Efficient Pitch Estimation on Natural Opera-Singing by a Spectral Correlation based Strategy

Fernando Villavicencio[1,a)] Jordi Bonada[2,b)] Junichi Yamagish[1,c)] Michel Pucher[3,d)]

Abstract: We present in this work a study for robust pitch estimation on signals presenting wide-range pitch content, as is the case of opera singing. Aiming to perform automatic features extraction for the further development of parametric opera singing synthesis technology we evaluate four state-of-the-art pitch estimators, reporting in particular technical details of one introduced in previous work, based in a technique called Spectral Amplitude Autocorrelation (SAC). The results issued from subjective and objetive evaluations show clear performance trends, denoting robust estimation performance for SAC without observing significant sensitivity to the pitch height.

Keywords: Pitch estimation, speech analysis, speech synthesis, singing voice

1. Introduction

Pitch estimation is an essential key technology in speech synthesis. In the speech synthesis fields, there are two dominants methodologies, that is, unit selection/waveform concatenation and statistical parametric speech synthesis. In the unit selection speech synthesis, pitch-synchronous overlap add is typically used to generate speech waveforms and in the statistical parametric speech synthesis, speech waveforms are typically generated using either source-filter vocoders with F0-adaptive spectral smoothing techniques ([1]), glottal-excitation based vocoders ([2]) or sinusoidal models such as harmonic plus noise model ([3]).

The robustness of the F0 (pitch) extraction task may represent an important factor of the performance of these techniques to properly model the periodic information of the speech signal. Note also the use of F0 information to achieve efficient estimation of the spectral envelope ([4]) as a way to improve the synthesis quality when modeling the voice timbre information ([5]).

There is a large list of algorithms proposed to estimate pitch accurately (*e.g.* [6]) well evaluated for normal read speech but poorly done for the case of singing voice despite the emergence of technologies where speech synthesis strategies are applied to musical purposes. As prominent examples, the HMM-based singing synthesis platform called *Sinsy* ([7]) and the technology developed by Yamaha Corporation called VOCALOID ([8]), of significant commercial success and popularity among general public. Note previous work in [9] and [10] presenting additional experimentation with this technology.

Pitch estimation on opera-singing appears to be particularly challenging if compared to spoken speech and pop singing (in which most of current singing synthesizers are mainly focused) due to the large pitch range frequently used and the required precision to reproduce the musical sequence. Moreover, although few recent work ([11]), opera-singing has not been well studied yet in the sense of the F0 estimation and speech synthesis.

The purpose of our study is to experimentally evaluate the performance of state-of-the-art strategies for automatic pitch-estimation of natural opera singing with a view to high-quality opera singing synthesis. For this purpose, we have recorded opera songs by professional singers. By using segments selected from these recordings we have evaluated four recently-proposed algorithms on an automatic basis (no adaptation of analysis parameters according to the pitch range of each segment), including one introduced in previous work and described more in detail in this paper.

This paper is organised as follows: In the section 2, we briefly explain our opera-singing data. In section 3, we introduce the selected pitch-estimation algorithms and explain the SAC algorithms proposed by one of the coauthors recently in details in section 4. Results of subjective and objective evaluation are shown in the section 5 and 6, respectively. We summarize our findings in section 7.

2. Singing-Voice Data

2.1 Opera Singing Collection

In speech synthesis, a corpus would be selected following a minimum set-cover basis and then be read by a speaker in a separate selection process. For selecting an opera corpus by following this way we would end up with a selection of opera songs that no available opera singer has in his/her repertoire. Therefore we decided to select a number of opera songs ($\approx$ 8-10), with the assistance a professional teacher, for four singer categories (bass, tenor, mezzo, and soprano) that are in the current repertoire of

[1] Sound and Media Group, National Institute of Informatics, Tokyo, Japan
[2] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[3] Telecommunications Research Center Vienna (FTW), Austria
[a)] villavicencio@nii.ac.jp
[b)] jordi.bonada@upf.edu
[c)] jyamagis@nii.ac.jp
[d)] pucher@FTW.at

