A Score-Informed Approach for Pitch Visualisation of *a Cappella* Vocal Quartet Performances

Rodrigo Schramm[1,2] **, Helena de Souza Nunes**[1] **, Emmanouil Benetos**[2]

[1]Department of Music, Universidade Federal do Rio Grande do Sul, Brazil

[2]Centre for Digital Music, Queen Mary University of London, UK

rschramm@ufrgs.br, helena.souza.nunes@ufrgs.br, emmanouil.benetos@qmul.ac.uk

Abstract

This paper presents a score-informed method for visualising the pitch content of polyphonic signals from audio recordings containing *a cappella performances* with multiple singers. A model based on and extending Probabilistic Latent Component Analysis (PLCA) is proposed for estimating the activations of multi-pitch candidates, with the support of a 4-dimensional dictionary built on spectral templates of singer vocalisations. The model is assisted through a soft masking mechanism built from the given music score during the vocal performance. Since the music score is prior knowledge of our system, the main contribution of this method is the potential frame-based visualisation of the fundamental frequencies of each vocal part, which can be further used for singing analysis including tuning, vibrato and portamento analysis. We evaluate our system on recordings of vocal quartets, including Bach Chorale and Barbershop styles. The evaluation process also takes into account possible discrepancies between the singing performance and the original music score. Experimental results show the influence of such mismatches on the final system accuracy.

1. Introduction

The visualisation of multi-pitch content (also known as pitch salience) of polyphonic music generated by multiple singers is useful information for singing analysis of tuning, vibrato, portamento and a variety of complex pitch contours. Often, multi-pitch salience is an intermediate step of automatic music transcription algorithms, which convert audio signals into a symbolic representation (such as a music score) and can further be used to support applications in music information problems, computational musicology, interactive music systems, and automatic music assessment.

A method for visualising the pitch content of polyphonic music signals was proposed in [1], where a pitch salience function was designed to produce continuous pitch values. With a similar aim, [2] presents an approach using the Fan Chirp Transform [3] for pitch visualisation. Despite the interesting results obtained in these two approaches, a robust method for visualisation of the multi-pitch content of polyphonic signals is still needed. In fact, unsupervised techniques without any additional prior information usually contain many errors since the mixture of the harmonic content from different sung notes tends to generate false positives.

Spectrogram factorisation algorithms have been extensively applied for automatic music transcription and source separation in the last decade, including approaches as non-negative matrix factorization (NMF) and probabilistic latent component analysis (PLCA) [4–7]. In these approaches, the input time-frequency representation (spectrogram) is decomposed into non-negative factors, consisting mainly of spectrum atoms and note activations. Techniques based on spectrogram factorisation have shown a straightforward framework for score-informed approaches [8, 9] since masking can be applied to the matrix's coefficients, guiding the convergence of the optimisation algorithm.

In this paper, we propose a frame-based system for visualising the pitch content from polyphonic audio recordings. Our system uses a variation of the spectrogram factorization method described in [10], and its scope focuses on audio recordings of *a cappella* performance with multi-

ple singers. The original method has shown good results for (blind) multi-pitch detection. However, it is not able to perform voice separation, i.e. assign each detected note to a specific voice type (e.g. soprano). In our new approach, we overcome the voice separation limitation by integrating the music score[1] information through a soft masking scheme. Thus, since the music score is considered prior knowledge, the central point of this new application is neither on pitch detection nor note transcription, but on the detailed visualisation of the pitch contour of each vocal part. Figure 1 shows an example of the output generated by our system. The top image in this figure shows the spectrogram estimated using the Variable-Q Transform representation [11] with 20 cent frequency resolution and frame with hop size of 20 ms. In the middle is shown the ground truth, where each colour means a vocal part (SATB), and on the bottom is shown the estimated multi-pitch visualisation obtained through the proposed score-informed model.

The remainder of this paper is organised as follows. Section 2 describes the proposed score-informed PLCA model with the soft masking procedure. Section 3 presents the experiments used to evaluate the system accuracy, including simulations of singing mistakes in order to quantify the impact caused by discrepancies between the notes that are out of the soft mask range. Section 4 draws conclusions and future work.

