EURASIP Journal on Applied Signal Processing 2004:7, 1007–1020

Vibrato in Singing Voice: The Link between Source-Filter and Sinusoidal Models

Ixone Arroabarren

Departamento de Ingeniería Eléctrica y Electrónica, Universidad Pública de Navarra, Campus de Arrosadia, 31006 Pamplona, Spain
Email: ixone.arroabarren@unavarra.es

Alfonso Carlosena

Departamento de Ingeniería Eléctrica y Electrónica, Universidad Pública de Navarra, Campus de Arrosadia, 31006 Pamplona, Spain
Email: carlosen@unavarra.es

Received 4 July 2003; Revised 30 October 2003

The application of inverse filtering techniques for high-quality singing voice analysis/synthesis is discussed. In the context of source-filter models, inverse filtering provides a noninvasive method to extract the voice source, and thus to study voice quality. Although this approach is widely used in speech synthesis, this is not the case in singing voice. Several studies have proved that inverse filtering techniques fail in the case of singing voice, the reasons being unclear. In order to shed light on this problem, we will consider here an additional feature of singing voice, not present in speech: the *vibrato*. Vibrato has been traditionally studied by sinusoidal modeling. As an alternative, we will introduce here a novel noninteractive source filter model that incorporates the mechanisms of vibrato generation. This model will also allow the comparison of the results produced by inverse filtering techniques and by sinusoidal modeling, as they apply to singing voice and not to speech. In this way, the limitations of these conventional techniques, described in previous literature, will be explained. Both synthetic signals and singer recordings are used to validate and compare the techniques presented in the paper.

Keywords and phrases: voice quality, source-filter model, inverse filtering, singing voice, vibrato, sinusoidal model.

1. INTRODUCTION

Inverse filtering provides a noninvasive method to study voice quality. In this context, high-quality speech synthesis is developed using a source-filter model, where voice texture is controlled by glottal source characteristics. Efforts to apply this approach to singing voice have failed, the reasons being not clear: either the unsuitability of the model, or the different range of frequencies, or both, could be the cause. The lyric singers, being professionals, have an efficiency requirement, and as a result, they are educated to change their formants position moving them towards the first harmonics position, what could also be another reason of the model's failure [1].

This paper purports to shed light on this problem by comparing two salient methods for glottal source and vocal tract response (VTR) estimation, with a novel frequency-domain method proposed by the authors. In this way, the inverse filtering approach will be tested in singing voice analysis. In order to have a benchmark, the source-filter model will be compared to sinusoidal model and this comparison will be performed thanks to the particular feature of singing voice: vibrato.

Regarding the voice production models, we can distinguish two approaches as follows.

(i) On the one hand, interactive models are closer to the physical features of the vocal system. This system is composed by two resonant cavities (subglottal and supraglottal) which are connected by a valve, the glottis, where vocal folds are located. The movement of the vocal folds provides the harmonic nature of the air flow of voiced sounds, and also controls the coupling between the two resonant cavities, which will be different during the open and closed phases. As a result of this effect, the VTR will change during a single fundamental period and there will be a relationship between the glottal source and the VTR. This physical behavior has been modeled in several ways, by physical models [2] or aerodynamic models [3, 4]. From the signal processing point of view, in [4] the VTR variation is related to the glottal area, which controls the coupling of the cavities, and this relationship is represented by a frequency modulation of the central frequency and bandwidth of the formants. Other effect of the source-tract interaction is the increase of the skewness of the glottal source [4], which emphasizes the difference between the glottal area and the glottal source [5].

(ii) On the other hand, Non Interactive Models separate the glottal source and the VTR, and both are independently modeled as linear time-varying systems. This is the case of the source-filter model proposed by Fant in [6]. The VTR is modeled as an all-pole filter, in the case of nonnasal sounds. For the glottal source several waveform models have been proposed [7, 8, 9], but all of them try to include some of the features of the source-tract interaction, typically the asymmetric shape of the pulse. These models provide a high quality synthesis framework for the speech with a low computational complexity. The synthesis is preceded by an analysis stage, which is divided into two steps: an inverse filtering step where the glottal source and the VTR are separated [9, 10, 11, 12, 13] and a parameterization step where the most relevant parameters of both elements are obtained [14, 15, 16].

