Pitch Analysis for Active Music Discovery

Justin Salamon JUSTIN.SALAMON@NYU.EDU
Music and Audio Research Laboratory and Center for Urban Science and Progress, New York University, NY, USA

Abstract

A significant proportion of commercial music is comprised of pitched content: a melody, a bass line, a famous guitar solo, etc. Consequently, algorithms that are capable of extracting and understanding this type of pitched content open up numerous opportunities for active music discovery, ranging from query-by-humming to musical-feature-based exploration of Indian art music or recommendation based on singing style. In this talk I will describe some of my work on algorithms for pitch content analysis of music audio signals and their application to music discovery, the role of machine learning in these algorithms, and the challenge posed by the scarcity of labeled data and how we may address it.

1. Introduction

Music discovery is commonly posed as a recommendation problem: given a user's listening history (and that of other users), recommend new music the user is likely to enjoy (Celma, 2010; Schedl et al., 2015). But music discovery goes beyond the classic recommendation scenario, in which other than giving the occasional "thumbs up" the user is often passive. Paradigms for active music discovery include query-by-humming (where a search is seeded by a user-generated, often sung, sample) and music-feature-based exploration (Porter et al., 2013) for example. While "classic" music recommendation is mostly driven by collaborative-filtering (Celma, 2010), the aforementioned active music discovery paradigms rely heavily on audio content analysis. Given that a large proportion of all music contains pitched content (such as melodies, chords or solos), building systems that facilitate discovery driven by these musical facets is an attractive prospect. To do this, we require algorithms that can extract and (ideally) understand information about the pitched content of a music recording.

A concrete example is Audio Melody Extraction (AME) algorithms, which attempt to estimate the pitch (fundamental frequency or f_0) of the predominant melodic line in a music recording (Salamon et al., 2014). Such algorithms are useful, e.g., for building a query-by-humming search engine, since the user is likely to sing part of the melody as the query and thus extracting a melodic representation of all samples in our database would facilitate the matching.

2. Machine learning for pitch analysis

To date, the majority of AME algorithms in the literature are comprised of a composition of signal processing blocks followed by some form of temporal tracking (e.g. HMM+Viterbi) to select the final melodic line, see (Salamon et al., 2014) for a review. While some approaches use different flavors of unsupervised matrix decomposition to enhance the melody signal (Durrieu et al., 2010; Tachibana et al., 2010), for a long time the only approach that relied heavily on supervised learning was (Ellis & Poliner, 2006) which trained an SVM to predict the melody pitch directly from the spectrogram. In (Salamon & Gómez, 2012) we proposed the Melodia algorithm which is based on contour characterization: a salience function is computed from the audio signal highlighting predominant pitches, and these are tracked into pitch contours representing salient trajectories in time/frequency. A set of features is computed for each contour (e.g. vibrato rate, pitch variance, loudness, etc.), and their distribution for melodic and non-melodic contours is used to derive a set of heuristics for selecting the contours that belong to the melody.

The natural next step was to replace these heuristics with a supervised classifier, and in (Salamon et al., 2012b) we proposed a simple generative model based on multivariate Gaussians. In (Bittner et al., 2015) we replaced the generative model with a discriminative model (a random forest). The model outperformed its generative counterpart and was able to perform comparably to the heuristic version, with the advantage of not requiring any manually derived rules and being easily adaptable to specific mu-

Attribution: Justin Salamon, "Pitch Analysis for Active Music Discovery" *Machine Learning for Music Discovery Workshop* at the 33^{rd} *International Conference on Machine Learning*, New York, NY, 2016.

sical genres through re-training. It is worth pointing out that these models are trained on features derived from relatively high-level constructs (in terms of abstraction from the raw signal), namely the pitch contours. They illustrate how even though employing ML to perform end-to-end melody extraction remains an unsolved problem, it can be incorporated in different stages of an algorithm's processing pipeline, replacing manually engineered components.

