UNSUPERVISED TRAINING OF DETECTION THRESHOLD FOR POLYPHONIC MUSICAL NOTE TRACKING BASED ON EVENT PERIODICITY

Tiago Fernandes Tavares[1*], *Jayme Garcia Arnal Barbedo*[2], *Romis Attux*[1], *Amauri Lopes*[1]

[1]University of Campinas
School of Electrical and Computer Engineering
Av. Albert Einstein, 400, Campinas-SP, Brazil
[2]Embrapa Agricultural Informatics
Av. Andre Tosello, 209, Campinas-SP, Brazil

ABSTRACT

A common approach to the detection of simultaneous musical notes in an acoustic recording involves defining a function that yields activation levels for each candidate musical note over time. These levels tend to be high when the note is active and low when it is not. Therefore, by applying a simple threshold decision process, it is possible to decide whether each note is active or not at a given time. Such a threshold, in general, is hard to set and has no physical meaning. In this paper, it is shown that the rhythmic characteristic of the musical signal may be used to obtain a suitable threshold. The proposed method for obtaining the threshold is shown to have a greater generalization capability over different databases.

Index Terms— Polyphonic note tracking, Transcription, Rhythm.

1. INTRODUCTION

Note tracking is the task of detecting musical notes in an acoustic signal. Note trackers may be used for several applications, such as query-by-content databases [1], automatic music tutoring software [2] and computer-assisted musical analysis [3]. Due to its usefulness, it has become an important problem in the context of Music Information Retrieval (MIR), and several approaches to this problem are evaluated yearly in the Music Information Retrieval Exchange (MIREX) [4].

Most recent note tracking systems are based on executing a function that yields activation values for each note over time. This may be understood as building an activation matrix $\boldsymbol{A}$, where $a_{p,q}$ has a high value if the note p is active in the time frame q, and a low value otherwise. After that, a simple threshold is applied so that high values are separated from low ones [5, 6, 7].

However, there has not been, so far, discussion on how to find a suitable value for the threshold. In general, it is set manually or using supervised training. Using a fixed threshold ensures that the decision system can be executed in real time, but, at the same time, it may yield sub-optimal results.

This paper proposes an unsupervised method for finding an optimal threshold for note detection in an activation matrix $\boldsymbol{A}$. The proposed method assumes that a piece of music probably presents rhythm, thus the detected events should be organized as to resemble time-wise periodicities. Therefore, the desired threshold should maximize periodicity of the detected events.

The experiments performed in this paper are based on a state-of-the-art note tracking algorithm proposed by Dessein *et al.* [5]. It was observed that using the periodicity for obtaining a threshold gives results that are equivalent to using an optimal threshold, obtained using supervised training. It is shown that the unsupervised method also works with datasets with significant timbre differences, without the need for running a new training process.

This paper is organized as follows. In Section 2, related work on note tracking is discussed. In Section 3, the proposed method for finding optimal thresholds is discussed. The experimental setup, the results and other discussions are shown in Section 4 and conclusive remarks are stated in Section 5.

2. RELATED WORK

Thresholding is a one-dimensional classification technique that has been used in many note-tracking systems. It consists of assigning an element to group 1 if its value is higher than a threshold α or to group 0 otherwise. This method is used to separate significant source activities from noise, such as thresholding the output of a detector to avoid false positives.

Many systems for note tracking rely on the factorization of a spectrogram $\boldsymbol{X}$ as a linear combination $\boldsymbol{A}$ of spectral templates $\boldsymbol{B}$ related to notes [8, 9, 6, 10, 7, 11, 5, 12, 13, 14], that is, $\boldsymbol{X} \approx \boldsymbol{BA}$. Although it is possible to obtain this approximation using many different techniques, all systems based on this approach currently use a threshold decision pro-

*The authors thank CNPq for the financial support.

cess to estimate which notes are active and which are not.

Other systems have used a nonlinear approximation of the spectrogram, rather than a linear one [15, 16, 17]. In this case, the note activation level is obtained by a nonlinear function, like a neural network. Again, activation levels for each note are yielded and systems often rely on thresholding for deciding over note activity.

