White Paper: Audio AI – Models, Architectures, and Use Cases

# 1. Introduction

Audio AI is revolutionizing how machines perceive and interact with human audio signals, including speech, music, and environmental sounds. This white paper presents a categorized overview of audio AI models, their architectural flows, and common use cases.

# 2. Automatic Speech Recognition (ASR)

Models:  
- Whisper (Transformer): Multilingual ASR and translation.  
- Wav2Vec 2.0 (CNN + Transformer): Self-supervised ASR.  
- DeepSpeech (RNN): Lightweight ASR.  
- Conformer (CNN + Transformer): High-accuracy ASR.  
- Jasper / QuartzNet (1D CNN): Efficient, fast inference.  
  
Architecture:  
Audio → Feature Extractor (CNN) → Transformer or RNN → Text Output (CTC or Decoder)  
Use Cases: Voice commands, captioning, transcription.Architecture Flow:  
Raw Audio → (Optional CNN Feature Extractor) → Transformer Encoder or RNN → Decoder or CTC → Text Output  
  
- Whisper: Log-Mel Spectrogram → Transformer Encoder → Transformer Decoder → Text  
- Wav2Vec 2.0: Raw Audio → CNN Encoder → Transformer → CTC Decoder → Text  
- DeepSpeech: Raw Audio → RNN (GRU) → CTC → Text  
- Conformer: Raw Audio → CNN + Transformer → CTC/Decoder → Text

# 3. Speaker Recognition (Verification & Diarization)

Models:  
- ECAPA-TDNN: Voice authentication and diarization.  
- x-vector: Speaker diarization and clustering.  
- Resemblyzer: Speaker embedding for similarity tracking.  
  
Architecture:  
Audio → DNN → Temporal Aggregation → Speaker Embedding → Similarity Calculation  
Use Cases: Voice login, meeting transcription, speaker tracking.Architecture Flow:  
Raw Audio → DNN (TDNN / LSTM) → Temporal Pooling → Embedding Vector → Cosine Similarity or Clustering  
  
- ECAPA-TDNN: Audio → TDNN + Attention → Embedding → Cosine Similarity  
- x-vector: Audio → Frame-Level DNN → Stats Pooling → Embedding → Clustering

# 4. Audio Classification / Event Detection

Models:  
- PANNs: Audio tagging and sound event detection.  
- YAMNet: MobileNet for on-device classification.  
- VGGish / OpenL3: Audio embedding and tagging.  
  
Architecture:  
Audio → Mel Spectrogram → CNN → Dense Classifier → Output Labels  
Use Cases: Surveillance, music genre detection, acoustic monitoring.Architecture Flow:  
Audio → Mel Spectrogram → CNN → Dense Classifier → Predicted Labels  
  
- PANNs: Spectrogram → CNN14 → Multi-label Classifier  
- YAMNet: Spectrogram → MobileNet → Labels

# 5. Text-to-Speech (TTS)

Models:  
- Tacotron 2: Seq2Seq + Attention.  
- FastSpeech 2: Non-autoregressive and fast.  
- VITS / Glow-TTS: Variational/GAN-based synthesis.  
- HiFi-GAN / MelGAN: Real-time neural vocoders.  
  
Architecture:  
Text → Encoder → Decoder → Mel Spectrogram → Vocoder → Audio  
Use Cases: Voice synthesis, audiobooks, accessibility.Architecture Flow:  
Text → Encoder → (Attention or Duration Predictor) → Decoder → Mel Spectrogram → Neural Vocoder → Audio  
  
- Tacotron2: Text → Encoder → Attention Decoder → Mel → WaveNet → Audio  
- FastSpeech2: Text → Duration Predictor → Mel → HiFi-GAN → Audio  
- VITS: Text → Variational Encoder + GAN → Audio

# 6. Speech Enhancement / Denoising

Models:  
- Demucs: U-Net with LSTM for music/speech.  
- SEGAN / VoiceFixer: GAN-based denoising.  
- DCCRN: RNN on complex spectrograms.  
  
Architecture:  
Noisy Audio → Feature Encoder → Denoising Model → Clean Audio Output  
Use Cases: VoIP, call centers, noisy recordings.Architecture Flow:  
Noisy Audio → CNN / RNN / GAN → Denoised Audio Output  
  
- Demucs: Audio → U-Net with LSTM → Clean Audio  
- SEGAN: Audio → Generator (GAN) → Clean Audio

# 7. Audio Generation / Diffusion

Models:  
- DiffWave: Diffusion model for audio synthesis.  
- AudioLM / MusicLM: Text-to-audio/music generation.  
- SoundStream: Audio compression and generation.  
  
Architecture:  
Random Noise → Denoising U-Net or Transformer → Generated Audio  
Use Cases: AI music, generative art, sound design.Architecture Flow:  
Gaussian Noise → Denoising Transformer / U-Net → Audio Waveform  
  
- DiffWave: Noise → U-Net → Audio  
- AudioLM: Audio → Quantization + Transformer → Regenerated Audio

# 8. Audio Embedding & Representation Learning

Models:  
- HuBERT / TERA: Transformer-based self-supervised learning.  
- BYOL-A / CLAP: Contrastive audio representation.  
  
Architecture:  
Raw Audio → CNN → Transformer → Embedding → Downstream Task  
Use Cases: Search, retrieval, multi-modal AI.Architecture Flow:  
Raw Audio → Feature Encoder (CNN) → Transformer → Embeddings → Downstream Tasks  
  
- HuBERT: Audio → CNN → Transformer → Embeddings  
- CLAP: Audio + Text → Contrastive Transformer → Aligned Embeddings

# 9. Tools & Libraries

- torchaudio: Preprocessing, models in PyTorch.  
- librosa: Feature extraction and visualization.  
- SpeechBrain / ESPnet / WeNet: Full pipeline toolkits.  
- OpenVINO / ONNX: Model optimization.  
- Audacity / SoX: Preprocessing and editing.

# 10. Conclusion

The rapid progress in audio AI has led to the development of diverse and powerful models for speech, music, and sound. With open-source toolkits and modular architectures, these models are now deployable in real-time and edge applications. The future of audio AI lies in multimodal fusion, low-resource adaptation, and real-time performance.