Automatic Transcription of a Cappella Recordings from Multiple Singers

Schramm, R; Benetos, E; 2017 AES International Conference on Semantic Audio

Audio Engineering Society

Conference Paper

Presented at the Conference on
Semantic Audio
2017 June 22 – 24, Erlangen, Germany

Automatic Transcription of *a Cappella* recordings from Multiple Singers

Rodrigo Schramm[1,2] and Emmanouil Benetos[2]

[1]*Computer Music Lab, Universidade Federal do Rio Grande Sul, Brazil*
[2]*Centre for Digital Music, Queen Mary University of London, UK*

Correspondence should be addressed to Rodrigo Schramm (r.schramm@qmul.ac.uk)

ABSTRACT

This work presents a spectrogram factorisation method applied to automatic music transcription of *a cappella* performances with multiple singers. A variable-Q transform representation of the audio spectrogram is factorised with the help of a 6-dimensional sparse dictionary which contains spectral templates of vowel vocalizations. A post-processing step is proposed to remove false positive pitch detections through a binary classifier, where overtone-based features are used as input into this step. Preliminary experiments have shown promising multi-pitch detection results when applied to audio recordings of Bach Chorales and Barbershop music. Comparisons made with alternative methods have shown that our approach increases the number of true positive pitch detections while the post-processing step keeps the number of false positives lower than those measured in comparative approaches.

1 Introduction

This paper is situated in the context of polyphonic music generated by multiple singers, which is a traditional form in Western music culture. Automatic music transcription is a process that converts audio signals into a symbolic representation (such as a music score) and can further be used to support applications in music informatics, musicology, interactive music systems, and automatic music assessment [1]. Despite recent advances in the field, automatic music transcription of multiple *a cappella* singers has not yet been extensively explored. This work focuses on recordings of singing performances by small groups of singers (usually vocal quartets) without instrumental accompaniment. Usually, the vocal parts are divided into specific voice types known as soprano, alto, tenor, and bass (SATB), corresponding to note ranges from high to low pitches, respectively. The main goal of this work is to analyse an audio recording, detecting and tracking multiple overlapping pitches, and finally generating an automatic transcription of the sung notes.

Spectrogram factorization algorithms, such as non-negative matrix factorization (NMF) and probabilistic latent component analysis (PLCA) have been extensively used in the task of audio source separation and multi-pitch estimation in the last decade [2, 3, 4, 5, 6, 7, 8]. The main idea of these approaches is to decompose an input time-frequency representation (such as a spectrogram) as a linear combination of non-negative factors, consisting mainly of spectral atoms and note activations. Independently of the mathematical framework, the model parameter estimation of NMF

or PLCA may suffer of *local optima* issues. A variety of approaches have been proposed to achieve more meaningful spectrogram decompositions. Kameoka et al. [2] exploited structural regularities in the music spectrograms, adding constraints and regularizations to reduce the degree of freedom of their model. Common music structural regularities are based on time-varying basis spectra (e.g. using sound states: "attack", "decay", "sustain", "release"), and are also included in other probabilistic models [9, 8]. Fuentes et al. [10] introduced the concept of brakes, stopping the convergence of the model parameters that are known to be properly initialized. Also avoiding undesirable parameter convergence, other approaches [8, 5, 4] use pre-learning steps, where the spectral atoms from specific instruments are extracted in a supervised manner. Using the Constant-Q transform (CQT) [11] as time-frequency representation, some approaches developed techniques using shift-invariant models [4, 6, 9], allowing the creation of dictionary templates with a sparse set of spectral envelopes. Shift-invariant features are also used in several recent approaches [7, 9, 12].

Despite the promising results of these template-based techniques, the high number of instruments and the notable variability of their timbre along the pitch range make the task of automatic transcription difficult. As the number of instruments in the dictionary increases, the accuracy of pitch and source assignment tends to deteriorate. Furthermore, dictionary templates for voice are hard to build due to the considerable variation in the spectral shape along several singers and the high number of phonemes. We attenuate this difficulty by proposing a method based on two steps. First, we have built a dictionary of "generic" templates from singing spectrograms in a similar way as the *eigeninstruments* concept proposed in [5]. Our goal in this step is seeking for a initial multi-pitch estimate, even containing several mistakes. We then use this initial estimate to improve the final multi-pitch transcription through a subsequent classification-based step, which implements a refinement procedure.

