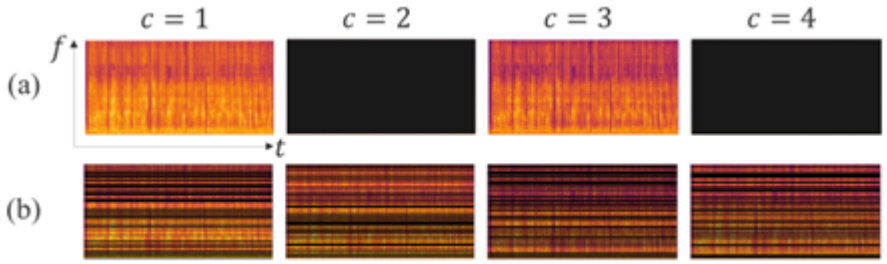


Figure 2 – Visualization of ChannelAugment on 4-channel speech signal (in black the masked data): (a) frequency independent masking, (b) frequency dependent masking

To validate our approach, we generated 460 hours of 16-channel training data by starting from the clean partition of the Librispeech and the English partition of the Mozilla Common Voice datasets, by simulating typical meeting room dimension and several random positions for the linear array microphone. For creating the test data, a subset of the clean test partition of Librispeech dataset was played back in a real room through an artificial mouth loudspeaker placed in two positions with different distance and angle with respect to the capture device and recorded through a linear microphone array. Two different kinds of neural beamforming were used: Spatial Filtering (SF) and MVDR.

To demonstrate the robustness to variations of the array geometry, we varied the number of microphones from 2 to 16 and the microphone spacing from 33 mm to 495 mm. Figure 3 shows the seven array configurations used in testing.

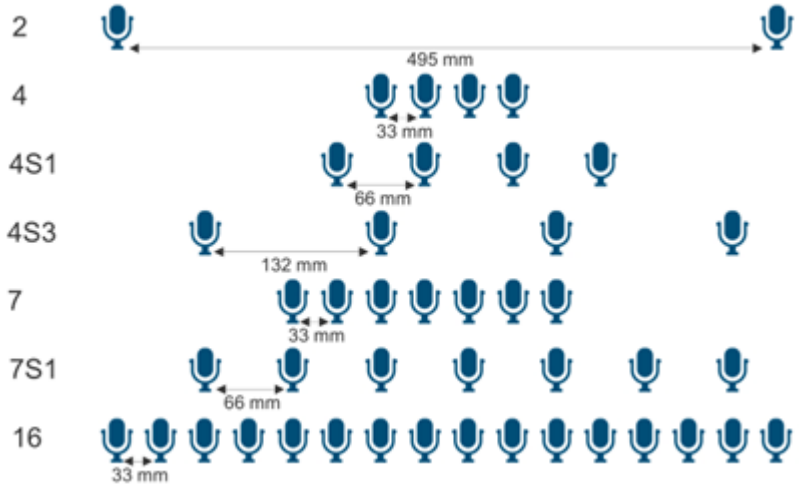


Figure 3 – Linear array configurations used in testing

When the SF beamformer is used, we found that by applying the Frequency Independent ChannelAugment technique it is possible to slightly improve the accuracy in matched condition by up to 2% in terms of relative word error rate reduction (WERR), while the robustness in several different array configurations improves by up to 10.6% WERR on average. Similar results were achieved by using the second set of test data, captured with a different microphone position; in this way we demonstrate that the robustness is independent from the device position. By applying the Frequency Dependent ChannelAugment with channel keep probability of 25%, we obtained the same result as the Frequency Independent ChannelAugment by keeping 4 channels when both models were tested on the configuration "4" in Figure 3.

When the MVDR beamformer is used, the model trained by enabling the Frequency Independent ChannelAugment technique by selecting only 4 random channels out of 16 performs similarly to the model trained with all the 16-channel input on average across test conditions (matched and mismatched). This yields a 75% reduction in training time, because unselected channels are simply dropped from Neural MVDR input, as opposed to masking them in the Spatial Filtering beamformer.

In summary, the ChannelAugment technique can be easily applied to multi-channel end-to-end ASR model during training, improving robustness to microphone array geometry variation and training efficiency and facilitating the deployment of ASR systems in real-world settings.

Tags: Speech recognition