MELODY EXTRACTION AND DETECTION THROUGH LSTM-RNN WITH HARMONIC SUM LOSS

Hyunsin Park and Chang D. Yoo

Korea Advanced Institute of Science and Technology (KAIST)

ABSTRACT

This paper proposes a long short-term memory recurrent neural network (LSTM-RNN) for extracting melody and simultaneously detecting regions of melody from polyphonic audio using the proposed harmonic sum loss. The previous state-of-the-art algorithms have not been based on machine learning techniques and certainly not on deep architectures. The harmonics structure in melody is incorporated in the loss function to attain robustness against both octave mismatch and interference from background music. Experimental results show that the performance of the proposed method is better than or comparable to other state-of-the-art algorithms.

Index Terms— Melody extraction, LSTM-RNN

1. INTRODUCTION

Melody is a key ingredient in music composition, and melody extraction from polyphonic audio is playing an important role in music information retrieval (MIR) [1]. Given a polyphonic audio, melody is defined as the pitch sequence that a listener might reproduce when asked to whistle or hum a piece of polyphonic music [1, 2, 3]. The pitch sequence of the leading instrument or singing voice is often considered as the melody. Based on this observation, the single most dominant pitch sequence is inferred from polyphonic audio as the melody.

Many melody extraction algorithms have been proposed over the last decade [3, 4, 5, 6, 7, 8, 9, 10]. Goto and Cao *et al.* extract melody in two steps [3, 4]. First, multipitch is extracted from each frame by either expectation-maximization (EM) [3] or subharmonic summation spectrum [4]. Thereafter, both methods construct melody line based on complicated heuristic rules. In Poliner *et al.* and Ryynänen *et al.* the melody extraction problem is casted as a classification problem, and use support vector machine (SVM) [5, 6] and hidden Markov model (HMM) [7] respectively used to predict quantized target pitch. Durrieu *et al.* and Tachibana *et al.* extract melody line based on source separation techniques such as non-negative matrix factorization (NMF) [8] and harmonic/percussive sound separation (HPSS) [9]. Deep neural network was also applied to melody extraction [10]. Dressler *et al.* and Salamon *et al.* extract melody based on pitch salience [11] and auditory streaming [12], and these algorithms are regarded as state-of-the art.

Despiting reaching certain success, previous melody extraction algorithms are still far from satisfactory. The following obstacles must be overcome:

1. Accompaniment interference: Harmonic and percussive components of accompaniment signal interfere with melody extraction.
2. Octave mismatch: Each melody note can be represented as the sum of harmonically modulated fundamental frequency. However a framed melody note can be erroneously estimated as the sum of sub-multiples of the true harmonic fundamental frequency and its harmonics.
3. Difficulty in predicting dynamic variation: Sudden change in melody pitch is very hard to predict.

Conventional recurrent neural network (RNN) and long short-term memory recurrent Neunural network (LSTM-RNN) have shown good performance to sequence to sequence problems. In handwriting recognition, bidirectional LSTM has shown to preform better than HMM based systems [13]. In speech recognition, deep LSTM models show better performance than hybrid acoustic models that consist of deep neural network and HMM [14, 15]. In language modeling, conventional RNN based language model (LM) [16] and LSTM-RNN LM [17] have significantly lower perplexity than standard n-gram models. LSTM-RNN has also been applied to image description [18] and video description [19]. In music analysis, RNNs have been used for modeling temporal dependency of polyphonic music [20] and signing voice detection [21].

In this paper, LSTM-RNN based architecture for extracting melody from polyphonic is proposed. Search space of melody pitch is quantized, such that each bin corresponds to a pitch class. Region without melody is assigned to zero pitch class. Therefore, the proposed approach simultaneously performs melody extraction and melody detection that is to decide whether a particular time frame contains a melody pitch or not. We believe that LSTM-RNN can represent the dynamic variations in melody pitch sequence and is robust against octave mismatch. We also propose a loss function for

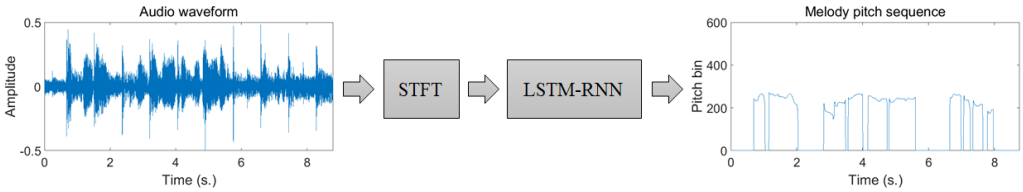

Fig. 1. Melody extraction procedure

considering harmonic structure of melody note in spectral domain. Minimizing the harmonic sum loss reduces the harmonics of the target melody note so that octave mismatch error may be decreased. The proposed model is trained using about 8 hours polyphonic audio dataset. Two well-known databases of ADC2004 and MIREX2005 are used for evaluation of the proposed model.