In the first step, we have formulated a new PLCA-based model tailored for the problem of singing voice transcription for multiple singers, following ideas from [8] which uses fixed dictionary templates, representing the log-spectrogram through a variable-Q transform. The primary purpose is to explore the factorization of the log-spectrogram into components that have a close connection with singing characteristics as voice type (soprano, contralto, tenor, baritone and bass) and the vocalization of vowels. We formulate the templates into a 6-dimensional tensor, representing pitch, voice type, vowel type, singer source, log-frequency index, and tuning deviation (with a 20 cent resolution). As a similar approach to [5], the singer source and vowel type parameters constrain the search space into a mixture-of-subspaces, clustering a large range of singers (instruments) into a small number of categories. This configuration allows us exploring temporal continuity and sparsity among these characteristics and, consequently, improving the final accuracy of the multi-pitch estimation. In this model, the voice type parameter represents each vocal part of the song (eg. SATB), with each vocal part linked to sets of human subjects (ie. singer source). Even though this work does not explicitly evaluate the model's capabilities with respect to voice assignment performance, voice type assignment is supported by this model and is left as future work.

Harmonic components of each singer source are likely to have several overlaps in the frequency domain, causing masking of fundamental frequencies and also the appearance of ghost pitch detections. This fact is a well known issue in the field of multi-pitch detection and has been addressed in algorithms for iterative multiple fundamental frequency estimation [13, 14]. Despite algorithms based on the spectral smoothness principle [15] have successfully been used to reduce issues caused by the overlapping of harmonics, this principle can not be directly applied in our case since a reasonable convergence of the PLCA algorithm depends on the matching between the spectra from the input signal and the dictionary templates. Thus, the second step of our algorithm addresses this issue by performing a binary classification of each single pitch estimate from the previous PLCA-based step. This classification is accomplished by a random forest classifier using as input overtone features extracted from the log-frequency spectrogram.

2 Model

In order to perform polyphonic transcription from multiple singers, we have formulated a variant of the spectrogram factorization-based model proposed in [8]. Our PLCA-based model is also based on a fixed dictionary of log-spectral templates and uses the normalized VQT spectrogram $V_{\omega,t} \in \mathbb{R}^{\Omega \times T}$ as input, where ω denotes frequency and t time.

The normalized log-frequency spectrogram $V_{\omega,t}$ is approximated by a bivariate probability distribution $P(\omega,t)$. $P(\omega,t)$ is in turn decomposed as:

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

where $P(t)$ is the spectrogram energy (known quantity). Variable $p \in \{21,...,108\}$ denotes pitch in MIDI scale, s denotes the singer source, o denotes the vowel type, v denotes the voice type, and f denotes tuning deviation from 12-tone equal temperament in 20 cent resolution ($f \in \{1,\ldots,5\}$, with $f = 3$ denoting ideal tuning). $\Phi = P(\omega|s,p,f,o,v)$ is the pre-extracted spectral template dictionary. $P_t(s|p)$ is the singer contribution per pitch over time, $P_t(f|p)$ is the tuning deviation per pitch over time, $P_t(o|p)$ is the time-varying vowel contribution per pitch, $P_t(v|p)$ is the voice type activation per pitch over time, and $P_t(p)$ is the pitch activation at frame t.