In general, inverse filtering techniques yield worse results as the fundamental frequency increases, as is the case of women and children in speech and singing voice. In the latter case, singing voice, the number of published works is very scarce [1, 17]. In [1], the glottal source features are studied in speech and singing voice by acoustic and electroglottographic signals [18, 19]. From these works, it is not apparent which is the main limitation of inverse filtering in singing voice. It might be possible that the source-tract interaction was more complex than in speech, what would represent a paradox in the noninteractive assumption [20]. Other reason mentioned in [1] is that perhaps the glottal source models used in speech are not suitable for singing voice. These statements are not demonstrated, but are interesting questions that should be answered.

On the other hand, in [17] the noninteractive source-filter model is used as a high-quality singing voice synthesis approach. The main contribution of that work is the development of an analysis procedure that estimates the parameters of the synthesis model [12, 21]. However, there is no evidence that could point to differences between speech and singing as it is indicated in [1].

One of the goals of the present work is to clarify whether the noninteractive models are able to model singing voice in the same way as high-quality speech, or on the contrary, the source-tract interaction is different from speech, and precludes this linear model assumption. If the noninteractive model could model singing voice, the reason of the failure of inverse filtering techniques would be just the high fundamental frequency of singing voice.

To this end, we will compare in this paper three different inverse filtering techniques, one of them novel and proposed recently by the authors in order to obtain the source-filter decomposition. Though they work correctly for speech and low-frequency signals, we will show their limitations as the fundamental frequency increases. This is described in Section 2.

Since fundamental frequency in singing voice is higher than in speech, it seems obvious that the above-mentioned methods fail, apparently due to the limited spectral information provided in high pitched signals. To compensate for that, we claim that the introduction of a feature such as vibrato may serve to increase the information available by virtue of the frequency modulated nature, and therefore wider bandwidth, of vibrato [22, 23, 24]. Frequency variations are influenced by the VTR, and this effect can be used to obtain information about it.

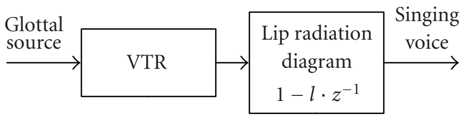


Figure 1: Noninteractive source-filter model of voice production system.

With this in mind, it is not surprising that vibrato has been traditionally analyzed by sinusoidal modeling [25, 26], the most important limitation being the impossibility to separate the sound generation and the VTR. In Section 3, we will take a step forward by introducing a source-filter model, which accounts for the physical origin of the main features of singing voice. Making use of this model, we will also demonstrate how the simpler sinusoidal model can serve to obtain a complementary information to inverse filtering, particularly in those conditions where the latter method fails.

2. INVERSE FILTERING

Along this section, the noninteractive source-filter model, depicted in Figure 1, will be considered and some of the possible estimation algorithms for it will be reviewed.

According to the block diagram in Figure 1, singing voice production can be modeled by a glottal source excitation that is linearly modified by the VTR and the lip radiation diagram. Typically, the VTR is modeled by an all-pole filter, and relying on the linearity of the model, the lip radiation system is combined with the glottal source, in such a way that the glottal source derivative (GSD) is considered as the vocal tract excitation.

In this context, during the last decades many inverse filtering algorithms to estimate the model elements have been proposed. This technique is usually accomplished in two steps. In the first one, the GSD waveform and the VTR are estimated. In the second one, these signals are parameterized in a few numerical values. This whole analysis can be practically implemented in several ways. For the sake of clarity, we can group these possibilities into two types.

(i) In the first group, the two identification steps are combined in a single algorithm, for instance in [9, 12]. There, a mathematical model for GSD and the autoregressive (AR) model for the VTR are considered, and then authors estimate simultaneously the VTR and the GSD model parameters. In this way, the GSD model parameterizes a given phonation type. Several different algorithms follow this structure, but all of them are invariably time domain implementations that require glottal closure instant (GCI) detection [27]. Therefore, they suffer from a high computational load, what makes them very cumbersome.