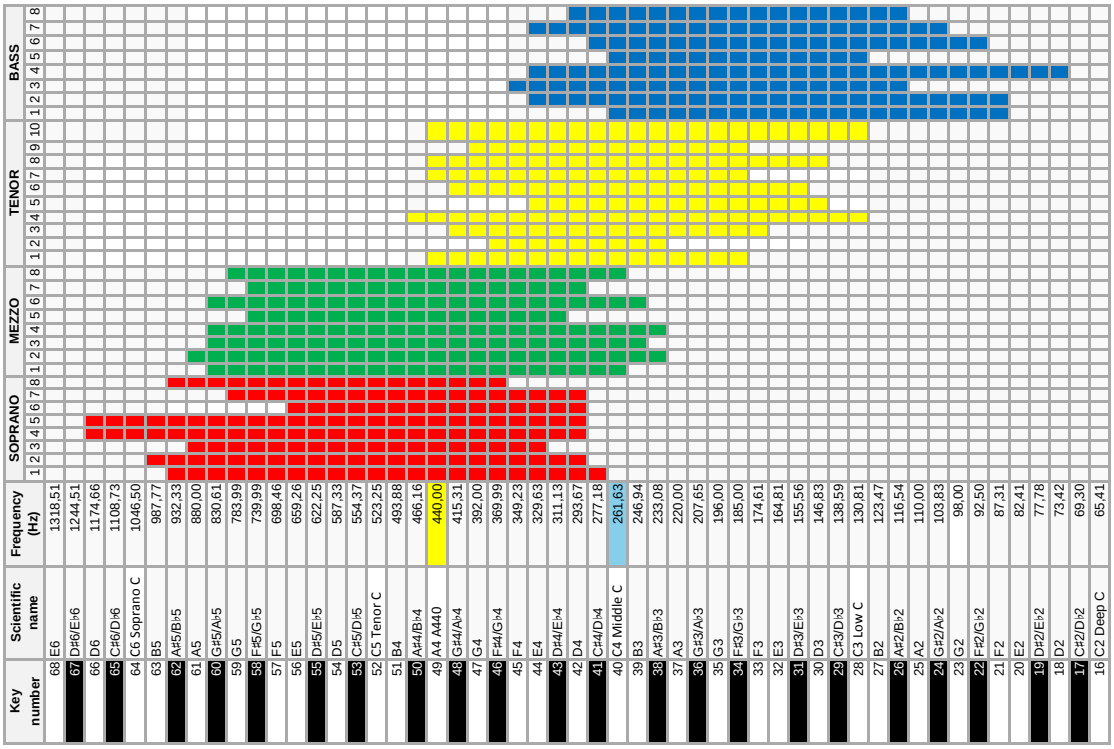

Fig. 1 Musical interval of the recorded opera songs with four different professional singers.

that singer and that cover the space of opera songs along a lyrical - dramatic and slow - fast axis. We also checked that these songs cover the F0 range of that singer category. A further restriction in the selection of songs was the fact that we were only looking at songs in the German language.

Fig. 1 shows the range of the music scores on the piano roll of all the recorded songs colored for each of the four singers, denoting that the total musical interval covered ranges from D2 (73 Hz) to D6 (1175 Hz). Note also that all songs show a range larger than one octave, suggesting us that an adaptation of the analysis range of the F0 estimator at a song level may not be enough to avoid octave errors that can be found on many estimators when working on these conditions.

2.2 Test Data

For simplicity, we extracted short excerpts of continuous singing as representative data of each singer to set up an experimental evaluation. More precisely, five segments of four different songs were chosen, resulting in a total of eighty test samples with an average duration of five seconds. Most of the resulting samples included vibrato phenomena, a recurrent feature of opera singing. There were no additional restrictions for the data selection since the fundamental interest is to perform robust pitch extraction on naturally sung signals, independently of the singer type and the singing style.

The samples were down sampled to $S_r = 44100Hz$ in order to keep high-quality conditions. The ranges and average values of the pitch on the test data of each singer (20 samples) are shown in Table 1. The resulting overall pitch range of the signals used for all our experiments was therefore found in $[70, 1000]Hz$, a range significantly larger and higher than the values in which speech based pitch extraction studies are typically carried out.

Table 1 Pitch statistics of the test data per singer (in Hertz).

