Audio alignment for improved melody transcription of Irish traditional music

Hannah Robertson
MUMT 621 — Winter 2012

In order to study Irish traditional music comprehensively, it is critical to work from recordings, whether by studying audio directly or by transcribing it. Computational musical-analysis tools such as Music21 work from symbolic scores, so transcription is required. In this paper, I propose a method for improving the accuracy of melody transcriptions from often noisy polyphonic recordings by taking advantage of the repetitive nature of Irish traditional music. Aligning and combining repetitions of a tune in the frequency domain before moving on to the transcription step will strengthen the presence of incidental but recurring musical events that might otherwise fall below the threshold of background accompaniment or environmental noise. In the following sections, rational for this alignment method is presented in more detail and then the method is implemented as a proof of concept.

1 Discussion

Much of the computational ethnomusicological work currently being done on Irish traditional music focuses on tune identification and segmentation. In the past couple of years, the Audio Research Group at the Dublin Institute of Technology has published several interesting papers focusing on segmentation, fingerprinting, and tune recognition (Duggan et al. 2009a, Duggan et al. 2009b, Kelly et al. 2009). In his PhD thesis "Machine Annotation of Traditional Irish Dance Music," Duggan applies transcription and alignment algorithms for metadata purposes: to identify and match recorded tunes to existing transcriptions (2009). A different set of questions, about the propagation of change in tunes over time, ornamentation evolution, and other sorts of dynamic change are better answered when more is know about each iteration of the tune. Rather than studying clusters of tunes as a matched set, it would be useful to study en masse transcriptions of each recording.

1.1 Transcription in ethnomusicology

In 1962, George List discussed transcription as a tool for ethnomusicology, calling it "a prerequisite to certain types of ethnomusicological studies," such as the "musical factors: mode, melody, form, etc." While the transcription method that he refers to is monophonic melograph recording, and the comparison of two transcriptions visual rather than computer–based, his overall summary is still very relavent:

> The value of a transcription, then, is that it facilitates immediate comparisons. It does not follow that the transcription is completely accurate, that it renders all detail, or that it represents all aspects of the musical event. ... The value of a transcription lies not in its complete reproduction of all aspects of a musical event

> but in the fact that it facilitates the comparison of a number of individual and separable elements or aspects of the musical event. (List 1962)

Questions that can be answered through such analysis include but are not limited to the roles that various musical elements or aspects play in defining styles or genres, particular musical cultures, and geographic-cultural musical areas, the vertical relations between pitches, and the musical patterns that occur in the music of a particular culture. Ultimately, he recommends approaching transcription approach with care, in great part because "transcribing is an arduous labor."

Thanks to computers, transcription is no longer so arduous: while automated transcription is far from perfect, especially for polyphonic audio, it has certainly improved since the 60's! In addition, it is now easier than ever to carry out large-scale analysis of a large musical corpus, with tools such as Music21.[1] With machine-assisted or even fully-automated transcription, the ethnomusicological questions List poses can now be tackled in earnest. Hillhouse's transcription-based musicology Master's thesis, "Tradition and innovation in Irish instrumental folk music" (2005), investigates the nature of tunes that become part of the common-practice repertoire, meaning tunes that are popular enough to enter the corpus of tunes commonly played at sessions and dances. His analysis is primarily by hand, but it would be fascinating to repeat the various musicological analysis on a full corpus. Once transcriptions have been made from recordings, can we truly ask questions about what notes are critical versus ornamental to the integrity of a tune, where in the tune pitch or rhythm substitutions are commonly accepted, and what sorts of variations lead to tunes being considered independent of one another.

1.2 Irish traditional music representations

In symbolic representations of a Western classical music, pitch and rhythm are held constant among each performance, with variations chalked up to emotive style (if not musician error!). Perhaps most important to this consistency is that western classical music is learned and performed through studying a symbolic score. Folk music, on the other hand, is often passed down through the oral tradition and learned by ear, such that over time differences in a tunes performance may be introduced and retaught. This means that, unlike in Western classical music, Irish music does not have a ground truth, symbolic or otherwise: any single transcription of a single tune cannot possibly encompass the tune's geographically and historically dynamic nature. This is compounded by the fact that Irish traditional music is highly ornamented, with te type and frequency of the ornaments both region- and instrument-based. For example, different types of ornaments are favoured by pipers and fiddlers due to the mechanics of the instruments, but the same tune will be played by both sets of musicians and recognized as the same.

