Parameter Optimization in Automatic Transcription of Music

Claus Weihs and Uwe Ligges

Fachbereich Statistik, Universität Dortmund, 44221 Dortmund, Germany*

Abstract. Based on former work on automatic transcription of musical time series into sheet music (Ligges et al. (2002), Weihs and Ligges (2003, 2005)) in this paper parameters of the transcription algorithm are optimized for various real singers. Moreover, the parameters of various artificial singer models derived from the models of Rossignol et al. (1999) and Davy and Godsill (2002) are estimated. In both cases, optimization is carried out by the Nelder-Mead (1965) search algorithm. In the modelling case a hierarchical Bayes extension is estimated by WinBUGS (Spiegelhalter et al. (2004)) as well. In all cases, optimal parameters are compared to heuristic estimates from our former standard method.

1 Introduction

The aim of this paper is the comparison of different methods for automatic transcription of vocal time series into sheet music by classification of estimated frequencies using minimal background information. Time series analysis leads to local frequency estimation and to automatic segmentation of the wave into notes, and thus to automatic transcription into sheet music (Ligges et al. (2002), Weihs and Ligges (2003, 2005)). The idea is to use as little information as possible about the song to be transcribed and the singer interpreting the song to be able to transcribe completely unknown songs interpreted by unknown singers.

For automatical accompaniment Raphael (2001) uses Bayes Belief Networks. Cano et al. (1999) use Hidden-Markov-Models (HMMs) for training along known sheet music. Rossignol et al. (1999) propose a model for pitch tracking, local frequency estimation and segmentation taking into account the extensive vibrato produced by, e.g., professional singers. Davy and Godsill (2002) are using an MCMC model for polyphonic frequency estimation. The MAMI (Musical Audio-Mining, cp. Lesaffre et al. (2003)) project has developed software for pitch tracking.

There are some software products available for transcription (or at least fundamental frequency tracking), such as AmazingMidi (`http://www.pluto.dti.ne.jp/~araki/amazingmidi`), Akoff Music Composer (`http://www.akoff.com`), Audio to score (logic) (`http://www.emagic.de`), Autotune

* This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

(http://www.antarestech.com), DigitalEar (http://www.digital-ear.com), Melodyne (http://www.celemony.com), IntelliScore (http://www.intelliscore.net), and Widi (http://www.widisoft.com). None of them produced satisfying results in our test with a professional soprano singer, either because of inability to track the frequency, or because of non-robustness against vibrato.

All our calculations are made in R (R Development Core Team (2004)).

2 Heuristic Automatic Transcription

For automatic transcription we assume that CD-quality recordings are available down sampled to 11025 Hz (in 16 bit). As an example we studied the classical song "Tochter Zion" (G.F. Händel) (see, e.g., Weihs and Ligges (2005)).

The heuristic transcription algorithm proposed in Weihs and Ligges (2003) has the following form:

- Pass through the vocal time series by *sections* of size 512.
- Estimate pitch for each section by the heuristics: $ff_{\text{heur}} = h + [(s-h)/2]\sqrt{ds/dh}$, where h = first peaking Fourier frequency, s = peaking neighbor, dh and ds corresponding spectral density values. This way, $|\text{error}| < 2$ Hz can be shown for pure sine waves in frequency range of singing.
- Check by means of higher partial tones whether estimated pitch relates to fundamental frequency.
- Classify note for each section using estimated fundamental frequencies (given well-tempered tuning, and the (estimated) concert pitch of a').
- Smooth the classified notes because of vibrato by means of doubled *running median* with window width 7.

An example of the result of this algorithm can be seen in Figures 1 and 2 for soprano singer S5. Note that a' corresponds to 0. Singer S5 has an intensive vibrato. Thus classification switches rapidly between 2 (b'), 3 (c''), and 4 ($c\#''$) in the first 2 rows before smoothing (Figure 1). Unfortunately,

```
NA NA -30 NA  2  2 15 15 15 15 15 15  2  3  3  3  3  3  2  2   2  2  4  4
 3  2   2  2  2  4  4  3  2  2  2  2  3  2  2  2  2  2  2  2 -29 NA NA NA
NA NA  NA -2 -1  0  0 -1  0  0  0  0  0  0  0  0  0  0  0  0   0 -1 -1 -1
 0  0   0  0 -1 -1 -1  0  0  0  0  0 -1  1  1  1  1  1  1  1   1  0 NA NA
```