Singer	Bass	Tenor	Mezzo	Soprano
range	[72, 477]	[107, 528]	[195, 823]	[130, 979]
mean(std)	206 (56)	283 (74)	542 (112)	575 (128)

3. Pitch Estimation Techniques

Most of pitch estimation techniques have been optimised for spoken voice, which represents a reduced challenge in terms of the expected pitch range to track (e.g. within $[80, 300]Hz$ for speech). In general, a proper adjustment of this range is an important factor of the performance, representing a problem for automatic features extraction of natural opera singing, due, as it was mentioned, to the expected wide-range of the pitch. In this work we experimentally compare for this task the state-of-the-art methods introduced in this and next sections.

3.1 Sawtooth waveform inspired pitch estimator (SWIPE)

Estimation based on the selection of the F0 of a sawtooth waveform template whose spectrum best matches that of the input signal considering a pitch-dependant optimal window size. The improved version of this technique (SWIPE'), using only information of the first and prime harmonics, has shown outperforming estimation performance on speech and musical instruments databases when compared against state-of-the-art algorithms [12].

3.2 Robust epoch and pitch estimator (REAPER)

Developed by D. Talkin (Google), this method is mainly based on the identification of voiced epochs and further estimation of

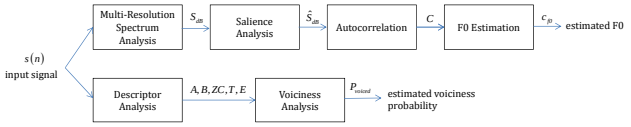

Fig. 2 Schema of SAC algorithm.

glottal closure instants (GCI) in order to compute the instantaneous F0 as the inverse between successive GCIs. This strategy uses time-domain information of a linear-prediction residual to find GCI candidates (furthermore processed by dynamic programming) and low-frequency content for voicing characterisation [13]. There is no available information concerning a performance evaluation by the authors of the technique.

3.3 Summation of residual harmonics (SRH)

This method also uses a residual signal after auto-regressive modeling of the input signal. The estimation is based on the $f0$ value maximizing locally a proposed SRH measure. A thresholding of this measure was also found useful for voicing boundaries decisions. This strategy was reported with comparable performance against state-of-the-art techniques and providing improved robustness to additive noise [14].

3.4 Spectral amplitude autocorrelation (SAC)

This new technique has been only generally described so far in [15], we include a more-in-depth description in the next section. For a mater of space we refer to the referenced works for a complete technical description of the other techniques.

Note that in our experimental study, focused on a straightforward application scenario we kept the default analysis settings (e.g. analysis window, hop-size) as found in the implementations from the authors that were found available ([16], [13], [17]). Exceptionally, the pitch search range was set to $[70, 1400]Hz$ for all estimators according to the conditions that can be expected on opera singing in general.

4. SAC: Spectral-Amplitude Autocorrelation

This algorithm, briefly summarized in Fig. 2 is based on the autocorrelation of the amplitude spectrum. It consists of several steps comprising resampling, frame-by-frame F0 candidate estimation, and post-processing.

4.1 Multiresolution Spectrum

First, the audio signal is resampled using polyphase filters to a sampling frequency of $SR = 11.025$ kHz, in order to reduce the computational cost. Next, the audio is segmented into a sequence of overlapping frames of $N = 640$ samples (58 ms) with a hop-size of 64 samples (5.8 ms). For each frame the DC is removed and its Discrete Fourier Transform (DFT) spectrum is computed with a rectangular window without zero-padding. Then, the spectrum is convolved with a variable convolution kernel that depends on the bin frequency and that corresponds to the transform of a Blackman-Harris window with a length of 640 samples for frequencies lower than 200 Hz, 256 for frequencies higher than 1000 Hz, and linearly interpolated in between. Only 41 bins of the window transform are used in the convolution to reduce the computational cost. This process generates a multi-resolution spectrum X with desirable properties. The amplitude spectrum is measured in decibels (X_{dB}).

