AlterDub – Conceptual Training Data Flow

This document provides a clean, architecture-level understanding of how audio data flows from raw WAV files to training-ready tensors in AlterDub, focusing on Modules 2, 3, and 4. It is intended as a visual and mental reference.

1. Raw Audio to Clips

Long recordings are first segmented into manageable audio clips, typically between 3 to 8 seconds. Each clip represents one training example candidate.

2. Framing (Time Axis Creation)

Each audio clip is divided into overlapping frames using a fixed hop size (e.g., 10 ms). This step introduces the time axis (T). The total number of frames T depends on clip duration and hop size.

3. Feature Extraction (Module 2)

For every frame in a clip, two primary features are computed:

- **Mel Spectrogram:** A time-frequency representation capturing phonetic content, formants, timbre, and articulation. Stored as a 2D structure [T, Mel_Bins].
- **Pitch (F0):** One fundamental frequency value per frame, capturing intonation and prosody. Stored as a 1D structure [T].

Each clip produces one mel file and one pitch file, both already containing all frames for that clip.

4. Dataset Indexing (Module 3)

Each clip is assigned a speaker ID. Metadata records link together mel path, pitch path, speaker ID, and frame length (T). At this point, one clip equals one sample.

5. VCDataset (Module 4)

VCDataset defines what one training sample is. It loads all features for a single clip and returns mel, pitch, speaker ID, and length as structured numerical containers.

6. Batching & T_max (Collate Function)

Training uses batches (groups of samples). Because clips have different lengths, the collate function pads samples to match the longest sequence in the batch, called T_max.

- Batch size (B): Number of samples grouped together (e.g., B = 4).
- T_max: Maximum frame length among samples in the batch.
- Mask: Identifies real frames versus padded frames.

The resulting tensors have shapes like: mel [B, T_max, 80], pitch [B, T_max], speaker_id [B].

7. Model Training

The model consumes these batched tensors. Masks ensure that learning happens only on real frames, not padding. This standardized tensor flow enables efficient GPU training and stable convergence.

Key Mental Model: One audio clip equals one sample. A batch is multiple samples stacked along a batch dimension. T_max exists to make variable-length speech trainable.