Comparative Analysis of Melodia and Time-Domain Adaptive Filtering based Model for Melody Extraction from Polyphonic Music

Ranjeet Kumar Anupam Biswas Pinki Roy Yeshwant Singh
Department of Computer Science and Engineering
National Institute of Technology Silchar, Assam, India
{ranjeet_rs, anupam, pinki, yeshwant_rs}@cse.nits.ac.in

Abstract

Among the many applications of Music Information Retrieval (MIR), melody extraction is one of the most essential. It has risen to the top of the list of current research challenges in the field of MIR applications. We now need new means of defining, indexing, finding, and interacting with musical information, given the tremendous amount of music available at our fingertips. This article looked at some of the approaches that open the door to a broad variety of applications, such as automatically predicting the pitch sequence of a melody straight from the audio signal of a polyphonic music recording, commonly known as melody extraction. It is pretty easy for humans to identify the pitch of a melody, but doing so on an automated basis is very difficult and time-consuming. In this article, a comparison is made between the performance of the currently available melody extraction approach that is state-of-the-art Melodia and the technique based on time-domain adaptive filtering for melody extraction in terms of evaluation metrics introduced in MIREX 2005. Motivating by the same, this paper focuses on the discussion of datasets and state-of-the-art approaches for the extraction of the main melody from music signals. Additionally, a summary of the evaluation matrices based on which methodologies have been examined on various datasets is also present in this paper.

1 Introduction

In recent times, the music business and music suppliers such as Google, Spotify, and others have seen significant growth. By that time, the music business had also been reorganized from the cylinder age to the digital era, resulting in the current scenario where consumers may acquire millions of songs on personal phones or via cloud-based services, as well as the future. It is necessary to cope with the enormous quantity of music to search for and recover the required record effectively. At the moment, the primary issue of music suppliers is to categorize the vast number of songs available on the market based on their many components, such as rhythm, pitch, melody, and so forth. When we need to identify a particular soundtrack, we often reproduce the melody. There is a great deal of continuous progress in audio processing, which may assist customers in interacting with the songs via their sound component. Music transcription is the act of translating an aural input into a detailed description of all the notes being performed (Gómez et al., 2012). It is a task that a competent music student should be able to do very efficiently. It has, on the other hand, long been the topic of computer research. Despite this, owing to musical harmony's intricate and intentionally overlapping spectral structure, it has proved to be very difficult to achieve (Dressler, 2011).

> "It is melody that enables us to distinguish one work from another. It is melody that human beings are innately able to reproduce by singing, humming, and whistling. It is melody that makes music memorable: we are likely to recall a tune long after we have forgotten its text.". (Hofmann-Engl, 1999)

The definition given by Poliner et al. (2007) is one of the most frequently cited in the literature and is one of the most widely used:

> "roughly speaking, the melody is the single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music, and that a listener would recognize as being the 'essence' of that music when heard in comparison".

The melody is restricted to a single sound source throughout the work being examined, which is deemed the most prominent instrument or voice in the mix (Yeh et al., 2012; Klapuri, 2004). When

Proceedings of the Workshop on Speech and Music Processing (SMP-2021), pages 24–32
co-located with ICON-2021, December 16-19, 2021, Silchar, India.

polyphonic music is played, the melody is the single or monophonic pitch grouping that an audience may replicate during any moment in time to whistle or hum a piece of the music, so a large number of listeners would perceive as the 'essence' of the music when the music is played in contrast (Reddy and Rao, 2018). This concept is now susceptible to a great deal of subjective interpretation since different members of an audience may hum other portions in the aftermath of listening to a comparable piece of music.

Because of these vast number of various interpretations of melody and polyphony available, it becomes easier to categorize melody retrieval as a signal processing task than it was before.: We wish to correctly predict the series of $f0$ values that correlate to the voices or devices that are prominently featured in a clip of polyphonic music Aside from that, we must approximate the periods during which this voice is absent from the mixture (a challenge also termed as the "voicing detection" issue) (Salamon et al., 2013). While this job may seem virtually insignificant to a human listener, many of us are capable of singing along to the melodies of our favorite songs even if we have no formal musical training.