Symbolic representations of Irish music do exist; over the years, ethnomusicologists have transcribed and compiled many tunes, a number of which have been digitized. Alan Ng's "Irish Traditional Music Tune Index"[2] contains a massive index of tunes featured in recordings, and links to many of these transcriptions in both digitized and print format.

While the transcriptions are useful for answering static musical questions, they are not a complete picture: symbolic representations of folk tunes are either the tune's generally accepted contours at the time of their notation, as in Figure 1, or a complete transcription of one player's unique rendition in a single performance, as in Figure 2. Neither case can be treated as a gold standard representation of the tune. The first ignores the significance of ornamentation

[1] http://mit.edu/music21/
[2] www.irishtune.info

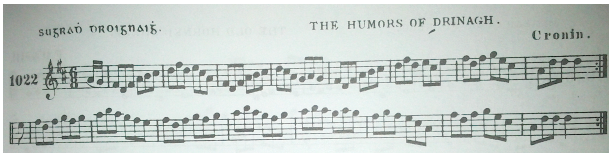

Figure 1: A basic transcription of the first half of "The Humors of Drinagh" (O'Neill and O'Neill 1903, #1022). Note the lack of extensive ornamentation.

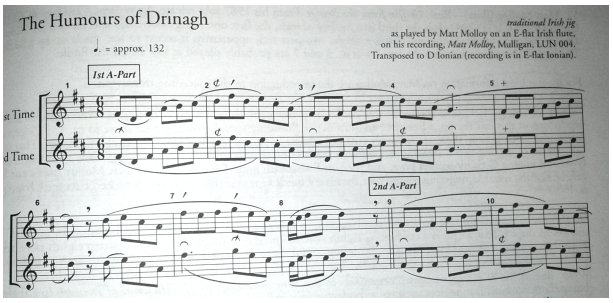

Figure 2: A transcription by Larsen of "The Humors of Drinagh" as played by Matt Molloy, 'A' section only (Larsen 2003).

and the potential for variation even at that one point in time; even if the tune is written out exactly how the composer intended, it has not necessarily been passed on to other musicians that precisely or accepted into the common-practice repertoire in that particular form. The second case is similarly a snapshot, and while it provides additional information such as ornamentation, it is also not the tune's Platonic ideal, as no note is made as to which of those ornaments and variations are unique to the player, unique to a set of players taught from a single player or recording, or a unique aspect of the tune that remains constant despite variation around it. Luckily, many audio recordings exist, archived in collections such as Taisce Cheol Dúchais Éireann (The Irish Traditional Music Archive).[3] Full and complete transcriptions of the individual performances do not yet exist.

1.3 Automated tune transcription

Recordings of Irish traditional music often involve several melodic instruments playing at the same time and in the same pitch range. Often, the musicians play the same overall melody but with different ornamentation. Backup instrumentation may include guitar, piano, or accordion, although these three instruments may also play melody. Often there is a drum or two. Oftentimes Irish traditional music is played at dances or in jam "sessions" in noisy settings such as pubs.

Polyphonic transcription is an ongoing research challenge in the field of music information retrieval (MIR), and musically dense recordings such as in the case presented here are difficult to separate into their component parts. Usefully, a full polyphonic transcription is not needed for the melody transcription presented here. While it would certainly be interesting to study exactly what each player is playing and when, the objective is to study an overall tune, i.e. general note onsets, pitch classes, and harmonic layers. As long as the transcription such a

[3] http://www.itma.ie/about/about-the-itma/

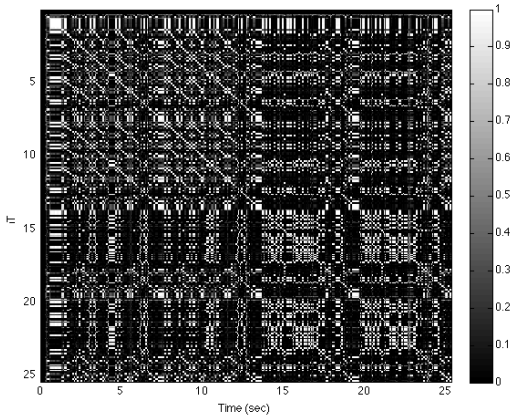

Figure 3: "The Humors of Drinagh" as played by Matt Molloy, first repetition only (Molloy and Lunny 1984).

recordings keeps track of the relative strength of musical events such as pitches played and note onsets, an "average" melody for the group can be obtained. Any future analyses done with this data can either investigate strictly the strongest pitches and onsets, or the range of pitches and onsets, depending on the question being studied. This transcription is still at risk of more spurious errors or omissions than usual, given the noisy recording space and multiple instruments. Since an average is all that is needed, we can take advantage of the repetitive nature of Irish music and combine multiple iterations of the same tune to get an average before actually transcribing it. This step is discussed in the following section.

