ỨNG DỤNG NHẬN DẠNG & TẠO BIÊN BẢN CUỘC HỌP (MEETING STT)

1. Tổng quan

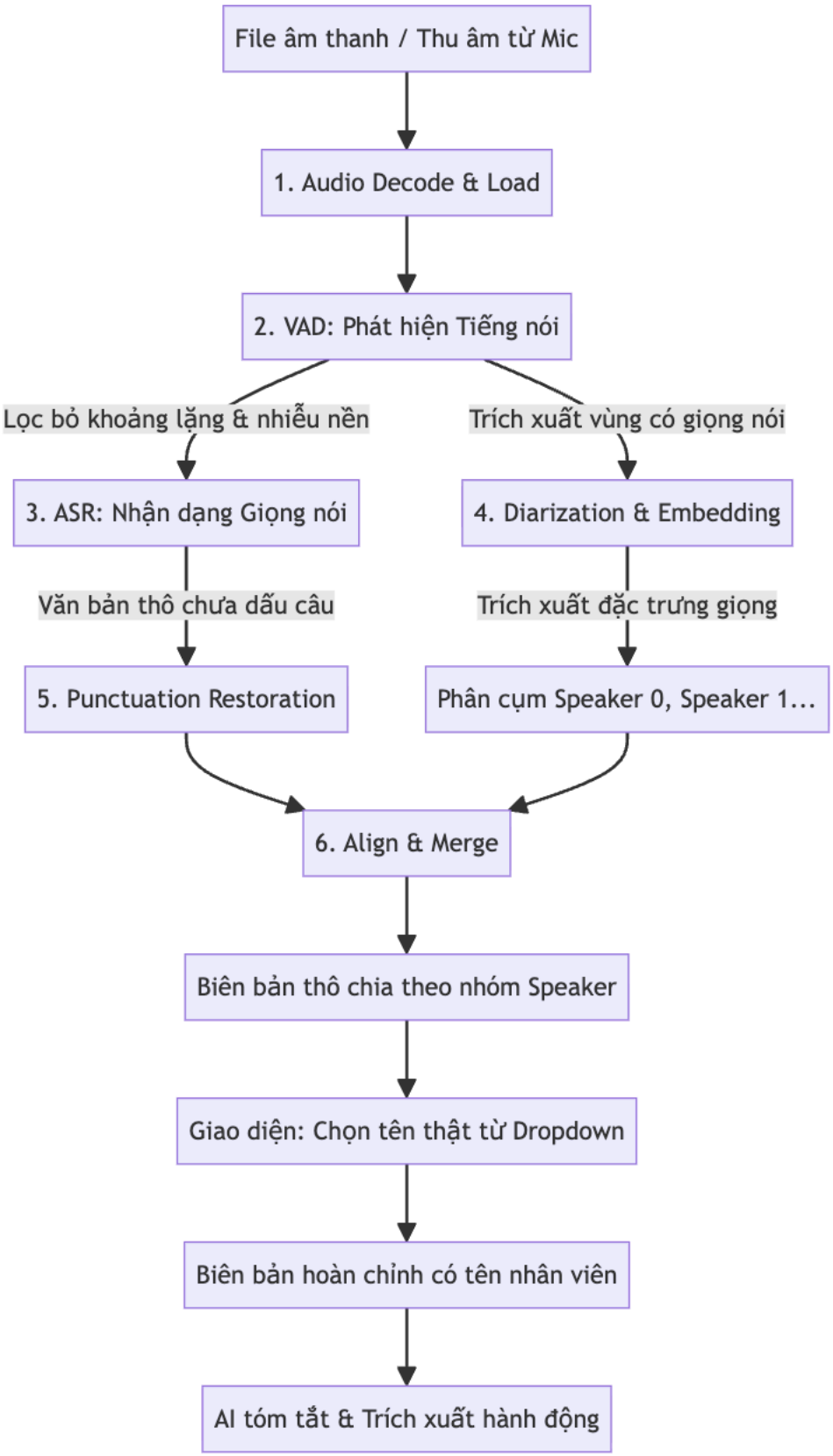
Ứng dụng **Meeting STT** được phát triển dưới dạng phần mềm Desktop chạy trên đa nền tảng (macOS & Windows), phục vụ mục đích tự động ghi nhận, nhận dạng giọng nói và chuyển đổi cuộc họp thành biên bản tiếng Việt một cách bảo mật và tiết kiệm chi phí.

