AUTOMATIC TRANSCRIPTION OF DIATONIC HARMONICA RECORDINGS

Filipe Lins[†] *Marcelo Johann*[†] *Emmanouil Benetos*[⋆] *Rodrigo Schramm*[†]

[†]Universidade Federal do Rio Grande do Sul / BR
[⋆]Queen Mary University of London / UK

ABSTRACT

This paper presents a method for automatic transcription of the diatonic Harmonica instrument. It estimates the multi-pitch activations through a spectrogram factorisation framework. This framework is based on Probabilistic Latent Component Analysis (PLCA) and uses a fixed 4-dimensional dictionary with spectral templates extracted from Harmonica's instrument timbre. Methods based on spectrogram factorisation may suffer from local-optima issues in the presence of harmonic overlap or considerable timbre variability. To alleviate this issue, we propose a set of harmonic constraints that are inherent to the Harmonica instrument note layout or are caused by specific diatonic Harmonica playing techniques. These constraints help to guide the factorisation process until convergence into meaningful multi-pitch activations is achieved. This work also builds a new audio dataset containing solo recordings of diatonic Harmonica excerpts and the respective multi-pitch annotations. We compare our proposed approach against multiple baseline techniques for automatic music transcription on this dataset and report the results based on frame-based F-measure statistics.

Index Terms— automatic music transcription, harmonic constraints, probabilistic latent component analysis

1. INTRODUCTION

The Harmonica instrument can emit melodic and polyphonic sounds (single notes, intervals and chords), having great appeal in the versatility of techniques and timbres [1]. The high degree of musical expressiveness allied with its disseminated presence in the musical world scene opens a wide range of possibilities in the Music Information Retrieval (MIR) field, including applications such as music performance analysis, musicology, and tools for music instrument learning.

Research on Automatic Music Transcription (AMT) has addressed multi-pitch detection for an extensive range of musical instruments [2]. In [3] the authors proposed a system that detects multi-pitch candidates among spectral peaks, and [4] calculates the strength of multi-pitch candidates as a weighted sum of the amplitudes of its harmonic partials. Non-Negative Matrix Factorisation (NMF) [5, 6, 7] and Probabilistic Latent Component Analysis (PLCA) [8, 9, 10] have also been widely used in polyphonic music transcription systems. Supervised implementations of spectrogram factorisation techniques allow these systems to learn spectral templates that represent the timbre of specific instruments [7, 9], improving multi-pitch detection accuracy. A language model is integrated into the PLCA-based approach described in [10], and [11] combines music key information in order to derive a more musically meaningful transcription. Both the language model and key information, respectively proposed in these techniques, are integrated into the spectrogram factorisation process and not implemented as a pre/post-processing step. Multi-pitch detection has also been addressed by recurrent and deep convolutional neural network approaches [12, 13], however, these methods typically focus on piano transcription.

In this work, we exploit the Harmonica instrument design to achieve a more accurate multi-pitch transcription. To the authors' knowledge, this is the first attempt to create a system for automatic transcription of harmonica, as well as the first approach in the broader field of MIR focusing on harmonica as an instrument. We propose a set of constraints based on the pitch range and on its layout (placement) in the instrument body. These constraints regard the instrument key and Harmonica playing techniques, such as blow, draw, pitch bend, single notes, and chords. Contrary to [11], the proposed system does not only integrate key information but also exploits the playability possibilities given the physiological limitations imposed to the performer by the Harmonica instrument. We map these aspects as a set of hard masks which are combined into a PLCA based pipeline. A new audio dataset with Harmonica recordings and respective annotations is built to evaluate this work.

2. DIATONIC HARMONICA

The standard ten hole major diatonic harmonicas come in all twelve keys of tonal music and allow music performances over a complete seven key note major scale in each possible key configuration of the harmonica. The pitch range of diatonic Harmonicas covers three octaves. Unlike other wind instruments where the player can only blow, the Harmonica

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. EB is supported by a RAEng Research Fellowship (RF/128).

instrument also allows drawing. Each hole in the harmonica has one specific pitch for the blowing technique and another related for the drawing technique. Many additional notes from outside the major scale can be acquired by bending certain draw and blow notes. Furthermore, the instrument has a set of possible chords and intervals that can be played concomitantly, depending on the hole and mouth (lips and tongue) positions. Figure 1 shows a diagram with these Harmonica characteristics. Possibilities of musical note emissions are present over and below each hole, enumerated from 1 to 10, together with the number of semitone intervals from the reference key.