It is necessary to automatically acquire a series of frequency values of the dominant melodic line for polyphonic audio signals in order to complete the melody extraction job successfully Fig. 1. As defined by the American Institute of Music, polyphonic music is music in which at least two notes may be played at the same time on a variety of instruments (for example, bass, voice, and guitar) or on a single instrument that can play numerous notes in a single period (for example, the piano). A listener may imitate the tunes even if he or she does not have any musical training. However, when we try to automate this process, things become a little more complicated primarily due to two reasons: First, a polyphonic music signal is generated up of all the sound waves from all the devices in the track superimposed on each other. In the spectral content of the signal, various sources' frequency components overlap, making it difficult to assign particular energy levels in specific frequency bands to separate instruments' notes. Second, even after obtaining a pitches-based representation of the audio stream, we must still determine the pitch values that correspond to the dominant melody in the audio stream.

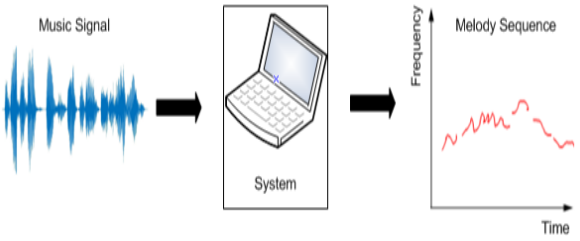

Figure 1: Melody extraction from audio signal of polyphonic music.

The task of automated melody extraction is common in the area of Music Information Retrieval (MIR). There have been a plethora of methods developed for the extraction of melodies from polyphonic music. Based on the methods used to develop them, these algorithms can be classified namely Source separation-based approach and salience-based approach (Salamon and Gómez, 2012). On the other hand, some methods do not fall under any of these categories. Algorithmic technique which is categorised as data-driven approaches, the power spectrum is directly send to deep neural network based machine learning system, which attempts to determine the melody frequency from each frame.

1.1 Salience-based approach:

Following the principles established by Scheirer Scheirer (2000), melody extraction approaches based on salience function are founded on the concept of "understanding without separation." Primarily, the following steps are required in melody extraction: The majority of the time, in preprocessing phase, to increase the melodic content of a composite signal, filtering is applied to it (**?**). Aspects of the music signal's time-domain samples are divided into frames of similar length and translated to the spectrum domain during the spectral representation and processing step. To follow the f_0 transitions in the dominant instrument, the selected window widths give sufficient frequency resolution to differentiate sinusoidal partials (Goto, 2004; Hsu and Jang, 2010). Most techniques handle the modified signal's raw spectral peaks. To put it simply, a salience function is just an evaluation of the salience of pitch values over time that is dependent on the recently identified partial peaks. Candidate melodies for the melody f_0 are considered to

be the peaks in the salience function (Klapuri, 2004). It is necessary to discover the salience peaks that correlate to actual melody peaks like the last stage in this process. The majority of algorithms directly monitor the melody peaks from the salience function.

1.2 Source separation-based approach:

It is feasible to distinguish the source responsible for the fundamental frequency from the remainder of the composite signal by using several source separation techniques (Ryynänen and Klapuri, 2008). By considering the polyphonic signal's power spectrum as the sum of lead and harmony voices, it was suggested to use source separation-based melody extraction to extract melodies (Durrieu et al., 2010). It is suggested to characterise lead vocals using a source-filter-based paradigm, and to describe accompaniment as a sum of arbitrary sources with different spectral shapes, respectively. For the source-filter model, two new models are proposed: the "Smooth-Instantaneous Mixture Model (SIMM)" and the "Smooth Gaussian-Scaled Mixture Model (SGSMM)". The SIMM is used to represent the dominating voices, while the SGSMM is used to represent the accompaniment. The expectation maximization approach is used to estimate the system model parameters. In order to determine the singer's f_0 contour from the tape, Tachibana et al. (2010), employed the temporal variability of the song.