Hệ thống được thiết kế theo mô hình lai (Hybrid Desktop App) kết hợp 3 lớp công nghệ:

- **Giao diện người dùng:** hiện đại, mượt mà và tương tác trực quan cho người dùng.
- **Hệ thống lõi:** chịu trách nhiệm bảo mật dữ liệu, quản lý cơ sở dữ liệu nội bộ, quản lý vòng đời tệp tin, giao tiếp với hệ thống phần cứng và khởi chạy tiến trình xử lý AI.
- **Bộ xử lý AI:** được đóng gói độc lập chạy ngầm bên dưới dạng offline mà không cần kết nối internet.
- Kết nối internet chỉ cần thiết trong trường hợp cần tích hợp khả năng phiên dịch các ngôn ngữ khác ngoài tiếng Việt trong cuộc họp

2. Mô hình xử lý

Quy trình xử lý âm thanh trong ứng dụng được thiết kế khép kín nhằm tối ưu hóa chất lượng văn bản đầu ra.


File âm thanh / Thu âm từ Mic
1. Audio Decode & Load
2. VAD: Phát hiện Tiếng nói
Lọc bỏ khoảng lặng & nhiễu nền
Trích xuất vùng có giọng nói
3. ASR: Nhận dạng Giọng nói
4. Diarization & Embedding
Văn bản thô chưa dấu câu
Trích xuất đặc trưng giọng
5. Punctuation Restoration
Phân cụm Speaker 0, Speaker 1...
6. Align & Merge
Biên bản thô chia theo nhóm Speaker
Giao diện: Chọn tên thật từ Dropdown
Biên bản hoàn chỉnh có tên nhân viên
AI tóm tắt & Trích xuất hành động

Khái niệm	Giải thích
VAD (Voice Activity Detection)	Công nghệ phát hiện hoạt động giọng nói. Bộ lọc này quét qua file âm thanh để cắt bỏ các khoảng lặng, tiếng thở, hoặc tiếng ồn môi trường, chỉ giữ lại các đoạn có giọng người nói thực tế. Điều này giúp giảm 30% - 50% khối lượng dữ liệu cần xử lý ở các bước sau.
ASR (Automatic Speech Recognition)	Quá trình chuyển đổi sóng âm thanh giọng nói của con người thành văn bản chữ viết
Speaker Diarization & Embedding	Trả lời câu hỏi *"Ai đang nói gì và vào lúc nào?"*. Hệ thống sẽ trích xuất đặc trưng giọng nói của từng người nói , sau đó phân cụm các đoạn tiếng nói giống nhau về cùng một mã định danh (Speaker 1, Speaker 2...). Cuối cùng, hệ thống phân cụm các đoạn thoại về cùng một định danh vô danh (Speaker 0, Speaker 1...). Trên giao diện kết quả, người dùng có thể lựa chọn gán nhanh các Speaker này với danh sách hồ sơ nhân viên đã lưu trong CSDL để tự động thay thế thành tên thật trên toàn bộ biên bản.
Punctuation Restoration	Phân tích ngữ cảnh để tự động thêm dấu phẩy, dấu chấm, dấu hỏi và viết hoa các chữ cái đầu câu hoặc tên riêng phù hợp, giúp văn bản dễ đọc và chuyên nghiệp hơn.

Hệ thống hỗ trợ cấu hình linh hoạt giữa phiên bản Nhỏ (Tốc độ) và phiên bản To (Độ chính xác cao):