In (Salamon et al., 2012c) we used the same set of contour features to classify the genre (or rather, singing style) of a music recording. This type of classification could be used for active music discovery where the user is interested in music with a specific singing style. Pikrakis et al. (2012) used melody extraction to identify recurring melodic patterns in Flamenco music, a type of "expert music discovery" and a common use case in computational musicology. Another example of discovery through pitch analysis is our algorithm for tonic identification in Indian classical music (Salamon et al., 2012a). After computing a histogram of the most commonly played pitches in the piece, the algorithm employs a tree classifier trained to derive a set of rules for selecting the tonic (previous approaches relied on matching of templates derived from Indian classical music theory). Both this algorithm and Melodia are used to facilitate active exploration and discovery of Indian classical music in the Dunya web platform (Porter et al., 2013).

Despite recent advances, to date no AME algorithm has significantly outperformed the purely-heuristic Melodia algorithm. I argue that this is purely an artefact of the severe lack of labeled data for melody extraction (and multiple f_0 estimation), which means existing algorithms are most likely over-fitting the evaluation sets (even the ones that have been kept private), and supervised learning approaches simply do not have sufficient data to construct a model that will generalize well to unseen data (not to mention data-hungry deep learning models).

3. Overcoming data scarcity

To train/evaluate a melody extraction algorithm we require an accurate annotation of the melodic f_0 contour on a very fine timescale (e.g. 10 ms). Annotating this contour manually is intractable, and the standard approach is to run a monophonic pitch tracker on the separate melody track (which requires access to a multitrack recording) and then manually correct any estimation errors. This correction step is still labor-intensive, and consequently most labeled datasets are on the order of tens of recordings, usually short excerpts. The small size of these datasets (and their homogeneity) makes evaluation problematic. This motivated the creation of MedleyDB (Bittner et al., 2014), a multitrack dataset of 122 (mostly) full-length songs. While this is still relatively small in ML terms, the dataset will grow as more recordings are collected from NYU's Dolan music recording studio. Multitrack recordings have also been made popular (and more accessible for research) via music games such as Guitar Hero and other research efforts such as the open multitrack testbed (De Man et al., 2014). Furthermore, since models for melody extraction are trained at the frame (or contour) level, a single recording actually corresponds to hundreds (or thousands) of training samples.

It thus appears that multitrack datasets are reaching sizes that are (not ideal but) at least sufficient for revisiting supervised approaches to pitch extraction. However, annotating continuous f_0 remains a significant barrier to the scalable generation of data for supervised learning. Alternative solutions to manual corrections include instruments outfitted with sensors that simultaneously generate audio and annotations (Emiya et al., 2010) and using MIDI-controlled instruments for annotation by playing (Su & Yang, 2015). However, such approaches are limited either in the type of sources they can use, e.g. piano, or in the annotations they can generate, e.g., discrete notes instead of continuous f_0.

In our latest work we present a method for continuous f_0 annotation based on analysis/synthesis that is fully automatic and requires minimal human effort (Salamon et al., 2016). Instead of manually correcting the pitch tracker's estimation errors we use the f_0 as the input to a wide-band harmonic modelling algorithm (Bonada, 2008) that estimates the frequency, amplitude and phase of every harmonic in the signal. We then use this information to re-synthesize the track and mix it back with the rest of the instruments in the multitrack recording, resulting in a polyphonic mixture for which we have a perfect f_0 ground truth. The method is validated by comparing the performance of several algorithms on the original and synthesized data, showing no statistically significant difference.

4. Summary

Active music discovery can be enhanced by the automatic extraction of pitched content from audio signals, examples including melody extraction for query-by-humming and tonic identification for exploring Indian art music. Due to the scarcity of labeled data, most AME algorithms do not exploit supervised learning, and those that do perform comparably but do not outperform unsupervised or heuristic methods. The recent advent of multitrack datasets and novel methods for automatic data labeling make it an exciting time to revisit supervised learning for pitch analysis in support of new (active) music discovery paradigms.