1.4 Audio alignment for improved transcription

Like the folk music in many cultures, Irish traditional music has a highly repetitive and highly patterned structure. While skilled players increasingly ornament the repetitions of each tune, a tune's initial melody maintains a strong presence throughout the entire duration of the tune. This is in contrast to music like jazz, for example, which includes improvisational sections that deviate from primary tune. This repetition is clearly seen in self-similarity matrices, as in Figure 3.

Once two or more repetitive sections are aligned, the most common elements can be taken as constants in the average tune: musical events such as onsets and pitches that are the same in every rendition will show up stronger. The comparison of two tune repetitions is reasonable from both a musical and a recording standpoint. Musically, if the musicians are constant between the two (or more) repetitions, any constant musical events represent that musician's understanding of the tune. From a recording standpoint, the environment and recording equipment remains generally constant, so any audio artifacts due to the environment are a constant between recordings and will not show up as a feature in a transcription. That is, the transcription of a recording (or an averaged recording, as here) can be fine-tuned to adjust for environmental factors such as subtle onsets or an overall buzz in the key of E, but if those factors are not constant across the entirety of the averaged set that buzz in the key of E might be taken as a desired musical feature in one recording but not in the other, rather than merely an air conditioner in the background.

This same alignment method could be applied to multiple renditions of a tune, to get an overall average over time. For a number of reasons, it is likely that without major adjustments in the alignment algorithms, however, transcribing the audio before aligning the melodies might give more reliable results. First, perhaps aligning the transcriptions would be cleaner: matching is itself a tricky business, and so might introduce additional unnecessary error. Second, the computations involved in aligning symbolic transcriptions are much less computationally expensive than aligning audio, so if there was not much to gain from the additional feature space it might not be worth the computational power. In this paper, however, the goal is aligning audio to strengthen noisy or poor quality recordings before doing any sort of analysis or transcription, so that notes with ambiguous importance build on one another and contribute to the overall average melody. If polyphonic transcription was perfect, this step would not be necessary.

1.4.1 Dynamic time warping

One way to align multiple repetitions of the same tune is by finding an optimal mapping path through a process called dynamic time warping (DTW). DTW was migrated to MIR from the speech recognition field by Berndt (1994). Some of the many applications of DTW in MIR include the aligning of cantillations and other vocal chant utterances in order to study variability across different vocal traditions (Ness 2009), audio alignment of multiple recordings of the same musical piece (Dixon and Widmer 2005), identification of unique musical patterns in traditional Greek music (Pikrakis et al. 2003), and alignment of audio to symbolic data without first transcribing the audio (Hu et al. 2003).

The DTW algorithm aligns the two time series $U = u_1, ..., u_m$ and $V = v_1, ..., v_n$ by finding a minimum cost path through a distance matrix comprised of the distance between each point in U and each point in V, where each of these points is a feature vector. The traditional DTW algorithm takes all possible path choices into consideration and calculates the minimum cost path in quadratic time. A live-warping algorithm introduced by Dixon in (2005) and made accessible through the MATCH toolkit (Dixon and Widmer 2005) modifies the original algorithm by linearly constraining the DTW to an optimal forward path. This means that the alignment path is calculated in linear time and space, even "on-line," i.e. in realtime, if need be. Figure 4 shows an example of this path. Once a best fit mapping between two musical lines is determined, their frequency domains can be merged together and a transcription made.

2 Implementation Details

This section steps through the workflow used in the alignment process. Rather than a full transcription, chroma vectors are used to visualize the results.