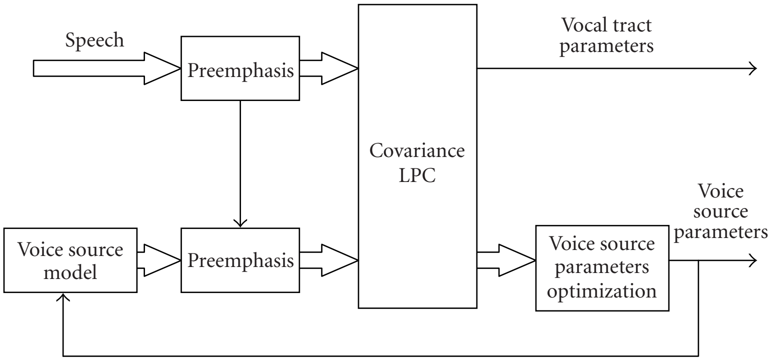


FIGURE 2: Block diagram of the AbS inverse filtering algorithm.

(ii) The procedures in the second group split the whole process into two stages. Regarding the first step, different inverse filtering techniques are proposed, [11, 13]. These algorithms remove the GSD effect from the speech signal and the VTR is obtained by linear prediction (LP) [28] or alternatively by discrete all-pole (DAP) modeling [29], which avoids the fundamental frequency dependence of the former.

For this comparative study three inverse filtering approaches have been selected. The first one is the *analysis by synthesis* (AbS) procedure presented in [9], the second one is the one proposed by the authors in [13], *Glottal Spectrum Based* (GSB) inverse filtering. In this way, both groups of algorithms mentioned above are represented. In addition, the *Closed Phase Covariance* (CPC) [10] has been added to the comparison. This approach is difficult to classify because it only obtains the VTR, as it is the case in the second group, but it is a time domain implementation as in the first one. The most interesting feature of this algorithm is that it is less affected by the formant ripple due to the source-tract interaction, because it only takes into account the time interval when the vocal folds are closed. In what follows, the three approaches will be shortly described, and finally compared.

2.1. Analysis by synthesis

This inverse filtering algorithm was proposed in [9]. It is based on covariance LPC [29], but the least squares error is modified in order to include the input of the system:

$$E = \sum_{n=0}^{N-1}\left(s(n)-\hat{s}(n)\right)^2 = \sum_{n=0}^{N-1}\left(s(n)-\left(\sum_{k=1}^{p}a_k s(n-k)+a_{p+1}g(n)\right)\right)^2, \tag{1}$$

where $g(n)$ represents the GSD, and

$$H(z) = \frac{a_{p+1}}{1-\sum_{k=1}^{p}a_k z^{-k}} \tag{2}$$

represents the VTR. Since neither VTR nor GSD parameters are known, an iterative algorithm is proposed and a simultaneous search is developed. The block diagram of the algorithm is represented in Figure 2.

As in covariance LP without source, this approach allows shorter analysis windows. However, the stability of the system is not guaranteed and a stabilization step must be included with this purpose. Also, and since it is a time domain implementation, the voice source model must be synchronized with the speech signal and a high sampling frequency is mandatory in order to obtain satisfactory results. As a result, the computational load is also high. Regarding the GSD parameter optimization, it is dependent on the chosen model. In the results shown in Section 2.4, the LF model is selected because it is one of the most powerful GSD models, and it allows an independent control of the three main features of the glottal source: open quotient, asymmetry coefficient and spectral tilt. The disadvantage of this model is its computational load. For more details on the topic readers are referred to [8].

Regarding fundamental frequency limits, it is shown in [1] that this algorithm provides unsatisfactory results for medium and high pitched signals.

2.2. Glottal spectrum based inverse filtering

This technique was proposed by the authors in [13] and will be briefly described here. Unlike the technique described in the previous section, it is essentially a frequency domain implementation. In the AbS approach, the GSD effect was included in the LP error, and the AR coefficients were obtained by Covariance LPC. In our case, a short term spectrum of speech is considered (3 or 4 fundamental periods), and the GSD effect is removed from the speech spectrum. Then, the AR coefficients of (2) are obtained by the DAP modeling [29].