The factorization can be achieved by the expectation-maximization (EM) algorithm [16], where the unknown model parameters $P_t(s|p)$, $P_t(f|p)$, $P_t(o|p)$, $P_t(v|p)$, and $P_t(p)$ are iteratively estimated. In the *Expectation* step we compute the posterior as:

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

Each unknown model parameter is then updated in the *Maximization* step, using the posterior from (2):

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

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

$$P_t(o|p) \propto \sum_{s,f,v,\omega} P_t(s,p,f,o,v|\omega)V_{\omega,t} \quad (5)$$

$$P_t(v|p) \propto \sum_{s,f,o,\omega} P_t(s,p,f,o,v|\omega)V_{\omega,t} \quad (6)$$

$$P_t(p) \propto \sum_{s,f,o,v,\omega} P_t(s,p,f,o,v|\omega)V_{\omega,t} \quad (7)$$

The EM algorithm iterates from equation (2) to (7); in our experiments we have used 35 iterations and random initialization of unknown parameters. Analogous to [8], we have applied sparsity constraints on $P_t(o|p)$ (vowel type) and $P_t(v|p)$ (voice type). These constraints are based on the assumption that a vowel utterance as well as a voice type for a specific pitch (also regarding distinct singers in the template dictionary) are very unlikely to be co-occurring at time t by distinct singers. Thus, these constraints help towards the estimation of a meaningful solution. As in [6], the output of the transcription model is a semitone-scale pitch activity matrix and a pitch shifting tensor, respectively given by $P(p,t) = P(t)P_t(p)$ and $P(f,p,t) = P(t)P_t(p)P_t(f|p)$. By stacking together slices of $P(f,p,t)$ for all values of p, we can create a 20 cent-resolution time-pitch representation:

$$P(f',t) = [P(f,21,t)...P(f,108,t)] \quad (8)$$

where $f' = 1,...,880$ denotes pitch in 20 cent resolution.

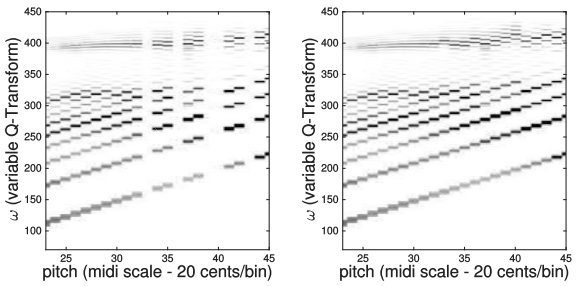
2.1 Dictionary extraction

Dictionary $P(\omega|s,p,f,o,v)$ with spectral templates from multiple singers is built based on English pure vowels (monophthongs), such as those used in the solfège system of learning music: Do, Re, Mi, Fa, Sol, La, Ti, and Do. The dictionaries use spectral templates extracted from solo singing recordings in the RWC audio dataset [17]. The recordings contain sequences of notes following a chromatic scale, where the range of notes varies accordingly to the tessitura of distinct vocal types: bass, tenor, alto and soprano. Each singer sings a scale in five distinct English vowels (/a/, /æ/, /i/, /ɒ/, /u/). In total, we have used 15 distinct singers (9 male and 6 female, consisting of 3 human subjects for each voice type: bass, baritone, tenor, alto, soprano).

The fundamental frequency (f_0) sequence from each monophonic recording is estimated using the pYIN algorithm [18]. After, the spectrogram representation is extracted using the VQT, with 60 bins/octave. A spectral template is extracted for each frame, regarding the singer source, vowel type, and voice type. In order to incorporate multiple estimates from a common pitch, the set of estimates that fall inside the same pitch bin are replaced by its metrically trimmed mean, discarding 20% of the samples as possible outliers. The set of spectral templates are then pre-shifted across log-frequency in order to support tuning deviations $\pm 20, 40$ cent and are stored into a 6-dimensional tensor matrix

$P(\omega|s,p,f,o,v)$. Since the recording sessions of the proposed chromatic scale are likely to capture only part of all possible pitches, the dictionary $P(\omega|s,p,f,o,v)$ is actually a sparse matrix in which several templates are missing along the pitch scale, as shown in Figure 1a.

We have investigated alternative ways to fill out the missing templates in the dictionary, including: spectrum estimation by replication [19, 12], linear and non-linear interpolation, and generative process based on Gaussian mixture model (inspired on [20, 21]). Following experimentation, we have chosen the linear replication approach, where existing templates belonging to a same dictionary are used to fill the missing parts of the pitch scale. In this approach, a spectral shape of a given pitch p_n is repeated (with the appropriated log-frequency shift) over all subsequent pitches $p \in [p_{n+1}, p_{m-1}]$ until another template is found (the pitch template p_m). Figure 1b illustrates the resulting dense dictionary templates of one singer example (vowel /a/) from our audio dataset.