1.3 Data-driven approach:

In contrast to data-driven strategies, which have only been examined seldom, most algorithms, as we have previously stated, are based on the salience function and source separation from music mixing. However, in recent years, this sort of method has emerged as a promising new field of investigation (Park and Yoo, 2017; Su, 2018). In order to visualise the distribution of energy in a music signal across time and frequency, spectrograms are used in preprocessing step. To minimise the leakage that happens during spectral transforms hanning window is used. The majority of researchers chose STFT because it gives time-based frequency information regarding signals whose frequency components fluctuate over time. When it comes to music recordings, a time-frequency representation is provided by the Constant—Q Transform (CQT). In compared to STFT, CQT is virtually the best fit, and the resultant representation is very low in dimensionality as a consequence. (Kum et al., 2016; Rao and Rao, 2010) devise the concept of multi-column deep neural networks for the extraction of musical notes As a classification-based technique, Using the aforementioned methodology, scientists trained each neural network how to correctly anticipate a pitch label. Author combined the output of networks and post-processed it using a hidden Markov model to deduce the melodic contour, which they labelled as a result of their efforts.

Some of the state-of-the-art approaches for extracting the melody from music signals are described in detail in this paper, which also demonstrates how these techniques are instantly applicable to MIR research. Further results of these models upon well-known datasets are also analyzed. The following is the outline for the rest of the paper. Section II describes the experimental setup in which melody extraction approach has been discussed and including dataset and performance measures are also being discussed here. Results of the assessment are reported in Section III, followed by a result analysis. finally conclusions in section IV.

2 Experimental setup

2.1 Models:

This section provides a quick overview of some of the state-of-art ways for extracting melody from a piece of music data.

2.1.1 Melodia

Salamon and Gómez (2012) proposed a model which is very popular in the filed of MIR in which he uses the Pitch Contour Characteristics to extract the melody from polyphonic music. In this model, Contour characterization and its use for melodic filtering are the most significant contributions. As seen in Fig. 2, this technique is composed of four major components that work together.

Sinusoid Extraction: Three states are present in this stage: filtering, spectral transform, and sinusoid frequency correction. In this case, an loudness filter (equal) has been applied to increase frequencies that the ear of human is more sensitive to. Then the ShortTime Fourier Transform (STFT) applied and taken small hop size to improve F0 tracking while creating pitch contours. The FFT's bin frequencies constrain the position of spectral peaks, resulting in high peak frequency estimate errors for low frequencies. For overcome this

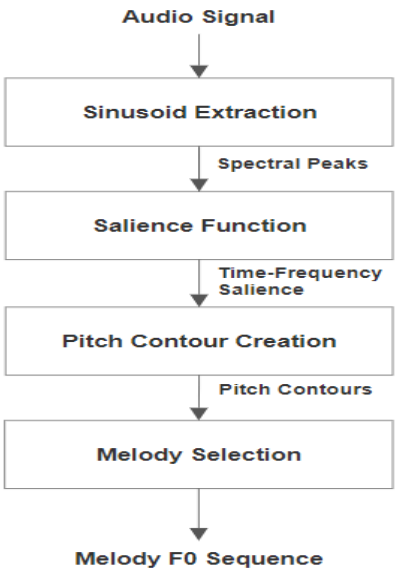

Figure 2: Block diagram of Melodia.

issue, they have calculated peak's instantaneous frequency (f_i) and amplitude by using phase spectrum.

$$\hat{f}_i = (k_i + \kappa(k_i))\frac{f_s}{N} \quad (1)$$

Salience Function: To illustrate the change in pitch salience over time, a salience function is constructed from the spectra that have been extracted and plotted against time. When this function is used, the peaks create the F0 candidates for the main melody. In this model, harmonic summation is used to calculate salience. An integer multiple (harmonic) of a frequency's salience is calculated as the sum of the weighted energies present there. The summing solely uses the spectral peaks, excluding spectral values with masking or noise. Salience function $S(b)$ at each frame can be evaluated using following definition:

$$S(b) = \sum_{h=1}^{h}\sum_{i=1}^{I} e(\hat{a}_i).g(b,h,\hat{f}_i).(\hat{a}_i)^{\beta} \quad (2)$$

where, β represents the parameter of magnitude compression and $g(b,h,\hat{f}_i)$ defines the weighting function.