For this spectral implementation, the KLGLOTT88 model [7] has been considered. It is less powerful than the LF model, but of a simpler implementation.

As it is shown in Figure 3, there is a basic voicing waveform controlled by the open quotient (O_q) and the amplitude of voicing (AV), the spectral tilt being included by a first-order lowpass filter.

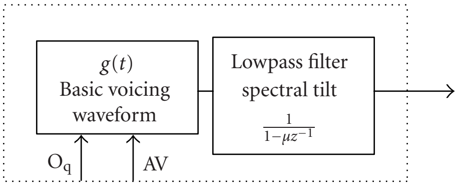


Figure 3: Block diagram of the KLGLOTT88 model.

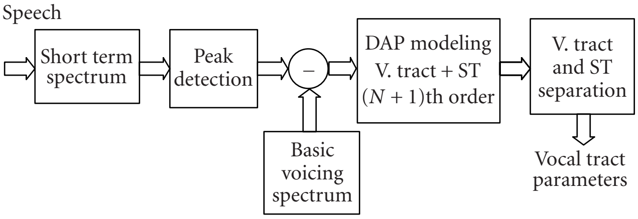


Figure 4: Block diagram of the GSB inverse filtering algorithm.

In our inverse filtering algorithm, once the short term spectrum is calculated, the glottal source effect is removed, by spectral division, by using the spectrum of the basic voicing waveform (3), which can be directly obtained by the Fourier transform of the basic voicing waveform [30]:

$$G(f) = \frac{27\,\mathrm{AV}}{2\,\mathrm{O_q}(2\pi f)^3}\left[\frac{je^{-j2\pi f\,\mathrm{O_q}\,T_o}}{2} + \frac{1+2e^{-j2\pi f\,\mathrm{O_q}\,T_o}}{2\pi f\,\mathrm{O_q}\,T_o} + 3j\frac{1-e^{-j2\pi f\,\mathrm{O_q}\,T_o}}{\left(2\pi f\,\mathrm{O_q}\,T_o\right)^2}\right]. \tag{3}$$

The spectral tilt (ST) and the VTR are combined in an $(N+1)$th order all-pole filter. The block diagram of the algorithm is shown in Figure 4.

Since DAP modeling is the most important part of the algorithm, we should explain its rationale. In classical autocorrelation LP [28], it is a well-known effect that as fundamental frequency increases the resulting transfer function is biased by the spectral peaks of the signal. This happens because the signal is assumed to be the impulse response of the system, and this assumption is obviously not entirely correct. In order to avoid this problem, an alternative proposed in [29] is to obtain the LP error based on the spectral peaks, instead of on the time domain samples. Unfortunately, this error calculation is based on an aliased version of the right autocorrelation of the signal, and this aliasing grows as the fundamental frequency increases. Then, the resulting transfer function is not correct again. To solve this problem, the DAP modeling uses the Itakura-Saito error, instead of the least squares error, and it can be shown that the error is minimized using only the spectral peaks information. The details of the algorithm are explained in [29]. This technique allows higher fundamental frequencies than classical autocorrelation LP, but for proper operation requires an enough number of spectral peaks in order to estimate the right transfer function. So, this inverse filtering algorithm will also have a limit in the highest achievable fundamental frequency.

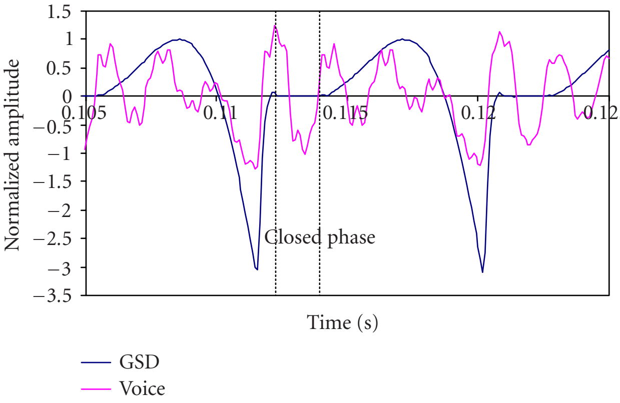


Figure 5: Closed phase interval in voice.