4.2 Salience Spectrum

Next, spectral peaks (local maxima) are estimated $\{p_i\}_{i=1}^{P}$, and the spectrum is segmented into P peak regions with the corresponding P+1 boundaries $\{s_i\}_{i=0}^{P}$ set at local minima.

$$s_i = \begin{cases} 0 & if \quad i = 0 \\ \arg\min_{k\in[p_i,p_{i+1}]} X_{dB}(k) & if \quad 0 < i < P \\ N/2 & if \quad i = P \end{cases} \tag{1}$$

Then, a salience spectrum S_{dB} is estimated: for each peak region $[s_i, s_{i+1}]$, a local weighted mean of X_{dB} is subtracted and negative values are set to zero. The local mean $\overline{X}$ is computed as

$$\overline{X}_i = \frac{\sum_{k=-r}^{r} w(p_i+k)\cdot X_{dB}(p_i+k)}{\sum_{k=-r}^{r} w(p_i+k)} \tag{2}$$

$$w(k) = \left(1 - \frac{p_i - k}{r}\right)^2 \tag{3}$$

$$r = \begin{cases} r_1 + \frac{p_i}{k_c} r_2 & if \quad p_i \le k_c \\ r_1 + r_2 & if \quad p_i > k_c \end{cases} \tag{4}$$

where p_i is the peak bin index, $k_c = \frac{700\cdot N}{SR}$, $r_1 = \frac{80\cdot N}{SR}$, and $r_2 = \frac{200\cdot N}{SR}$. $\overline{X}_p$ is subtracted to all the bins in the peak region.

$$S_{dB}(k) = \begin{cases} 0 & if \quad X_{dB}(k) < \overline{X}_i \\ X_{dB}(k) - \overline{X}_i & if \quad X_{dB}(k) \ge \overline{X}_i \end{cases} \quad \forall k \in [s_i, s_{i+1}] \tag{5}$$

Since the constant values subtracted to each region are likely to be different, we expect that discontinuities will appear at region boundaries. In order to avoid them, for each region boundary s_i, the corresponding local means $\overline{X}_{i-1}$ and $\overline{X}_i$ are linearly interpolated along 4 bins around the boundary bin.

The salience spectrum is non-linearly scaled so that values above 20 dB to the maximum S_{max} are raised, while values lower than a frequency dependent threshold are lowered. This scaling is written as follows.

$$\hat{S}_{dB}(k) = \begin{cases} S_{dB}(k)\cdot\left(1 + \frac{S_{dB}(k) - th_h}{20}\right) & \text{if } S_{dB}(k) \ge th_h \\ \frac{S_{dB}(k)}{1 + \frac{1}{2}(th_l - S_{dB}(k))} & \text{if } S_{dB}(k) < th_l \\ S_{dB}(k) & \text{otherwise} \end{cases} \tag{6}$$

$$th(k) = \begin{cases} 50 & if \quad k < k_1 \\ 40 + 10\cdot\frac{k-k_1}{k_2-k_1} & if \quad k_1 < k < k_2 \\ 40 & if \quad k \ge k_2 \end{cases} \tag{7}$$

where $th_h = S_{max} - 20$, $th_l = S_{max} - th(k)$, $k_1 = \frac{150\cdot N}{SR}$, $k_2 = \frac{300\cdot N}{SR}$ and $S_{max} = \max_{k\in[0,N/2]} S_{dB}(k)$.

4.3 Spectrum Correlation

A peak is artificially added at zero frequency to help fundamental frequency detection of signals with few or just one harmonic: $\hat{S}_{dB}(0) = \hat{S}_{max}$, $\hat{S}_{dB}(1) = 0.9\cdot\hat{S}_{max}$, and $\hat{S}_{dB}(2) = 0$, where $\hat{S}_{max} = \max_{k\in[0,N/2]} \hat{S}_{dB}(k)$. The next step is to compute the correlation of $\hat{S}_{dB}$ multiplied by a zero-centered Hann window by means of the Fast Fourier Transform. The peaks of the correlation function C correspond to spectral bin distances likely to

explain a harmonic structure of the spectrum. A subset of peaks is computed $\{c_i\}_{i=1}^{C}$ with the following constraints

$$C(c_i) > C(k) \qquad \forall k \subset \{c_i - 2, c_i - 1, c_i + 1, c_i + 2\}$$

$$C(c_i) > 0.57 \cdot C_{max}$$

$$0.1 \cdot c_{max} \leq c_i \leq N/2$$

where $c_{max} = \arg\max_{k \in [0, N/2]} C(k)$ and $C_{max} = C(c_{max})$. The estimated fundamental frequency bin is the peak with lowest index $c_{f0} = c_1$. Finally, the fundamental frequency estimation is refined with a 2^{nd} order polynomial interpolation. The estimated pitch in Hz is $f_0 = \frac{SR}{N} c_{f0}$.

