



Transcribing Oral Folk Repertoire With Audio Alignment Tools: A Workflow for Field Researchers

Andres Avetisyan^{1*}, Zarina Peralta², Mateo Uche³

¹Department of Music, Volta Center for Music, Serbia

²Department of Music, Pearl River Research Institute of Music, Taiwan

³Department of Music, Loire Center for Music, Argentina

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijm.2026.v1.i1.pp13-18](https://doi.org/10.62715/ijm.2026.v1.i1.pp13-18)

Received: 16 April 2026

Accepted: 18 June 2026

Published: 29 August 2026

Abstract

Transcription remains the slowest stage of fieldwork on orally transmitted song and tune repertoires, and the one most exposed to the habits of the transcriber's own ear. Free software can now track the pitch of a solo voice or instrument, mark note onsets, and align one rendition of a melody with another, or a sung text with its recording. Field researchers without training in signal processing rarely use these tools together, partly because the literature describes them one at a time and for Western concert music. This article proposes a workflow in eight steps that takes a field recording from the microphone to an annotated, time-aligned transcription: recording and documentation, archiving, segmentation, pitch and onset tracking, estimation of the tonal reference and scale, alignment of strophes and of text, notation of a melodic skeleton with its variants, and review with performers before deposit. For each step the article names suitable open tools, the decision that remains with the researcher, and the file that should be kept. The workflow is a proposal built from published methods and archival guidelines. It has not been tested on a field collection here, and the article sets out what such a test would need to show and where the approach is likely to fail.

Keywords: *Music transcription, Audio alignment, Dynamic time warping, Pitch tracking, Ethnomusicological fieldwork, Oral tradition*

© 2026 Andres Avetisyan et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

A singer in a kitchen gives twelve verses of a ballad, each to the same tune and none of them quite alike. The researcher who recorded her will later sit with headphones and a pencil, replaying two seconds at a time, and will have to decide for every note what pitch it was, how long it lasted, and whether the difference between verse three and verse seven is a variant worth writing down or the ordinary give of a living performance. The labour is considerable, and its results depend heavily on who is listening.

Computational tools can now take over part of the measuring. Pitch trackers return the fundamental frequency of a monophonic line frame by frame; onset detectors propose note boundaries; alignment algorithms find which moment in one recording corresponds to which moment in another. Surveys of

* Corresponding author: Andres Avetisyan, Department of Music, Volta Center for Music, Serbia. E-mail: andres.avetisyan@volta.edu

automatic transcription are frank that fully automatic notation of real recordings is still unreliable, particularly outside the piano and outside equal temperament (Benetos *et al.*, 2019). The opportunity for field researchers lies elsewhere: in semi-automatic work, where software measures and aligns and the researcher interprets.

This article proposes a workflow of that kind for solo or lightly accompanied oral repertoire, such as unaccompanied song, fiddle or flute tunes, and chant. It is written for researchers who know their repertoire well and their software less well. The workflow is put forward as a design. It has not been validated on a field collection in this article, and a later section states what validation would involve.

2. Background

Two discussions frame the design: the long argument within ethnomusicology about what a transcription is for, and the technical literature on pitch tracking and alignment.

2.1. *Transcription as an interpretive act*

Charles Seeger distinguished prescriptive notation, a set of instructions for performance, from descriptive notation, a record of what was performed, and argued that Western staff notation serves the second purpose poorly (Seeger, 1958). His remedy was a machine-drawn graph of pitch against time. Bartók's transcriptions of the Parry collection took the opposite route, loading the staff with microtonal signs and exact durations until the result recorded nearly everything and could be read by almost no one (Bartók and Lord, 1951). Ellingson's survey shows how each generation of transcribers has chosen a different point between those extremes (Ellingson, 1992).

The discipline has since grown more sceptical. Nettl treats transcription as one representation among several, shaped by the question being asked (Nettl, 2015), and others have asked who needs it at all once recordings can be published alongside the analysis. The workflow below accepts that critique. It treats notation as a reading aid tied to the audio by time stamps, so that every written note can be checked against the sound it summarises.

2.2. *Pitch tracking and alignment*

For a single melodic line, the YIN algorithm and its probabilistic successor pYIN estimate fundamental frequency reliably enough for analytical use (de Cheveigné and Kawahara, 2002; Mauch and Dixon, 2014). pYIN is built into Tony, a program designed for correcting pitch tracks and note segmentations by hand (Mauch *et al.*, 2015), and is available as a plugin for Sonic Visualiser, a general viewer and annotator for music audio (Cannam *et al.*, 2010).

