WHITE PAPER

Ambient Audio Processing: From Sound to Structured Note

How raw audio becomes a structured clinical document — speaker diarisation, medical NER, section classification, and note assembly.

Published:	August 2025
Document ID:	WP-2025-006
Classification:	Public
Author:	ClinixSummary Research Team

1. Executive Summary

The transformation of a live clinical conversation into a structured, accurate clinical note is the core technical challenge in ambient AI documentation. This white paper traces the complete processing pipeline from the moment audio enters the system to the final structured note, explaining each processing stage and the clinical intelligence that makes it possible.

2. The Processing Pipeline

Ambient audio processing in ClinixSummary follows a six-stage pipeline, each stage building on the outputs of the previous one:

- Stage 1: Audio Capture & Preprocessing
- Stage 2: Speech Recognition & Transcription
- Stage 3: Speaker Diarisation
- Stage 4: Medical Named Entity Recognition (NER)
- Stage 5: Section Classification & Clinical Reasoning
- Stage 6: Document Assembly & Formatting

Total end-to-end latency from final audio segment to complete note is typically under 30 seconds for a standard 15-minute consultation.

3. Stage 1: Audio Capture & Preprocessing

Raw audio captured from the clinician's device undergoes immediate preprocessing:

- Adaptive noise reduction tuned for clinical environments (keyboard typing, door sounds, equipment beeps, HVAC).
- Automatic gain control to normalise volume differences between speakers at varying distances from the microphone.
- Voice Activity Detection (VAD) to segment audio into speech and non-speech regions, reducing unnecessary processing.
- Audio quality assessment to flag recordings that may produce unreliable transcriptions (excessive noise, very low volume).

4. Stage 2: Speech Recognition

Preprocessed audio is streamed to the ASR engine in real time. The medical ASR model produces a preliminary transcript with word-level timestamps and confidence scores. Low-confidence segments are flagged for downstream verification against clinical context.

Key differentiators of ClinixSummary's medical ASR:

- 250,000+ medical term vocabulary with specialty-specific pronunciation variants.

- Contextual language model that biases recognition toward clinically probable terms (e.g., ‘dyspnoea’ over ‘Disney’ in a respiratory consultation).
- Real-time code-switching support for multilingual consultations.
- Robust handling of medical abbreviations, acronyms, and numeric expressions (dosages, lab values, vital signs).

5. Stage 3: Speaker Diarisation

Speaker diarisation identifies who said what — critical for accurate clinical documentation. ClinixSummary’s diarisation system:

- Supports 2-4 speakers per session (clinician, patient, family member, interpreter).
- Uses voice embedding models trained on clinical dialogue to distinguish speakers based on vocal characteristics.
- Assigns clinical roles (clinician vs. patient) based on speech patterns, terminology usage, and conversational dynamics.
- Handles overlapping speech and interruptions common in clinical consultations.

6. Stage 4: Medical Named Entity Recognition

A specialised NER model extracts clinical entities from the diarised transcript:

- Diagnoses and conditions (mapped to ICD-10).
- Medications with dosage, route, frequency, and duration.
- Procedures and interventions (mapped to CPT where applicable).
- Anatomical references and laterality.
- Laboratory values with units and reference ranges.
- Temporal markers (onset, duration, frequency of symptoms).
- Negation detection (‘no chest pain’, ‘denies fever’).

7. Stage 5: Section Classification & Clinical Reasoning

Extracted entities and transcript segments are classified into clinical note sections. ClinixSummary’s section classifier goes beyond simple keyword matching — it understands clinical conversational flow:

For example, when a patient says ‘I’ve been having this pain in my chest for about two weeks, it gets worse when I lie down,’ the system recognises this as History of Present Illness content, extracts ‘chest pain’ as a symptom with ‘two weeks’ duration and ‘worse when supine’ as a modifying factor — even though the patient never used those clinical terms.

8. Stage 6: Document Assembly

The final stage assembles classified, entity-rich content into a structured clinical document:

- Template selection based on specialty, note type, and clinician preferences.
- Section ordering per clinical documentation standards.
- Clinical language normalisation (converting colloquial patient language to appropriate clinical terminology while preserving meaning).
- Completeness checking — flagging potentially missing sections or information.
- Code suggestions (ICD-10, CPT) based on extracted clinical entities.

9. Performance Characteristics

Typical performance metrics for a 15-minute general practice consultation:

- Audio-to-transcript latency: <500ms (streaming).
- End-to-end processing: <30 seconds from final audio to complete note.
- Speaker diarisation accuracy: >95% on 2-speaker consultations.
- Medical NER F1 score: >0.93 across core entity types.
- Section classification accuracy: >91% on standard SOAP sections.

10. Conclusion

The journey from ambient audio to structured clinical note is a complex, multi-stage process that requires deep integration of speech technology, medical NLP, and clinical domain knowledge. ClinixSummary's purpose-built pipeline handles this complexity invisibly, allowing clinicians to focus on their patients while the system handles the documentation.