ACCURATE ANALYSIS AND VISUAL FEEDBACK OF VIBRATO IN SINGING

*José Ventura, Ricardo Sousa and Aníbal Ferreira**

University of Porto - Faculty of Engineering -DEEC
Porto, Portugal

ABSTRACT

Vibrato is a frequency modulation effect of the singing voice and is very relevant in musical terms. Its most important characteristics are the vibrato frequency (in Hertz) and the vibrato extension (in semitones). In singing teaching and learning, it is very convenient to provide a visual feedback of those two objective signal characteristics, in real-time. In this paper we describe an algorithm performing vibrato detection and analysis. Since this capability depends on fundamental frequency (F0) analysis of the singing voice, we first discuss F0 estimation and compare three algorithms that are used in voice and speech analysis. Then we describe the vibrato detection and analysis algorithm and assess its performance using both synthetic and natural singing signals. Overall, results indicate that the relative estimation errors in vibrato frequency and extension are lower than 0.1%.

1. INTRODUCTION

A voice signal is multidimensional in the sense that it conveys information allowing to answer to at least three questions: who said, what, and how ? These questions regard the identity of the speaker, the contents of the voice message, i.e. the semantics, and the speaking style, for example, it can be normal, breathy or pressed [1]. Similarly, the singing voice conveys different types of information including the sound signature of the singer, the lyrics and musical characteristics such as melody and its variations. In singing teaching and learning, the dialogue between student and instructor frequently includes subjective terms such as 'vibrant' or 'focused', as well as metaphors such as 'dark' or 'bright'. In order to make the dialogue between student and instructor more objective, it is important to provide an objective visual representation of musically relevant voice characteristics, namely fundamental frequency (i.e. pitch or F0) trajectories taking as a reference a given musical scale of notes such as the equal tempered musical scale. This particular representation allows very objectively to assess if the singing is in tune and if the melody transitions are as desired. This basic functionality is supported by several commercial software products including Sing&See, Music Master Works, Singing Coach, Singing Tutor, Singing SuperStar and Music Tutor.

Several singing voice characteristics have been studied and several estimation algorithms have been developed [2, 3]. One particular musical characteristic of singing is vibrato which may be defined as periodic variations of the fundamental frequency of the singing voice around an average value. Technically, vibrato is a sinusoidal-like frequency modulation effect. The two most important parameters describing vibrato are the vibrato frequency in Hertz, which corresponds to the rate of the periodic variation of the fundamental frequency, and the vibrato extension in semitones which corresponds to the amplitude of the frequency variation of the fundamental frequency. Most singing voice analysis/visual-feedback software products such as those indicated previously, do not support automatic detection and parametric characterization of vibrato. Also, research addressing vibrato detection has typically focused on applications other than real-time visual feedback of explicit vibrato parameters [4, 5].

In this paper, we first address in section 2 the problem of fundamental frequency estimation in singing since it determines the subsequent stage of vibrato estimation. We compare three algorithms used in speech and voice analysis, we assess their performance using synthetic and natural singing signals, and considering also the influence of noise. In section 3 we address the problem of automatic detection and parametrization of vibrato in singing. We describe the algorithmic approach for vibrato analysis and discuss its performance using both synthetic and natural singing signals. Section 4 concludes this paper.

2. FUNDAMENTAL FREQUENCY ESTIMATION

The estimation of the fundamental frequency of a signal consisting of a harmonic structure of sinusoids, notably of voiced speech or singing, has been a topic of intense research for many decades [6].