2. Proposed Method

Our system for multi-pitch visualisation is a simplified variant of the spectrogram factorisation process described in [10]. In this model, the input time-frequency signal representation is decomposed into several components denoting the activations of pitches, voice types, and singer-timber atoms. This factorisation is supported by the use of a fixed dictionary of log-spectral templates, which are extracted from solo singing recordings in the RWC audio dataset [12]. In order to build the dictionary, we used recordings from subjects of distinct voice types: bass, baritone, tenor, alto, soprano. The dictionary has spectral templates from 15 distinct singers

[1] We use a MIDI representation in the case of this work.

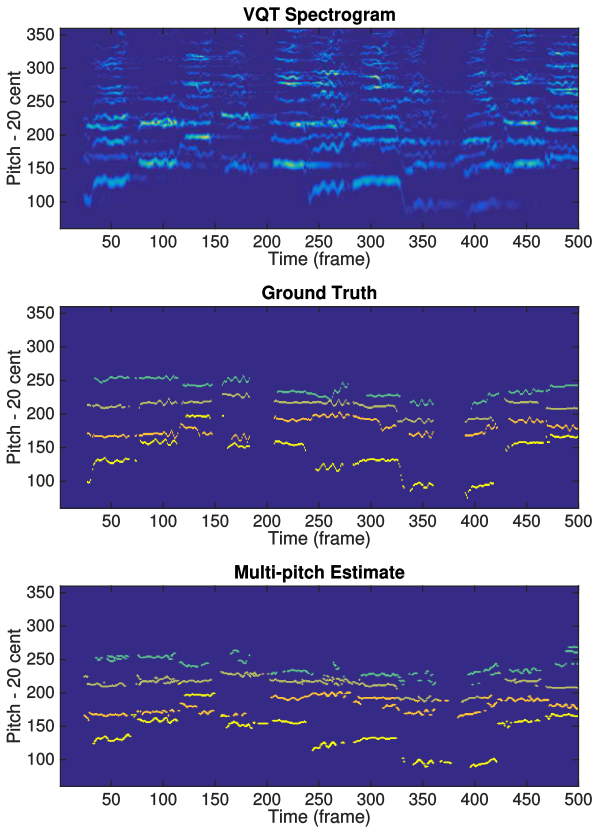


Figure 1: Multi-pitch visualization

(9 male and 6 female), that have sung sequences of notes following a chromatic scale and distinct vowels (/a/, /æ/, /i/, /ɒ/, /u/).

This collection of pre-extracted spectral templates is represented by $\Phi = P(\omega|s,p,v)$, where variable $p \in \{105, ..., 540\}$ denotes pitch in log-frequency scale (12-tone equal temperament in 20 cent resolution scale from MIDI pitch 21 to 88), s denotes the singer-timber atom index (15 distinct singers), v denotes the voice type (e.g. soprano, alto, tenor, bass). Both the input signal and the spectral templates use a normalised variable-Q transform (VQT) representation [13]. Details on the procedure for the construction of a similar dictionary are available in [10].

2.1. Multi-pitch estimation

The input VQT spectrogram is denoted as $X_{\omega,t} \in \mathbb{R}^{\Omega \times T}$, where ω denotes log-frequency and t time. In the model, $X_{\omega,t}$ is approximated by a bivariate probability distribution $P(\omega,t)$, which is in turn decomposed as:

$$P(\omega,t) = P(t)\sum_{s,p,v}\Phi P_t(s|p)P(v)P_t(p|v) \quad (1)$$

where $P(t)$ is the spectrogram energy (known quantity).

This model decomposes the probabilities of pitch and voice type as $P(v)P_t(p|v)$. $P(v)$ is the mixture weight that denotes the overall contribution of each voice type presenting in the input recording and $P_t(p|v)$ denotes the pitch activation for a specific voice type (eg. SATB) over time. The contribution of specific singer subjects from the training dictionary is modelled by $P_t(s|p)$, i.e. the singer-timber contribution per pitch over time. All unknown model parameters $P_t(s|p)$, $P_t(p|v)$, and $P(v)$ are estimated through the iterative expectation-maximization (EM) algorithm [14].