This paper is organized as follows. Section 2 describes the melody extraction method using LSTM-RNN with harmonic sum loss. Section 3 presents the melody extraction experiments, and finally Section 4 concludes the paper.

2. SEQUENTIAL CLASSIFICATION USING LSTM-RNN FOR MELODY EXTRACTION

2.1. Background

The LSTM-RNN has special units called memory blocks in the recurrent hidden layer. The memory block contains memory cells with input, output, and forget gates. The memory cell stores past information and the gates control the flow.

In this paper, an LSTM takes an input sequence $\mathbf{x}$ and generate hidden state sequence $\mathbf{h}$ by using the following equations,

$$\begin{aligned} i_t &= \sigma(\mathbf{W}_{ix}x_t + \mathbf{W}_{ih}h_{t-1} + \mathbf{W}_{ic}c_{t-1} + b_i), \\ f_t &= \sigma(\mathbf{W}_{fx}x_t + \mathbf{W}_{fh}h_{t-1} + \mathbf{W}_{fc}c_{t-1} + b_f), \\ c_t &= f_t \odot c_{t-1} + i_t \odot tanh(\mathbf{W}_{cx}x_t + \mathbf{W}_{ch}h_{t-1} + b_c), \\ o_t &= \sigma(\mathbf{W}_{ox}x_t + \mathbf{W}_{oh}h_{t-1} + \mathbf{W}_{oc}c_t + b_o), \\ h_t &= o_t \odot tanh(c_t), \end{aligned}$$

where the $\mathbf{W}$ terms denote weight matrices (e.g. $\mathbf{W}_{ix}$ represent the weight matrix from the input x_t to the input gate i_t), σ is the logistic sigmoid function, and i, f, o, and c are the input gate, forget gate, output gate and cell activation, respectively. Operator $\odot$ denotes element-wise product.

To reduce the computational complexity of learning LSTM, long short-term memory projected (LSTMP) was proposed in [15]. This architecture projects the output of LSTM on $\mathbf{W}_{rh}$ to reduce the size of the recurrent input to LSTM such that,

$$r_t = \mathbf{W}_{rh}h_t.$$

The projection is feed into the LSTM. Thus, h_{t-1} in above LSTM equations should be replaced with r_{t-1}.

2.2. Melody extraction using LSTM-RNN

The procedure for melody extraction and detection is represented in Figure 1. Any void in the melody estimation represents region of no melody. Given the short-time Fourier transform (STFT) $\mathbf{X} \in \mathcal{R}^{D\times T}$ of an audio, an LSTM-RNN architecture predicts the melody pitch sequence of the audio. T is the number of frames and D is the dimension of the extracted feature. The architecture simultaneously detect region without melody. Then, LSTM-RNN architecture is used to predict melody pitch sequence $\mathbf{y} \in \mathcal{B}^T$ from the $\mathbf{X}$, where $\mathcal{B}$ is the melody pitch space.

Predicted melody pitch is linearly quantized in cent scale. Without loss of generality, region without melody is assigned to zero pitch. The proposed architecture simultaneously extracts and detects melody. To predict y_t from LSTM output h_t or LSTMP output r_t, softmax output layer is used. Let the LSTM-RNN parameter set be Θ. In learning of LSTM-RNN, cross-entropy loss function of LSTM-RNN output and melody pitch label is used for the main objective function $L(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{X}, \Theta))$. Here, $\mathbf{Y} \in \{0,1\}^{|\mathcal{B}|\times T}$ and $\hat{\mathbf{Y}} \in [0,1]^{|\mathcal{B}|\times T}$ are one-hot representation sequences of label sequence $\mathbf{y}$ and posterior distribution sequence of prediction $\hat{\mathbf{y}}$, respectively.