In this paper we assume the equal temperate scale as the reference musical organization of notes. In this scale, the centre frequency in Hertz of a note with index n is given by

$$f = 2^{n/12} F_{\text{ref}}, \tag{1}$$

*This work was supported by the Portuguese Foundation for Science and Technology, an agency of the Portuguese Ministry for Science, Technology and Higher Education, under research project PTDC/SAU-BEB/14995/2008.

where F_{ref} is a reference frequency, e.g. 440 Hertz (corresponding to note A4). The musical note corresponding to a specific index is known as semitone, or ST. The frequency doubles when the index n increases by 12, i.e. an octave consists of 12 semitones. For practical purposes, it is common to use the rule given by eq. (2) to convert the frequency of a musical note into an index in the MIDI scale (MIDI stands for Musical Instrument Digital Interface and consists of a protocol specifying a symbolic notation of music; MIDI is used by electronic musical instruments, computer and other musical devices):

$$P = 69 + \log_2 \frac{f}{F_{\text{ref}}} \ . \tag{2}$$

The index P is coded as a binary word by the MIDI protocol and denotes the musical note that is synthesised by a MIDI synthesizer using an appropriate mathematical model of a musical instrument.

The first task of a fundamental frequency estimator, or pitch detector, is to reliably and accurately estimate the frequency of the lowest partial in a harmonic structure of sinusoids and this type of structure is naturally generated by most string and wind musical instruments, including voice. The task is strongly affected by the influence of noise, discontinuities in the harmonic structure, including the problem of missing fundamental, as well as by the simultaneous occurrence of competing harmonic structures corresponding to different musical notes. Pitch estimation methods can be broadly classified as time-based and spectral-based methods [6]. The former are usually more simple and less demanding computationally that the latter.

In our comparative evaluation we have included two time-based methods and one frequency-domain method. An important common feature is that all three have been tailored for speech or voice analysis and for real-time applications. The time-based methods are based on the autocorrelation function and have been proposed by Boersma [7] and Cheveigné and Kawahara [8]. The method proposed by Boersma is a part of a popular voice analysis software known as Praat[1]. The Cheveigné and Kawahara [8] method is known as Yin and has been acknowledged by several authors as an accurate and robust method. The frequency-based method is based on our previous research results [9, 10, 11]. This method is based on a two step approach. First, using a cepstral analysis, the most likely eight fundamental frequency candidates are identified. Secondly, for each candidate, the magnitude spectrum is analysed in detail so as to compute the likelihood of that candidate considering such aspects as harmonic discontinuities, total number and power of the existing harmonic partials. Finally, all candidates are ranked and the candidate reaching the highest likelihood score is selected.

In order to test the different algorithms we have considered three types of test signals with no vibrato:

- synthetic singing,
- synthetic singing affected by noise,
- natural singing.

The synthetic singing signals have been generated using a publicly available synthesizer (MADDE[2]) developed at the Royal Institute of Technology (KTH). In MADDE a significant number of parameters may be adjusted as desired to control the synthetic singing such as the F0 frequency, the formant frequencies, the spectral tilt of the glottal source, and the main vibrato parameters: vibrato frequency (in Hertz) and extension (in ST).

In the following three sub-sections we describe the different tests and we discuss the main results and conclusions.

2.1. F0 Estimation using synthetic singing

Using the MADDE synthesizer, we synthesized 22 files of singing voice comprising the semitones from G2 ($F_0 = 98,0$ Hz) till G5 ($F_0 = 784,0$ Hz). The sampling frequency is 22050 Hz, the duration of each test file is 2 seconds and the singing is 'flat', i.e. it has no vibrato. We evaluated the relative error of the F0 estimation for each algorithm and test file. The statistics obtained for each algorithm averaging results over all test files are shown in Table 1. It can be concluded

Table 1. Relative F0 estimation errors obtained for the SearchTonal, Yin and Boersma algorithms in the absence of noise.

	SearchTonal	Yin	Boersma
Max	0,79%	0,79%	1,94%
Min	0,11%	0,11%	0,19%
Mean	0,30%	0,30%	0,69%
STD	0,20%	0,19%	0,37%

that while the SearchTonal and Yin algorithms perform quite similarly, the Boersma algorithm exhibits relative estimation errors which are in average two times as much as those of the other two algorithms. A detailed analysis of the results has revealed that the performance of the Boersma algorithm degrades for F0 higher than 200 Hz.