Bayesian classifiers, which are base on statistical properties of the elements to be classified, have also been used for note tracking [18, 19, 20]. Implicitly, they also use the idea of thresholding [21], which is preceded by a nonlinear mapping to a space where thresholding can be applied. This is similar to the case of neural networks.

All of the systems cited above map the spectral content of the analyzed audio, over time, to a space of note activations $\boldsymbol{a}$. After that, a threshold-based decision process determines which notes are active and which are not. However, to the authors' knowledge, there has been no systematic discussion on how to obtain the threshold.

It is usual to obtain the threshold manually, which is interesting for real-time applications, as the detection sensitivity may be configured according to the user's style and needs regarding false positives and false negatives. However, this can greatly harm the system's final results for offline applications, as the manually-set threshold may not generalize for different instruments, different musicians or simply different data acquisition conditions. Furthermore, the detected activation levels often do not have any physical meaning, making it hard to set them manually without considerable testing.

This paper proposes using an expected property of the output, namely the rhythm, to automatically calculate a suitable value for the threshold. The system is designed to operate offline, using a long-term audio property to obtain its results. In the next section, the method will be described in further detail.

3. THE PROPOSED METHOD

Periodic signals are those that may be described by the expression $x(t) = x(t - \tau)$, where τ is the fundamental period of $x(t)$. In many applications, however, signals have only an approximately periodic behavior, due to either noise or the natural dynamics of the system. For this reason, it is important to deal with the concept of *periodicity level*.

The concept of periodicity level is employed by many methods for fundamental frequency estimation in audio. Two emblematic cases are the Yin method [22], which consists of estimating a value of τ that minimizes $\|x(t) - x(t - \tau)\|^2$, and the multiple fundamental frequency estimation method developed by Klapuri [23], which consists of calculating the weighted sum of Discrete Fourier Transform coefficients to detect which fundamental frequency candidates are more prominent. In both cases, a definition for the periodicity level is built based on either time-domain or frequency-domain properties of periodic signals.

The method discussed in this section aims at finding the threshold α that, when applied to the activation matrix $\boldsymbol{A}$, yields a binary detection matrix $\boldsymbol{D}$ that presents the greatest periodicity. The periodicity of $\boldsymbol{D}$ is related to the timewise organization of the detected events, and the exact nature of such events may be ignored for this purpose. To represent these structures, the vector $\boldsymbol{c} = \sum_{\forall i} d_{i,j}$ is calculated, and c_j is the number of active events in the j-th signal frame.

It is clear that the periodicity characteristics of $\boldsymbol{c}$ will be affected by changing α, as true positives are expected to be organized in a more periodic structure than events detected as false positives. The periodicity level is closely related to how much of the signal can be described using a Fourier series. It is calculated using the spectrum of $\boldsymbol{c}$, as follows.

The spectral representation $\boldsymbol{\gamma}$ of $\boldsymbol{c}$ is calculated by subtracting the mean value of $\boldsymbol{c}$ to avoid the DC component, multiplying it by a Hanning window to reduce spectral leakage, and then calculating its DFT. Both magnitude and power representations for $\boldsymbol{\gamma}$ were tested in order to determine which is more representative. The prominence of the Fourier series that can describe the spectrum is given by its Harmonic Sum Spectrum (HSS), calculated as follows:

1. $\boldsymbol{\gamma}_H \leftarrow \boldsymbol{\gamma}$,
2. $w \leftarrow 2$,
3. $\boldsymbol{u} \leftarrow \boldsymbol{\gamma}$ downsampled by a factor of w,
4. $\boldsymbol{\gamma}_H \leftarrow \boldsymbol{\gamma}_H + \boldsymbol{u}$,
5. $w \leftarrow w + 1$,
6. If $w < 9$, go back to step 3. Else end.

The maximum value of the HSS corresponds to the prominence of the strongest Fourier series that can be used to describe $\boldsymbol{c}$. As a result, the ratio $y = \max \boldsymbol{\gamma}_H / (\sum_{\forall j} \gamma_j)$ represents how well $\boldsymbol{c}$ can be described as a periodic signal, thus y is the periodicity level of $\boldsymbol{c}$. As the periodicity level has been defined, it is necessary to define a search algorithm for α.