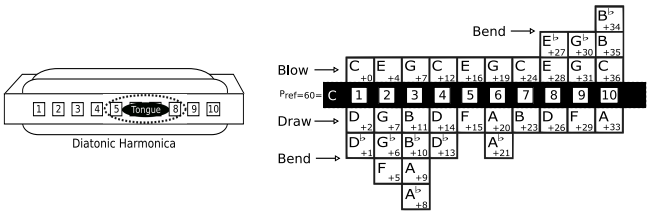

Fig. 1. Playing possibilities on the diatonic Harmonica accordingly to the blow, draw and bend regions.

2.1. Harmonic Constraints

Let p denote pitch, m denote a specific harmonic constraint and r denote the Harmonica key. $\Psi_{p,m,r}$ are fixed harmonic constraints implemented as hard masks, regarding distinct pitch combinations and the respective Harmonica key indexed by r. The weights (priors) at each possible p in these masks are chosen using empirical analysis of the Harmonica playing possibilities and aim to avoid unreachable pitch combinations. For example, let us define p_{ref} as the reference tonic of a specific Harmonica key (e.g., Harmonica in C has $p_{ref} = 60$). If the player is blowing the Harmonica, then there are possible pitches over $p \in \{p_{ref}+\{0, 4, 7, 12, 16, 19, 24, 28, 31, 36\}\}$ (blow pitch region in Figure 1). In this case, any pitch activation over the draw region (i.e. $p \in \{p_{ref} + \{2, 7, 11, 14, 17, 21, 23, 26, 29, 33\}\}$ should not be allowed. Moreover, only pitches inside the spanned pitch range over one octave are allowed to be active[1].

The proposed sets of constraints can be organised in distinct weighting levels L. For this research, we have tested three options: 1) L_1 – chords; 2) L_2 – melodic notes over the blow and the draw regions; and 3) L_3 – single notes over the bend regions. These weights are used in the construction of a mask as shown in Figure 2. The model allows the prioritisation among these levels by choosing relative weights (priors) for each pitch region. For the results of Section 4, we found $L_1 = 0.46$, $L_2 = 1.0$ and $L_3 = 0.46$ as good options based on a grid search (cross-validation scheme) varying the balance among these three weighting levels. Figure 2 illustrates one example of a set of masks for the diatonic Harmonica tuned in C ($p_{ref} = 60$). Each column in this grid represents one mask m. The entire set of masks (matrix) can be adjusted for different instrument keys by shifting these prototypes (columns) horizontally, accordingly to the desired Harmonica key.

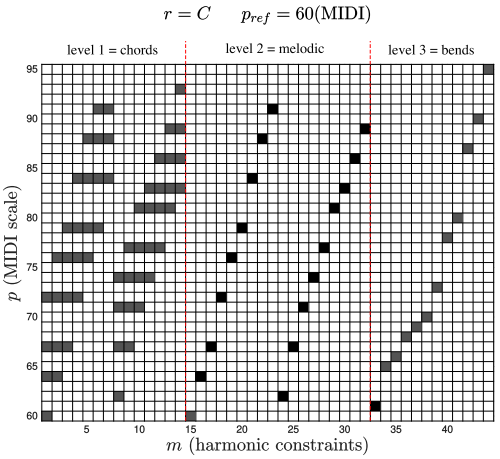

Fig. 2. Hard masks for Harmonica in C. Each column represents one mask and the weights are illustrated in gray scale (darker colours mean higher values and white means zero).