2.2. F0 Estimation using synthetic singing affected by noise

Regarding the second type of test signals, we have added white Gaussian noise to each one of the previously generated clean signals such as to reach different SNRs: 5, 10, 15, 20, 25 and 30 dBs. Thus, the number of test signals has increased by a factor of 6. Due to its relative poor performance,

[1] http://www.fon.hum.uva.nl/praat/, last accessed on Oct 28th 2011.

[2] Madde software, available at http://www.speech.kth.se/music/downloads/smptool/. Accessed on June 26th 2011.

the Boersma algorithm has been excluded from this test. In addition, we concluded that for F0 values higher that about 250 Hz, estimation errors increase significantly, particularly for the SearchTonal algorithm. A detailed study of the problem led to the conclusion that the reason is the fact that using the default parameters concerning the spectral tilt of the glottal source, MADDE synthesises singing voice using essentially the lowest ten harmonic partials. By increasing the noise level, some of the partials become overwhelmed by the noise and therefore can not be detected. This is a problem for SearchTonal because it identifies all harmonics that are at least 5 dBs above the noise floor and it requires that at least three harmonic sinusoids be detected for a harmonic structure to be recognized. For these reasons, only F0 values ranging from G2 ($F_0 = 98,0$ Hz) till F4 ($F_0 = 349,2$ Hz), in total 14 different pitches, were included in the test. Given the SNR range, in total 84 test files were generated and tested.

Proceeding as in the previous section and averaging results over F0 and SNR for each algorithm, we obtained the relative estimation errors which are shown in Table 2. These

Table 2. Relative F0 estimation errors obtained for the SearchTonal and Yin algorithms when the tests signal are contaminated with AWGN.

	SearchTonal	Yin
Max	2,14%	0,79%
Min	0,24%	0,18%
Mean	0,68%	0,40%
STD	0,56%	0,18%

results involving synthetic singing signals reveal that the Yin algorithm is quite robust since the noise influence does not degrade significantly its performance relative to the case of clean test signals. On the other hand, results suggest that because the SearchTonal algorithm expects a minimum number or 'surviving' partials in the harmonic structure, this constrains its performance under strong noise influence. It should be noted however that the test signals were synthesized using MADDE which may differ significantly from natural singing.

2.3. F0 Estimation using natural singing

In this test twelve natural singing voices (not necessarily 'flat') were involved, six male voices and six female voices. The male voices consist of vowels /a/, /e/ and /u/ sung at average pitch C4, and the same vowels sung at average pitch G3. The female voices consist of the same vowels sung at average pitch A4, and the same vowels sung at average pitch D5. The average duration of each test file is 8 s. Since no ground truth exists for the exact F0 contour of each test file, the comparison between the outputs of the algorithms was assessed visually by inspecting such aspects as smoothness and consistency of the results.

Figure 1 illustrates the F0 estimation results on the MIDI scale (according to eq. (2)) due to the three algorithms under test and using one particular test file. For this particular example, it can be concluded that the SearchTonal algorithm delivers the smoothest F0 contour, that F0 estimation using Yin gives rise to several F0 discontinuities, and that the Boersma algorithm tends to deliver coarse F0 estimation results (i.e. top-flattened), particularly in regions of the singing signal exhibiting very low pitch variations. Although Fig. 1 represents one example, the associated conclusions have been confirmed for most test files. The results obtained for the SearchTonal and Yin algorithms were very consistent for a reduced number of test files.