Pitch Contours: It is then determined which peaks at each frame are probable melody F0 possibilities based on the salience function that was produced. Firstly, non-salient peaks are filtered out to minimize the noise contours creation. In order to determine the most appropriate parameters for contour formation, they compared contours created from various excerpts to the melodic ground truth of the excerpts and assessed them in terms of pitch accuracy and voicing accuracy. After contours creation, the main challenge is to finding the specific contours which belongs to pitch. It is necessary to establish a set of contour attributes that will be utilised to assist the system in picking melodic contours in order to do this.

Melody selection and extraction: As an alternative to picking melody contours, they formulate this issue as a contour filtering problem, with objective being to filter out any contours that are not melodic. The job of detecting when the melody is there and when it is not is referred to as voicing detection. In the last stage, choose the peaks that are associated with the primary melody from among the remaining shapes.When a frame has more than one contour, the melody is selected as the peak of the most salient contour. A frame without a contour is considered unvoiced.

2.1.2 Model based on Time-Domain Adaptive Filtering

Model developed by Reddy and Rao (2018) is basically based on time-domain adaptive filtering. The suggested approach extracts the voice melody in phases from polyphonic music. The difference in excitation intensity between the vocal and non-vocal regions of the music signal distinguishes the vocal from the non-vocal regions. The vocal regions are then split into a sequence of notes by detecting their onsets in the composite signal's frequency representation, which is then used to segment the sequence of notes further. Individual voice note melodic contours may be obtained by using adaptive zero-frequency filtering in the time domain.

In order to distinguish vocal and non-vocal areas, the music signal is first passed through a zero-frequency filter (ZFF), after which the vocal regions are segmented into a series of notes is created. According to the original ZFF approach, the monaural speech signal with a single excitation source is utilised to extract the F0 from the signal before further processing. The mean subtraction filter is designed using the time domain autocorrelation function to produce the average pitch period, which is obtained using the time domain autocorrelation function Music, on the other hand, is a composite signal that is made up of a number of

different pitched sources. The autocorrelation function cannot be used to determine the singer's resonance frequency or average pitch period because it is too complex. The next step is to detect the voiced and unvoiced segments.

Voiced and Unvoiced segment detection: Because the ZFR attenuates the vocal tract resonances to a large extent, passing the signal through it twice has considerably highlighted the source signal. When comparing the vocal source to the other sources in a polyphonic music signal with a lead voice, it is the vocal source that is most prominent. It is thus possible to identify the vocalic areas by analysing the strength of excitation (SOE) (Salamon et al., 2014). A consequence of the vocals' dominance feature is that the ZFF signal contains a significant amount of energy in the voiced areas and a very low amount of energy in the regions which is unvoiced. In order to determine the intensity of the excitation contour, the ZFF signal's slope at the instants of zero crossings of the ZFF signal is calculated.

Voiced Note Onset detection: The melody source's fundamental frequency fluctuates greatly between notes. In order to produce an accurate F0 for the lead voice, a simple mean subtraction filter is insufficient. By recognising note onsets, the voiced segments discovered before may be further split into voiced note-like regions. Signal parameters such as short-time energy, spectral magnitude, phase spectrum, etc. exhibit considerable changes at an onset. Using a low-pass filtering technique, the difference between the current frame and prior frames of a detection function, which are exponentially weighted, is calculated by

$$y(n) = F(n) - \sum_{a=1}^{A} \frac{F(n-a)}{a} \qquad (3)$$

Where the onset detection functions are represented by $F(n)$ and a represents the weighting factor.