Fig. 1. Unsmoothed classification for sing S5

```
15 15  15 15 15 15 15 15 15 15 15 15  3  3  3  3  3  3  2  2   2  2  2  2
 2  2   2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2   2 NA NA NA
NA NA  NA  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0   0  0  0  0
 0  0   0  0  0  0  0  0  0  0  0  0  0  1  1  1  1  1  1  1   1  1 NA NA
```

Fig. 2. Smoothed classification for singer S5

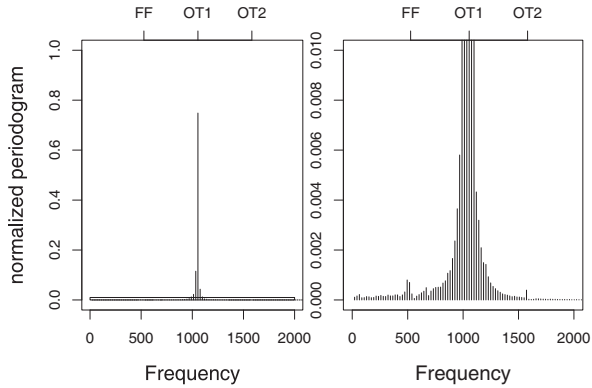

Fig. 3. Periodogram: only first overtone easily visible (right: zoomed in)

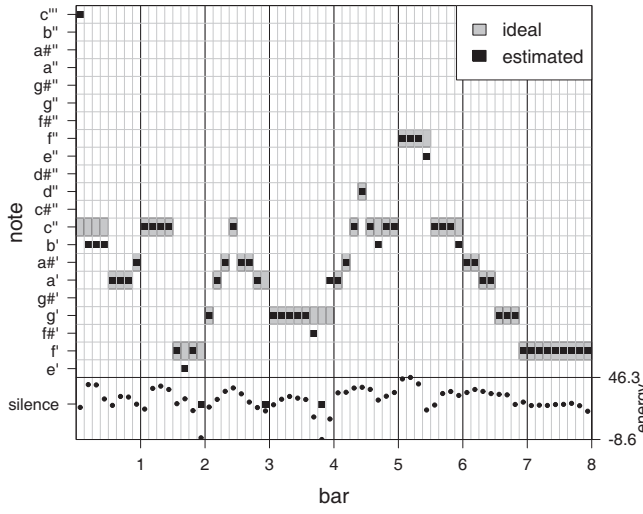

Fig. 4. Outcome of the heuristics

smoothing does not lead to the correct note 3 (c'') (Figure 2). E.g., classification leads to a note one octave too high in the beginning. To demonstrate the reason consider the corresponding periodogram (based on 512 observations) in Figure 3, where the first overtone (c''') has the only high peak, and neither the fundamental (c'') nor the second overtone is reasonably present.

In order to produce sheet music, the blocks of 512 observations corresponding to eighths are combined assuming constant tempo, and the mode of the corresponding classes is taken as the pitch estimator. Figure 4 compares the outcome of this heuristics with the correct sheet music (grey horizontal bars) for singer S5. Note that energy indicates the relative amplitude of the local wave, see Weihs and Ligges (2005) for a definition. Very low energy indicates rests, consonants or breathing.

3 Parameter Optimization of Heuristics

The idea of this paper is to try to improve the heuristics in various ways. First, the parameters of the heuristics will be optimally adjusted individually to the singer whose wave should be transcribed. This is carried out by means of a Nelder-Mead (1965) optimization of the error rate based on the third part of the example song, i.e. on the last 8 measures of "Tochter Zion". Note that such optimization needs training with known scores before application. Thus, this analysis just indicates to what amount the heuristics could be improved by means of a-priori learning.

The parameters of the heuristics are, defaults in parentheses ():

- *pkhght*: Indicates that "high peaks" need to have a peak height $>$ a percentage of maximum peak height (1.5%).
- *slnc*: Indicates that "Low energy periods" are a certain percentage of periods with lowest energy (20%).
- *minp*: Indicates that "Silence" is defined as low energy periods with more than minimum no. of high peaks (noise) (7).
- *srch1-4*: Parameters deciding about meaningfulness of candidate fundamental frequency *cff* also based on overtones (*ots*).
- *srch1*: Multiplier m (1.8) of frequency of first high peak (*fp*) so that $cff \in [fp,\ m \cdot fp]$.
- *srch2*: No. of unimportant smallest Fourier frequencies (10)
- *srch3,4*: Multipliers ml, mr (1.35, 1.65) of *cff* so that high peak $\in [ml \cdot cff,\ mr \cdot cff]$ only if 1st overtone was found instead of fundamental frequency *ff*.
- *mdo*: Order of median smoother (3) so that window width $= 2 \cdot mdo + 1$
- *mdt*: No. of median smoother replications (2)
- *htthr*: Halftone threshold from where on the next halftone is classified: displacement from 50 cents = 0.5 halftone (0)