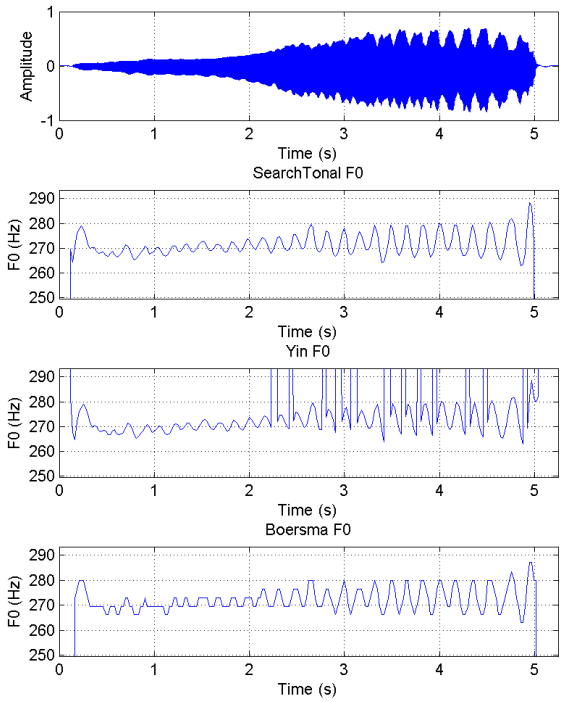

Fig. 1. Pitch estimation of natural singing (vowel /a/, the time representation is in the top panel) by three different estimation algorithms.

Taking into consideration the results presented in subsections 2.1, 2.2, and 2.3, we have decided to use the SearchTonal algorithm as the basic F0 estimation algorithm for the vibrato detection tests. This decision represents a choice mainly due to the smooth behaviour of SearchTonal, as suggested in Fig. 1, and strictly not a selection since the Yin algorithm could also be taken as a fair choice.

3. VIBRATO ESTIMATION

Vibrato is characterized according to four parameters: frequency (in Hertz), extension (in semitones or ST), duration

(in seconds) and regularity. Of these only regularity has not a clear definition, it can be broadly described as an oscillation pattern around a center frequency. In most cases the expected simplest pattern is just a single sinusoidal variation. This is illustrated in Fig. 2 which represents the graphical environment of a real-time singing analysis software (SingingStudio) we have developed in the context of an academic spin-off company[3]. In addition to the sinusoidal variation of the vibrato, Fig. 2 also helps to readily identify the center frequency of the musical note (A3 in the illustrated example) as well as the extension of vibrato which, in the illustrated example, is bounded by the limits of a semitone, i.e. the extension is 0.5 ST. In addition, Fig. 2 also illustrates the spectral structure of a region in the signal which is signalled by means of a subtle vertical bar on the piano keyboard panel of the SingingStudio GUI. The spectral representation highlights the harmonic structure of the signal whose harmonic partials are all signalled, namely the first ten, and which are used to perform accurate frequency estimation according to the SearchTonal algorithm [9, 10]. The displayed singing signal has been generated using the MADDE synthesizer.

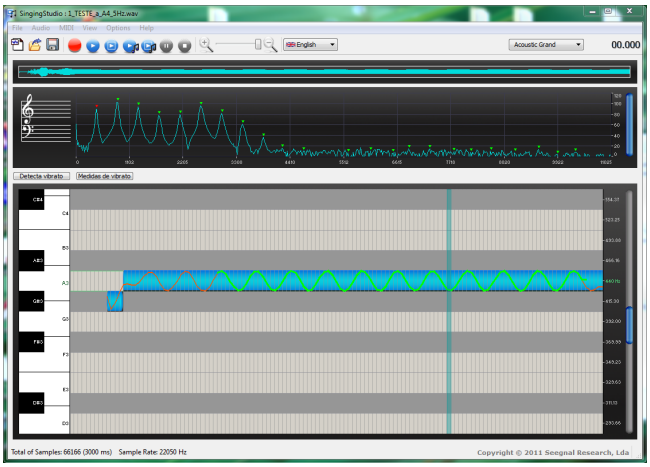

Fig. 2. GUI of the SingingStudio singing analysis software.

According to Sundberg [12], an aesthetically reasonable range for the vibrato frequency is between 5.5 Hertz and 7.5 Hertz. Similarly, an aesthetically reasonable upper limit for the vibrato extension is 2 ST, which corresponds to a frequency variation relative to the center frequency by about 12%.