	Phiên bản (Cấu hình)	Vai trò trong hệ thống (Pipeline)
Lọc nhiễu (VAD)	Mặc định	Phát hiện tiếng người nói, loại bỏ khoảng lặng và các loại nhiễu môi trường trước khi nhận dạng.
Nhận dạng giọng nói (ASR)	**Bản Tốc độ (Speed)**	Chuyển đổi giọng nói thành văn bản thô. Tối ưu tốt cho CPU máy tính văn phòng thông thường.
Nhận dạng giọng nói (ASR)	**Bản Chính xác (Large)**	Tối ưu nhận dạng độ chính xác cao cho phòng họp có tạp âm hoặc giọng địa phương nặng.
Phân đoạn giọng nói	Mặc định	Phân tích đặc trưng âm học thời gian thực để nhận diện chính xác thời điểm thay đổi người nói.
Định danh người nói	**Bản To (Large)**	Trích xuất đặc trưng và phân cụm người nói độ chính xác cao dựa trên giải thuật VBx.
Định danh người nói	**Bản Nhỏ (Small)**	Trích xuất Voiceprint và so khớp Cosine Similarity trong SQLite để hiển thị tên thật nhân viên.
Khôi phục dấu câu	Mặc định	Tự động thêm các dấu câu (chấm, phẩy, hỏi...) và viết hoa chữ đầu câu/danh từ riêng.
Đánh giá âm thanh	Mặc định	Đo chất lượng tín hiệu âm thanh đầu vào. **Lưu ý:** Chỉ dùng nội bộ ghi log. Có thể thay thế bằng RMS/Centroid nếu thương mại thương mại hóa 100%.

Ngoài cơ chế để người thư ký cuộc hop tự định danh người nói trên giao diện, hệ thống còn cho phép Đăng ký Vân tay Giọng nói & Định danh (Bán tự động) để giảm công sức của thư ký.

Chú ý: Hệ thống chưa hỗ trợ hiển thị chữ chạy theo giọng nói thời gian thực (Real-time Streaming).

3. Yêu cầu phần cứng tối thiểu

Máy MacBook: 14-core CPU, 18GB Unified Memory, 16-core Neural Engine
Hoặc máy windows: Intel core i7 (24 core 28 thead), 16GB RAM

Cấu hình server tối thiểu khi dùng cơ chế cài đặt trên server: Intel Xeon Silver 4514Y 2.0Ghz 16C/32T 30M Cache, 128G RAM, RTX PRO 5000 48GB GDDR7

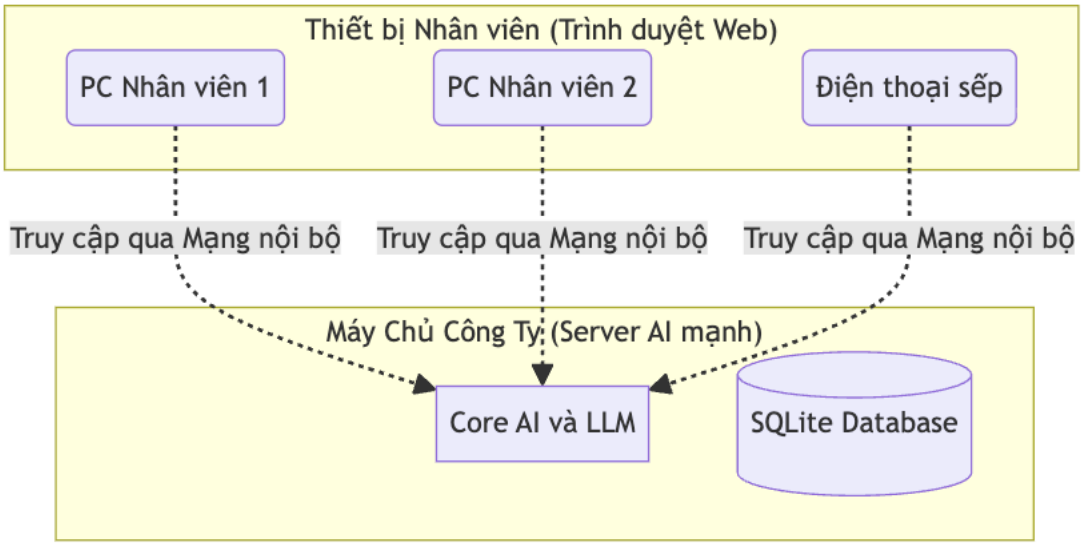


Figure 1: Mô hình hệ thống với server

4. Giao diện mẫu

Bên dưới là 1 số giao diện cơ bản của hệ thống, được xây dựng bằng công nghệ hiện đại, trực quan

