[Preprocessing and feature extraction](https://review.mlplatform.org/plugins/gitiles/ml/ethos-u/ml-embedded-evaluation-kit/+/c930ad9dc189d831ac77f716df288f70178d4c10/docs/use_cases/kws.md" \l "preprocessing-and-feature-extraction)

The DS-CNN keyword spotting model that is supplied with the Code Samples expects audio data to be preprocessed in a specific way before performing an inference. This section aims to provide an overview of the feature extraction process used.

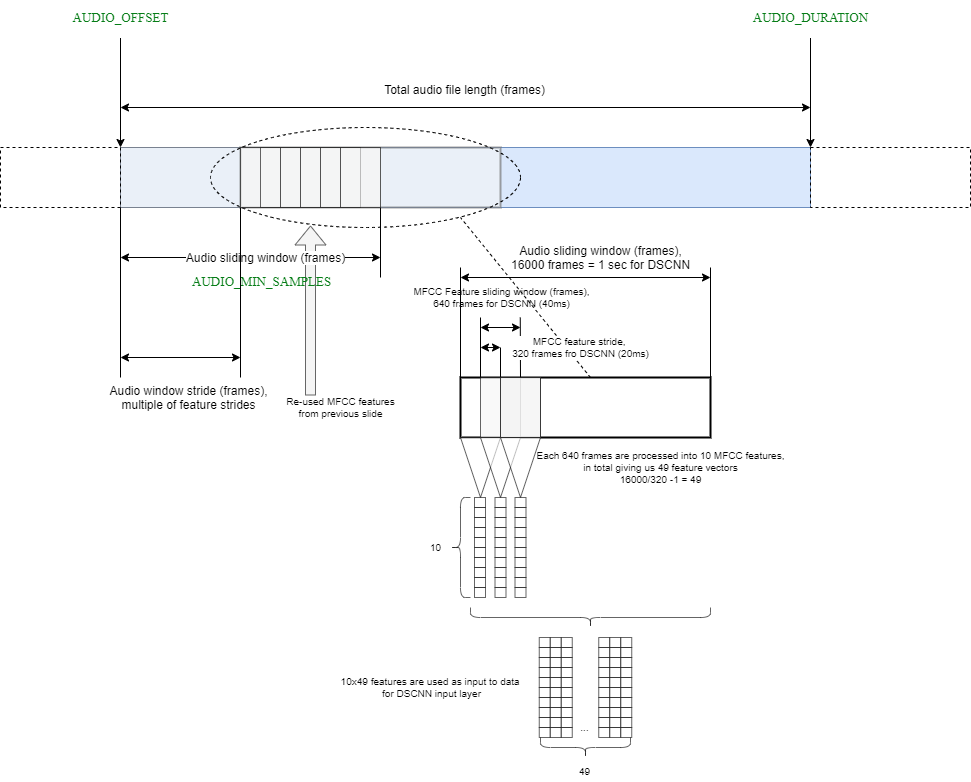
First the audio data is normalized to the range (-1, 1).

**Note:** Mel-frequency cepstral coefficients (MFCCs) are a common feature extracted from audio data and can be used as input for machine learning tasks like keyword spotting and speech recognition. See source/application/main/include/Mfcc.hpp for implementation details.

Next, a window of 640 audio samples is taken from the start of the audio clip. From these 640 samples we calculate 10 MFCC features.

The whole window is shifted to the right by 320 audio samples and 10 new MFCC features are calculated. This process of shifting and calculating is repeated until the end of the 16000 audio samples needed to perform an inference is reached. In total this will be 49 windows that each have 10 MFCC features calculated for them, giving an input tensor of shape 49x10.

These extracted features are quantized, and an inference is performed.



If the audio clip is longer than 16000 audio samples then the initial starting position is offset by 16000/2 = 8000 audio samples. From this new starting point, MFCC features for the next 16000 audio samples are calculated and another inference is performed (i.e. do an inference for samples 8000-24000).

**Note:** Parameters of the MFCC feature extraction such as window size, stride, number of features etc. all depend on what was used during model training. These values are specific to each model and if you try a different keyword spotting model that uses MFCC input then values are likely to need changing to match the new model. In addition, MFCC feature extraction methods can vary slightly with different normalization methods or scaling etc. being used.

Postprocessing

After an inference is complete the highest probability detected word is output to console, providing its probability is larger than a threshold value (default 0.9).

If multiple inferences are performed for an audio clip, then multiple results will be output.