Acknowledgments

I would like to thank my co-authors on the aforementioned studies and Brian McFee for his helpful feedback.

References

Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., and Bello, J. P. MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *15th Int. Soc. for Music Info. Retrieval Conf.*, pp. 155–160, Taipei, Taiwan, Oct. 2014.

Bittner, R. M., Salamon, J., Essid, S., and Bello, J. P. Melody extraction by contour classification. In *16th Int. Soc. for Music Info. Retrieval Conf.*, Malaga, Spain, Oct. 2015.

Bonada, J. Wide-band harmonic sinusoidal modeling. In *11th Int. Conf. on Digital Audio Effects (DAFx-08)*, pp. 265–272, Espoo, Finland, Sep. 2008.

Celma, O. *Music Recommendation*. Springer, 2010.

De Man, B., Mora-Mcginity, M., Fazekas, G., and Reiss, J.D. The open multitrack testbed. In *137th Convention of the Audio Engineering Society*, Los Angeles, USA, Oct. 2014.

Durrieu, Jean-Louis, Richard, Gaël, David, Bertrand, and Févotte, Cédric. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Trans. on Audio, Speech, and Language Processing*, 18(3):564–575, March 2010.

Ellis, D. P. W. and Poliner, G. E. Classification-based melody transcription. *Machine Learning*, 65(2-3):439–456, 2006.

Emiya, V., Badeau, R., and David, B. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Trans. on Audio, Speech, and Language Processing*, 18(6):1643–1654, Aug. 2010.

Pikrakis, A., Gómez, F., Oramas, S., Días-Báñez, J. M., Mora, J., Escobar, F., Gómez, E., and Salamon, J. Tracking melodic patterns in flamenco singing by analyzing polyphonic music recordings. In *13th Int. Soc. for Music Info. Retrieval Conf.*, pp. 421–426, Porto, Portugal, Oct. 2012.

Porter, A., Sordo, M., and Serra, X. Dunya: A system for browsing audio music collections exploiting cultural context. In *14th Int. Soc. for Music Info. Retrieval Conf.*, pp. 101–106, Curitiba, Brazil, Nov. 2013.

Salamon, J. and Gómez, E. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, Aug. 2012. doi: 10.1109/TASL.2012.2188515.

Salamon, J., Gulati, S., and Serra, X. A multipitch approach to tonic identification in indian classical music. In *13th Int. Soc. for Music Info. Retrieval Conf.*, pp. 499–504, Porto, Portugal, Oct. 2012a.

Salamon, J., Peeters, G., and Röbel, A. Statistical characterisation of melodic pitch contours and its application for melody extraction. In *13th Int. Soc. for Music Info. Retrieval Conf.*, pp. 187–192, Porto, Portugal, Oct. 2012b.

Salamon, J., Rocha, B., and Gómez, E. Musical genre classification using melody features extracted from polyphonic music signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 81–84, Kyoto, Japan, Mar. 2012c.

Salamon, J., Gómez, E., Ellis, D. P. W., and Richard, G. Melody extraction from polyphonic music signals: Approaches, applications and challenges. *IEEE Signal Processing Magazine*, 31(2):118–134, Mar. 2014.

Salamon, J., Bittner, R. M., Bonada, J., Bosch, J.J., Gómez, E., and Bello, J. P. Automatic F0 annotation of multitrack datasets using an analysis/synthesis framework. *Submitted*, 2016.

Schedl, M., Knees, P., McFee, B., Bogdanov, D., and Kaminskas, M. *Recommender Systems Handbook*, chapter Music Recommender Systems, pp. 453–492. Springer US, Boston, MA, 2015.

Su, L. and Yang, Y.-H. Escaping from the abyss of manual annotation: New methodology of building polyphonic datasets for automatic music transcription. In *Int. Symp. Computer Music Multidisciplinary Research (CMMR)*, pp. 221–233, Plymouth, UK, Jun. 2015.

Tachibana, H., Ono, T., Ono, N., and Sagayama, S. Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 425–428, Mar. 2010.