4.4 Trajectory Locking

Let it be c_{f0}^m and C^m the estimated fundamental frequency bin and correlation function of the m^{th} frame. If the previous five frames were estimated as voiced with frame-to-frame differences lower than 20 bins, then the estimated fundamental frequency is updated (if it exists) as the largest peak of C around the previous frame estimation c_{f0}^{m-1} that fulfills the following constraints

$$c_{f0}^m = \arg\max_k C^m(k) \qquad \forall k \left| \begin{array}{c} \left\lfloor 0.8 \cdot c_{f0}^{m-1} \right\rfloor \leq k \leq \left\lfloor 1.25 \cdot c_{f0}^{m-1} \right\rfloor \\ C^m(k-1) < C^m(k) < C^m(k+1) \\ C^m(k) > 0.3 \cdot C^m\left(c_{f0}^m\right) \end{array} \right.$$

4.5 Voiceness Probability

A voiceness probability is computed by combining the voiciness probability of several descriptors. One descriptor is related to the variance of the correlation function A around the estimated fundamental frequency. it is computed as

$$A = \frac{C\left(c_{f0}\right) - C\left(c_{minr}\right)}{C_{max}} \cdot \frac{C\left(c_{maxr}\right) - C\left(c_{minr}\right)}{C_{max}} \cdot \frac{C\left(c_{f0}\right) - C\left(c_{minl}\right)}{C_{max}} \quad (8)$$

where

$$c_{minr} = \arg\min_{k \in [c_{f0}, 2 \cdot c_{f0}]} C(k)$$

$$c_{maxr} = \arg\max_{k \in [c_{minr}, 2.3 \cdot c_{f0}]} C(k)$$

$$c_{minl} = \arg\min_{k \in [0.2 \cdot c_{f0}, c_{f0}]} C(k)$$

Another descriptor relates to the variance of the correlation function in the 2-5kHz frequency band B, computed as

$$B = \frac{\sigma_B \cdot C\left(c_{f0}\right)}{{\mu_B}^2} \quad (9)$$

where

$$k_{B1} = \frac{2000 \cdot N}{SR}, k_{B2} = \frac{5000 \cdot N}{SR}$$

$$\mu_B = \frac{\sum_{k=k_{B1}}^{k_{B2}} C(k)^2}{\sum_{k=k_{B1}}^{k_{B2}} C(k)}$$

$$\sigma_B = \sqrt{\frac{\sum_{k=k_{B1}}^{k_{B2}} \left(C(k) - \mu_B\right)^2}{k_{B2} - k_{B1}}}$$

Other descriptors used are time-domain zero-crossing ZC, energy E and waveform tremolo $\hat{T}$. The waveform tremolo is modified depending on A as follows

$$T = \begin{cases} \hat{T} & if \quad A \leq 0.015 \\ \hat{T} \cdot \left(1 - \frac{A - 0.015}{0.008}\right) & if \quad 0.015 < A < 0.023 \\ 0 & if \quad A \geq 0.023 \end{cases} \quad (10)$$

The voiceness probability for each descriptor is computed as follows

$$P_A = \begin{cases} 1 & if \quad A \geq 0.2 \\ e^{-\frac{1}{2}\frac{(A-0.2)^2}{0.15^2}} & if \quad A < 0.2 \end{cases}$$

$$P_B = \begin{cases} 1 & if \quad B \geq 0.52 \\ e^{-\frac{1}{2}\frac{(B-0.52)^2}{0.04^2}} & if \quad B < 0.52 \end{cases}$$

$$P_{ZC} = \begin{cases} 1 & if \quad ZC \leq 0.1 \\ e^{-\frac{1}{2}\frac{(ZC-0.1)^2}{0.2^2}} & if \quad ZC > 0.1 \end{cases}$$

$$P_T = \begin{cases} 1 & if \quad T \leq 0.2 \\ e^{-\frac{1}{2}\frac{(T-0.2)^2}{0.16^2}} & if \quad T > 0.2 \end{cases}$$

$$P_E = \begin{cases} 1 & if \quad E \leq 0.00002 \\ e^{-\frac{1}{2}\frac{(E-0.00002)^2}{0.00001^2}} & if \quad E > 0.00002 \end{cases}$$

Finally, the resulting frame voiceness probability is estimated multiplying the voiciness probability of each descriptor as

$$P_{voiced} = P_A \cdot P_B \cdot P_{ZC} \cdot P_T \cdot P_E. \quad (11)$$

4.6 Voicing decision and post-processing

Let it be P_{voiced}^m the voiciness probability of the m^{th} frame. If the previous five frames were estimated as voiced with frame-to-frame c_{f0} differences lower than 20 bins, then the frame is considered to be voiced if $max\left(P_{voiced}^m, P_{voiced}^{m-1}\right) > 0.44$. Otherwise, the frame is considered voiced if $P_{voiced}^m > 0.44$.