Audio

The tune used to test this procedure is the reel "The Mason's Apron," as performed live by the Chieftains in 1991 on the album "An Irish Evening: Live at the Grand Opera House, Belfast" (Chieftains 1992). A transcription of "The Mason's Apron" published in 1903 is presented in Figure 5 (O'Neill and O'Neill 1903); no composition date or composer is known. The tune is a 32 bar reel in A major, with 8-bar phrases that repeat in the form AABB (Ng 2002). This recording was chosen because it contains both ensemble and solo playing. The recording starts off with the whole ensemble playing the tune through once. Flautist Matt Molloy then repeats the tune multiple times, increasing in speed and variation complexity. Finally, the

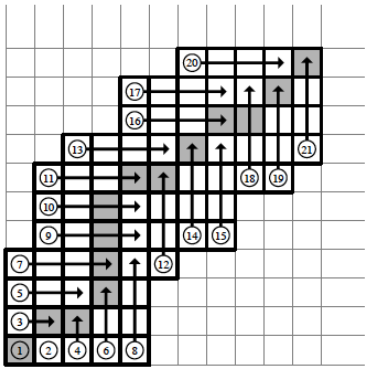

Figure 4: "An example of the on-line time warping algorithm with band width $w = 4$, showing the order of evaluation for a particular sequence of row and column increments. The axes represent time in the two files. All calculated cells are framed in bold, and the optimal path is coloured grey" (Dixon and Widmer 2005).

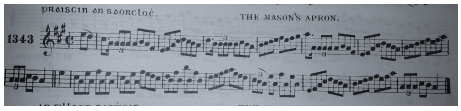

Figure 5: A transcription of "The Mason's Apron" (O'Neill and O'Neill 1903, #1343).

whole ensemble returns to play the tune a final time, but in a different key and over top of the thunderous applause for Molloy.

For analysis, four pairs of the 8-bar 'A' section were chosen: the two A repetitions by the full ensemble at the beginning; two monophonic flute solo A repetitions at the beginning of the solo section and two faster repetitions near the end; and the full ensemble A sections at the end, somewhat buried in audience applause. In addition to these live recordings, slow and fast midi recordings were made from the O'Neill transcription, transposed to the key of G as in the Chieftains' recording. The two different sections of audio compared at any two times will be called "reference track (RT)" and "aligned track (AT)," and the segments of recordings used can be heard online.[4]

Pre-processing

First, the full track was split into individual repetition segments by hand, using Audacity.[5] Next, each pair of segments was saved to a single file, with RT in the left track and AT in the right track, as required for input to the alignment software.

[4]www.music.mcgill.ca/ hannah/MUMT621/robertson_mumt6_project.html
[5]http://audacity.sourceforge.net/

Alignment

To align the two segments, the file containing the RT/AT pair was loaded into Sonic Visualizer[6] and alignment performed with the MATCH Vamp Plugin.[7] The "B-A Align" transform with a frame size of 2048 and a hop size of 512 was used, and gave as output paired alignment times in seconds, saved to a .csv file.

Merging

In Matlab, the original .wav segment files were loaded and converted to the frequency domain. For this paper, the files were further reduced to chromagrams using the LABROSA MATLAB script.[8] Chromagrams were used here for visual confirmation pitch class presence, given that no final transcription step was performed in this paper. In the future, any frequency domain representation is effective for this step as long as the window and hop sizes are the same as in the the MATCH algorithm (2048, 512).

Because the window size for both the alignment algorithm and the frequency domain conversion are the same, the alignment path matches locations in AT to each window of RT. AT was segmented at each alignment point listed in the paired path, and each AT segment was matched to a single frame in RT. To merge the two files without favouring the longer frame, the merged amplitudes were first normalized, with each segment in AT inversely scaled by its length in relation to frame. For example, in an AT segment with a duration equivalent to 1.5 frames (1.5*2048 samples), each amplitude value in the chroma vector is divided by 1.5 before being summed with the amplitude of the chroma vector of the corresponding RT frame, which by definition always has a length of 1.

Once every frame in AT has been scaled and summed to RT, the entire merged audio file is renormalized.

3 Results

The results so far are limited but seem promising. Chroma vectors for the original and merged audio are presented in Figures 6 (solo), 7 (ensemble), and 8 (ensemble with audience noise).

Pitch histograms are presented in Figure 9-11 and show the overall presence of each pitch class present in each of the three audio files: RT, AT, and merged. It is clear from Figures 9 and 10 that even though the same tune was played in the same key by each, the pitch representations vary between when played by soloist vs. ensemble. This cannot be attributed to accompaniment, as all instruments present are playing the melody, and demonstrates why it is important to capture each performance independently.