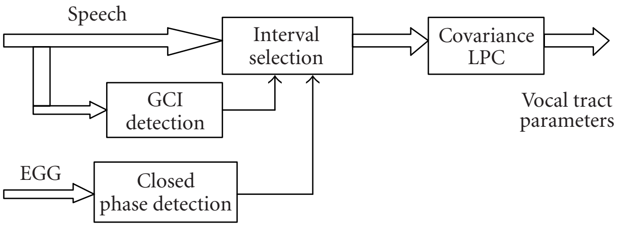


Figure 6: Closed phase covariance (CPC).

2.3. Closed phase covariance

This inverse filtering technique was proposed in [31]. It is also based on covariance LP, as the AbS approach explained above. However, instead of removing the effect of the GSD from a long speech interval, the classical covariance LP takes only into account a portion of a single cycle where the vocal folds are closed. In this way, and in the considered time interval, there is no GSD information to be removed, and the application of covariance LP will lead to the right transfer function. Considering the linearity of the model shown in Figure 1, the closed phased interval will be the time interval where the GSD is zero. This situation is depicted in Figure 5.

The most difficult step in this technique is to detect the closed phase in the speech signal. In [10], a two-channel speech processing is proposed, making use of electroglottographic signals to detect the closed phase. Electroglottography (EGG) is a technique used to indirectly register laryngeal behavior by measuring the electrical impedance across the throat during speech. Rapid variation in the conductance is mainly caused by movement of the vocal folds. As they approximate and the physical contact between them increases, the impedance decreases, what results in a relatively higher current flow through the larynx structures. Therefore, this signal will provide information about the contact surface of the vocal cords.

The complete inverse filtering algorithm is represented in Figure 6.

