```
                              ┌─────────────────┐
                              │   Audio Input   │
                              └─────────────────┘

        ┌──────────────────┐              ┌──────────────────┐
        │  Mel Spectrogram │              │       MFCC       │
        │   128 mel bands  │              │  13 coefficients │
        │   Log transform  │              │                  │
        └──────────────────┘              └──────────────────┘

              ┌──────────────────────────────────────────┐
              │   Concatenate Features (141 channels)     │
              └──────────────────────────────────────────┘

    ┌──────────────────────────────────────────────────────────────┐
    │               Multi-Scale Feature Extractor                   │
    │  ┌────────┐  ┌────────┐  ┌────────┐  ┌────────┐                │
    │  │ Conv1D │  │ Conv1D │  │ Conv1D │  │ Conv1D │  → Concat (256)│
    │  │  k=3   │  │  k=5   │  │  k=7   │  │  k=11  │                │
    │  └────────┘  └────────┘  └────────┘  └────────┘                │
    └──────────────────────────────────────────────────────────────┘

    ┌──────────────────────────────────────────────────────────────┐
    │                     Residual Blocks                           │
    │  ┌─────────────┐  ┌─────────────┐  ┌─────────────┐            │
    │  │ ResBlock 256│  │ ResBlock 256│  │ResBlock 512↓│            │
    │  └─────────────┘  └─────────────┘  └─────────────┘            │
    │  ┌─────────────┐  ┌─────────────┐  ┌──────────────┐           │
    │  │ ResBlock 512│  │ ResBlock 512│  │ResBlock 1024↓│           │
    │  └─────────────┘  └─────────────┘  └──────────────┘           │
    │  ┌──────────────┐  ┌──────────────┐                           │
    │  │ResBlock 1024 │  │ResBlock 1024 │                           │
    │  └──────────────┘  └──────────────┘                           │
    └──────────────────────────────────────────────────────────────┘
```

Audio Input

Mel Spectrogram — 128 mel bands, Log transform

MFCC — 13 coefficients

Concatenate Features (141 channels)

Multi-Scale Feature Extractor
- Conv1D k=3
- Conv1D k=5
- Conv1D k=7
- Conv1D k=11
- → Concat (256)

Residual Blocks
- ResBlock 256
- ResBlock 256
- ResBlock 512↓
- ResBlock 512
- ResBlock 512
- ResBlock 1024↓
- ResBlock 1024
- ResBlock 1024

Multi-Head Self-Attention
8 heads, 1024 dimensions

Q
K
V

self-attention

Global Pooling
- Avg Pool
- Max Pool
- → Concat (2048)

Classification Head
- Dropout(0.5) → Linear(2048→512) → ReLU
- Dropout(0.3) → Linear(512→256) → ReLU
- Dropout(0.2) → Linear(256→5)