(a) (b)

Fig. 1: Example from an /a/ vowel utterance (one singer) templates: (a) original sparse templates from the VQT spectrogram; (b) dense estimates by replication.

2.2 Post-processing

The 20 cent resolution pitch activation matrix $P(f',t)$ obtained from the output of the PLCA algorithm contains non-binary values between 0 and 1. A binary version of this matrix is usually obtained by thresholding. However, the PLCA output estimates might contain certain erroneous pitch candidates. These erroneous estimates often appear in the presence of harmonically-related pitches [22]. Aiming to diminish this issue and as a consequence to improve the final pitch estimates, we propose a classification procedure as a post-processing step.

Our approach is based on the hypothesis that the overtone components from the expected sung notes have distinct characteristics when compared with the overtone content from the false positive pitch detections. Based on this assumption, a set of features based on the harmonic content of each pitch candidate can be easily extracted from the VQT spectrogram. These features are used to train a random decision forest classifier [23] with binary output: $0 =$ false positive; $1 =$ valid pitch. This classifier aggregates a large ensemble of weak learners (bagging procedure) based on the harmonic features extracted at each time frame t. The list of features evaluated in this work is listed in Table 1. For notational simplicity, we omit the time index t. **F1** $= f'$ is the detected pitch itself (from $P(f',t)$) which we want to classify into a valid estimate or into a false positive. It is valued in the same log-frequency scale used in V, and serves as reference for the computation of the the rest of the feature set; **F2** $= V(f')$ is the magnitude of the fundamental frequency at pitch estimate f' in the log-frequency spectrogram V. Analogously, we estimate the features **F3**, **F4**, **F5**, and **F6** as the individual magnitudes related to the first four overtones $H_{1:4}$, such that $H(V_t,Q,f',k) = \frac{1}{|Q|}\sum_{q\in Q} V_t(f'+q+h(k))$, and $h(k) = \lceil b \times 12\log_2(k+1) \rfloor$ is the overtone log-frequency index. b is the number of bins per semitone used to compute the spectrogram, Q is a neighbourhood around the overtone pitch, and $|.|$ is the cardinality operator. We have used $Q = \{-3,\ldots,3\}$ for the first 3 overtones and $Q = \{-1,0,1\}$ for the rest since the energy distribution around the first three harmonics is considerably wider than around the higher harmonics.

F7 is the summation of the first r_1 harmonic magnitudes. **F8** is similar to **F7**, but it also aggregates harmonic content from a larger temporal window. The rationale behind this feature is that pitched sounds tend to describe a similar f_0 contour at their overtones. In our experiments we have used $w = 7$ (aggregation over 1.4 seconds). Based on empirical experiments, we have set $r_1 = 10$ and $r_2 = 4$. All overtone based features are normalized by the f_0 magnitude in order to keep the system less sensitive to sound dynamics.

The output of this classification step is a binary matrix $B(f',t)$, containing validated pitch activations. The proposed classifier is designed to prioritize the true positives in detriment of false positives. This means that our classifier keeps most of the true pitch candidates but in counter part remove less false positives. Nonetheless, keeping this goal during the training stage, the

Table 1: Feature set

ID	Feature	Description
F1	f'_t	f_0
F2	$V_t(f')$	f_0 magnitude
F3	$H(V_t,Q,f',k=1)$	f_1 magnitude
F4	$H(V_t,Q,f',k=2)$	f_2 magnitude
F5	$H(V_t,Q,f',k=3)$	f_3 magnitude
F6	$H(V_t,Q,f',k=4)$	f_4 magnitude
F7	$\sum_{k=1}^{r_1} H(V_t,Q,f',k)$	Sum of r_1 first overtones
F8	$\sum_{j=t-w}^{t}\sum_{k=1}^{r_2} H(V_j,Q,f',k)$	Sum of r_2 first overtones along w window time

proposed classifier reduces around 50% the number of false pitch candidates from the original PLCA output, while reduces the true positives in around 10%. More details about the training and measurements using testing datasets will be discussed in Section 3.