4 Future work

Several adjustments might improve alignment. Kirchhoff and Lerch point out that the choice of features used for alignment depend on the use case and audio type (2011). The MATCH algorithm's feature vectors consist of frequency bins: spectral features mapped to a linear scale at low frequencies and logarithmic scale at high frequencies (Dixon 2005). It is possible that including additional features, such as MFCCs, in the feature vector could improve alignment

[6] http://www.sonicvisualiser.org/
[7] http://www.vamp-plugins.org/download.html
[8] http://labrosa.ee.columbia.edu/matlab/chroma-ansyn/

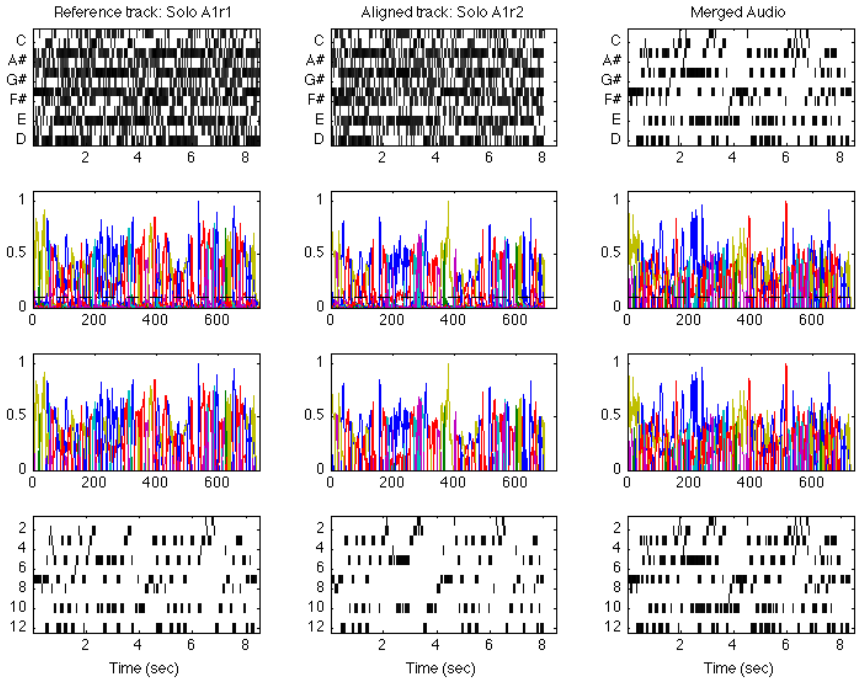

Figure 6: Chroma vectors of the original and then merged audio files. The audio in this example is from the solo flute, first two repetitions.

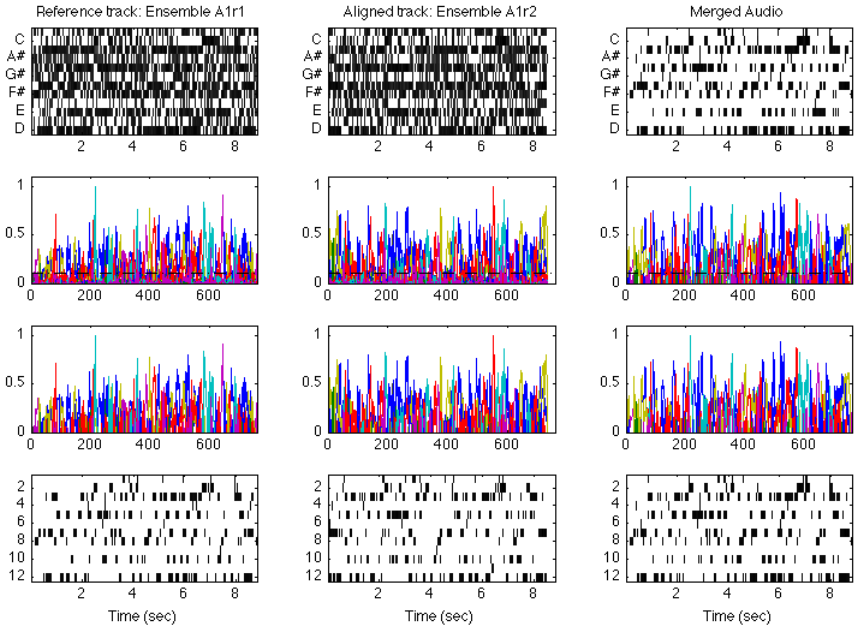

Figure 7: Chroma vectors of the original and then merged audio files. The audio in this example is from the ensemble, first repetitions.