The last step consists of a 5 frames post-processing that allows octave jumps typical of speech signals (e.g. creaky voice) while avoiding octave up or down segments of less than 3 frames. Also it avoids irregular F0 sequences with several large jumps of a few or more semitones, as well as avoids voiced segments shorter than 4 frames.

5. Perceptual Evaluation

Following our informal observations we found distinctive performance across the different techniques, with SAC, SWIPE, REAPER and SRH in order of performance . As expected, the most of errors were found as octave or random jumps within short time intervals, principally in vibrato regions and high-pitch singing (mezzo and soprano voices). SAC and SWIPE appeared to be the most robust and less sensitive to these phenomena. Moreoever, pitch oscillations (e.g. vibrato) were commonly not extracted smoothly by REAPER and SRH, as shown in Fig. 3 on a segment of a soprano sample. This could be perceived as a degradation after performing STRAIGHT resynthesis [18] using the pitch estimates as input. A MUSHRA test [19] reported in the next section, was conducted using these synthetic signals seeking to confirm our findings by perceptual evaluation.

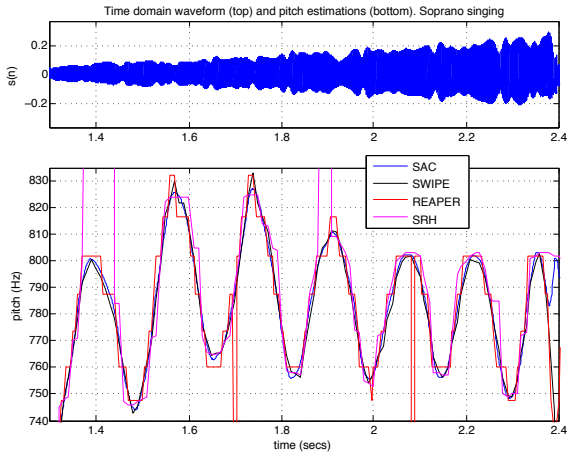

Fig. 3 Example of pitch estimation on high-pitched singing (soprano singer). Time-domain signal (top), estimates comparison (bottom).

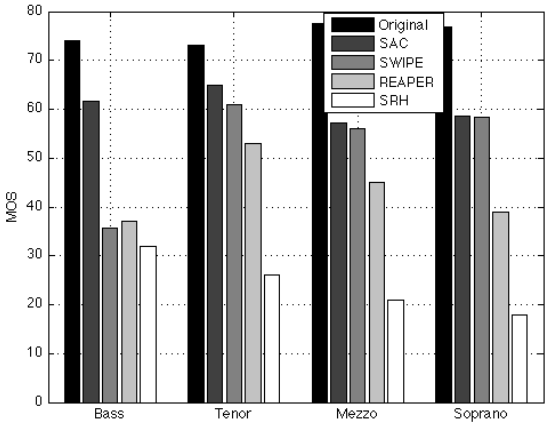

Fig. 4 Perceptual evaluation results (MOS scaled to [0, 100]).

5.1 MUSHRA Test

Ten samples of each of the four singers were used for resynthesis using the four different pitch estimations, resulting in a total of two hundred samples (including the originals). Ten listeners with music or singing background participated in the test. The test interface presented the original sample as reference and the five samples to evaluate randomly selected (including again the original). The listeners were asked to score in a MOS basis and to give the maximal value (in a scale up to 100) to the assumed hidden original. The participants were asked to focus principally in the quality of the pitch reproduction. It was allowed to replay any sample as many time as necessary until feel comfortable with the scores.

The results are shown in Fig. 4, organized by singer type. The tests confirmed the performance trends in both techniques and pitch range aspects. Resynthesis from SAC estimates obtained the highest scores, showing low sensibility to the pitch range. As expected, SWIPE was found the second best, followed by REAPER and SRH, in which were perceived noticeable artefacts due to jumps.