As a final procedure in the post-processing step, we refine the binary matrix output B in order to obtain more accurate pitch activations, also regarding temporal continuity. In this process, we scan each frame of the matrix B and replace the pitch candidates by spectrogram peaks detected in V that are validated by a minimum pitch distance rule:

$$(\Delta_{peaks}(V_t,B_t) < T_1) \vee (\Delta_{peaks}(V_t,B_{t-1}) < T_2) \quad (9)$$

where the function Δ_{peaks} in (9) indicates the minimum pitch distance between the selected peak candidates in V_t and the pitch candidate in B_t and B_{t-1}, respectively. In our experiments we have used $T_1 = 1$ and $T_2 = 3$, based on density distributions of $|\Delta_{peaks}|$, that were estimated from measurements in our datasets using the pitch ground truth.

3 Experiments

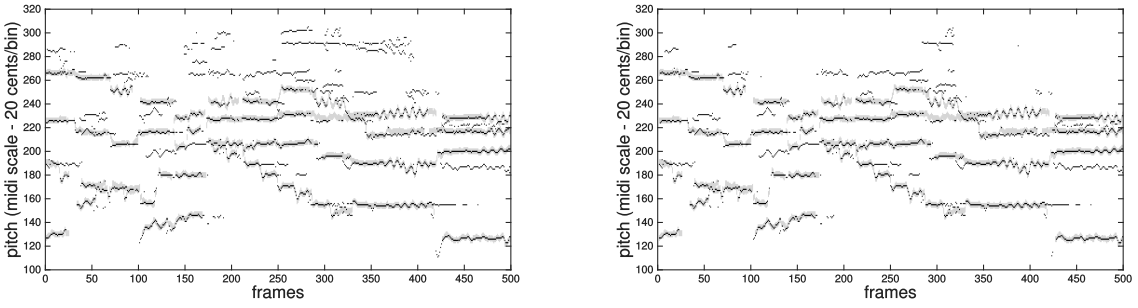
We performed experiments regarding the first step only (PLCA model) and the first + second step (PLCA + pruning of false positives by binary classification).

3.1 Dataset

To estimate the accuracy of the multi-pitch detection we evaluate the proposed model using recordings of Bach Chorales and Barbershop music with performances of vocal quartets. The set of Bach Chorale recordings contains mixtures of two males voices and two females voices, whereas the Barbershop recordings contain only male voices. Each recording in both audio datasets has four distinct vocal parts in the style SATB. These audio datasets were built using recordings available in http://pgmusic.com. A complete list of the music pieces used in our experiments is available in the supporting materials of this paper, see Section 5. Each recording in these datasets is a mixed multitrack wave file containing four vocal parts (one per channel, with sampling rate of 22.05 kHz and 16 bits per sample). We also generate a frame-based pitch ground truth pitch estimate for each vocal part (before the mix down) through the help of the pYIN (single) pitch detection algorithm [18]. From visual inspection, we confirmed the pitch estimates for the isolated voice tracks are reliable in the vast majority of cases. We did not make any further manual corrections, thus the ground truth might eventually contain errors. There are a total of 26 recordings in the Bach Chorales dataset and a total of 22 recordings in the Barbershop dataset, with a total duration of 104 minutes. The multi-pitch estimates were extracted from all files from these two datasets, using the PLCA model proposed in Section 2 and the respective 6-dimensional tensor described in Section 2.1.