when tailored towards specific ensemble groupings and recording environments. In addition, more testing needs to be done to determine and customize optimal window and hop sizes for the MATCH algorithm as applied to Irish traditional music. Shorter windows would enable the capturing of more short ornamentation details, as long as the window is not so short as to lose pitch information. Müller et et al. have devised a multi-scale DTW to audio synchronization that incorporates both the larger melodic contours of a melody and the smaller ornamentation (2006); it is possible that incorporating aspects of this algorithm into MATCH could improve alignment as well.

In terms of the overall alignment-to-transcription project, there are several next steps. A transcription method needs to be chosen; even though the ultimate melody is monophonic it is possible that the best type of transcription for keeping track of secondary and ornamental pitch and onsets is polyphonic, or involves some sort of weighted note array. In addition, segmentation of the original tunes into their component sections could be automated with the aid of self-similarity matrices, as is done by the Audio Research Group at the Dublin Institute of Technology (Kelly et al. 2010). As in many MIR tasks, a lack of ground truth test data is a problem for this endeavour, but perhaps transcriptions like those of Larsen (2003) can be digitized as a starting point.

Finally, it is possible that this alignment method can be used for more than just pairs of repetitions taken from the same tune. This would be very useful, as it would prevent loss of information about audio features until the very last transcription step across any audio comparison, and would also allow for fast aural comparisons. In the current form of the algorithm, however, matching between tunes with different ornamentation was successful between MIDI and solo flute but not between solo flute and the full ensemble, which indicates that the

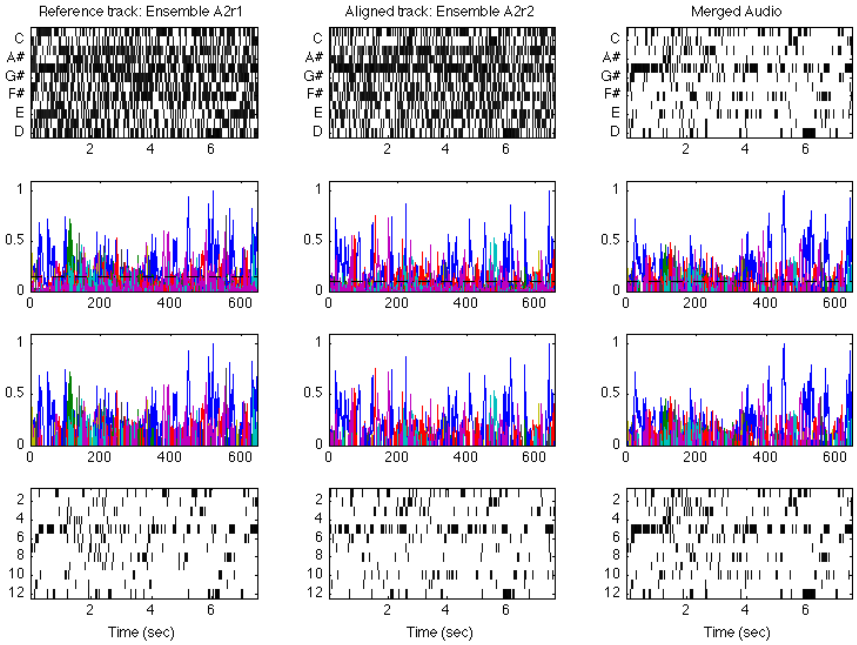

Figure 8: Chroma vectors of the original and then merged audio files. The audio in this example is from the ensemble, last repetitions.