In the *Expectation* step we compute the posterior as:

$$P_t(s,p,v|\omega) = \frac{\Phi P_t(s|p)P(v)P_t(p|v)}{\sum_{s,p,v}\Phi P_t(s|p)P(v)P_t(p|v)} \quad (2)$$

In the *Maximization* step, each unknown model parameter is then updated by:

$$P_t(s|p) \propto \sum_{v,\omega} P_t(s,p,v|\omega)X_{\omega,t} \quad (3)$$

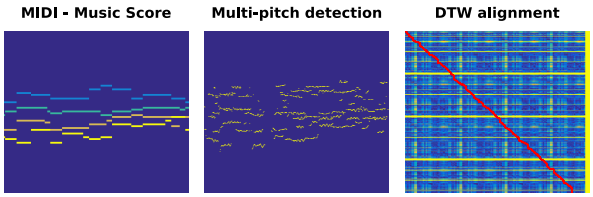
$$P_t(p|v) \propto \sum_{s,\omega} P_t(s,p,v|\omega)X_{\omega,t} \quad (4)$$

$$P(v) \propto \sum_{s,p,\omega,t} P_t(s,p,v|\omega)X_{\omega,t} \quad (5)$$

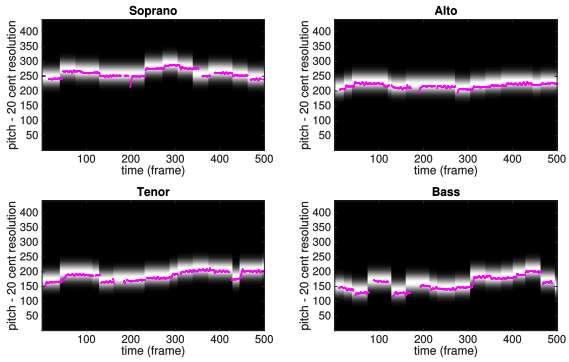
The model parameters are randomly initialised, and the EM algorithm iterates over Eqns (2)-(5). In our experiments we use 25 iterations. The output of the PLCA model is a 20 cent-resolution time-pitch representation for each voice type, given by $P(p,v,t) = P(t)P(v)P_t(p|v)$.

2.2. Soft Masking

This multi-pitch estimation model without any additional information does not perform well the voice separation. Aiming to overcome this drawback we introduce a soft masking mechanism. The soft mask is generated from the music score which was used as the reference for the singing performance.



(a)



(b)

Figure 2: Soft mask generation: a) multi-pitch detection and score alignment; b) generated soft mask per vocal part.

For automatically aligning the reference MIDI score with the audio recording made by the vocal quartet, we employ a dynamic time warping (DTW) algorithm [15]. Throughout this process, each note from the music score is time-aligned with the sung notes present in the audio recording, such that an optimal match between two given sequences (multi-pitch detection over time and the frame-based representation of the MIDI score) is found.

The proposed DTW algorithm in this paper uses a particular local cost function:

$$C(\mathbf{m}_{t_i}, \mathbf{h}_{t_j}) = \min(\sum_{p_v \in \{\mathbf{h}_{t_j}\}} \min(|p_v - \mathbf{m}_{t_i}|^2), \beta), \quad (6)$$

for measuring the distance between the list of multi-pitch estimates $\mathbf{m}_{t_i}$ at frame time t_i and the list of notes $\mathbf{h}_{t_j}$ from the music score at frame time t_j. β is a constant that imposes a limit in the cost contribution when there is no good match between points t_i and t_j.

After the time alignment the notes from the music score, at each time frame t, are used to generate the soft mask such that

$$M_t(p|v) \sim \mathcal{N}(p_t^v, \sigma_m) \quad (7)$$

follows a normal distribution $\mathcal{N}$ centred at pitch p_i^v, from the music note p at voice v, and with standard deviation $\sigma_m = 20$ bins (equivalent to 4 semitone). The soft mask is normalised along the frequency bins. Figure 2 illustrates this process.

2.3. Joint multi-pitch visualisation and voice separation

The PLCA model is initialised with random parameters, without using the soft mask scheme. The algorithm iterates until the model convergence (usually after 15 iterations). At this stage, we extract the multi-pitch detection and perform the score time alignment to generate the soft mask. After this point, the spectrogram factorisation continues over Eqns (2)-(5). However, Eqn (4) is replaced by

$$P_t(p|v) \propto \alpha(\sum_{s,\omega} P_t(s,p,v|\omega)X_{\omega,t}) + (1-\alpha)\phi_t(p|v)) \quad (8)$$

where α is a weight parameter controlling the effect of the soft mask (we have set $\alpha = 0.5$ based on our experiments) and ϕ is a hyperparameter defined as:

$$\phi_t(p|v) \propto M_t(p|v)P_t(p|v). \quad (9)$$

The hyperparameter of Eqn (9) acts as a soft mask, reweighing the pitch contribution of each voice regarding only the pitch neighbourhood previously defined by the aligned notes in the music score.