3. MODEL

The proposed system computes the pitch activations of the Harmonica instrument over time by implementing a spectrogram factorisation method based on PLCA [9]. A major point of our proposed factorisation is the exploitation of the harmonic constraints that are inherent to the instrument note layout or are caused by specific playing techniques. The inclusion of these constraints in the model avoids undesirable concurrent pitch activations. The input of our method is the variable-Q transform (VQT) [14] spectrogram $V_{\omega,t} \in \mathbb{R}^{\Omega \times T}$, where ω denotes frequency (60 bins per octave) and t time. This model uses a fixed dictionary of log-spectral templates extracted from Harmonicas regarding distinct tonal keys. As in [9], the non-negative frame level normalised log-frequency spectrogram $V_{\omega,t}$ is approximated by a bivariate probability distribution $P(\omega, t)$. Our model factorises it as:

$$P(\omega, t) = P(t) \sum_{p,f,k,m,r} \Phi P_t(f|p) P_t(k|p) P_t(p, m, r). \quad (1)$$

In this model, variable $p \in \{55, ..., 96\}$ indicates pitch in MIDI scale (12-tone equal temperament). $f \in \{1, ..., 5\}$ denotes the tuning deviation in 20 cent resolution ($f = 3$ for ideal tuning)[2]. Variable k indexes multiple spectral templates for each pitch estimate regarding distinct Harmonicas' timbres (eg. specific brands) and also different playing styles for a

[1] This restriction is based on the fact that blowing or drawing more than four Harmonica holes is very unlikely.

[2] Pitch bend technique allows the harmonica's player to create portamento.

given pitch. Variable $r \in \{\text{G, Bb, C}\}$ denotes the key of the diatonic Harmonica (among 3 possibilities[3]), and $m \in \{1, ..., 44\}$ denotes the inherent harmonic constraints.

The first term $P(t)$ in the right side of Eq. (1) is the normalised spectrogram energy (known quantity). $\Phi = P(\omega|p, f, k)$ is the set of fixed pre-built spectral templates. $P_t(f|p)$ estimates the pitch deviation and $P_t(k|p)$ gives the likely timbre contribution of several Harmonicas' spectral templates present in the dictionary. The joint distribution $P_t(p, m, r)$ is in turn decomposed as:

$$P_t(p, m, r) \propto P_t(p|m, r)P_t(m|r)P(r)\Psi_{p,m,r} \quad (2)$$

where the constrained pitch activations are estimated through $P_t(p|m, r)$. $P_t(m|r)$ gives the activation at each modelled constraint for each Harmonica key r, and $P(r)$ is the overall Harmonica instrument key probability across the audio recording. $\Psi_{p,m,r}$ denotes the set of hard masks which are built based on the inherent harmonic constraints of the instrument.

We use the expectation-maximization (EM) algorithm [15] to factorise $P(\omega, t)$ by iteratively estimating the unknown model parameters $P_t(f|p)$, $P_t(k|p)$, $P_t(p|m, r)$, $P_t(m|r)$, and $P(r)$. The derivation of Eqs. (3)-(8) follows the procedure stated in [16]. In the *Expectation* step we compute the posterior as:

$$P_t(p, f, k, m, r|\omega) \propto \Phi P_t(f|p)P_t(k|p)P_t(p, m, r) \quad (3)$$

In the *Maximization* step, each unknown model parameter is then updated using the posterior from Eq. (3):

$$P_t(f|p) \propto \sum_{k,m,r,\omega} V_{\omega,t}P_t(p, f, k, m, r|\omega) \quad (4)$$

$$P_t(k|p) \propto \sum_{f,m,r,\omega} V_{\omega,t}P_t(p, f, k, m, r|\omega) \quad (5)$$

$$P_t(p|m, r) \propto \Big(\sum_{f,k,\omega} V_{\omega,t}P_t(p, f, k, m, r|\omega) \Big)^{\alpha_1} \quad (6)$$

$$P_t(m|r) \propto \Big(\sum_{f,k,p,\omega} V_{\omega,t}P_t(p, f, k, m, r|\omega) \Big)^{\alpha_2} \quad (7)$$

$$P(r) \propto \Big(\sum_{p,f,k,m,\omega,t} V_{\omega,t}P_t(p, f, k, m, r|\omega) \Big)^{\alpha_2} \quad (8)$$

where $\alpha_1 = 1.1$ and $\alpha_2 = 4$ are based on heuristics [17] to enforce sparsity over $P_t(p|m, r)$, $P_t(m|r)$ and $P(r)$, respectively. Temporal continuity is applied over $P_t(p|m, r)$ and $P_t(m|r)$ estimates, before each normalisation step at each EM iteration through a median filter span of 150ms. Model parameters are randomly initialised, and the EM algorithm iterates over Eqs. (3)-(8). In our experiments, we use 30 iterations.