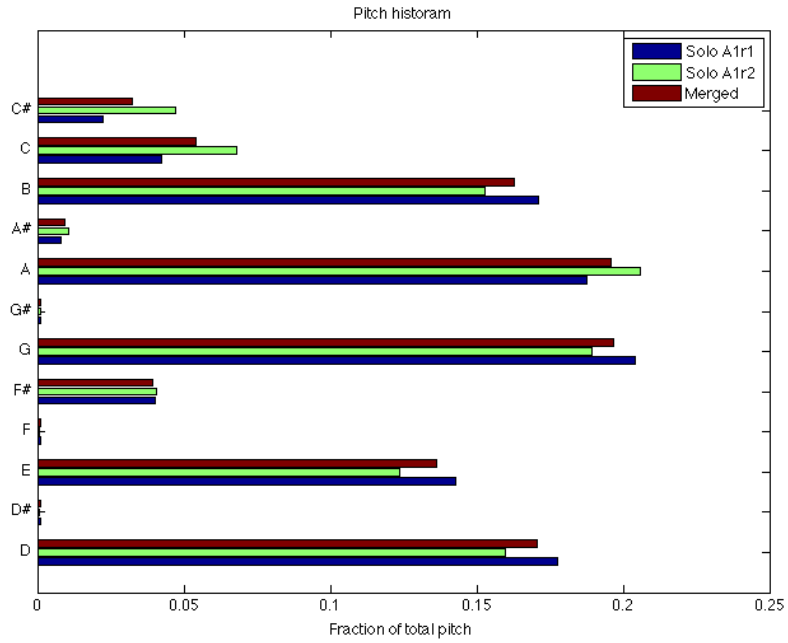

Figure 9: Chroma histogram of solo flute, first two repetitions.

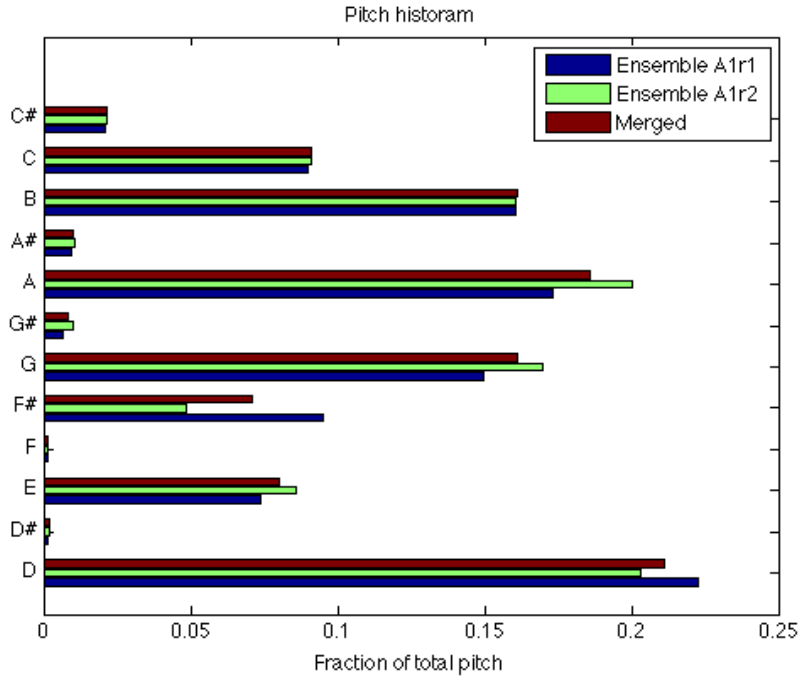

Figure 10: Chroma histogram of ensemble, first repetitions.

