Motivation & Effects of data augmentation techniques

Time Mask

Goal : Make the model more robust to **partial loss of small segments of speech**

Principle : Mask **certain time ranges** with the mean value of the spectrogram or zero (we used zero here) :
1. Draw a random moment
`t` in the time sample
2. Draw a random mask range
`t_mask` with `T` the maximum value
3. The mask is full of 1, with only 0 in the time range of
`[t, t+t_mask]`

Disclaimer : if the time range between `t` and the end of the time sample is less than `t_mask`, the mask will not cover a range of `t_mask`

Frequency Mask

Goal : Make the model more robust to **partial loss of frequency information**

Principle : Mask **certain frequency bands** with either the mean value of the spectrogram or zero (we used zero here) :
1. Draw a random frequency
`f` in the frequency range of the sample
2. Draw a random mask range
`f_mask` with `F` the maximum value
3. The mask is full of 1, with only 0 in the frequency range of
`[f, f+f_mask]`

Disclaimer : if the frequency range between `f` and the maximum frequency is less than `f_mask`, the mask will not cover a range of `f_mask`

Time Shift

Goal : Make the model more robust to delay in the audio (the signal is not centered in the time range)

Principle :
1. Draw a random number of values
`shift` to be shifted
2. Shift
`shift` values to the right (the values shifted out of the range of the sample are re-injected to the left/at the beginning of the sample)

Disclaimer : The shift can cut the speech command in half, and reverse the order, causing the signal to be nonsensical (ex : “Yes” will be heard “Sye”)