2.3. Harmonic sum loss

The learning criterion of the LSTM-RNN for melody extraction is given as,

$$L(\Theta) = L(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{X}, \Theta)) + \alpha_h L_h(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{X}, \Theta)), \qquad (1)$$

where α_h and $L_h(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{X}, \Theta))$ are balancing weight and harmonic sum loss, respectively. The harmonic sum loss is motivated by the fact that harmonics of melody pitch appear in octave intervals in the spectral domain and reduces error due to octave mismatch problem. The harmonic sum loss function is given as

$$L_h(\mathbf{Y}, \hat{\mathbf{Y}}(\mathbf{X}, \Theta)) = \sum_t \mathbf{p}^T \hat{\mathbf{y}}_t \qquad (2)$$

where $\mathbf{p} \in \{0,1\}^{|\mathcal{B}|}$ is a column vector to sum all the harmonics of the melody pitch y_t. The j-th element of $\mathbf{p}$ is given

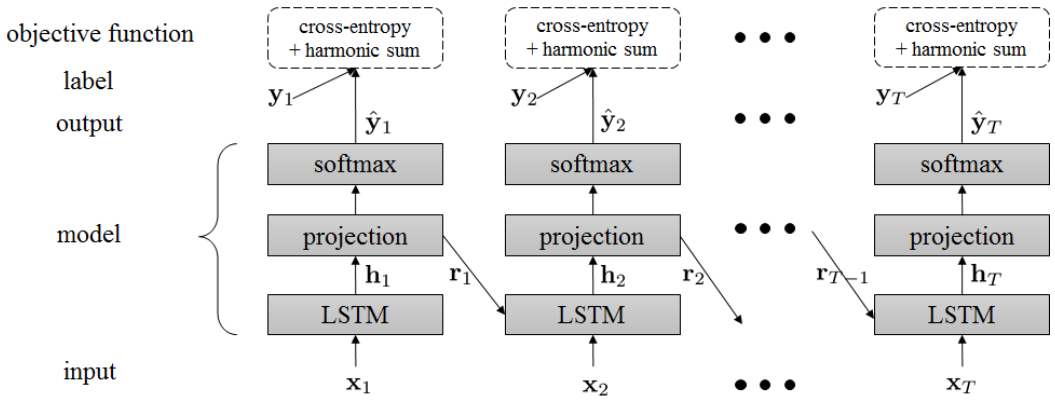

Fig. 2. LSTM-RNN for melody extraction

as

$$p_j = \begin{cases} 1 & \text{if } \mathrm{mod}(j,w) = \mathrm{mod}(y_t,w) \text{ and } j \neq y_t, \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

where $\mathrm{mod}(j,w)$ is the remainder after division of j by w. Here, $w < |\mathcal{B}|$ is the length of one octave. In the time frame without melody pitch, $\mathbf{p}$ becomes zero vector so that harmonic structure is not considered. We think that minimizing $\mathbf{p}^T \hat{\mathbf{y}}_t$ may reduce the octave mismatch errors.

Generally, in a label space of classification task, there may be groups that have similar instances. Increasing the discriminative capability for labels that belong to same group is a key issue for decreasing the classification error. If we can know the label groups in advance, we can increase the discriminative capability by explicitly decreasing the output score of model for the labels in the group in which the true label is included.

For training LSTM-RNN, the standard mini-batch stochastic gradient descent with error back-propagation algorithm can be used to minimize the objective function with respect to the network parameters Θ. The proposed model is represented in Figure 2. In this paper, LSTMP that contains recurrent projection layer is used for modeling dynamics of the hidden states.

3. EXPERIMENTS

3.1. Database and evaluation metrics

The proposed algorithm was evaluated and compared with other melody extraction algorithms using the following databases: the ISMIR 2004 Audio Description Contest (A4) database [22], the MIREX 2005 Audio Melody Extraction Competition training (M5T) database [22], the Real World Computing Music (RWC) database [23], the Multimedia Information Retrieval lab 1000 song clips (MIR-1K) [24], and the MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research [25]. Please refer to the references for details. In training LSTM, MIR-1K and MedleyDB are used for training set and RWC is used for validation set. When using the RWC database, we randomly select 20 audio clips of vocal songs as described in [26]. For test set, A4 and M5T are used.

The performance of the proposed algorithm was evaluated in terms of raw pitch accuracy (RPA) and raw chroma accuracy (RCA). The RPA is defined as the proportion of the correct frames. The RCA is defined in the same manner as the RPA; but, it allows octave transpositions. The estimated melody is considered "correct" when the absolute value of the difference between the ground-truth frequency and estimated frequency is less than or equal to 50 cents (5 bins in this paper). Moreover, melody detection accuracy (MDA) is also used as evaluation measure.

3.2. Training

For learning the LSTM-RNN architecture, audio are resampled to 16kHz. The search range for melody pitch is set between 55Hz and 1760Hz corresponding to the 1st and 600th bin in the cent scale. The Hanning window was used with 48ms frame length and 10ms frame hop-size for spectral analysis. The number of DFT points is set to 2048. By taking absolute value of the DFT coefficients, 1025 dimensional vector is used to input vector at each frame. Each input vector is normalized to have zero mean and unit variance.