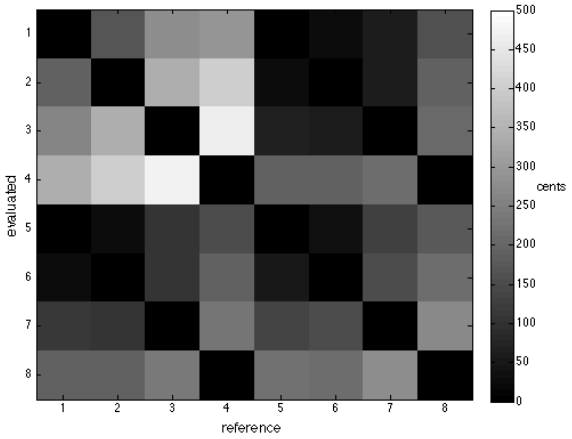

Fig. 5 Cross pitch-deviation comparison matrices (in cents).

5.2 Crossed Comparison

Since we lack of information of the glottal activity (i.e. EGG signals) an objective evaluation it is not straightforward. Nevertheless, we wanted to measure the degree of deviation between the different techniques and their sensibility to the pitch range (singer type) as informative of the stability of the techniques and the performance trend related to the height of the voice (singer). Therefore, we did a crossed comparison measuring the average standard deviation between the different estimates per sung excerpt. The computation was done also in cents as a musically motivated measure (100 cents represent a distance of a semitone).

For a matter of space we only show the result in cents in Fig. 5. The matrix shown correspond to four sub-matrices ordered as follows: Bass (left-top), Tenor (right-top); Mezzo (left-bottom); and Soprano (right-bottom). The horizontal axis denotes the technique used as reference and the vertical one the evaluated estimates following this order: SAC, SWIPE, REAPER, SRH. For each comparison, the time-axis of the reference estimation was fixed and the pitch of the evaluated one was linearly interpolated accordingly. All pitch deviations were calculated exclusively on segments identified as voiced by both SAC and REAPER voicing decisions. There were not found visible voicing mismatches across the different estimators that could significantly impact the performance comparison.

The results confirmed the distinctive performance in the expected order in terms of the degree of deviation between estimates. This was specially observed for low and high voices (bass and soprano) when measuring in hertz. SAC and SWIPE were systematically found stable and close, specially for mid and high pitch cases. It was interesting to find that, in general, the performance did not drop significantly on high or very high pitch (soprano), which is commonly considered a major challenge in the bibliography. The results measured in cents (Fig. 5) showed similar trends, with slightly closer rates for miid and high pitch voices but larger differences appearing on low pitch, suggesting that the impact of pitch errors may be more noticeable, in a musical or perceptual basis, on bass voices.

6. Objective Evaluation

We carried out an evaluation using synthetic signals. Following the subjective results we used SAC estimates as pitch templates that were furthermore scaled one semitone up and down to produce two synthetic versions. Then, we measured again the average standard deviations using the scaled templates as reference. Note that the characteristics of STRAIGHT resynthesis may have an impact on the performance due to: 1) a potential emphasizing/degradation of the information of the signal used by a particular algorithm, or 2) limitations when processing signals with extremely low or high pitch. We claim however that the performance trends might follow, globally, those of measurements using actual pitch information and natural signals.

The results are shown in Table 2 (hertz) and Table 3 (cents). The order of performance remains (excepting a switch between REAPER and SRH for the bass voice), denoting SAC as the best and more stable estimator with larger benefits at the extreme pitch ranges (low, high). Also, note that the effect of the pitch range does not seem to be as significant as it would be expected. Moreover, excepting SAC technique, there were confirmed larger error intervals on low-pitched singing. The overall results allow us to claim improved accuracy and high robustness of SAC over the other techniques for automatic wide-range pitch extraction.