Melody detection: A polyphonic music signal's lead voice melody may be found by removing the trend in the ZFR output of each note segment adaptively using a mean subtraction window length that corresponds to average pitch period of the lead voice in the segment. As a final step, each segmented note is subjected to Zero Frequency Filtering with a trend elimination window based on its average pitch period. In order to get the melody of the lead voice, the inverse of the difference between consecutive GCI's is calculated using the note segments that represent the GCI's.

Table 1: Dataset description.

Name	*Sample Rate (in KHZ)*	*Number of clips*
MIREX 2005	*44.1*	*13*
ADC 2004	*44.1*	*20*
IITKGP HPMD	*44.1*	*28*

2.2 Dataset:

The state-of-the-art to evaluating the melody of an audio clip have been described in previous section. In this part of article, we will cover datasets that are used to analyze the aforementioned approaches. In the form of time–frequency pairings, the datasets include music snippets as well as the accompanying melodic ground truth. Specifically, the ADC2004, Mirex05TrainFiles, and IITKGP HPMD which each included 20, 13 and 28 excerpts, were employed, respectively.

ADC 2004: This dataset contains four clips from each of the following genres: pop, jazz, daisy, opera, and MIDI (Musical Instrument Digital Interface). This dataset comprises of twenty audio clips were captured at a sample rate of 44,100 Hz for about 20 seconds each using pulse code modulation (16-bit) and a length of around 20s.

MIREX05: MIDI datasets with genres such as rock, pop, jazz, and classical piano are the most often utilised in melody extraction. This database contains 20–30 s segments of single channel 16-bit 44,100 Hz sampling .

IITKGP HPMD: (Reddy and Rao, 2018) Hindustani Classical Polyphonic Music recorded by professional musicians, which are known as IITKGP HPMD. The dataset contains 28 music clips, each of which has an average length of 30 seconds and is performed by both male and female musicians.

Table 1 lists all the datasets that were utilised in the assessment process.

2.3 Performance measures:

In order to extract the melody, techniques must perform two objectives: first to estimate which part of audio has melody and which part does not contain melody (voicing detection) and secondly, to

predict the proper predominant fundamental frequency as melody (pitch estimation). A melody extraction method usually outputs two columns, the first with fixed interval timestamps generally of 10ms and with f_0 values indicating the algorithm's pitch estimate for the melody at each timestamp in the second column. Additionally, for each frame, the algorithm specifies whether or not it believes the melody is present or missing in that particular frame. For frames when the melody is judged to be missing, this is usually expressed in a third output column or by returning an f_0 value with a negative sign. It is possible for algorithms to report a pitch label even in frames where algorithm assume the pitch is missing i.e., unvoiced frames, which is helpful for evaluating the performance of the algorithm. The accuracy of a pitch estimation algorithm may be evaluated independently of the quality of its voice detection method in this way. In another word, voicing detection mistakes do not affect pitch estimation accuracy.
The output of an algorithm is compared with the ground truth of an audio excerpt in order to assess its performance for a particular audio clip. Ground truth files are identical to output files, but they include the proper sequence of f_0 values indicating the melody of the audio clip. A monophonic pitch tracker is used to create the ground truth on the excerpt's solo melody track. In other word, every song we evaluate requires a multi track recording. In order to evaluate an algorithm, it is necessary to compare its output on a frame-by-frame basis to the ground truth file supplied by the ground truth file. The algorithm should report that it has identified the lack of melody in unvoiced frames in the ground truth. It is anticipated that the method will provide a frequency value that is identical to the one found in the ground truth for voiced frames. Some of the performance metrics frequently employed for melody extraction methods have been addressed in this section.

We calculate five global metrics based on this frame-by-frame comparison that evaluate various elements of the algorithm's performance for the audio sample in the issue. These metrics were introduced in MIREX 2005 and are now often used to assess melody extraction methods.