The model output is given by the pitch activation matrix $B(p, t) = P(t)P_t(p, \hat{m}, \hat{r})$, where $\hat{m} = \arg\max_m P_t(m|\hat{r})$, and $\hat{r} = \arg\max_r P(r)$. A fixed threshold ($\rho = 0.05$), estimated during our experiments with leave-one-out cross-validation, is applied to $B(p, t)$ in order to get the final binary multi-pitch detection.

[3] This model can be theoretically extended to 12 keys.

3.1. Dictionary extraction

Dictionary $\Phi = P(\omega|p, f, k)$ with spectral templates is built based on recordings of melodic scales over the entire pitch range of the diatonic Harmonica. These recordings are obtained from the RWC dataset [18]. A spectral template is extracted using the variable-Q transform [14] with 60 bins per octave, for each time frame, pitch and key. The fundamental frequency (pitch) of each template is estimated by processing the audio signal in the time domain with an autocorrelation function (based on the YIN algorithm [19]). The set of spectral templates is then pre-shifted across log-frequency in order to support tuning deviations $\pm[20, 40]$ cent and is stored into a 4-dimensional tensor $P(\omega|p, f, k)$. Similar to [10], we incorporate multiple estimates from a common pitch by replacing the set of template estimates that fall inside the same pitch bin by its metrically trimmed mean, discarding 20% of the samples as possible outliers. Eventual missing pitch candidates are filled by a linear replication process [20].

4. EXPERIMENTS

We have conducted experiments in order to evaluate the multi-pitch detection and key identification capabilities of this proposed model. The results are presented using the F-measure, and the complete pipeline is compared with other six public available algorithms for multi-pitch detection. Model parameters ρ, α_1, α_2, L_1, L_2 and L_3 are estimated using grid search through a leave-one-out cross-validation scheme.

4.1. Dataset

To the authors's knowledge, there is no public available audio dataset containing solo performances of Harmonica plus annotations of the respective multi-pitch activations. To measure the performance of our proposed model we have built a new dataset containing 42 solo recordings: 15 in key C, 13 in key Bb, and 14 in key G. These performances have melodic sequences and chords in varied combinations and styles. The respective recordings were captured with both high and low quality audio systems, and all have a sample rate of 44.1kHz and 16 bit per sample. Annotations were manually added by an expert.

4.2. Evaluation

We evaluate the multi-pitch detection capabilities of our system in a frame-based fashion, using an evenly-spaced time-grid with hop size of 9.5ms (VQT time resolution).

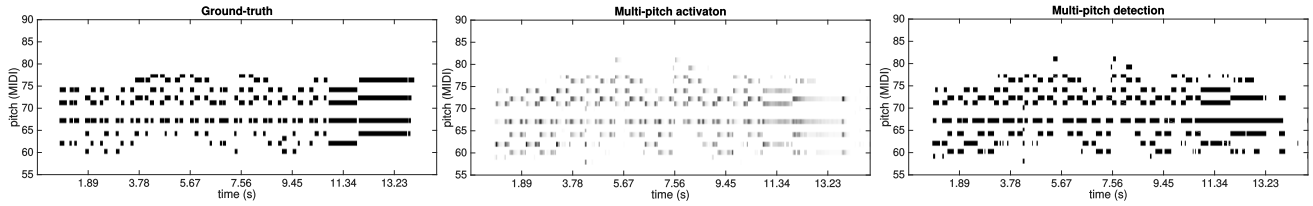

Fig. 3. Multi-pitch detection applied to *harmonicasolo001* audio file: ground-truth (left); multi-pitch activations obtained with HARM-B (middle); final multi-pitch detections (right).