The search range of melody pitch is set from 55Hz to 1760Hz (five octaves), and it is quantized into 600 bins on a cent scale similar to [11]. Frequency p is converted to bin number y as

$$y(p) = \mathrm{floor}\left(\frac{1200 \cdot \log_2\left(\frac{p}{55}\right)}{10} + 1 \right).$$

All networks are trained using stochastic gradient descent with learning rate 10^{-5}, momentum 0.9, and random initial

weights are drawn uniformly from [-0.05 0.05]. The maximum number of iterations was set to 20. When the validation loss increases in each iteration, the learning rate is decreased by a factor of 0.618. The behavior of the gradient update is controlled by RmsProp algorithm [27]. We used the Microsoft Research CNTK [28] software for the following experiments. The training time of one LSTM-RNN model with 2000 cells and 1000 hidden units took about 8 hours with a machine which has i7-4820K CPU, 64GB RAM and GTX 980 Ti GPU with 6GB memory.

3.3. Experimental results

Given a test polyphonic audio, predicted melody pitch sequence is obtained by taking pitch bin which has the maximum output in frame-by-frame manner. In selecting the maximum bin, the non-melody output is omitted. This predicted melody pitch sequence is used for measuring RPA and RCA. Besides, binary sequence for melody detection by comparing is extracted for measuring MDA. When the non-melody bin has the maximum value among the outputs, the time frame is considered to contain no melody.

Table 1. Melody extraction and detection results by adjusting the sizes of cell (C) and projected hidden layer (P)

	(C, P)	$(1000, 500)$	$(2000, 500)$	$(2000, 1000)$
A4	RPA	80.4	**80.9**	80.4
	RCA	83.3	**84.0**	83.9
	MDA	65.8	**66.8**	63.2
M5T	RPA	84.7	**87.2**	86.8
	RCA	85.6	**87.6**	87.3
	MDA	72.4	74.4	**74.5**

Table 1 shows the melody extraction and detection results by adjusting the sizes of cell and projected hidden layer. In this experiments, LSTM-RNN is trained without harmonic sum loss. From the results, we use the model with 2000 cells and 500 recurrent projection states for the follwing experimetns.

Table 2. Melody extraction and detection results by adjusting the balancing weight α_h

	α_h	0	0.01	0.1	1.0
A4	RPA	80.9	81.2	80.8	**82.2**
	RCA	84.0	83.3	82.9	**84.5**
	MDA	66.8	**69.5**	65.6	64.6
M5T	RPA	87.2	86.3	**87.6**	86.8
	RCA	87.6	86.9	**87.8**	87.4
	MDA	74.4	73.2	**74.7**	74.6

Table 2 shows the melody extraction and detection results by adjusting harmonic loss weight α_h. When $\alpha_h = 0$, LSTM-RNN is trained without harmonic sum loss. From the results, we can find the effectiveness of the harmonic sum loss. The effectiveness of the harmonic sum loss is confirmed from the results

Table 3 shows RPA and RCA results of the proposed algorithm compared with other melody extraction algorithms by Poliner et al. [5], Ryynänen et al. [7], Cao et al. [4], Durrieu et al. [8], Dressler [12], Tachibana et al. [9], Joo et al. [29], Salamon et al. [30], and Jo et al. [26]. Their performances were cited from the results of the MIREX Audio Melody Extraction Contests or their papers. In terms of the RPA and RCA, the proposed algorithm showed comparable results for A4 database and outperformed other algorithms for M5T database.

Table 3. Result comparison of melody extraction algorithms

DB	Algorithms	RPA	RCA
A4	Poliner et al. [5]	73.2	76.4
	Ryynänen et al. [7]	82.4	83.5
	Cao et al. [4]	85.1	86.3
	Durrieu et al. [8]	85.7	86.2
	Dressler [31]	**87.1**	**87.6**
	Tachibana et al. [9]	62.9	73.4
	Joo et al. [29]	79.6	85.3
	Salamon et al. [11]	79.0	81.0
	Jo et al. [26]	80.7	**87.6**
	LSTM-RNN	80.9	84.0
	LSTM-RNN-HSL	82.2	84.5
M5T	Ryynänen et al. [7]	67.3	69.1
	Cao et al. [4]	82.2	
	Durrieu et al. [8]	74.5	79.6
	Tachibana et al. [9]	74.0	76.7
	Salamon et al. [11]	83.0	84.0
	Jo et al. [26]	81.2	83.7
	LSTM-RNN	87.2	87.6
	LSTM-RNN-HSL	**87.6**	**87.8**

4. CONCLUSION

A melody extraction algorithm based on the LSTM-RNN is proposed in this paper. This paper assumes that LSTM-RNN can represent latent dynamics of melody pitch sequence. To train LSTM-RNN, cross entropy loss with harmonic sum loss is used for objective function. Experimental results show that performance of the proposed algorithm is better than or comparable to other famous melody extraction algorithms in terms of the RPA and RCA.