Error rates are calculated based on eighths as follows:

$$\frac{\#\text{ erroneously classified eighth notes (without counting rests)}}{\#\text{ all eighth notes} - \#\text{ eighth rests}}.$$

In our example 64 eighth notes in 8 measures are considered. To create real sheet music equal sequential notes are joined. Note that this rule should be improved by identification of onset times of new notes. Table 1 shows the optimization results for sopranos S1, S2, S4, S5, and tenors T3, T6, T7. The first row indicates the defaults, rows 2 and 3 the starting values for optimization. Obviously, the only professional (S5) is the most outstanding, and the worst case concerning error rate at the same time. Figures 5 and 6 compare the original sheet music with the optimized outcome of S5. Note that parameter optimization overall leads to an error estimate *opte* roughly halving of heuristic error rates *heue*.

Further studies will have to show whether optimized parameters are general enough to be used for different performances of the same singer.

Table 1. Results of Nelder-Mead optimization in R (2004)

	pkhght	*slnc*	*minp*	*srch1*	*srch2*	*srch3*	*srch4*	*mdo*	*mdt*	*htthr*	*opte*	*heue*
default	1.50	20.0	7	1.80	10	1.35	1.65	3	2	0.0000		
start1	1.60	15.0	10	1.80	22	1.30	1.65	5	3	0.0000		
start2	1.20	25.0	6	1.80	9	1.36	1.70	3	2	0.0000		
S1	1.30	24.7	*4*	1.81	10	1.37	1.71	3	2	0.0026	5.7	13.1
S2	1.66	25.4	6	1.80	9	1.36	1.70	4	2	0.0035	3.9	7.7
S4	1.20	25.0	6	*1.97*	9	1.36	1.70	3	2	*0.0000*	7.5	10.9
S5	1.57	23.9	*10*	1.81	*23*	1.31	1.66	*5*	*3*	*0.0441*	*7.8*	*16.4*
T3	1.67	25.4	6	1.81	9	1.45	1.70	3	2	0.0089	1.7	1.7
T6	1.39	23.2	8	1.80	9	1.38	1.72	2	2	0.0194	7.0	12.1
T7	*2.23*	23.6	6	1.82	11	1.38	1.68	3	2	0.0182	1.7	1.8

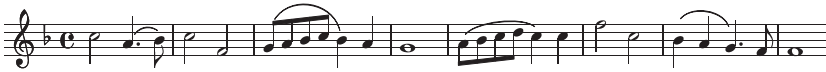

Fig. 5. Original sheet music of "Tochter Zion"

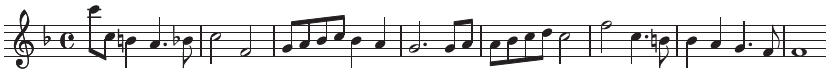

Fig. 6. Optimized outcome of the example's data, singer S5

4 Model based Automatic Transcription

Another way of improving pitch estimation might be the use of a wave model. Therefore, we combine two models in the literature, one of Rossignol et al. (1999), modelling vibrato in music, one of Davy and Godsill (2002). Based on this model we carry out a controlled experiment with artificial data, and estimate the unknown parameters of the model in two ways, one time based on periodograms, the other time based on the original wave data. In the first case a frequentist model is used, in the second case a Bayesian model.

In the used frequentist model vibrato is modelled as sine oscillation around heard frequency. Moreover, phase displacements are modelled as well as frequency displacements of overtones:

$$y_t = \sum_{h=1}^{H} B_h \cos\left[2\pi(h+\delta_h)f_0 t + \phi_h + (h+\delta_h)A_v \sin(2\pi f_v t + \phi_v)\right] + \epsilon_t,$$

where t = time index, f_0 = fundamental frequency, H = no. of partial tones (fundamental frequency $+H-1$ overtones), B_h = amplitude of h^{th} partial tone, δ_h = frequency displacement of h^{th} partial tone, $\delta_1 := 0$, ϕ_h = phase displacement of the h^{th} partial tone, f_v = frequency of vibrato, A_v = amplitude of vibrato, ϕ_v = phase displacement of vibrato, and ϵ = model error.