3. Evaluation

The proposed multi-pitch visualisation system is evaluated on two datasets of *a capella* recordings[2]. These datasets contain audio recordings of 26 Bach Chorales and 22 Barbershop quartets, respectively. All files are in four-channel wave format with a sample rate of 22.05 kHz and 16 bits per sample. Each channel corresponds to a particular vocal part (SATB). The Barbershop dataset contains only male voices, while the Bach Chorale dataset contains a mixture of two male and two female voices. We have extracted a frame-based pitch ground truth for each vocal part by using a monophonic pitch tracking algorithm [16] on each monophonic track. Experiments are conducted using the mix down of each audio file (i.e. polyphonic content), not the individual tracks.

We evaluate the multi-pitch visualisation and the respective voice separation capabilities of the proposed system by using metrics commonly used for multi-pitch detection and automatic transcription evaluation [7]. In these experiments, we estimate the frame-based precision, recall and F-measure as defined in the MIREX multiple-F0 estimation evaluations [17], with a frame hop size of 20 ms. For this, we use the individual voice ground truths and define voice-specific F-measures of F_s, F_a, F_t, and F_b for each respective SATB vocal part. We also define an overall voice assignment F-measure F_{va} for a given recording as the arithmetic mean of its four voice-specific F-measures.

3.1. Results

Our system generates joint multi-pitch visualisation and voice separation. For comparison with other benchmark techniques, which are originaly measured in semitone scale, we have down-sampled our system output into 88 MIDI semitones. The joint multi-pitch detection (after the binarization of the pitch activations) and voice separation output is named as MASK4-VA and MASK4-VA-20 for the semitone representation and for the 20 cent resolution, respectively.

Table 1 shows the F-measure comparisons between our proposed method and other three baseline techniques: MSINGERS-VA [10], VOCAL4-MP, and VOCAL4-VA [18]. All the benchmark techniques are PLCA-based models, but they are not score informed approaches. In addition, VOCAL4-VA also implements a language model based on Hidden Markov models to improve the voice separation results.

From the multi-pitch detection results shown in Table 1, it can be seen that MASK4-VA achieves very high F_{va} on both datasets. This

[2]Original recordings available at `http://pgmusic.com`.

Model	Bach Chorales				
	F_{va}	F_s	F_a	F_t	F_b
MSINGERS-VA	18.02	15.37	17.59	26.32	12.81
VOCAL4-MP	21.84	12.99	10.27	22.72	41.37
VOCAL4-VA	45.31	26.07	37.63	49.61	67.94
MASK4-VA	82.93	72.16	80.22	90.65	88.70
MASK4-VA-20	55.93	56.59	53.50	58.14	55.48
Model	Barbershop Quartets				
	F_{va}	F_s	F_a	F_t	F_b
MSINGERS-VA	12.29	9.70	14.03	27.93	9.48
VOCAL4-MP	18.35	2.40	10.56	16.61	43.85
VOCAL4-VA	46.92	40.01	35.57	29.76	82.34
MASK4-VA	78.75	69.14	71.97	88.87	85.02
MASK4-VA-20	51.31	47.16	50.59	57.77	49.75

Table 1: Voice assignment results.

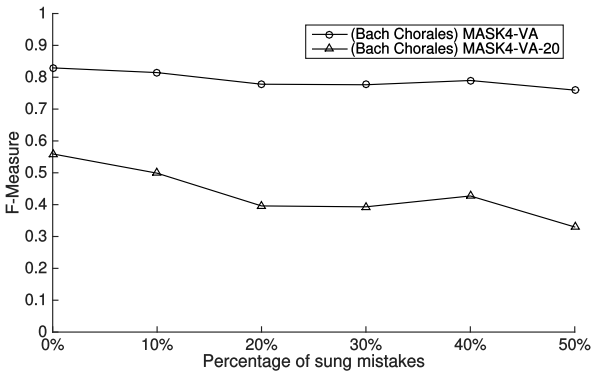


Figure 3: F-measure evolution as a function of sung mistakes.

good performance is already expected since our approach is score-informed. The F_{va} measure for the MASK4-VA-20 (20 cent resolution) is substantially lower. This mainly occurs because of small discrepancies between the ground truth and the pitch tracking in regions with vibrato.

