



The Problem

Dr. Lyle Campbell's 50+ years of linguistic fieldwork produced audio recordings of Indigenous languages across Latin America. Over the decades, recordings were digitized, migrated, and copied across formats, creating unknown duplicates across six collections totaling 724 audio files with no systematic inventory.

The Archive of the Indigenous Languages of Latin America (AILLA) needed to identify duplicates and prepare a clean, ingestion-ready package for their digital repository.

Approach

1. Acoustic Fingerprinting. Two independent methods: MFCC + Dynamic Time Warping and wav2vec2 neural embeddings.
2. Metadata Pairing. Linked CD folder names, PCL tape numbers, and MESO identifiers to narrow ~180,000 possible pairs to ~3,500 targeted comparisons (98% reduction).
3. Manual Review & Delivery. All pairs above a confidence threshold manually reviewed. De-duplicated package with full provenance, AILLA metadata workbook, and BagIt checksums.

The Collection



Figure 1: Cassettes, CDs, and reel-to-reel tapes from Campbell's fieldwork.

Key Insight

Byte-identical files were caught perfectly. For near-exact copies, acoustic methods achieved roughly 50% precision (MFCC 54%, wav2vec2 50%), correctly identifying duplicates digitized through different pathways but also flagging false positives from same-speaker recordings.

For lower-confidence matches, audio analysis alone found zero true positives. Researcher metadata (tape numbers, CD identifiers, Campbell's annotations) raised precision to 90%. Beyond that threshold, neither acoustic analysis nor metadata could reliably distinguish duplicates from unrelated recordings.

Acoustic Analysis

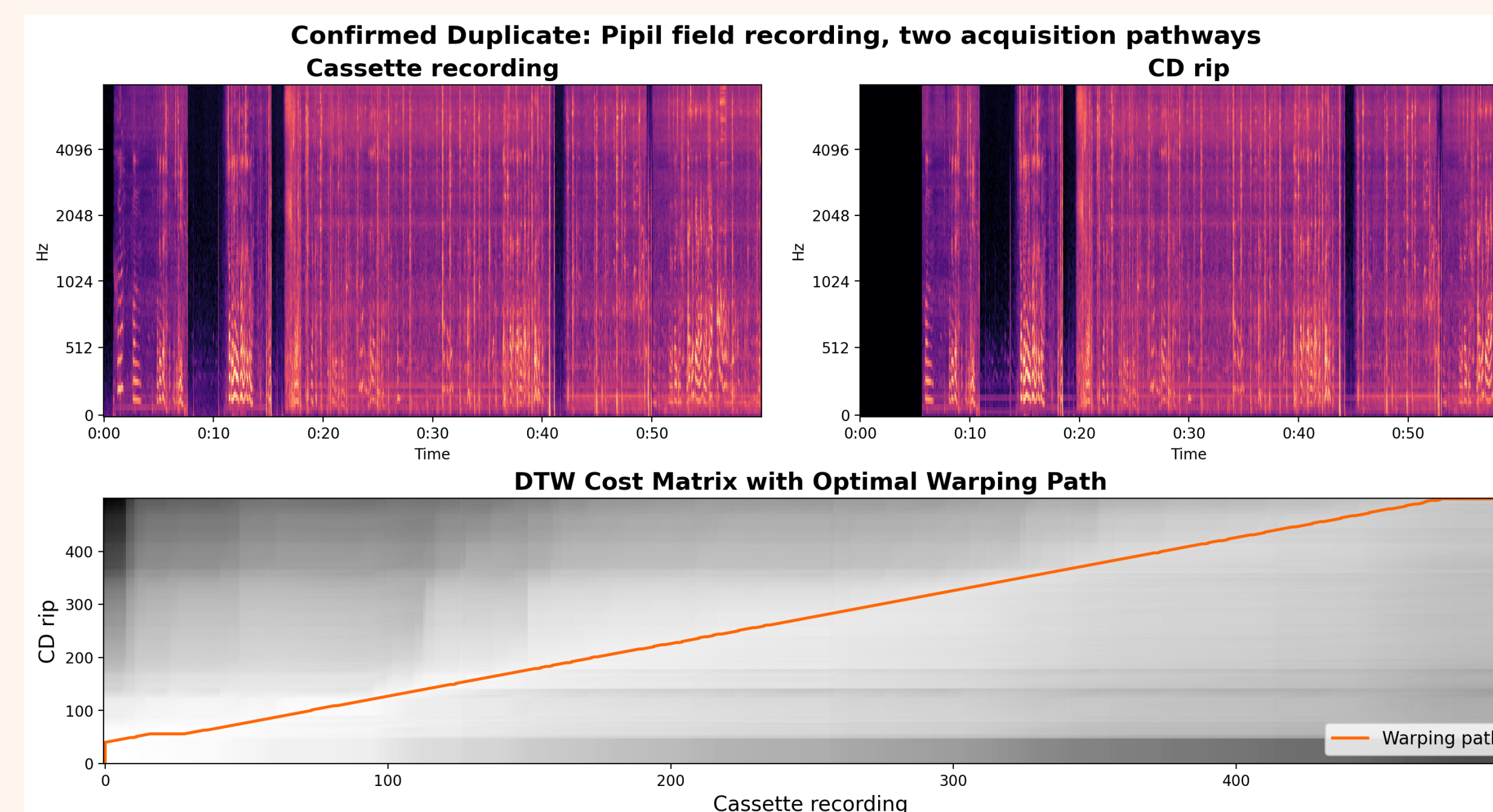


Figure 2: A confirmed duplicate: Pipil field recording acquired from two sources, one on cassette and one ripped from CD. Spectrograms show matching content; DTW path aligns despite analog drift.

Results

Table 1: Key outcomes

Metric	Value
Input files	724
Confirmed duplicate pairs	156
Unique recordings preserved	571
Languages documented	29
Speakers identified	35
Final bag size	236.7 GB

How Duplicates Were Found

Table 2: By discovery method

Discovery Method	Pairs
Audio analysis alone	57
Both audio + metadata	53
Metadata alone	46
Total	156

The de-duplicated collection is now ready for ingest into AILLA's digital repository.

Languages

Mam, Tzeltal, Tojolabal, Pipil, Chorote, Nivacle, Lenca, Popti, Chuj, Mocho, Wichi, Huastec, Jicaque, Xinkan, Yucatec, and others.

Acknowledgments

This work would not exist without the Indigenous communities who generously shared their knowledge over decades of fieldwork across Latin America.

Capstone Supervisor: Jeremy Thompson, Digital Processing Archivist, UT Libraries

AILLA Coordinator: Dr. Susan Kung, Benson Latin American Collection

Indigenous Language Field Researcher: Dr. Lyle Campbell

Course Instructor: Andrea Baer, School of Information

LLILAS Benson Latin American Studies and Collections, UT Libraries