Preliminary tests have shown that calculating the periodicity y is considerably faster than calculating the activation matrix $\boldsymbol{A}$. For this reason, the search for α may be conducted using exhaustive search, in a range in which the optimal threshold is likely to be found. In this work, the search was carried out between the mean and the maximum values of $\boldsymbol{A}$.

Experiments showing how the proposed method behaves in real applications will be shown in the next section.

4. EXPERIMENTS AND RESULTS

The experiments performed in this section were based on the automatic piano transcriber proposed by Dessein *et al.* [5]. This transcriber calculates the linear approximation

$\boldsymbol{X} \approx \boldsymbol{BA}$ by first obtaining the prototype vectors $\boldsymbol{B}$ using supervised training with pre-recorded samples, and then calculating $\boldsymbol{A}$ minimizing the beta-divergence [24] between the measured spectrogram $\boldsymbol{X}$ and the approximation $\boldsymbol{BA}$. The exact details of this method are not important, as it could be substituted by any other approach that uses the activation matrix as an intermediate step.

The database used in the following experiments consisted of 24 pieces for piano solo used for testing purposes by Polliner and Ellis [25]. The samples for obtaining $\boldsymbol{B}$ were downloaded from the Iowa University Musical Instrument Samples database [26]. Audio files were rendered from the MIDI files using the Iowa University samples and were labeled as the IOWA dataset. These audio files were later processed by adding a digital chorus effect, generating the CHORUS set, which aims at simulating an analysis over a different timbre.

In all experiments, the performance measurements, as adopted in MIREX, were calculated as follows. A note is considered correct if its pitch matches (has the same MIDI number) and its onset is within 50 ms of the onset of a note in the ground truth. The recall is defined as the ratio between the number of correct notes and the total number of notes in the ground truth. The precision is defined as the ratio between the number of correct notes and the total number of notes yielded by the automatic system. The F-Measure is defined as the harmonic mean between the recall and the precision.

The first test aimed at evaluating the correlation between the F=Measure and the periodicity level. For this purpose, the $\boldsymbol{A}$ matrix of a piece selected at random from the IOWA set was calculated, and the detection threshold was progressively increased. For each threshold value, the performance measurements and the periodicity level were measured, yielding the values seen in Figure 1.

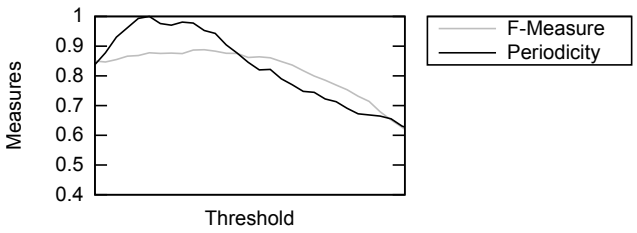

Fig. 1. Periodicity of a detection matrix $\boldsymbol{D}$ and the related F-Measures, for different threshold levels considering a single piece. The periodicity value was normalized to provide a better visualization.

As can be seen, the F-Measure curve presents a close-to-maximum value when the periodicity value is at its the global maximum. The shape of the periodicity curve is reasonably similar to the shape of the F-Measure curve, which indicates that it can be useful in finding an optimal threshold. This hypothesis must be assessed by means of tests over the whole database.

These tests were performed as follows. First, using the IOWA set, the value for α that yields the greatest F-Measure was calculated by exhaustive search. This is the value that maximizes the performance of the system when using a fixed threshold, considering the best case scenario. After that, new thresholds were obtained considering the periodicity maximization, using both the magnitude and the power spectrum representations for $\boldsymbol{\gamma}$. The mean value[1] of each performance measure, considering each one of these methods, is shown in Figure 2.

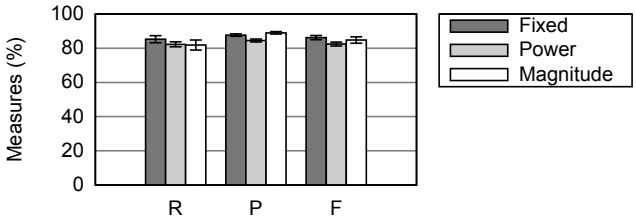

Fig. 2. Recall (R), Precision (P) and F-Measure (F) obtained for the IOWA database when using a fixed threshold and using the periodicity approach, considering both the power and the magnitude spectral representations.