Alignment rests on dynamic time warping, a dynamic-programming method first developed for matching spoken words uttered at different speeds (Sakoe and Chiba, 1978). Given two sequences of features, it finds the path of correspondences that minimises their accumulated difference. Müller's textbook explains its use for synchronising music recordings (Müller, 2015), and the librosa library offers it to anyone able to run a short Python script (McFee *et al.*, 2015). For sung words, forced aligners trained on speech can place each syllable in time where an acoustic model for the language exists (McAuliffe *et al.*, 2017).

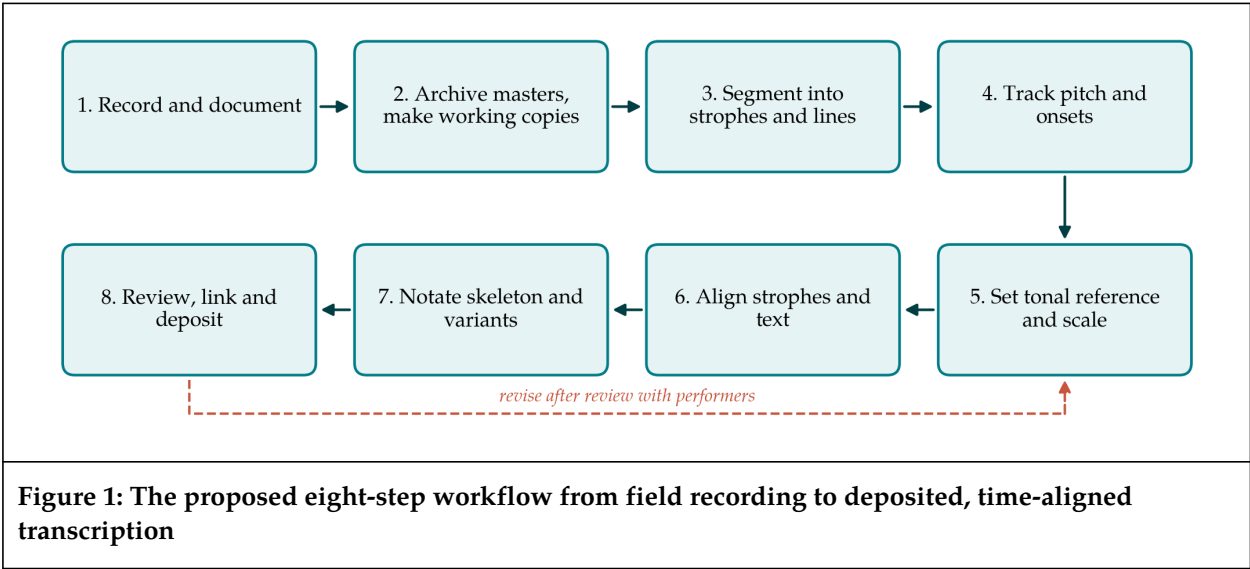
3. The Proposed Workflow

The workflow has eight steps, shown in sequence in Figure 1. Table 1 lists the tools suggested for each and the file that each step should leave behind. The principle throughout is that the software proposes and the researcher decides, and that each decision is saved in a form another person can open.

3.1. *Recording, archiving and segmentation*

Step one is the recording itself, made with informed consent that covers analysis, publication of excerpts and deposit. A close microphone on the singer or player, and as little room noise as the setting allows, does more for later pitch tracking than any processing. Spoken identification of performer, place, date and

item at the head of each take costs seconds and saves confusion. Step two follows the guidelines of the International Association of Sound and Audiovisual Archives: masters are kept as uncompressed WAV files at 24-bit depth and a sampling rate of at least 48 kHz, never edited, with working copies made for analysis (IASA, 2009).



Step three divides each item into strophes and lines. This is done by ear in Sonic Visualiser, or in ELAN, an annotation tool built for time-aligned tiers of text (Wittenburg et al., 2006). A tier for strophes, one for lines and one for the words is enough to begin. Segmentation is an analytical claim about how the performer structures the piece, and where the performer has terms for the units, those terms should name the tiers.