For the multi-pitch detection task we analyse the results of two model configurations, named here as HARM-A and HARM-B. HARM-A implements the spectrogram factorisation without any harmonic constraints. HARM-B uses the complete model with the three proposed weight levels for the harmonic constraints (L_1, L_2, L_3).

Our results are compared with other publicly available algorithms for multi-pitch detection: Pertusa & Iñesta [3], Vincent et al. [21], Klapuri [4], Salamon et al. [22], Böck & Schedl [12], and Hawthorne et al. [13]. Pertusa & Iñesta detect multi-pitch candidates selected among spectral peaks. Vincent et al. perform multi-pitch detection using an adaptive spectral decomposition based on unsupervised NMF. Klapuri calculates the salience of F0 candidates as a weighted sum of the amplitudes of its harmonic partials. Salamon et al. estimate time continuous sequences of pitch candidates grouped using auditory streaming cues. Böck & Schedl and Hawthorne et al. are systems designed for piano transcription using recurrent and deep convolutional neural networks. All benchmark algorithms were used as provided by the respective authors, in self-contained scripts/libraries. In these experiments, we kept the original trained neural network from Böck & Schedl, but we retrained the system proposed by Hawthorne et al., using our harmonica dataset and 10-fold cross-validation.

4.3. Results

The obtained frame-based pitch estimations are evaluated through measures of precision ($\boldsymbol{P}$), recall ($\boldsymbol{R}$) and F-measure ($\boldsymbol{F}$) [23]. Figure 3 presents one example of multi-pitch detection output estimated from one audio recording in the Harmonica audio dataset. Table 1 shows the comparative results when applying all the aforementioned techniques and model configurations to the Harmonica dataset. As can be seen in these results, our proposed method outperforms in most cases the comparative approaches. HARM-B outperforms HARM-A substantially, indicating that the proposed harmonic constraints can drive the acoustic model to a more meaningful factorisation. The final F-measure obtained with the HARM-B model is very close to the one achieved by the method proposed in Pertusa & Iñesta [3]. A visual inspection of the output activation matrices and the respective ground-truth shows that a major part of missed pitch detections in our models (HARM-A, HARM-B) is caused by false positives at the related first harmonic positions. These missed pitch detection indicate some possible poor spectral representation in the dictionary of templates (Φ), which incorrectly enhances the first overtone.

Algorithm	$\boldsymbol{P}$		$\boldsymbol{R}$		$\boldsymbol{F}$	
	avg.%	std.%	avg.%	std.%	avg.%	std.%
Pertusa & Iñesta [3]	66.04	16.62	71.33	16.48	67.84	15.08
Vincent et al. [21]	39.85	12.98	**77.14**	13.45	51.35	11.03
Klapuri [4]	38.41	13.38	66.31	18.34	47.69	14.68
Salamon et al. [22]	56.25	20.48	59.30	17.11	55.24	15.96
Böck & Schedl [12]	22.11	15.79	29.16	18.80	23.60	14.87
Hawthorne et al. [13]	**72.55**	8.77	58.08	17.56	63.23	13.06
HARM-A $\rho = 0.05$	62.16	19.28	67.62	17.35	60.75	11.38
HARM-B $\rho = 0.05$	69.02	15.95	69.63	13.73	**68.13**	12.52

Table 1. Multi-pitch detection results.

5. CONCLUSIONS

In this paper we presented a multi-pitch detection system for the diatonic Harmonica instrument[4]. A set of harmonic constraints are integrated into a PLCA-based model, guiding the factorisation process until the convergence into meaningful multi-pitch pitch activations. These constraints are based on the playability possibilities given the physiological limitations imposed to the performer by the Harmonica instrument. Experimental results regarding multi-pitch detection have shown that the integrated model (PLCA+harmonic constraints) achieved better performance when compared with the version of our system built using only the acoustic model. Also, our model outperforms most of the baseline multi-pitch detection systems, achieving the state-of-the-art performance. Our model is faster to train if compared with deep learning approaches, and shows a potential machine learning alternative in the case of small training datasets.