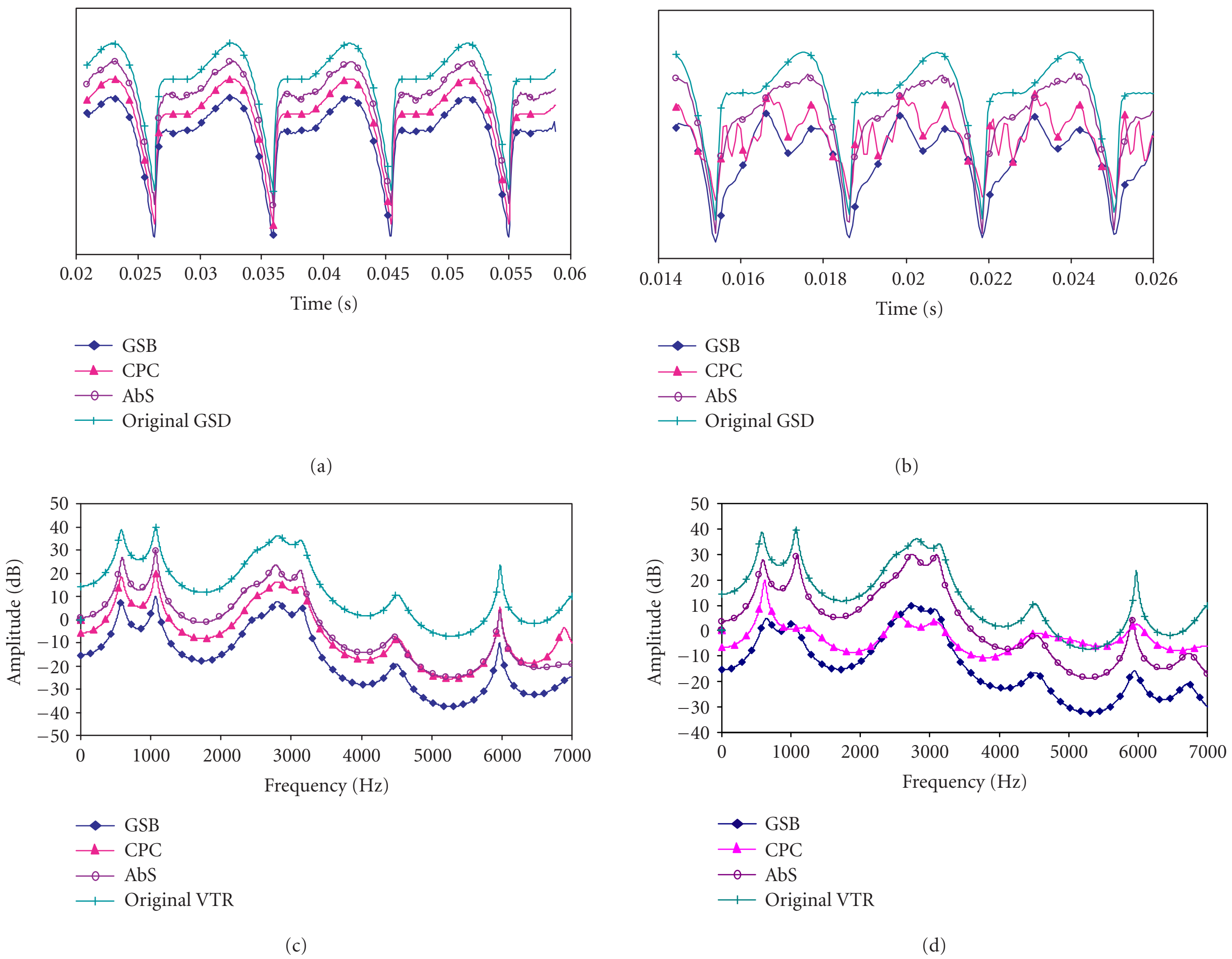


Figure 7: (a) Estimated GSD. $F_0 = 100$ Hz, vowel "a." (b) Estimated GSD. $F_0 = 300$ Hz, vowel "a." (c) Estimated VTR. $F_0 = 100$ Hz, vowel "a." (d) Estimated VTR. $F_0 = 300$ Hz, vowel "a."

In Figure 6, a GCI detection block [27] is included, because, even though both acoustic and electroglottographic signals are simultaneously recorded, there is a propagation delay between the acoustic signal recorded on the microphone and the impedance variation at the neck of the singer. Thus, a precise synchronization is mandatory.

Since this technique is based on the covariance LP, it may work with very short window lengths. However, as the fundamental frequency increases, the time length of the closed phase gets shorter, and there is much less information left for the vocal tract estimation. This fact imposes a fundamental frequency limit, even using the covariance LP.