Table 1: Steps of the proposed workflow, suggested open tools and the file each step leaves			
Step	Suggested tools	Decision left to the researcher	File kept
1. Record and document	Field recorder, close microphone	Scope of consent; what counts as one item	Consent record, field notes
2. Archive	Any audio editor for copies	File naming and identifiers	Unedited WAV masters
3. Segment	Sonic Visualiser or ELAN	Where strophes and lines begin	Annotation tiers
4. Track pitch and onsets	Tony or the pYIN plugin	Which frames are voiced; note boundaries	Corrected pitch track, note layer
5. Tonal reference and scale	Pitch histogram in cents	Which peaks are scale degrees	Histogram, degree table
6. Align	Dynamic time warping in librosa; forced aligner	Whether an alignment is musically right	Warping paths, aligned text
7. Notate	music21 or a notation editor	Level of detail; how variants are shown	Score file with time stamps
8. Review and deposit	Playback with performers	What performers accept or correct	Revised files, deposit record

3.2. Pitch, onsets and the tonal reference

In step four each strophe is passed through a pitch tracker and the result is corrected. Typical errors are octave jumps, pitches reported during consonants or breaths, and a second voice or drone captured in

place of the melody. Tony was designed for this correcting pass and lets the user redraw a contour or merge and split notes while listening. Ornaments need a policy stated in advance: whether a slide into a note is one event or two, and how short a pitch must be before it is treated as a grace note.

Step five avoids forcing the pitches onto a piano keyboard. All corrected pitch values for a performance are converted to cents relative to an arbitrary reference and plotted as a histogram. The peaks show the pitches the performer returns to; the most stable, usually the final, is taken as the tonal reference and the others are expressed as intervals above or below it. A third that sits between major and minor then appears as what it is, instead of being rounded to the nearer semitone. The degree table produced here governs the notation in step seven.

3.3. Alignment, notation and review

Step six is where alignment earns its place. Each strophe is aligned by dynamic time warping to a reference strophe, using the corrected pitch contour as the feature. The warping path shows which notes correspond, so that differences of pitch at corresponding points can be listed as melodic variants and stretches or compressions of the path as rhythmic ones. Renditions by other performers, or by the same performer on another day, are aligned to the same reference. Where an acoustic model exists for the language, the text is aligned to the audio with a forced aligner and corrected by hand; where none exists, syllables are placed manually against the note layer.

In step seven the researcher writes a melodic skeleton for the tune and attaches the variants to it, strophe by strophe. A toolkit such as music21 can build the score from the note layer and keep the time stamp of each note (Cuthbert and Ariza, 2010). The level of rhythmic detail is a choice: for free-rhythm song, proportional spacing with onset times may be more honest than bar lines. Step eight returns the result to the performers, played back with the notation or the pitch graph on screen, and their corrections are entered before the audio, annotations and score are deposited together under the terms agreed in step one.

4. Fitting the Workflow to Field Conditions

Field recordings are seldom clean. Where several people sing together, where the melody instrument plays over a drone or a guitar, or where the room is loud, pitch tracking of the kind described degrades quickly, and the researcher should expect to do step four largely by hand or to limit the workflow to the solo items in a collection. Heterophonic and polyphonic textures are beyond it. The steps are separable, however, and segmentation, tonal reference and deposit remain useful when tracking is not.

Questions of ownership and control run through every step. Anthony Seeger has shown how poorly copyright law fits music held in common or by lineage (Seeger, 1992), and work on the ethics of music information retrieval points out that turning a community's songs into data creates new ways for them to be reused without consent (Holzapfel et al., 2018). The workflow therefore places consent at the first step and review at the last, and treats the decision about what is deposited openly as belonging to the performers. Tools serve fieldwork only for as long as the people recorded remain partners in it.

A practical point concerns where the work is done. Every tool named here runs offline on an ordinary laptop, so segmentation and a first pitch track can be made in the evening after a session while the performer is still nearby to answer questions. Doing so changes the next day's recording, because gaps and doubtful passages are noticed in time to ask for a repeat. It also means that files multiply quickly. A fixed naming scheme that carries the item identifier, the step number and a version through every annotation, pitch track and score file is dull to maintain and is the only thing that keeps a collection of several hundred items usable a year later.

5. Limitations and What Remains to Be Shown

The workflow is a proposal and has not been applied or validated here. Three things would need to be shown. The first is agreement: whether transcriptions produced this way by different researchers from the same recordings differ less than transcriptions made by ear alone. The second is effort, since correcting a

poor pitch track can take longer than transcribing from scratch, and the balance will vary with recording quality. The third is acceptance, meaning whether performers recognise their own music in the result and whether the variants the alignment reports are differences they consider meaningful.