The uni-dimensional estimated melodic pitch frequency sequence and ground truth frequency sequence, represented by the vectors f and F, respectively (Kumar et al., 2020, 2019). The voicing indication vector is denoted by the v, whose i^{th} element $v_i = 1$ when the i^{th} frame is judged to be voiced (i.e., when a melody is present in the frame), with matching ground truth values V for the other elements in the vector. Unvoicing indications are expressed by the notation $\bar{v_i} = 1 - v_i$.

Voice Recall (VR): The algorithm's estimated voiced frame ratio to the ground truth melodic frame ratio. i.e., Frames that are really labeled as melodic/melodic frame based on ground truth.

$$VR = \frac{\sum_i v_i V_i}{\sum_i v_i} \tag{4}$$

Voicing False Alarm (VFA): The ratio of frames that were incorrectly assessed as melodic frames by the algorithm to frames that were labeled as non-melodic frames in ground truth.

$$VFA = \frac{\sum_i v_i \bar{v}_i}{\sum_i \bar{v}_i} \tag{5}$$

Raw Pitch Accuracy (RPA): The proportion of properly pitched frames compared to frames that are judged to be unpitched.

$$RPA = \frac{\sum_i v_i \tau[\zeta(f_i) - \zeta(F_i)]}{\sum_i v_i} \tag{6}$$

where, threshold feature is describe by τ and can be defined as:

$$\tau[a] = \{\, 1\ if \mid a \mid < 500 if \mid a \mid > 50 \tag{7}$$

Function ζ maps a frequency (Hz) to a perceptually motivated axis in which each semitone is split into a hundredth of a cent. A significant value number of cents may be used to indicate frequency over a reference frequency f_{ref}.

$$\zeta(f) = 1200 \log_2(\frac{f}{f_{ref}}) \tag{8}$$

Raw Chroma Accuracy (RCA): RCA works in the same way as the RPA, except it doesn't take into account the octave mistake (a common error made during melody extraction). i.e., The ground truth and approximated f_0 sequences are both assigned to a single octave.

$$RCA = \frac{\sum_i v_i \tau[\langle \zeta(f_i) - \zeta(F_i) \rangle_{12}]}{\sum_i v_i} \tag{9}$$

Table 2: Evaluation result achieved by Melodia for various testset.

Testset	VR	VFA	RPA	RCA	OA
ADC 2004	0.83	0.18	0.64	0.80	0.74
MIREX05	0.76	0.24	0.57	0.70	0.61
IITKGP HPMD	0.77	0.27	0.75	0.86	0.76

Table 3: Evaluation result achieved by Time-Domain AdaptiveFiltering-Based Method for various testset.

Testset	VR	VFA	RPA	RCA	OA
ADC 2004	0.87	0.11	0.65	0.83	0.79
MIREX05	0.80	0.20	0.62	0.73	0.65
IITKGP HPMD	0.83	0.30	0.71	0.86	0.73

Where,

$$\langle a \rangle_{12} = a - 12\lfloor \frac{a}{12} + 0.5 \rfloor \qquad (10)$$

Overall Accuracy (OA): Overall Accuracy is the percentage of frames properly identified with both pitch and voicing based on the combination of voicing detection and pitch estimation. In terms of L, OA may be characterised as:

$$OA = \frac{1}{L} \sum_{i} V_i \tau[\zeta(f_i) - \zeta(F_i)] + \bar{V}_i \bar{v}_i \qquad (11)$$

3 Result analysis

In this section we are comparing the result evaluation for Melodia and time domain adaptive filtering based model. In table 2 we can see the evaluation metrices performed on the Melodia for melody extraction and table 3 represents the result achieved by the time domain adaptive filtering based model. With the exception case of (VFA), which runs from 0 for best case to 1 for worst case scenarios, and all other measures range from worst (0) to best (1). The algorithm's efficiency is calculated by averaging the evaluation score of all music excerpts for the measure in consideration across the entire music dataset.