As can be seen, the performance measurements, when using the proposed approach, have only a small difference in comparison with the supervised training approach. This result is important, since the periodicity approach is unsupervised, and the optimal value for the fixed threshold can only be obtained if the ground-truth result in known *a priori*.

The final test analyzed the effects of using the same threshold for different datasets. For this purpose, two thresholds were used to detect notes in the CHORUS dataset: a fixed threshold that maximizes the average F-Measure over the dataset, and a trained threshold, which was obtained by maximizing the average F-Measure over the IOWA dataset. The results obtained using both thresholds were compared to those obtained using the periodicity method, as shown in Figure 3.

The comparison shows that using the trained threshold for a different database significantly harmed the F-Measure of the system. This is a consequence of the sensible decrease in the Recall, which could not be balanced by the increase in the Precision. Thus, the optimal detection threshold may be different for different databases.

These differences in Recall and Precision were smaller when using the periodicity with a power spectrum representation. As a consequence, a higher average F-Measure was observed. This shows that, although the periodicity cannot

[1]The confidence intervals were calculated using the expression $\sigma/\sqrt{N}$, where σ is the standard deviation of the measure over the dataset and N is the number of pieces in the dataset.

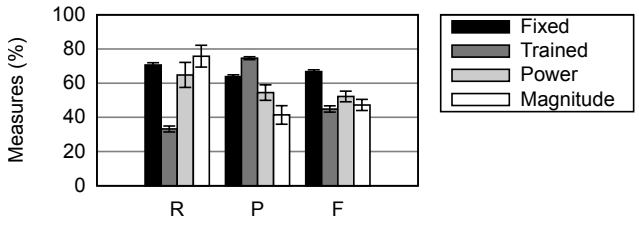

Fig. 3. Recall (R), Precision (P) and F-Measure (F) obtained for the CHORUS database when using a fixed threshold, a threshold trained in the IOWA database (Trained) and using the periodicity approach, considering both the power and the magnitude spectral representations.

yield an optimal average result, it is more robust to variations in the database than performing supervised training in a fixed threshold.

Periodicity has shown to be effective for finding a thresholds that give good average results over a database, especially when using the power spectrum representation. Also, it is obtained without the need for ground truth or manual setup, and is flexible towards different databases, which are desirable characteristics for music transcription applications.

5. CONCLUSION

This paper presented a novel training algorithm for the detection threshold of note-tracking systems. The presented method uses an important aspect of music, the rhythm, to search for an optimal threshold by means of unsupervised learning. It has been shown that the average results yielded when using the proposed method are comparable to those obtained by the best theoretical optimal threshold.

The method was tested using the note tracker proposed by Dessein *et al.* [5], chosen because of its simplicity. Nevertheless, it is applicable to any note tracking algorithm that uses an activation matrix $\boldsymbol{A}$, where $a_{i,j}$ has high values if the i-th note is active in the j-th frame and low values otherwise. This is a common step in many note tracking algorithms, which means that the periodicity-based threshold optimization is highly applicable in future research.

By using the proposed method, it is possible to run transcription algorithms over different databases, without requiring a training corpus for each one of them. As it has been shown, the unsupervised training outperforms the scenario where a database is used for training and another one is used for testing. This shows that the periodicity is an important feature to consider in note-tracking algorithms.

Future work points at finding other features that, combined to the periodicity, may result in an improvement to the system. Among those, the sparsity of the representation [27], as well as the expectancy on the time-domain progression the activation of notes [28], have shown to be interesting features. However, other musically-related features remain to be discovered.

6. REFERENCES

[1] J. Li, J. Han, Z. Shi, and J. Li, "An efficient approach to humming transcription for query-by-humming system," in *Image and Signal Processing (CISP), 2010 3rd International Congress on*, vol. 8, Oct. 2010, pp. 3746 –3749.

[2] J. Yin, Y. Wang, and D. Hsu, "Digital violin tutor: an integrated system for beginning violin learners," in *Proceedings of the 13th annual ACM international conference on Multimedia*, ser. MULTIMEDIA '05. New York, NY, USA: ACM, 2005, pp. 976–985. [Online]. Available: http://doi.acm.org/10.1145/1101149.1101353

[3] C. C. Liem, A. Hanjalic, and C. S. Sapp, "Expressivity in musical timing in relation to musical structure and interpretation: a cross-performance, audio-based approach," in *AES 42nd International Conference*, Jul. 2011.