Our vibrato detection algorithm takes as input the F0 contour as obtained from the fundamental frequency estimation stage, as discussed in section 2. The first step after the F0 contour data is available, is to convert it to the logarithmic scale according to eq. (2). The rationale is that vibrato perception is strongly linked to the natural frequency organization of the human auditory system which follows a log-like rule [13]. This step highlights that smoothness in the F0 contour information (i.e. absence of discontinuities) is very important. A convenient side-effect of this scale mapping is that the sinusoidal profile of natural vibrato is more faithfully represented in the log-based MIDI scale than in the linear frequency scale (in this scale the wave shape of vibrato is rather skewed).

Since vibrato analysis involves spectral analysis of the F0 contour, we use non-iterative accurate frequency estimation according to the algorithm presented in [10]. The main idea is to submit the F0 contour data to an FFT, and then to perform accurate frequency and magnitude estimation using the spectral peaks in the FFT magnitude spectrum denoting the vibrato effect.

First, we should address the theoretical accuracy of this approach. The F0 contour data is obtained from the SearchTonal algorithm using a 1024-point FFT analysis with 50% overlap on audio signals sampled at 22050 Hz. This means that the time resolution of the F0 contour data is 23.2 ms which corresponds to a sampling frequency of 43 Hz. If we admit an extended vibrato frequency range from 4 Hz till 8 Hz, one important condition constrains the size N of the FFT analysing the F0 contour data. In fact, in order to avoid leakage due to windowing prior to FFT analysis, the lowest vibrato frequency (4 Hz) should give rise to a spectral peak sufficiently distant from the first FFT bin (where all the DC component of the F0 contour signal falls). Considering that the main lobe width of the frequency response of popular windows such as the Rectangular (which is 2 DFT bins) and the Hanning window (which is 4 DFT bins), one easily concludes that the lowest vibrato frequency should give rise to a spectral peak falling on an FFT bin higher than 4, meaning that N should be larger than $(43 \times 5/4 =)\ 53.8$ or $N = 64$ if we choose the next power-of-two number.

On the other hand, the time resolution of the vibrato information must be commensurate to or less than the period of the highest vibrato frequency (8 Hz). This means that the shift in samples between adjacent FFTs should be in the order of $(43/8 =)\ 5.4$ samples; we adjust this number to 6 samples in order to facilitate real-time constraints. This analysis leads to the conclusion that our vibrato analysis algorithm should be based on a 64-point FFT running with about 91% overlap on the F0 contour data.

A final but important issue regards the accuracy of the vibrato frequency estimation. Obtaining the vibrato frequency by just rounding the frequency of a spectral peak in the FFT magnitude spectrum implies a maximum estimation error corresponding to 50% of the bin width, or $(0.5 \times 43/64 =)\ 0.34$ Hz. Using accurate frequency estimation as in [10] the maximum estimation error reduces to about 0.1% of the bin width which means that the maximum estimation error can be as low as 6.7E-4 Hz, i.e. less than 1/1000 of 1 Hz.

Thus, our vibrato estimation algorithm can be briefly described as follows:

- take a segment of 64 samples from the F0 contour data,

[3] http://www.seegnal.com, last accessed on Oct 28th 2011.

- remove the DC component from the segment such as to minimize leakage effects,
- compute the magnitude spectrum as specified in [10],
- obtain the spectral envelope model by short-pass liftering the real cepstrum (by preserving just the first four cepstrum bins as well as their replicas on the negative frequency axis),
- detect the largest peak in the magnitude spectrum within the expected vibrato frequency range (i.e. between bins 6 and 12),
- evaluate the dB difference between this local maximum and the noise floor as well as the floor defined by the spectral envelope model (for additional reliability),
- if the dB difference is larger than a predefined threshold (set to 3.8 dB), then declare the current segment exhibits vibrato and compute the accurate vibrato frequency using the frequency interpolation method described in [10],
- if vibrato has been declared, then compute its extension in ST by computing the average difference between F0 maxima and minima within the current segment of the F0 contour.