For analysis of these models lets check for its best possible outcome. Assuming that we have a flawless contour filtering strategy, we run tests to evaluate the best possible outcome our state-of-the-art algorithm could obtain. Taking a look at the findings that our system produced, we can make some observations. The total accuracy of the ideal contour filtering simulation, for starters, is less than

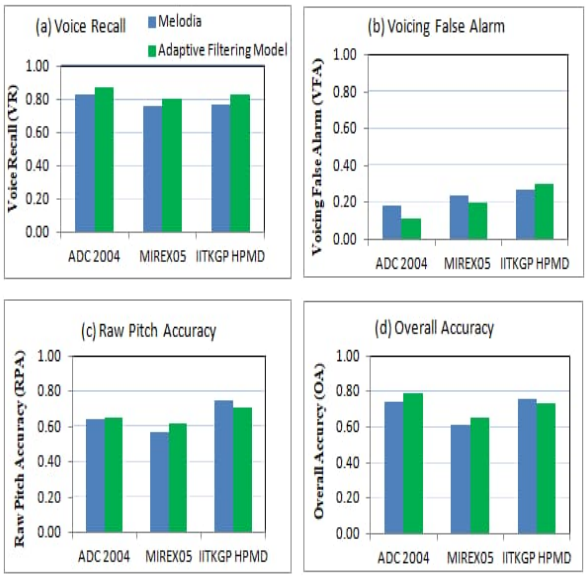

Figure 3: Performance comparison of Melodia and time domain adaptive filtering model over various test set. (a) Voicing Recall(VR) for Melodia and time domain adaptive filtering model. (b) Voicing False Alarm (VFA) for Melodia and time domain adaptive filtering model. (c) Raw Pitch Accuracy (RPA) for Melodia and time domain adaptive filtering model. (d) Overall Accuracy (OA) for Melodia and time domain adaptive filtering model.

one hundred percent, as shown in table. When comparing the datasets ADC2004 and Mirex05, we can see in Fig. 3, that the adaptive filtering based technique performs much better than Melodia in terms of RPA and OA. TWM is able to provide a resonance frequency that falls inside the ZFF's invariance range because of the predominance of the voices. On the IITKGP HPMD dataset, the time domain adaptive filtering technique achieves RP and OA results that are equivalent to those obtained with the Melodia method. It follows from this that the adaptive filtering based technique works better when dealing with music signals that have a high concentration of voices. Furthermore, owing to the impulsive nature of the percussion instrument's source, ZFF was unable to extract the proper GCI placements of the voices. In the datasets ADC2004, Mirex05, and IITKGP HPMD, an overall increases for adaptive model in VR is found, which may be ascribed to the broad dynamic range of the SoE contour used for threshold. SoE and misclassification of non-vocals into vocals have grown in IITKGP HPMD owing to the frequent stimulation of the Tabla, as well as the Drum, which causes an

increase in VFA performance.

4 Conclusion

For the purpose of automatically extracting the primary melody from a polyphonic piece of music, we investigated the performance of Melodia and a time domain adaptive filtering based model in this study. In Melodia, pitch contours were formed by combining the melodic pitch candidates that were obtained via various signal processing procedures. It is possible to identify melodic and non-melody contours by analysing these pitch contours and their distributions. In time domain adaptive filtering model, The ZFF's bandpass filtering properties are taken advantage of to create a hybrid time- and frequency-domain melody extraction approach. In polyphonic music, the SoE contour is thresholded to discern vocal and non-vocal parts. The note segment sequence is produced by sensing their frequency onsets. TWM method obtains the mean subtraction filter resonance frequency. Finally, the melody contour is retrieved by time-domain adaptive zero-frequency filtering each note segment. When using this approach, the lowered results are mostly due to the mean subtraction window length being identified often outside of the invariance range.

Acknowledgments

"This research was funded under grant number: ECR/2018/000204 by the Science & Engineering Research Board (SERB)."