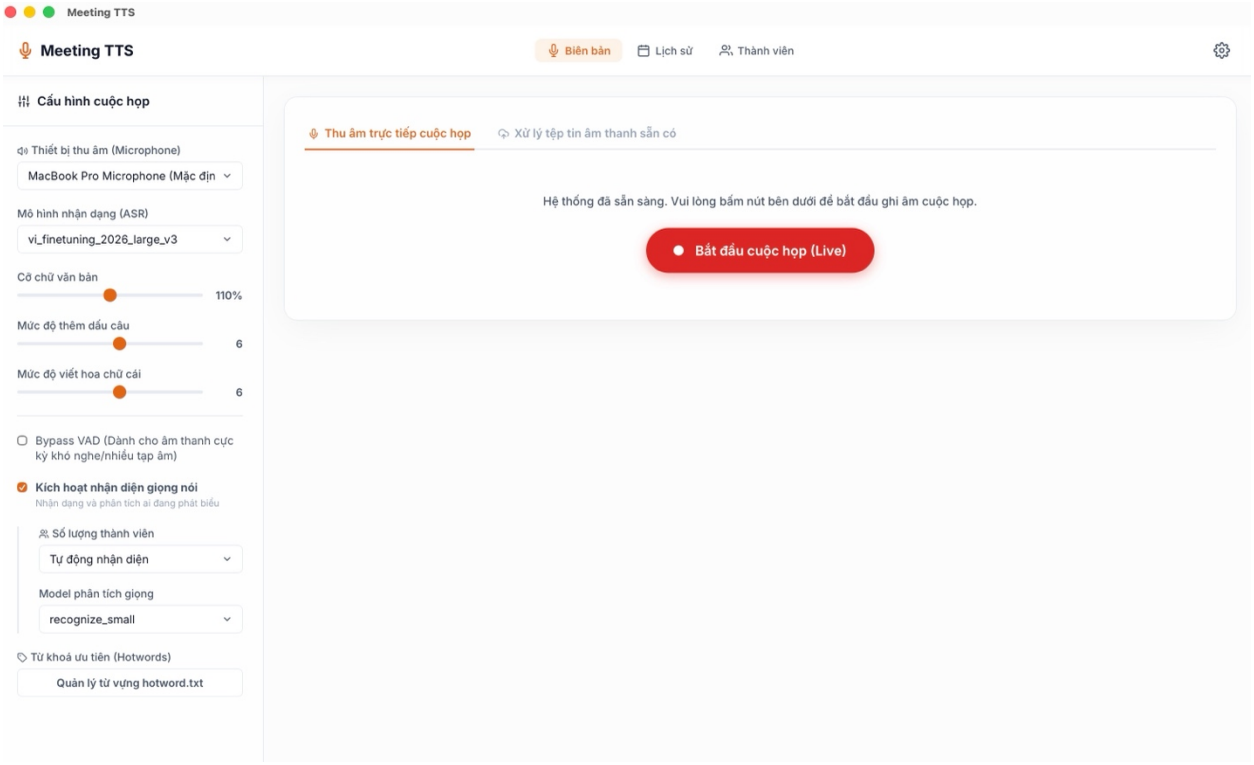


Figure 2: Trực tiếp cuộc họp

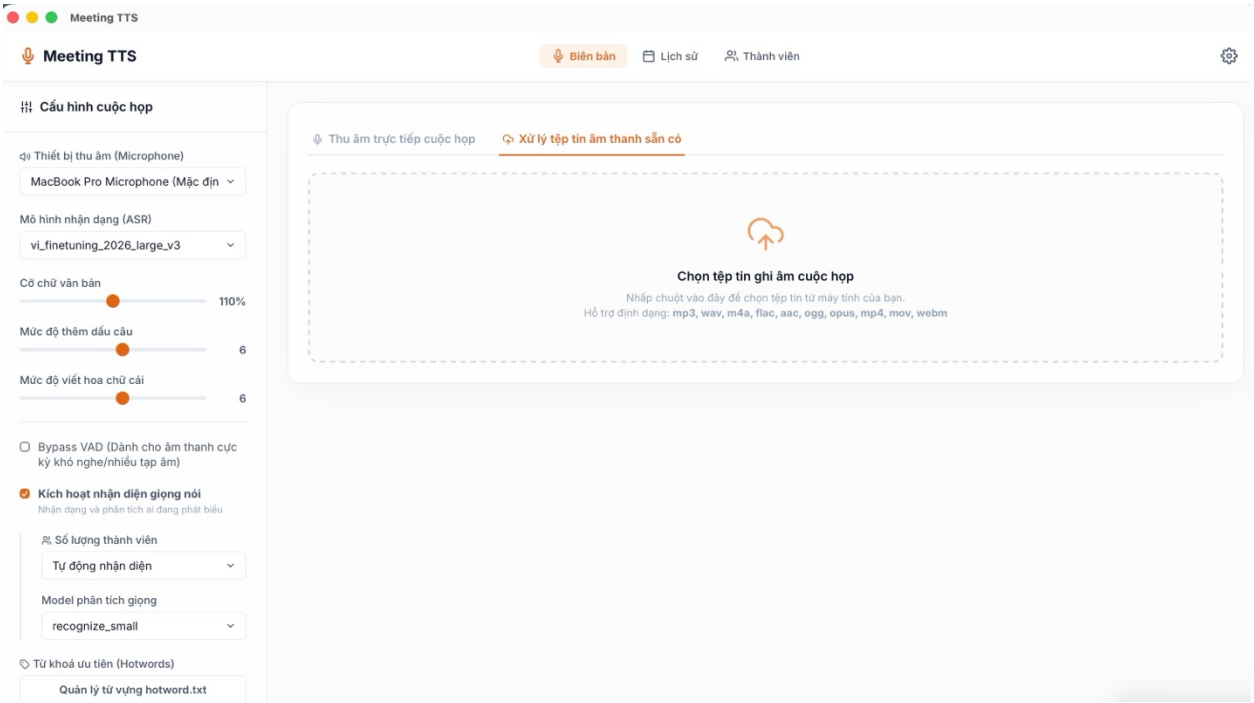