There are also limits that no test will remove. Dynamic time warping assumes that two renditions contain the same events in the same order; it copes badly when a singer inserts or omits a phrase, and such cases must be segmented first. Pitch histograms presuppose a stable tonal reference, which unaccompanied singers do not always keep across a long song. And a time-aligned score is still a reduction. It records pitch and timing and leaves out timbre, gesture, words unspoken and the occasion, which fieldnotes and the recording itself must carry.

6. Conclusion

Pitch tracking and audio alignment will not transcribe a field collection, but they can measure what the ear estimates and keep every written note attached to its sound. The eight-step workflow proposed here arranges freely available tools so that each supports a decision the researcher must still make and leaves a file that others can inspect. Its usefulness for a given repertoire is an open question that field researchers are best placed to answer, ideally by trying it on a small set of recordings alongside their usual practice and reporting where it helped and where it did not.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Bartók, B. and Lord, A.B. (1951). *Serbo-Croatian Folk Songs*. Columbia University Press, New York.
- Benetos, E., Dixon, S., Duan, Z. and Ewert, S. (2019). *Automatic music transcription: An overview*. *IEEE Signal Processing Magazine*, 36(1): 20-30. <https://doi.org/10.1109/MSP.2018.2869928>
- Cannam, C., Landone, C. and Sandler, M. (2010). *Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files*. *Proceedings of the 18th ACM International Conference on Multimedia*, 1467-1468. <https://doi.org/10.1145/1873951.1874248>
- Cuthbert, M.S. and Ariza, C. (2010). *music21: A toolkit for computer-aided musicology and symbolic music data*. *Proceedings of the 11th International Society for Music Information Retrieval Conference*, 637-642.
- de Cheveigné, A. and Kawahara, H. (2002). *YIN, a fundamental frequency estimator for speech and music*. *The Journal of the Acoustical Society of America*, 111(4): 1917-1930. <https://doi.org/10.1121/1.1458024>
- Ellingson, T. (1992). *Transcription*. In Myers, H. (ed.), *Ethnomusicology: An Introduction*. Macmillan, London, 110-152.
- Holzapfel, A., Sturm, B.L. and Coeckelbergh, M. (2018). *Ethical dimensions of music information retrieval technology*. *Transactions of the International Society for Music Information Retrieval*, 1(1): 44-55. <https://doi.org/10.5334/tismir.13>
- IASA Technical Committee (2009). *Guidelines on the Production and Preservation of Digital Audio Objects (IASA-TC 04, 2nd ed.)*. Bradley, K. (ed.). International Association of Sound and Audiovisual Archives.
- Mauch, M. and Dixon, S. (2014). *pYIN: A fundamental frequency estimator using probabilistic threshold distributions*. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 659-663. <https://doi.org/10.1109/ICASSP.2014.6853678>

- Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., Bello, J. and Dixon, S. (2015). [Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. *Proceedings of the First International Conference on Technologies for Music Notation and Representation \(TENOR\)*.](#)
- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. and Sonderegger, M. (2017). [Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of Interspeech 2017*, 498-502.](#)
<https://doi.org/10.21437/Interspeech.2017-1386>
- McFee, B., Raffel, C., Liang, D., Ellis, D.P.W., McVicar, M., Battenberg, E. and Nieto, O. (2015). [librosa: Audio and music signal analysis in Python. *Proceedings of the 14th Python in Science Conference*, 18-25.](#)
<https://doi.org/10.25080/Majora-7b98e3ed-003>
- Müller, M. (2015). [Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications.](#) Springer, Cham. <https://doi.org/10.1007/978-3-319-21945-5>
- Nettl, B. (2015). [The Study of Ethnomusicology: Thirty-Three Discussions \(3rd ed.\).](#) University of Illinois Press, Urbana.
- Sakoe, H. and Chiba, S. (1978). [Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26\(1\): 43-49.](#)
<https://doi.org/10.1109/TASSP.1978.1163055>
- Seeger, A. (1992). [Ethnomusicology and music law. *Ethnomusicology*, 36\(3\): 345-359.](#)
- Seeger, C. (1958). [Prescriptive and descriptive music-writing. *The Musical Quarterly*, 44\(2\): 184-195.](#)
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. and Sloetjes, H. (2006). [ELAN: A professional framework for multimodality research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation \(LREC 2006\)*, 1556-1559.](#)