Table 2 Average estimation deviation (in hertz) per excerpt.

algorithm	Bass	Tenor	Mezzo	Soprano
SAC	**4.77**	**2.61**	**3.41**	**4.32**
SWIPE	21.22	2.80	3.86	7.82
REAPER	47.68	9.71	11.09	22.32
SRH	29.73	17.63	37.52	37.07

Table 3 Average estimation deviation (in cents) per excerpt.

algorithm	Bass	Tenor	Mezzo	Soprano
SAC	**36.99**	**17.95**	**12.86**	**14.35**
SWIPE	137.05	19.66	13.64	27.71
REAPER	249.20	53.64	38.65	86.90
SRH	207.53	96.68	117.64	110.21

7. Conclusions

We presented in this work a robust strategy for automatic pitch estimation denoting high performance on wide pitch-range signals independently of the voice height. The proposed technique, called SAC, was evaluated and compared with three state-of-the-art techniques on natural opera singing observing rich pitch content. SAC estimation clearly showed the best and more robust estimation results according to subjective and objective evaluations when the F0 search interval in all methods is set wide enough to cover the overall expected range of opera singing.

Further study should be done to consider a global optimisation of the analysis settings in the other techniques that may improve their results seeking to complement the findings reported in this study. Also, a potential impact of STRAIGHT resynthesis on the performance of the different techniques might be clarified.

References

[1] Kawahara, H., Katsuse, M. and Cheveigne, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds, *Speech Communication*, Vol. 27, pp. 187–207 (1999).

[2] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. and Alku, P.: HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering, *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 19, No. 1, pp. 153–165 (2011).

[3] Erro, D., Sainz, I., Navas, E. and Hernaez, I.: Improved HNM-based Vocoder for Statistical Synthesizers, *In Proc. of Interspeech* (2011).

[4] Röbel, A., Villavicencio, F. and Rodet, X.: On cepstral and all-pole based spectral envelope modelling with unknown Model order, *Pattern Recognition Letters*, Vol. 28, No. 11, pp. 1343–1350 (2007).

[5] Villavicencio, F., Röbel, A. and Rodet, X.: Applying Improved Spectral Modeling for High-Quality Voice Conversion, *Proc. of ICASSP'09*, Vol. 1 (2009).

[6] Cheveigne, A. and Kawahara, H.: YIN, a fundamental fre- quency estimator for speech and music, *J. Acoust. Soc. Am.*, Vol. 111, No. 4 (2002).

[7] Saino, K., Zen, H., Nankaku, Y., Lee, A. and Tokuda, K.: An HMM-Based Singing Voice Synthesis System, *In Proc. of INTERSPEECH-ICSLP* (2006).

[8] Kenmochi, H. and Oshita, H.: VOCALOID Commercial singing synthesizer based on sample concatenation, *Proc. of INTERSPEECH'07*, Antwerp, Belgium (2007).

[9] Villavicencio, F. and Bonada, J.: Applying Voice Conversion to Concatenative Singing-Voice Synthesis, *Proc. of INTERSPEECH'10*, Vol. 1, Tokyo, Japan, pp. 2162–2165 (2010).

[10] Villavicencio, F. and Kenmochi, H.: Non-Parallel Singing-Voice Conversion by Phoneme-based Mapping and Covariance Approximation, *In Proc. of DAFx'11*, Paris, France, pp. 241–248 (2011).

[11] Babacan, O., Drugman, T., d'Alessandro, N., Henrich, N. and Dutoit, T.: A comparative study of pitch extraction algorithms on a large variety of singing sounds, *In Proc. of ICASSP* (2013).

[12] Camacho, A. and Harris, J.: A sawtooth waveform inspired pitch estimator for speech and music, *The Journal of the Acoustical Society of America*, Vol. 124 (2008).

[13] : [Online]. Available: https://github.com/google/REAPER (May 2015).

[14] Drugman, T. and A., A.: Joint robust voicing detection and pitch estimation based on residual harmonics, *In Proc. of Interspeech* (2011).

[15] Gómez, E. and Bonada, J.: Towards Computer-Assisted Flamenco Transcription: An Experimental Comparison of Automatic Transcription Algorithms from a Cappella Singing, *Computer Music Journal*, Vol. 37, pp. 73–90 (2013).

[16] : [Online]. Available: https://github.com/kylebgorman/swipe (May 2015).

[17] : [Online]. Available: https://github.com/covarep/covarep (May 2015).

[18] Kawahara, H. and Morise, M.: Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework, *SADHANA-Academy Proceedings in Engineering Science*, Vol. 36 (2011).

[19] : [Online]. Available: http://sourceforge.net/projects/matlabmushra/ (May 2015).