Figure 3: Từ file ghi âm

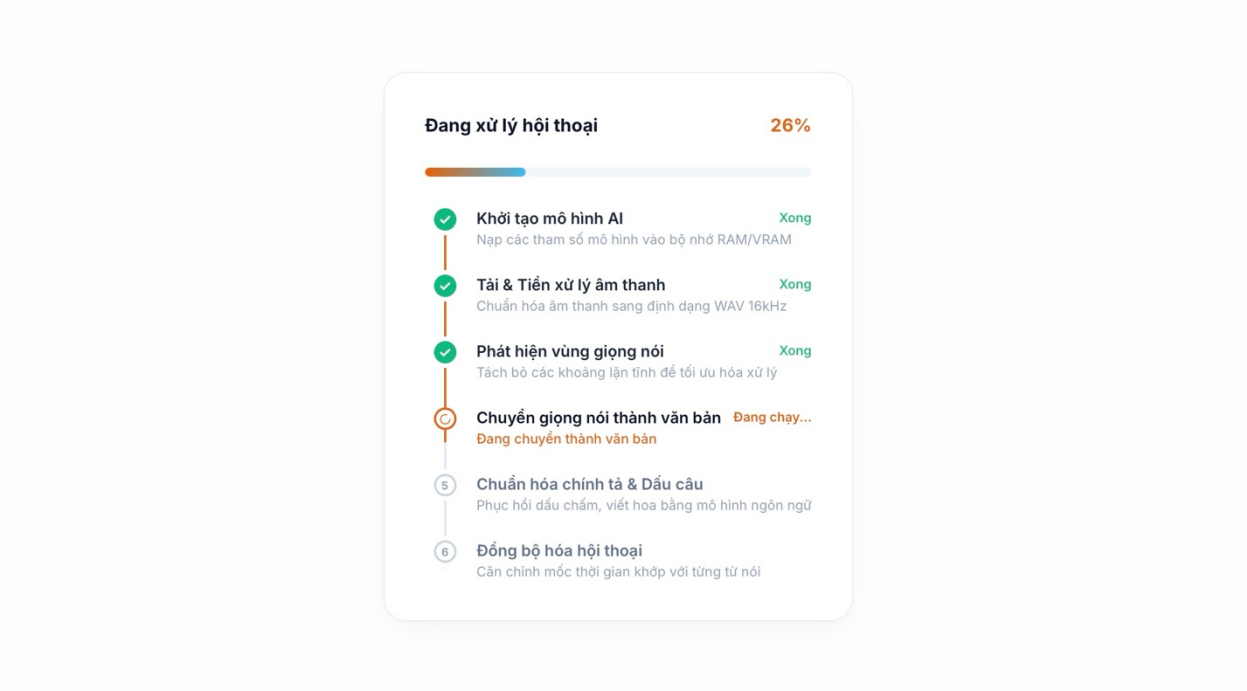


Figure 4: Giao diện xử lý