Figure 3 illustrates the result of the algorithm for a short segment of natural singing. The top panel in this figure represents the F0 contour data obtained from the SearchTonal algorithm, the other panels represent the estimated vibrato frequency in Hertz as a function of time, and the estimated vibrato extension in ST as a function of time.

In order to evaluate the performance of our vibrato detection algorithm, we have used two types of test signals containing vibrato: synthetic singing voices and natural singing voices.

The synthetic singing voices have been obtained using the MADDE software as mentioned in section 2. This software is very convenient because the frequency and extension of the vibrato can be independently adjusted.

In order to assess the performance of the vibrato frequency estimation, we set the F0 pitch frequency to A4 (440 Hz) and we generated several test signals by varying the vibrato frequency and using two vibrato extension values. A second set of test signals was specifically generated to assess the performance of the vibrato extension estimation. In this case, we set the F0 pitch frequency to C4 (261.6 Hz) and set the vibrato frequency to 5 Hz. Then, we varied the vibrato extension. The natural singing voices used in this test are the same as those mentioned in section 2.3. The following subsections describe the tests in detail and discuss the results.

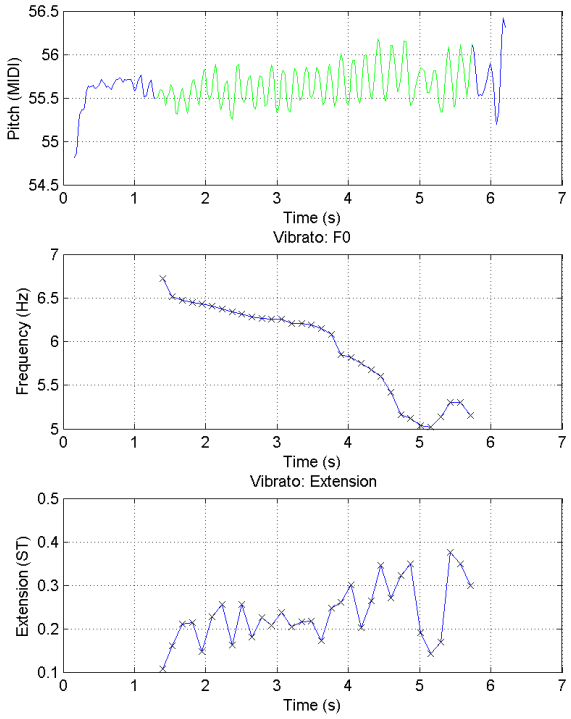

Fig. 3. Vibrato estimation from natural singing. The top panel represents the F0 contour data and the region where vibrato has been declared is highlighted. The other panels represent the vibrato frequency and extension as a function of time.

3.1. Vibrato frequency estimation using synthetic singing

Using the MADDE synthesizer, we synthesized two sets of 9 test files by varying the vibrato frequency from 4 Hz till 8 Hz in steps of 0.5 Hz (F0=440 Hz), and by setting the vibrato extension to 0.5 ST and 1.0 ST. The sampling frequency is 22050 Hz and the duration of each test signal is 2 s.

We evaluated the relative error of the vibrato frequency estimation for each test file. The statistics obtained by averaging results over all 9 test files for each setting of the vibrato extension, are shown in Table 3. The results reveal that the relative estimation errors increase slightly when the vibrato extension increases, and a detailed analysis of the results also

Table 3. Relative estimation errors of the vibrato frequency when the vibrato extension is set to 0.5 ST and to 1.0 ST. F0=440 Hz.

Extension	0.5 ST	1.0 ST
Max	0,080%	0,227%
Min	0,027%	0,019%
Mean	0,057%	0,082%
STD	0,020%	0,068%

reveals the same tendency applies for higher vibrato frequencies. This last aspect is a bit unexpected given that leakage effects are milder but is probably explained by the fact that the FFT overlap should increase for higher vibrato frequencies. In any case, the estimation errors are extremely modest indicating that the estimation algorithm is very accurate.