3.2 Classifier training

For training the binary classifier we have generated an augmented dataset using recordings from the RWC dataset and from [24]. This augmented dataset contains the mix down of recordings with several combinations of multiple singers, regarding distinct vowels, voice types and also varied pitch range. In addition to this augmented dataset, we have also used the Bach Chorales and the Barbershop datasets, however, to avoid overfitting during the training phase, we have excluded the barbershop dataset when testing the Bach Chorales recordings and vice versa. A random forest classifier [23] was trained for each case. We set the minimal leaf size to 1 and select the square root of the total number of features for each decision split at random. We also evaluated distinct subset combinations of the features listed in the Table 1. The classifier trained with all features achieved the best results in our experiments, and also created a decision forest with shallow trees, thus requiring less memory. Based on the analysis of



(a) (b)

Fig. 2: Pruning of false multi-pitch detections. The gray shaded area shows the pitch ground truth and black dots illustrate detected pitches. (a) Pitch activation (PLCA) + pitch refinement; (b) Pitch activation (PLCA) + false positive pruning + pitch refinement.

the error for out-of-bag observations and memory demand, we kept each final decision forest with 30 trees. The final training dataset contains 751741 feature vectors (excluding Bach Chorales recordings) and 882746 feature vectors (excluding Barbershop recordings). Figure 2 shows a comparison between the pitch detection procedure without and with the classification step. Both outputs have been processed by the refinement rule of Equation 9, however only Figure 2(b) was pruned by the random forest classifier. A considerable reduction of false positives can be seen, mainly at higher pitches.

4 Evaluation

Our evaluation is frame-based. Pitch estimates obtained through the proposed algorithm are extracted at each 20 ms step. The ground truth of the test datasets was generated by extracting the f_0 from the original monophonic recording tracks through the pYIN pitch tracking algorithm, as described in Section 3.1. We compared our approach with two benchmark multi-pitch detection methods, named here as PERT [25] and VINC [26], respectively. VINC performs multi-pitch detection using an adaptive spectral decomposition based on unsupervised NMF. PERT detects multi-pitch candidates selected among spectral peaks, which are further validated taking into account the sum of their hypothetical harmonic amplitudes and smoothness measures.

The obtained frame-based pitch estimations were evaluated through measures of precision (P), recall (R) and F-measure (F) [27]. Tables 2 and 3 show the comparative results when applying these techniques to the Bach Chorales and Barbershop datasets, respectively. Results from our algorithm using only the first step are labeled as MSINGERS. Results with the refinement procedure (Eq. 9) are shown as MSINGERS+, and results with the complete pipeline (including the classification step) are shown as MSINGERS†+. We also have applied the refinement procedure to the benchmark approaches (labeled as VINC+ and PERT+), since they were originally designed for a semitone resolution output. For each experiment, we evaluated the precision, recall and F-measure regarding two cases: 1) semitone pitch resolution (MIDI scale); and 2) 20 cent pitch resolution, same as in the VQT spectrogram. These two cases are labeled in the tables as “88” and “440”, respectively. In “440”, a pitch estimate is considered correct only if its distance from the expected pitch in the ground truth is less than or equal to ± 20 cent.

As can be seen in the results, the proposed method outperforms in most cases the comparative approaches. As expected, the classification step reduces the number of false positives, increasing the precision. The post-processing procedure enhances the precision on an average of 11% and 14% when applied to the Bach Chorales and the Barbershop dataset, respectively. On the other hand, it decreases the recall. This is caused because a few true positives are erroneously rejected by the post-processing step. However, the recall reduction is substantially smaller than the precision improvement, keeping a similar ratio ($1/4$ and $1/5$, respectively) as the one achieved during the training stage of the classifier. This shows that the post-processing procedure is working as expected, despite a slight loss of accuracy when compared with the out-of-bag estimates at training stage. Since we want to transcribe singing,

Table 2: Multi-pitch detection results: Bach Chorales

Algorithm	Pitch Resol.	P		R		F	
		avg.%	std.%	avg.%	std.%	avg.%	std.%
VINC	88	42.14	5.28	54.42	4.01	47.35	4.18
PERT	88	52.41	2.83	48.67	3.17	50.43	2.69
MSINGERS	88	**59.59**	4.70	**76.06**	3.74	**66.73**	3.75
VINC+	88	37.89	3.91	**78.75**	3.78	51.04	3.75
PERT+	88	56.22	3.55	74.01	3.98	63.85	3.36
MSINGERS+	88	**74.76**	3.82	60.31	3.75	66.68	3.08
MSINGERS†+	88	71.30	4.06	70.91	4.13	**71.03**	3.33
VINC	440	33.21	5.49	32.24	3.67	32.62	4.18
PERT	440	**43.42**	5.65	23.80	3.25	30.72	4.04
MSINGERS	440	37.81	2.77	**32.25**	2.68	**34.76**	2.39
VINC+	440	34.88	4.45	**67.74**	6.36	**45.95**	4.86
PERT+	440	39.11	3.63	44.93	4.70	41.78	3.90
MSINGERS+	440	44.87	4.40	37.29	3.63	40.66	3.60
MSINGERS†+	440	**54.76**	4.39	35.06	3.66	42.68	3.73