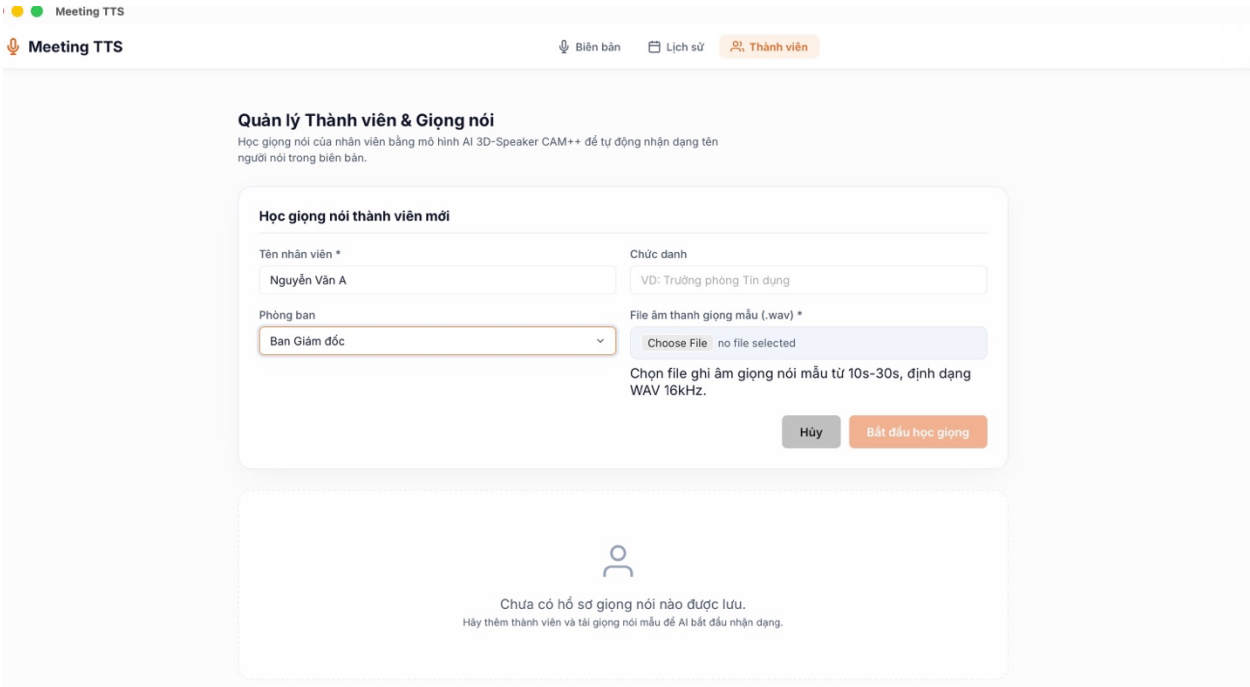


Figure 5: Đăng ký vân tay giọng nói

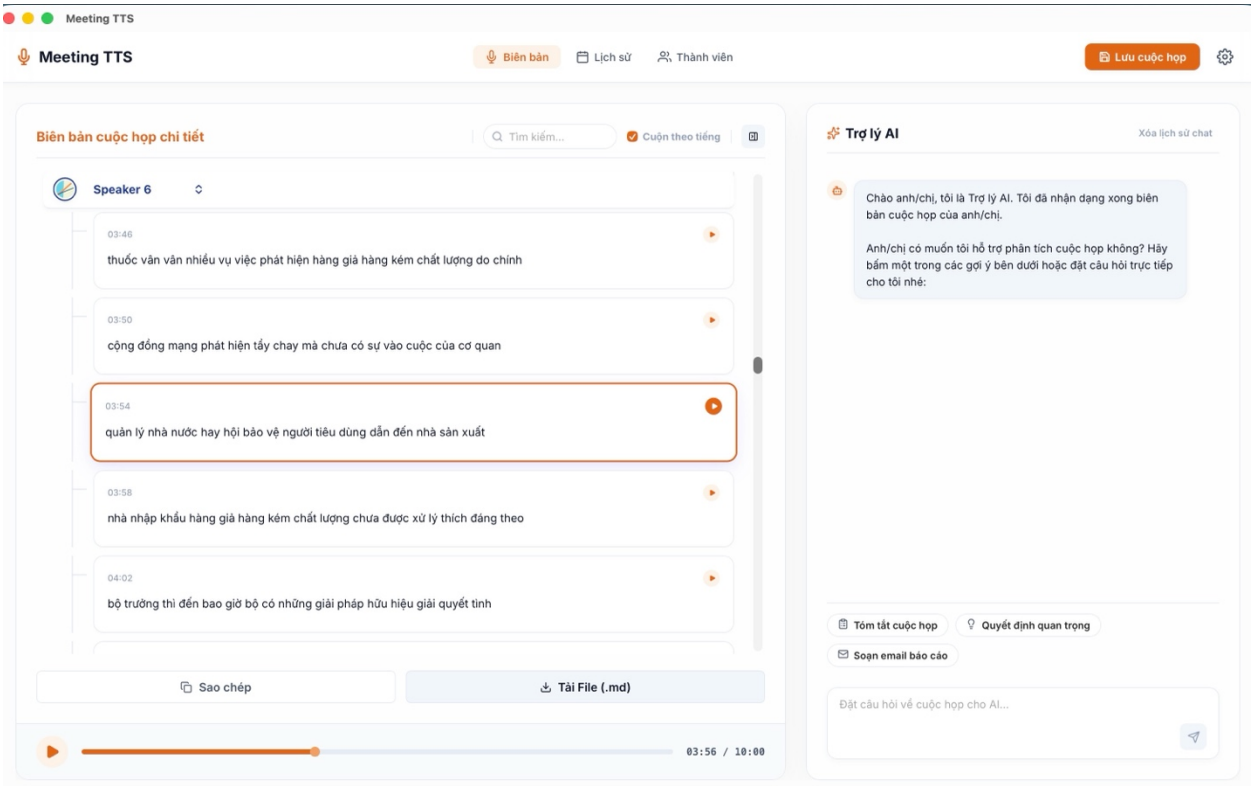