In the (hierarchical) Bayes MCMC variant of the same model the following stochastic model extensions are used: f_0, the fundamental frequency, is uniformly distributed in $[0, 3000]$ Hz, $H-1$, the no. of overtones, is truncated

Poisson distributed with a maximum of 11, the expected value of which is Gamma$(H, 1)$ distributed, B_h, the amplitudes, are normally distributed with a Gamma$(0.01, 0.01)$ distributed precision (= invers variance), δ_h, the frequency displacements, are normally distributed with a big Gamma$(100, 1)$ distributed precision, ϕ_h, the phase displacements, are uniformly distributed in $[-\pi/2, \pi/2]$, f_v, the vibrato frequency, is uniformly distributed in $[0, 12]$ Hz, A_v, the vibrato amplitude, is normally distributed with a general Gamma$(0.01, 0.01)$ distributed precision, ϕ_v, the vibrato phase displacement, is uniformly distributed in $[-\pi/2, \pi/2]$, ϵ, the model error, is normally distributed with a Gamma$(0.5, 2)$ distributed precision.

The design of experiments used is a full factorial in 5 variables, namely type of singer (professional female vs. amateur female), pitch (high vs. low, i.e. 1000 vs. 250 Hz), vibrato frequency (5 vs. 9 Hz), vibrato amplitude / vibrato frequency (5 vs. 15), vibrato phase displacement (0 vs. 3). In 4 additional experiments the vibrato amplitude was set to 0 with vibrato frequency and vibrato phase deliberate, set to 0 here. For data generation, professionals were modelled by $\mathit{fff} + 2\,ots$ with $B_1 = 3$, $B_2 = 2$, and $B_3 = 1$, amateurs with $\mathit{fff} + 1\,ot$ and $B_1 = 3$, and $B_2 = 1$, displacements and noise set to 0.

For the estimation of unknown parameters 512 or 2048 observations are used, respectively. Heuristic estimates of the fundamental frequency are taken from one 512 observations block or as the median over the estimates in 7 half overlaying blocks of 512 observations (without any smoothing). Estimations based on spectral information are based on periodograms of the 7 half overlaying blocks of 512 observations. Then the resulting $1792 = 256 \cdot 7$ Fourier frequencies built the basis for Nelder-Mead optimization of the unknown parameters using the following three starting vectors: $\mathit{fff} = \text{median}(\mathit{fff}_{\text{heur}}) + 2, 0, -2$ Hz, $B_h = 0.5$ for $h > 1$, $f_v = 7$, $A_v = 5$, $\phi_v = 0$. A model with fff and 2 overtones was used for estimation in any case. Note that standardized periodograms are used so that $B_1 = 1$ was fixed for identification. Estimated amplitudes B_h for $h > 1$ are thus relative to B_1. The default stopping criteria of the R function `optim` were used with a maximum of 5000 iterations.

For the estimation of the hierarchical Bayes model WinBUGS optimization (Spiegelhalter et al. (2004)) is used (the WinBUGS model is available from the authors). 512 observations are used and starting values are the same as for the above optimization based on periodograms, except that B_1 is free to be estimated now, and the number of overtones $H - 1$ is estimated as well. As a stopping criterion every 100 iterations it is checked whether linear regression of the last 50 residuals against iteration number delivers a slope significant at the 10% level with a maximum of 2000 iterations.

An overall comparison of the results by means of mean absolute deviation (MAD), and root mean squares deviation (RMSD) of the estimated fundamental frequency as well as run time (see Table 2) leads to the conclusion that the heuristics are as good as the more complicated estimation procedures, but much, much faster. Only an increase of the number of observations leads to

Table 2. Deviations of the estimated fundamental frequency for each method

	Heur. (1)	Heur. (median)	NM (spectral)	WinBUGS
ff MAD (cent)	5.06	2.38	1.29	4.88
ff RMSD (cent)	6.06	2.74	3.35	6.44
run time	< 1 sec	2 sec	4 h	31 h