Table 3: Multi-pitch detection results: Barbershop

Algorithm	Pitch Resol.	P		R		F	
		avg.%	std.%	avg.%	std.%	avg.%	std.%
VINC	88	45.40	7.14	57.10	5.28	50.36	5.75
PERT	88	53.31	3.86	47.87	5.46	50.34	4.30
MSINGERS	88	**70.71**	4.52	**64.42**	3.84	**67.35**	3.50
VINC+	88	40.06	6.11	81.97	4.44	53.58	5.77
PERT+	88	59.61	4.88	77.26	5.48	67.19	4.44
MSINGERS+	88	54.64	5.51	**80.12**	5.75	64.86	5.14
MSINGERS†+	88	**68.52**	6.54	73.66	7.54	**70.84**	6.17
VINC	440	39.67	7.63	41.20	6.53	40.25	6.56
PERT	440	**48.42**	10.29	28.74	7.40	36.00	8.54
MSINGERS	440	38.01	3.76	**43.01**	4.90	**40.29**	3.96
VINC+	440	40.87	7.23	**86.99**	11.68	**55.38**	8.26
PERT+	440	43.47	6.28	55.36	10.44	48.59	7.66
MSINGERS+	440	48.17	6.51	51.23	7.37	49.52	6.35
MSINGERS†+	440	**62.03**	7.39	47.72	8.06	53.74	7.29

a resolution of 20 cent can be more appropriate for certain applications (e.g. for measuring tuning deviations). In this case, regarding the F-measure, the PLCA model is outperformed by the VINC+ method using the proposed refinement procedure, for both test datasets. Despite a good reduction of false positives after the classification step, missed pitches from the PLCA-based pitch detection together with the remaining false positives (mainly related to overtones) are still the principal cause of errors in our final pitch detection estimates. It is worth noting that our approach using the "88" resolution achieves an F-measure over 70%, which is promising for semitone-level transcription applications (e.g. for music typesetting).

5 Conclusions

In this paper, we proposed an automatic transcription system for *a cappella* audio recordings of performances with multiple singers. The proposed system is designed in two steps. First a model based on probabilistic latent component analysis employing a fixed 6-dimensional dictionary of sound templates is used to extract initial estimates from the mixed audio signal. A binary random forest classifier was introduced into a second step, were initial pitch estimates can be refined. This classifier was built mainly using overtone-based features. Experiments were performed using real recordings from two vocal quartets. The proposed multi-pitch detection model shows promising results, achieving better performance in most of the test cases when compared to two benchmark multi-pitch detection approaches. These results were further enhanced trough the proposed classification-based post-processing step. Future work will focus on improving high-pitch resolution transcription performance and on voice assignment evaluations, i.e. assigning each detected pitch to a specific voice. Supporting material for this work are available at `http://inf.ufrgs.br/~rschramm/projects/music/msingers`.

6 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] Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., and Klapuri, A., "Automatic music transcription: challenges and future directions," *J. Intell. Inf. Syst.*, 41(3), pp. 407–434, 2013.

[2] Kameoka, H., Nakano, M., Ochiai, K., Imoto, Y., Kashino, K., and Sagayama, S., "Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints," in *ICASSP*, pp. 5365–5368, 2012.

[3] Venkataramani, S., Nayak, N., Rao, P., and Velmurugan, R., "Vocal Separation using Singer-Vowel Priors Obtained from Polyphonic Audio," in *ISMIR*, pp. 283–288, 2014.