3.2. Vibrato extension estimation using synthetic singing

We have also used the MADDE synthesizer to generate 11 test files by varying the vibrato extension from 0.2 ST till 1.2 ST in steps of 0.1 ST (F0=261.6 Hz), and by setting the vibrato frequency to 5 Hz. As before, the sampling frequency is 22050 Hz and the duration of each test signal is 2 s.

After obtaining the relative error of the vibrato extension estimation for each test file, statistics were obtained by averaging results over all 11 test files. The main results are shown in Table 4. A detailed analysis of the results reveals that the

Table 4. Relative estimation errors of the vibrato extension. The vibrato frequency is set to 5 Hz and F0=261.6 Hz.

Max	0,049%
Min	0,011%
Mean	0,029%
STD	0,012%

estimation errors decrease slightly when the vibrato extension increases, a result which is expected since there is less noise influence. Overall and as in the previous test, the estimation errors confirm that the estimation algorithm is very accurate.

3.3. Vibrato frequency and extension estimation using natural singing

In this test we have used all natural singing files already described in section 2.3. One example of the vibrato frequency and extension estimation has already been illustrated in Fig. 3. Because in this type of test there is no ground truth to asses the accuracy of the results, the assessment is made by evaluating the continuity and smoothness of the frequency and extension estimation contours. Overall, results reveal that the algorithm is able to track fast variations in frequency and extension, thus providing great detail to the analysis of the singing performance. In a reduced number of test files, small transitions in the frequency and extension contours of the vibrato estimation are observed but those take place is regions of very fast variations of the F0 contour, which looks consistent.

4. CONCLUSION

In this paper we have discussed the importance of automatic estimation of the vibrato frequency and extension in singing, we have compared several F0 estimation algorithms, and we have described a vibrato detection and analysis algorithm whose performance has been assessed using both synthetic and natural singing signals. Results are very encouraging as the estimation relative errors are less than 0.1%. The proposed algorithm will be included in the SingingStudio platform as a new real-time vibrato analysis functionality.

5. REFERENCES

[1] Paavo Alku, "An automatic method to estimate the time-based parameters of the glottal pulseform," in *IEEE ICASSP*, 1992, pp. II–29–II–32.

[2] K. Murphy, *Digital signal processing techniques for application in the analysis of pathological voice and normophonic singing voice*, Ph.D. thesis, Facultad de Informtica (UPM), Spain, 2008.

[3] A. Loscos, *Spectral Processing Of The Singing Voice*, Ph.D. thesis, Universidad Pompeu Fabra, Spain, 2007.

[4] Tin Nwe and Haizhou Li, "Exploring vibrato-motivated acoustic features for singer identification," *IEEE TASLP*, vol. 15, no. 2, pp. 519–530, February 2007.

[5] S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J.-L. Collette, "Vibrato: Detection, estimation, extraction, modification," in *In Proc. DAFx*, 1999.

[6] Wolfgang Hess, *Pitch Determination of Speech Signals -algorithms and devices*, Springer-Verlag, 1983.

[7] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. of the Inst. of Phonetic Sciences*, 1993, vol. 17, pp. 97–110, University of Amsterdam.

[8] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, no. 4, pp. 1917–1930, April 2002.

[9] Aníbal J. S. Ferreira, "Tonality detection in perceptual coding of audio," *98th AES Conv.*, February 1995, Preprint n. 3947.

[10] Ricardo Sousa and Aníbal J. S. Ferreira, "Non-iterative frequency estimation in the DFT magnitude domain," in *4th ISCCSP*, March 2010.

[11] Aníbal Ferreira, Filipe Abreu, and Deepen Sinha, "Stereo acc real-time audio communication," *125th AES Convention*, October 2008, Paper 7502.

[12] J. Sundberg, *The Science of the Singing Voice*, Northern Illinois University Press, 1987.

[13] Brian C. J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, 1989.