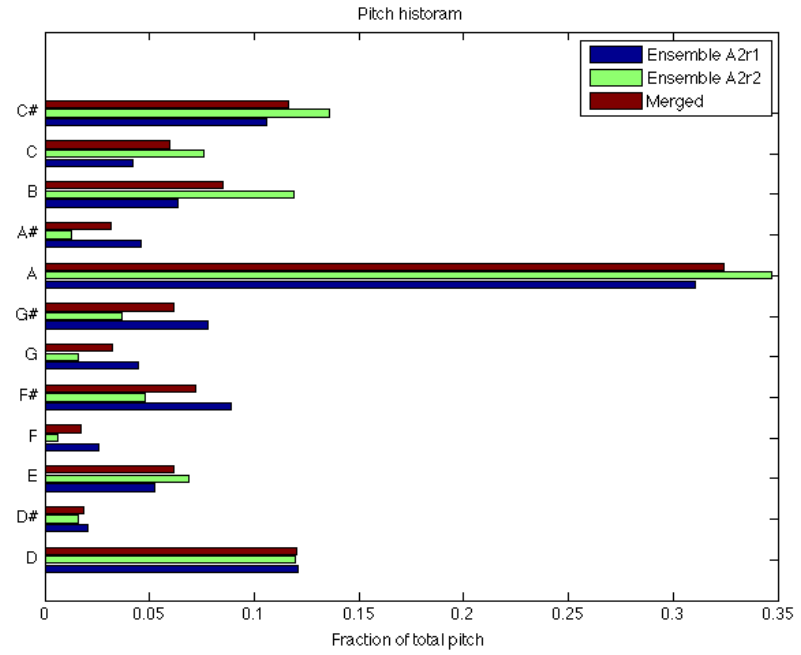

Figure 11: Chroma histogram of ensemble, last repetitions.

frequency bin features used in the MATCH algorithm are not be optimal features when comparing tunes from different recordings. An adjustment of the feature vectors used may improve this. In addition, basing alignment on features other than frequency bin onset could allow for key invariance in tune alignment. Müller and Clausen deal with transposition-invariant self-similarity matrices (2007), and perhaps these techniques could be applied at some stage in the pre-transcription alignment process.

Bibliography

Berndt, D. J., and J. Clifford. 1994. Using dynamic time warping to find patterns in time series. In *The National Conference on Artificial Intelligence (AAAI) Workshop on Knowldege Discovery in Databases*, 359–70.

Chieftains, T. 1992. *An Irish Evening: Live at the Grand Opera House, Belfast*, Chapter The Mason's Apron. BMG 09026 60916-2.

Dixon, S. 2005. Live tracking of musical performances using on-line timewarping. In *Proceedings of the International Conference on Digital Audio Effects*.

Dixon, S., and G. Widmer. 2005. MATCH: A music alignment tool chest. In *Proceedings of the International Conference on Music Information Retrieval*, 492–7.

Duggan, B. 2009. Machine annotation of traditional Irish dance music. PhD Thesis, Dublin Institute of Technology School of Computing.

Duggan, B., M. Gainza, B. O'Shea, and P. Cunningham. 2009. Compensating for expressiveness in queries to a content based music information retrieval system. In *Proceedings of International Computer Music Conference*, 33–6.

Duggan, B., B. O'Shea, M. Gainza, and P. Cunningham. 2009. The annotation of traditional Irish dance music using matt2 and tansey. In *Proceedings of the Information Technology & Telecommunication Conference, Galway Mayo Institute of Technology*.

Hillhouse, A. N. 2005. Tradition and innovation in irish instumental folk music. MA Thesis, The University of British Columbia Faculty of Music.

Hu, N., R. B. Dannenberg, and G. Tzanetakis. 2003. Polyphonic audio matching and alignment for music retrieval. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*: 185–8.

Kelly, C., M. Gainza, D. Dorran, and E. Coyle. 2010. Audio thumbnail generation of Irish traditional music. In *Proceedings of the Irish Signals and Systems Conference*.

Kirchhoff, H., and A. Lerch. 2011. Evaluation of features for audio-to-audio alignment. *Journal of New Music Research* 40 (1): 27–41.

Larsen, G. 2003. *The Essential Guide to Irish Flute and Tin Whistle*. Mel Bay Publications, Inc.

List, G. 1963. The musical significance of transcription. *Ethnomusicology* 7 (3): 193–7.

Molloy, M., and D. Lunny. 1984. *Matt Molloy*, Chapter The Humors of Drinagh. Green Linnet GLCD 3008.

Müller, M., and M. Clausen. 2007. Transposition-invariant self-similarity matrices. In *Proceedings of the International Society for Music Information Retrieval Conference*, 47–50.

Müller, M., H. Mattes, and F. Kurth. 2006. An efficient multiscale approach to audio synchronization. In *Proceedings of the International Society for Music Information Retrieval*, 192–7.

Ness, S. 2009. Content-aware visualizations of audio data in diverse contexts. PhD Thesis, University of Victoria.

Ng, A. 2002. irishtune.info for the mason's apron. Last accessed 28 April 2012, `irishtune.info`.

O'Neill, F., and J. O'Neill. 1903 (1996 Reprint). *O'Neill's Music of Ireland*. Mel Bay Publications.

Pikrakis, A., S. Theodoridis, and D. Kamaroto. 2003. Recognition of isolated musical patterns using context dependent dynamic time warping. *IEEE Transactions on Speech and Audio Processing*: 175–83.