5. ACKNOWLEDGEMENT

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.B0101-16-0307, Basic Software Research in Human-level Lifelong Machine Learning (Machine Learning Center))

6. REFERENCES

[1] G. E. Poliner, D. P. W. Ellis, and A. F. Ehmann, "Melody transcription from music audio: approach and evaluation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1247–1256, 2007.

[2] R. P. Paiva, T. Mendes, and A. Cardoso, "Melody detection in polyphonic musical signals: exploiting perceptual rules, note salience, and melodic smoothness," *Computer Music Journal*, vol. 30, no. 4, pp. 80–98, 2006.

[3] 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.

[4] C. Cao, M. Li, J. Liu, and Yonghong Yan, "Singing melody extraction in polyphonic music by harmonic tracking," in *ISMIR*, 2007.

[5] G. E. Poliner and D. P. W. Ellis, "A classification approach to melody transcription," in *ISMIR*, 2005.

[6] D. P. W. Ellis and G. E. Poliner, "Classification-based melody transcription," *Machine Learning*, vol. 65, pp. 439–456, 2006.

[7] M. P. Ryynänen and A. P. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Computer Music Journal*, vol. 32, no. 3, pp. 72–86, 2008.

[8] J. L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for main melody extraction from polyphonic audio signals," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 564–575, 2010.

[9] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source," in *ICASSP*, 2010, pp. 425–428.

[10] Sangeun Kum, Changheun Oh, and Juhan Nam, "Melody extraction on vocal segments using multi-column deep neural networks," in *ISMIR*, 2016.

[11] Justin Salamon and Emilia Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[12] Karin Dressler, "An auditory streaming approach for melody extraction from polyphonic music.," in *ISMIR*, 2011, pp. 19–24.

[13] Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[14] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013, pp. 6645–6649.

[15] Hasim Sak, Andrew W Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling.," in *INTERSPEECH*, 2014, pp. 338–342.

[16] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur, "Recurrent neural network based language model.," in *INTERSPEECH*, 2010, vol. 2, p. 3.

[17] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney, "Lstm neural networks for language modeling.," in *INTERSPEECH*, 2012, pp. 194–197.

[18] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[19] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville, "Describing videos by exploiting temporal structure," in *ICCV*, 2015, pp. 4507–4515.

[20] Nicolas Boulanger-lewandowski, Yoshua Bengio, and Pascal Vincent, "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription," in *ICML*, 2012, pp. 1159–1166.

[21] Simon Leglaive, Romain Hennequin, and Roland Badeau, "Singing voice detection with deep recurrent neural networks," in *ICASSP*, 2015, pp. 121–125.

[22] "http://labrosa.ee.columbia.edu/projects/melody/," .

[23] M. Goto, H. Hashiguchi, T. Nishimura, and Ryuichi Oka, "Rwc music database: Popular, classical, and jazz music databases," in *ISMIR*, 2002, pp. 287–288.

[24] "https://sites.google.com/site/unvoicedsoundseparation/mir-1k," .

[25] Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research.," in *ISMIR*, 2014, pp. 155–160.

[26] S. Jo, C. D. Yoo, and A. Doucet, "Melody tracking based on sequential bayesian model," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1216–1227, 2011.

[27] Tijmen Tieleman and Geoffrey Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, vol. 4, pp. 2, 2012.

[28] Amit Agarwal et al., "An introduction to computational networks and the computational network toolkit," Tech. Rep. MSR-TR-2014-112, August 2014.

[29] S. Joo, S. Jo, and C. D. Yoo, "Melody extraction from polyphonic audio signal mirex2010," in *MIREX Audio Melody Extraction Contest Abstracts*, 2010.

[30] J. Salamon and E. Gómez, "Melody extraction from polyphonic music audio," in *MIREX Audio Melody Extraction Contest Abstracts*, 2010.

[31] Karin Dressler, "Audio melody extraction for mirex 2014," in *MIREX Audio Melody Extraction Contest Abstracts*, 2014.