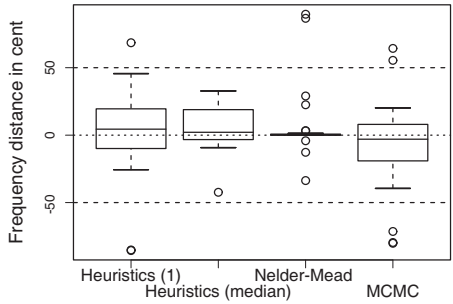

Fig. 7. Boxplots of deviations of the estimated fundamental frequencies

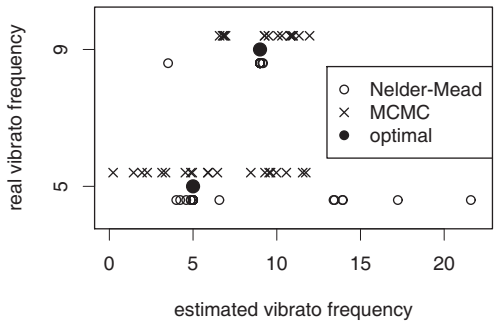

Fig. 8. Estimates of the vibrato frequency

a distinct improvement. Note in particular that already with 512 observations WinBUGS optimization needs 31 hours for the 36 experiments. Simpler methods programmed in C are in development.

The results of the optimizations are compared with the results of heuristic pitch estimation in more detail in boxplots corresponding to the estimated fundamental frequencies in Figure 7, where the horizontal lines of ± 50 cents $= \pm 0.5$ halftone correspond to the natural thresholds to the next halftone above or below, correspondingly. Note that the heuristic based on 2048 observations lead to perfect note classification, whereas (spectral) Nelder-Mead is most often much more exact, but in some cases even wrong in classification. The WinBUGS results are comparable with the results from the heuristic based on 512 observations. Estimates of the vibrato frequency in the model are compared as well (see Figure 8). Here (spectral) Nelder-Mead is nearly perfect in examples with 9 Hz, but unacceptable for 5 Hz. Also the WinBUGS results vary less with 9 Hz.

5 Conclusion

From the experiments in this paper it is learned that Heuristic Transcription can be individually improved by training, that a wave model is not better than the heuristics concerning *ff* classification, and that the estimation procedure is not good enough for vibrato frequency determination, except for high vibrato frequency and the spectral data estimator. Next steps will include experiments in the polyphonic case as well.

References

CANO, P., LOSCOS, A., and BONADA, J. (1999): Score-Performance Matching using HMMs. In: *Proceedings of the International Computer Music Conference*. Beijing, China.

DAVY, M. and GODSILL, S.J. (2002): *Bayesian Harmonic Models for Musical Pitch Estimation and Analysis*. Technical Report 431, Cambridge University Engineering Department.

LESAFFRE, M., TANGHE, K., MARTENS, G., MOELANTS, D., LEMAN, M., DE BAETS, B., DE MEYER, H., and MARTENS, J.-P. (2003): The MAMI Query-By-Voice Experiment: Collecting and annotating vocal queries for music information retrieval. In: *Proceedings of the International Conference on Music Information Retrieval*. Baltimore, Maryland, USA, October 26-30.

LIGGES, U., WEIHS, C., and HASSE-BECKER, P. (2002): Detection of Locally Stationary Segments in Time Series. In: W. Härdle and B. Rönz (Eds.): *COMPSTAT 2002 - Proceedings in Computational Statistics - 15th Symposium held in Berlin, Germany*, Physika, Heidelberg, 285–290.

NELDER, J.A. and MEAD, R. (1965): A Simplex Method for Function Minimization. *The Computer Journal, 7, 308–313*.

R DEVELOPMENT CORE TEAM (2004): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

RAPHAEL, C. (2001): A Probabilistic Expert System for Automatic Musical Accompaniment. *Journal of Computational and Graphical Statistics, 10, 487–512*.

ROSSIGNOL, S., RODET, X., DEPALLE, P., SOUMAGNE, J., and COLLETTE, J.-L. (1999): Vibrato: Detection, Estimation, Extraction, Modification. *Digital Audio Effects Workshop (DAFx'99)*.

SPIEGELHALTER, D.J., THOMAS, A., BEST, N.G. and LUNN, D. (2004): *WinBUGS: User Manual*. Version 2.0, Cambridge: Medical Research Council Biostatistics Unit.

WEIHS, C. and LIGGES, U. (2003): Automatic Transcription of Singing Performances. *Bulletin of the International Statistical Institute, 54th Session, Proceedings, Volume LX, Book 2*, 507–510.

WEIHS, C. and LIGGES, U. (2005): From Local to Global Analysis of Musical Time Series. In: K. Morik, A. Siebes and J.-F. Boulicault (Eds.): *Local Pattern Detection, Springer Lecture Notes in Artificial Intelligence, 3539*, Springer, Berlin, 217-231.