References

Karin Dressler. 2011. An auditory streaming approach for melody extraction from polyphonic music. In *ISMIR*, pages 19–24.

Jean-Louis Durrieu, Gaël Richard, Bertrand David, and Cédric Févotte. 2010. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE transactions on audio, speech, and language processing*, 18(3):564–575.

Emilia Gómez, Francisco J Cañadas-Quesada, Justin Salamon, Jordi Bonada, Pedro Vera-Candeas, and Pablo Cabañas Molero. 2012. Predominant fundamental frequency estimation vs singing voice separation for the automatic transcription of accompanied flamenco singing. In *ISMIR*, pages 601–606.

Masataka Goto. 2004. A real-time music-scene-description system: Predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43(4):311–329.

Ludger Hofmann-Engl. 1999. Review of wb hewlett & e. selfridge-field, eds., melodic similarity: Concepts, procedures, and applications (cambridge, massachusetts: Mit press, 1999). *Music Theory Online*, 5(4).

Chao-Ling Hsu and Jyh-Shing Roger Jang. 2010. Singing pitch extraction by voice vibrato/tremolo estimation and instrument partial deletion. In *ISMIR*, pages 525–530.

Anssi Klapuri. 2004. *Signal processing methods for the automatic transcription of music*. Tampere University of Technology Finland.

Sangeun Kum, Changheun Oh, and Juhan Nam. 2016. Melody extraction on vocal segments using multi-column deep neural networks. In *ISMIR*, pages 819–825.

Ranjeet Kumar, Anupam Biswas, and Pinki Roy. 2019. Melody extraction from polyphonic music using deep neural network: A literature survey. *Journal of Software Engineering Tools & Technology Trends*, 6(3):16–21.

Ranjeet Kumar, Anupam Biswas, and Pinki Roy. 2020. Melody extraction from music: A comprehensive study. In *Applications of Machine Learning*, pages 141–155. Springer, Singapore.

Hyunsin Park and Chang D Yoo. 2017. Melody extraction and detection through lstm-rnn with harmonic sum loss. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2766–2770. IEEE.

Graham E Poliner, Daniel PW Ellis, Andreas F Ehmann, Emilia Gómez, Sebastian Streich, and Beesuan Ong. 2007. Melody transcription from music audio: Approaches and evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1247–1256.

Vishweshwara Rao and Preeti Rao. 2010. Vocal melody extraction in the presence of pitched accompaniment in polyphonic music. *IEEE transactions on audio, speech, and language processing*, 18(8):2145–2154.

M Gurunath Reddy and K Sreenivasa Rao. 2018. Predominant melody extraction from vocal polyphonic music signal by time-domain adaptive filtering-based method. *Circuits, Systems, and Signal Processing*, 37(7):2911–2933.

Matti P Ryynänen and Anssi P Klapuri. 2008. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86.

Justin Salamon and Emilia Gómez. 2012. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770.

Justin Salamon, Emilia Gómez, Daniel PW Ellis, and Gaël Richard. 2014. Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, 31(2):118–134.

Justin Salamon, Joan Serra, and Emilia Gómez. 2013. Tonal representations for music retrieval: from version identification to query-by-humming. *International Journal of Multimedia Information Retrieval*, 2(1):45–58.

Eric D Scheirer. 2000. Machine-listening systems. *Unpublished Ph. D. Thesis, Massachusetts Institute of Technology*.

Li Su. 2018. Vocal melody extraction using patch-based cnn. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 371–375. IEEE.

Hideyuki Tachibana, Takuma Ono, Nobutaka Ono, and Shigeki Sagayama. 2010. Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 425–428. IEEE.

Tzu-Chun Yeh, Ming-Ju Wu, Jyh-Shing Roger Jang, Wei-Lun Chang, and I-Bin Liao. 2012. A hybrid approach to singing pitch extraction based on trend estimation and hidden markov models. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 457–460. IEEE.