[4] Mysore, G. J. and Smaragdis, P., "Relative pitch estimation of multiple instruments," in *ICASSP*, pp. 313–316, 2009.

[5] Grindlay, G. and Ellis, D. P. W., "Transcribing Multi-Instrument Polyphonic Music With Hierarchical Eigeninstruments," *IEEE J. Selected Topics in Signal Processing*, 5(6), pp. 1159–1169, 2011.

[6] Benetos, E. and Dixon, S., "A Shift-Invariant Latent Variable Model for Automatic Music Transcription," *Computer Music J.*, 36(4), pp. 81–94, 2012.

[7] Fuentes, B., Badeau, R., and Richard, G., "Blind Harmonic Adaptive Decomposition applied to supervised source separation," in *EUSIPCO*, pp. 2654–2658, 2012.

[8] Benetos, E. and Weyde, T., "An Efficient Temporally-Constrained Probabilistic Model for Multiple-Instrument Music Transcription," in *ISMIR*, pp. 701–707, 2015.

[9] Benetos, E. and Dixon, S., "Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model," *J. Acoustical Society of America*, 133(3), pp. 1727–1741, 2013.

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

[11] Brown, J., "Calculation of a constant Q spectral transform," *J. Acoustical Society of America*, 89(1), pp. 425–434, 1991.

[12] Benetos, E., Badeau, R., Weyde, T., and Richard, G., "Template Adaptation for Improving Automatic Music Transcription," in *ISMIR*, pp. 175–180, 2014.

[13] Zhou, R., Reiss, J. D., Mattavelli, M., and Zoia, G., "A Computationally Efficient Method for Polyphonic Pitch Estimation," *EURASIP J. Adv. Sig. Proc.*, 2009.

[14] Emiya, V., Badeau, R., and David, B., "Multipitch Estimation of Piano Sounds Using a New Probabilistic Spectral Smoothness Principle," *IEEE Trans. Audio, Speech & Language Processing*, 18(6), pp. 1643–1654, 2010.

[15] Klapuri, A. P., "Multipitch estimation and sound separation by the spectral smoothness principle," in *ICASSP*, pp. 3381–3384, 2001.

[16] Dempster, A. P., Laird, N. M., and Rubin, D. B., "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society*, 39(1), pp. 1–38, 1977.

[17] Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R., "RWC Music Database: Music Genre Database and Musical Instrument Sound Database," in *ISMIR*, pp. 229–230, 2004.

[18] Mauch, M. and Dixon, S., "pYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions," in *ICASSP*, pp. 659–663, 2014.

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

[20] Goto, M., "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), pp. 311–329, 2004.

[21] Kameoka, H., Nishimoto, T., and Sagayama, S., "A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering," *IEEE Trans. Audio, Speech and Language Processing*, 15(3), pp. 982–994, 2007.

[22] Yeh, C., Röbel, A., and Rodet, X., "Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals," *IEEE Trans. Audio, Speech and Language Processing*, 18(6), pp. 1116–1126, 2010.

[23] Breiman, L., "Random Forests," *Mach. Learn.*, 45(1), pp. 5–32, 2001.

[24] Schramm, R., Nunes, H. D. S., and Jung, C. R., "Audiovisual Tool for Solfège Assessment," *ACM Trans. Multimedia Comput. Commun. Appl.*, 13(1), pp. 9:1–9:21, 2016.

[25] Pertusa, A. and Iñesta, J. M., "Efficient methods for joint estimation of multiple fundamental frequencies in music signals," *EURASIP Journal on Advances in Signal Processing*, p. 2012:27, 2012.

[26] Vincent, E., Bertin, N., and Badeau, R., "Adaptive Harmonic Spectral Decomposition for Multiple Pitch Estimation," *IEEE Trans. Audio, Speech, and Lang. Processing*, 18(3), pp. 528–537, 2010.

[27] Bay, M., Ehmann, A. F., and Downie, J. S., "Evaluation of Multiple-F0 Estimation and Tracking Systems," in *ISMIR*, pp. 315–320, Kobe, Japan, 2009.