Another important evaluation is done regarding the vulnerability of our system when there is the presence of singing mistakes, i.e., when the sung notes mismatch the target notes in the music score. To simulate this situation, we randomly change the pitch value of a percentage of notes from the music score. The plot in the Figure 3 shows the F-measure evolution as the percentage of wrong notes increases for the Bach Chorales dataset. The decrease in accuracy is caused by two main factors: 1) direct mismatch between the ground truth and the sung note; 2) incorrect time alignment from the DTW. The second factor is a consequence of the first. This result implies that our technique is more suitable for singing recordings that are reliable performances of the respective reference music scores.

4. Conclusion

In this paper, we have presented a score-informed method for visualising the pitch content of polyphonic signals from audio recordings containing *a cappella performances* with multiple singers. The proposed system uses a spectrogram factorisation model for multi-pitch detection and a soft mask scheme for aiding voice assignment. We have evaluated our system on two datasets (Bach Chorales and Barbershop styles), comparing results with baseline approaches for multi-pitch detection and voice assignment.

Experimental results have shown that the soft-masking scheme improved the multi-pitch visualisation, ensuring good voice assignment. However, our system is vulnerable to singing mistakes since the soft-mask depends on the alignment between the singing performance and the reference music score. Thus, there is certainly room for improvement. Avenues for future work include a better handling of singing mistakes during the music score alignment and the search for alternative and robust masking approaches.

5. Acknowledgement

RS is supported by a UK Newton Research Collaboration Programme Award (grant no. NRCP1617/5/46). EB is supported by a UK Royal Academy of Engineering Research Fellowship (grant no. RF/128).

References

[1] Anssi Klapuri. A method for visualizing the pitch content of polyphonic music signals. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009, Kobe International Conference Center, Kobe, Japan, October 26-30, 2009*, pages 615–620, 2009.

[2] Luis Jure, Ernesto López, Martín Rocamora, Pablo Cancela, Haldo Sponton, and Ignacio Irigaray. Pitch content visualization tools for music performance anal-

ysis. In *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012, Mosteiro S.Bento Da Vitória, Porto, Portugal, October 8-12, 2012*, pages 493–498, 2012.

[3] Martín Rocamora Pablo Cancela, Ernesto López. Fan chirp transform for music representation. In *In Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10). Graz, Austria*, pages 330–337, 2010.

[4] Shrikant Venkataramani, Nagesh Nayak, Preeti Rao, and Rajbabu Velmurugan. Vocal separation using singer-vowel priors obtained from polyphonic audio. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, pages 283–288, 2014.

[5] Gautham J. Mysore and Paris Smaragdis. Relative pitch estimation of multiple instruments. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan*, pages 313–316, 2009.

[6] G. Grindlay and D. P. W. Ellis. Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1159–1169, Oct 2011.

[7] Emmanouil Benetos and Tillman Weyde. An efficient temporally-constrained probabilistic model for multiple-instrument music transcription. In *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015*, pages 701–707, 2015.

[8] Emmanouil Benetos, Anssi Klapuri, and Simon Dixon. Score-informed transcription for automatic piano tutoring. In *Proceedings of the 20th European Signal Processing Conference, EUSIPCO 2012, Bucharest, Romania, August 27-31, 2012*, pages 2153–2157, 2012.

[9] S. Ewert, B. Pardo, M. Muller, and M. D. Plumbley. Score-informed source separation for musical audio recordings: An overview. *IEEE Signal Processing Magazine*, 31(3):116–124, May 2014.

[10] R. Schramm and E. Benetos. Automatic transcription of a cappella recordings from multiple singers. In *AES International Conference on Semantic Audio*, June 2017.

[11] C. Schörkhuber and A. Klapuri. Constant-q transform toolbox for music processing. In X. Serra, editor, *Proceedings of 7th Sound and Music Computing Conference*, Barcelona (Spanien), 12 2010. procedure: peer-reviewed.

[12] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Music genre database and musical instrument sound database. In *ISMIR*, pages 229–230, 2004.

[13] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler. A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution. In *AES 53rd Conference on Semantic Audio*, January 2014.

[14] 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*, 39(1):1–38, 1977.

[15] Meinard Müller. *Information Retrieval for Music and Motion*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[16] Matthias Mauch and Simon Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*, 2014.

[17] M. Bay, A. F. Ehmann, and J. S. Downie. Evaluation of multiple-F0 estimation and tracking systems. In *ISMIR*, pages 315–320, October 26-30 2009.

[18] Rodrigo Schramm, Andrew McLeod, Mark Steedman, and Emmanouil Benetos. Multi-pitch detection and voice assignment for a cappella recordings of multiple singers. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, 2017.