Avenues for future work include a better handling of overtones in the acoustic model and incremental development of the Harmonica dataset. We also plan extensive experiments to evaluate the key detection capabilities over the twelve Harmonica keys and the concomitant multi-pitch detection. Additionally, we aim to exploit temporal constraints based on chord progressions and improve the dictionary of spectral templates, regarding varied Harmonica types.

[4]Supporting material for this work is available at <http://inf.ufrgs.br/lcm/projects/amt_harmonica/>.

6. REFERENCES

[1] Julian Vogels, *Harmonica-inspired digital musical instrument design based on an existing gestural performance repertoire*, Ph.D. thesis, McGill University, 2014.

[2] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[3] Antonio Pertusa and José M. Iñesta, "Efficient methods for joint estimation of multiple fundamental frequencies in music signals," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 27–40, 2012.

[4] Anssi Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *ISMIR, Victoria*, 2006, pp. 216–221.

[5] Romain Hennequin, Roland Badeau, and Bertrand David, "Time-dependent parametric and harmonic templates in non-negative matrix factorization," in *DAFx, Graz*, 2010, pp. 246–253.

[6] Gautham J. Mysore and Paris Smaragdis, "Relative pitch estimation of multiple instruments," in *ICASSP, Taipei*, 2009, pp. 313–316.

[7] G. Grindlay and D. P. W. Ellis, "Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1159–1169, 2011.

[8] B. Fuentes, R. Badeau, and G. Richard, "Controlling the convergence rate to help parameter estimation in a plca-based model," in *EUSIPCO, Lisbon*, 2014, pp. 626–630.

[9] Emmanouil Benetos and Tillman Weyde, "An efficient temporally-constrained probabilistic model for multiple-instrument music transcription," in *ISMIR, Málaga*, 2015, pp. 701–707.

[10] R. Schramm and E. Benetos, "Automatic transcription of a cappella recordings from multiple singers," in *AES International Conference on Semantic Audio, New York*, 2017, pp. 1–8.

[11] Emmanouil Benetos, Andreas Jansson, and Tillman Weyde, "Improving automatic music transcription through key detection," in *AES International Conference on Semantic Audio, Los Angeles*, 2014, pp. 3–7.

[12] Sebastian Böck and Markus Schedl, "Polyphonic piano note transcription with recurrent neural networks.," in *ICASSP, Kyoto*, 2012, pp. 121–124.

[13] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck, "Onsets and frames: Dual-objective piano transcription," in *ISMIR, Suzhou*, 2018.

[14] Christian Schörkhuber, Anssi Klapuri, Nicki Holighaus, and Monika Dörfler, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *AES 53rd International Conference on Semantic Audio, Los Angeles*, 2014, pp. 1–8.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal Of The Royal Statistical Society, SERIES B*, vol. 39, no. 1, pp. 1–38, 1977.

[16] Madhusudana Shashanka, Bhiksha Raj, and Paris Smaragdis, "Sparse overcomplete latent variable decomposition of counts data," in *Advances in neural information processing systems*, 2008, pp. 1313–1320.

[17] Graham Grindlay and Daniel P. W. Ellis, "A probabilistic subspace model for multi-instrument polyphonic transcription," in *ISMIR, Utrecht*, 2010.

[18] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *ISMIR, Barcelona*, 2004, pp. 229–230.

[19] 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.

[20] C. de Andrade Scatolini, G. Richard, and B. Fuentes, "Multipitch estimation using a plca-based model: Impact of partial user annotation," in *ICASSP, Brisbane*, 2015, pp. 186–190.

[21] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, and Lang. Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[22] J. Salamon, G. Peeters, and A. Röbel, "Statistical characterisation of melodic pitch contours and its application for melody extraction," in *ISMIR, Porto*, 2012, pp. 187–192.

[23] Mert Bay, Andreas F. Ehmann, and J. Stephen Downie, "Evaluation of multiple-f0 estimation and tracking systems," in *ISMIR, Kobe*, 2009, pp. 315–320.