Figure 6: Chỉnh sửa biên bản họp

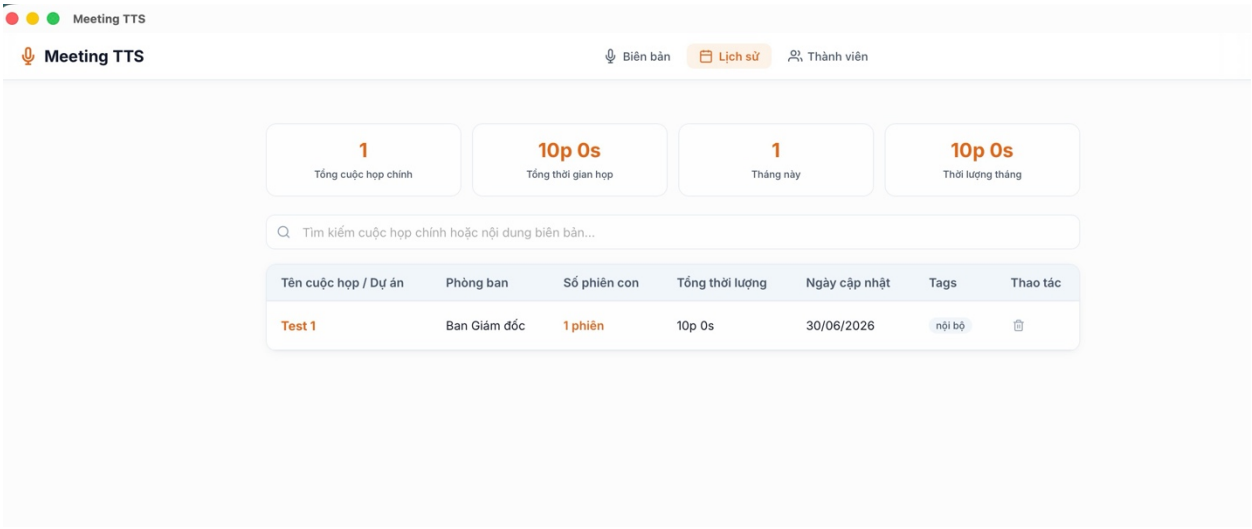


Figure 7: Lịch sử sử dụng