[4] S. J. Downie, "The music information retrieval evaluation eXchange (MIREX)," *D-Lib Magazine*, vol. 12, no. 12, Dec. 2006.

[5] A. Dessein, A. Cont, and G. Lemaitre, "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands, Aug. 2010, pp. 489–494.

[6] B. Niedermayer, "Non-negative matrix division for the automatic transcription of polyphonic music," in *Proceedings of the ISMIR*, 2008.

[7] G. Grindlay and D. Ellis, "Multi-voice polyphonic music transcription using eigeninstruments," in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09. IEEE Workshop on*, Oct. 2009, pp. 53 –56.

[8] S. Sophea and S. Phon-Amnuaisuk, "Determining a suitable desired factors for nonnegative matrix factorization in polyphonic music transcription," in *Information Technology Convergence, 2007. ISITC 2007. International Symposium on*, Nov. 2007, pp. 166 –170.

[9] E. Vincent, N. Berlin, and R. Badeau, "Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, Apr. 2008, pp. 109 –112.

[10] P. D. O'Grady and S. T. Rickard, "Automatic hexaphonic guitar transcription using non-negative constraints," in *Signals and Systems Conference (ISSC 2009), IET Irish*, Jun. 2009, pp. 1 –6.

[11] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 538 –549, Mar. 2010.

[12] G. Costantini, M. Todisco, R. Perfetti, R. Basili, and D. Casali, "Svm based transcription system with short-term memory oriented to polyphonic piano music," in *MELECON 2010 - 2010 15th IEEE Mediterranean Electrotechnical Conference*, Apr. 2010, pp. 196 –201.

[13] S. Phon-Amnuaisuk, "Transcribing bach chorales using non-negative matrix factorisation," in *Audio Language and Image Processing (ICALIP), 2010 International Conference on*, Nov. 2010, pp. 688 –693.

[14] J. Han and C.-W. Chen, "Improving melody extraction using probabilistic latent component analysis," in *Proceedings of the ICASSP*, 2011.

[15] M. Marolt, "Transcription of polyphonic piano music with neural networks," in *10th Mediterranean Electrotechnical Conference, MEleCon 2000, Vol. 11*, 2000.

[16] J. P. Bello, G. Monti, M. Sandler, and M. S, "Techniques for automatic music transcription," in *in International Symposium on Music Information Retrieval*, 2000, pp. 23–25.

[17] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *Multimedia, IEEE Transactions on*, vol. 6, no. 3, pp. 439 – 449, Jun. 2004.

[18] M. Goto, "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311 – 329, 2004, special Issue on the Recognition and Organization of Real-World Sound. [Online]. Available: http://www.sciencedirect.com/science/article/B6V1C-4D07TBJ-6/2/b4f333a919e72e99569f3997499c349a

[19] A. Kobzantsev, D. Chazan, and Y. Zeevi, "Automatic transcription of piano polyphonic music," in *Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on*, Sep. 2005, pp. 414 – 418.

[20] H. Thornburg, R. Leistikow, and J. Berger, "Melody extraction and musical onset detection via probabilistic models of framewise stft peak data," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1257 –1272, May 2007.

[21] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, Oct. 2000.

[22] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[23] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*. Springer-Verlag, 2006.

[24] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *CoRR*, vol. abs/1010.1763, 2010.

[25] G. Poliner, D. Ellis, A. Ehmann, E. Gomez, S. Streich, and B. Ong, "Melody transcription from music audio: Approaches and evaluation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1247 –1256, May 2007.

[26] University of Iowa, "Musical Instrument Samples," "http://theremin.music.uiowa.edu/MIS.html", 2005.

[27] S. A. Abdallah and M. D. Plumbley, "Polyphonic music transcription by non-negative sparse coding of power spectra," in *Proc. 5th Int. Conf. Music Inf. Retrieval (ISMIR'04)*, Barcelona, Spain, 2004.

[28] J. Bello, L. Daudet, and M. Sandler, "Automatic piano transcription using frequency and time-domain information," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 2242 –2251, Nov. 2006.