Pretreatments and data augmentation

▼ Data augmentation definitions

Goal :

Enhance the performance and robustness of the models (if we don't have a lot of data, it could help the model to avoid overfitting)

Steps :

1. **Preprocess the audio data (Pretreatments):** Before applying any augmentations, preprocess the audio data to ensure consistency and quality.
2. **Choose augmentation techniques :** Select augmentation techniques suitable for speech data. Common techniques include:
    - **Speed Perturbation:** Speed up or slow down the audio without changing the pitch. This can simulate variations in speaking rate.
    - **Pitch Shifting:** Alter the pitch of the speech while maintaining the duration. This can simulate variations in voice pitch.
    - **Noise Injection:** Add background noise to simulate different environmental conditions, such as street noise, office noise, or crowd noise.
    - **Time Stretching/Compression:** Stretch or compress the duration of the speech without changing the pitch. This can simulate variations in speech duration.
    - **Room Reverberation:** Add reverberation effects to simulate different room acoustics. This can include variations in room size, echo levels, and reflection characteristics.
    - **Dynamic Range Compression:** Apply compression to the audio to reduce the dynamic range, simulating different recording conditions or microphone characteristics.
    - **Clipping:** Introduce clipping distortion by clipping the audio waveform at specific levels. This can simulate low-quality recording equipment or transmission artifacts.
    - **Bandpass Filtering:** Apply bandpass filtering to emphasize or suppress specific frequency bands in the audio. This can simulate variations in microphone characteristics or transmission channels.
    - **Data Dropout:** Randomly remove short segments of the audio waveform to simulate dropout or missing data. This can help the model learn to handle missing or corrupted segments.
    - **Time Warping:** Apply time warping to stretch or compress different parts of the audio waveform independently. This can introduce subtle temporal variations while preserving the overall structure of the speech.
    - **Mixing Speech with Background Sounds:** Combine speech with various background sounds, such as music, environmental sounds, or other speech segments. This can simulate complex acoustic environments or overlapping speech.
3. **Implement augmentation pipeline:** Develop a pipeline to apply the chosen augmentation techniques to the audio data. This pipeline should be flexible and allow for easy customization and combination of different augmentation methods.
4. **Apply augmentation:** Apply the augmentation techniques to the audio data. This may involve generating new augmented versions of each original audio recording by applying one or more augmentation methods with varying parameters.

5. **Quality control:** After augmenting the data, perform quality control checks to ensure that the augmented data remains intelligible and retains its semantic content. Listen to a sample of the augmented recordings to verify their quality and suitability for training purposes.

6. **Create augmented dataset:** Once the augmented data has been generated and quality-checked, combine it with the original dataset to create an augmented dataset. Ensure that the augmented dataset maintains a balance between the original and augmented samples to prevent bias.

7. **Train machine learning models:** Use the augmented dataset to train machine learning models for tasks.

▼ Methodology

	Doc1	Doc2	Doc3	Doc4
Dataset	Google dataset v1 & v2	AudioSet	Google dataset v1 & v2	Google dataset v1 & v2
Length of sample	1s		1s	
Sampling rate	16kHz			
Pretreatments			40 MFCC features from a speech from of length 40ms with a stride of 20ms	* 64 MFCC from 25ms windows with a 10ms overlap * Symmetric padding of the temporal dimension with zeros to fixed length of 128 features vectors per sample
Training data augmentation	* time shift in range -100ms... 100ms * signal resampling with resampling factor in range 0.85...1.15 * background noise * frequency/time masking, based on SpecAugment	* mixup ratio=0.5 * spectrogram masking with max time mask length = 192 frames ; max frequency mask length = 48 bins	* background noise * random time shift of up to 100ms	* time shift perturbations in the range of [-5,5] ms * white noise with magnitude [-90,-46] dB * SpecAugment with 2 continuous time mask of size [0, 25] time steps ; and 2 continuous frequency mask of size [0, 15] frequency bands