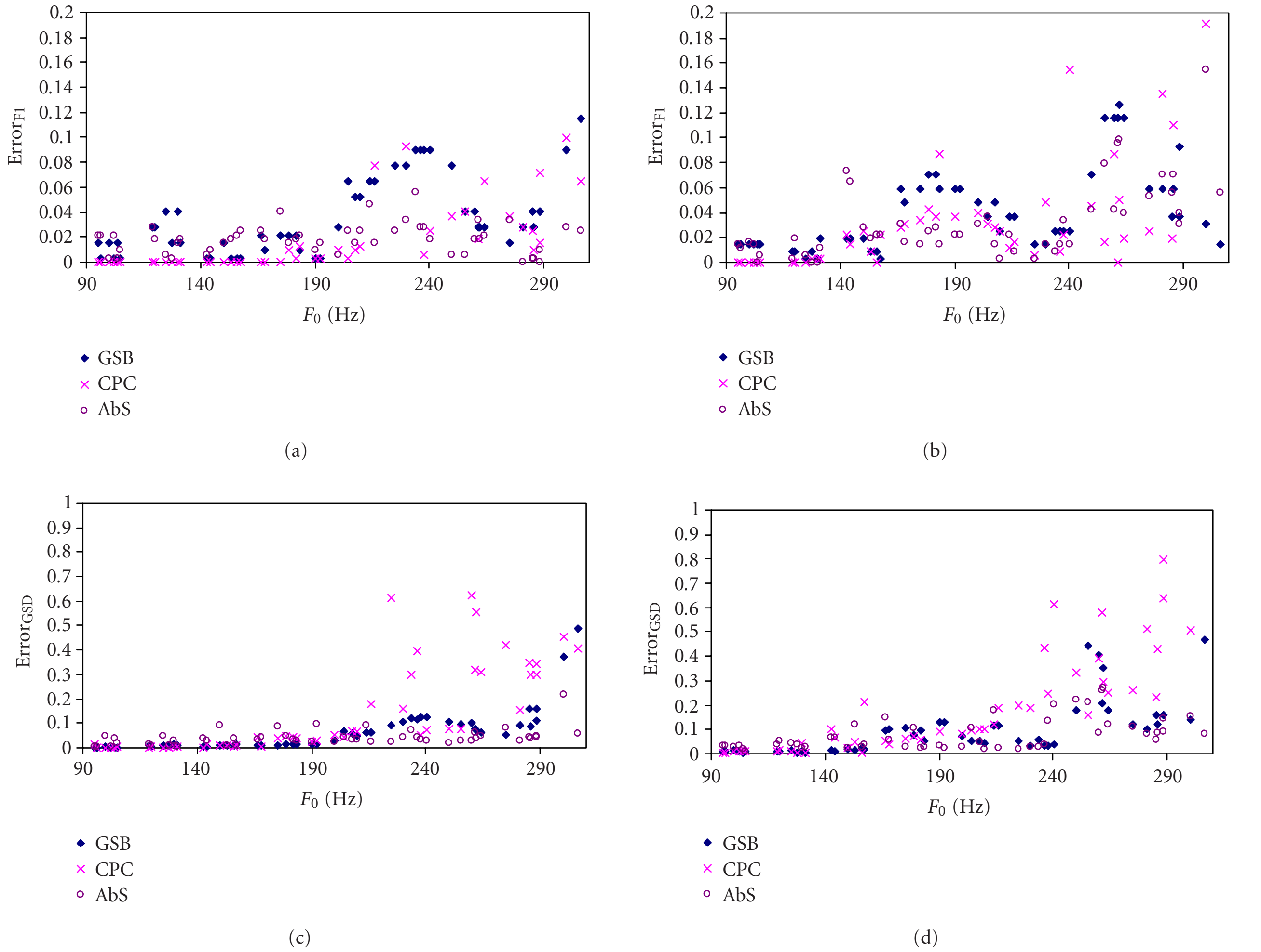
2.4. *Practical results*

Once the basics of three inverse filtering techniques have been presented and described, they will be compared by simulations and also by making use of natural singing voice records. The main goal of this analysis is to see how the three techniques are compared in terms of their fundamental frequency limitations.

2.4.1. *Simulation results*

First, the non interactive model for voice production shown in Figure 1 will be used in order to synthesize some artificial signals for test. The lip radiation effect and the glottal source are combined in a mathematical model for the GSD, also making use of the LF model. It is well known [1, 17] that the formant position can affect inverse filtering results. In [3], it is also shown that the lower first formant central frequency is, the higher is the source-tract interaction. So, the interaction is higher in vowels where the first format central frequency is lower. Therefore, and in order to cover all possible situations, two vocal all-pole filters have been used for synthesizing the test signal: one representing Spanish vowel "a," and the other one representing Spanish vowel "e." In this latter case, the first formant is located at lower frequencies.

In order to see the fundamental frequency dependence of inverse filtering techniques, this parameter has been varied from 100 Hz to 300 Hz in 25 Hz steps. For each fundamental frequency, the three algorithms have been applied and the GSD as well as the VTR have been estimated. In Figures 7a to



(a)

(b)

(c)

(d)

Figure 8: Fundamental frequency dependence. (a) Error$_{F1}$ in vowel “a.” (b) Error$_{F1}$ in vowel “e.” (c) Error$_{GSD}$ in vowel “a.” (d) Error$_{GSD}$ in vowel “e.”

7d, the glottal GSD and the VTR estimated by the three approaches are shown for two different fundamental frequencies. Note that in them, and in other figures, DC level has been arbitrarily modified to facilitate comparisons.

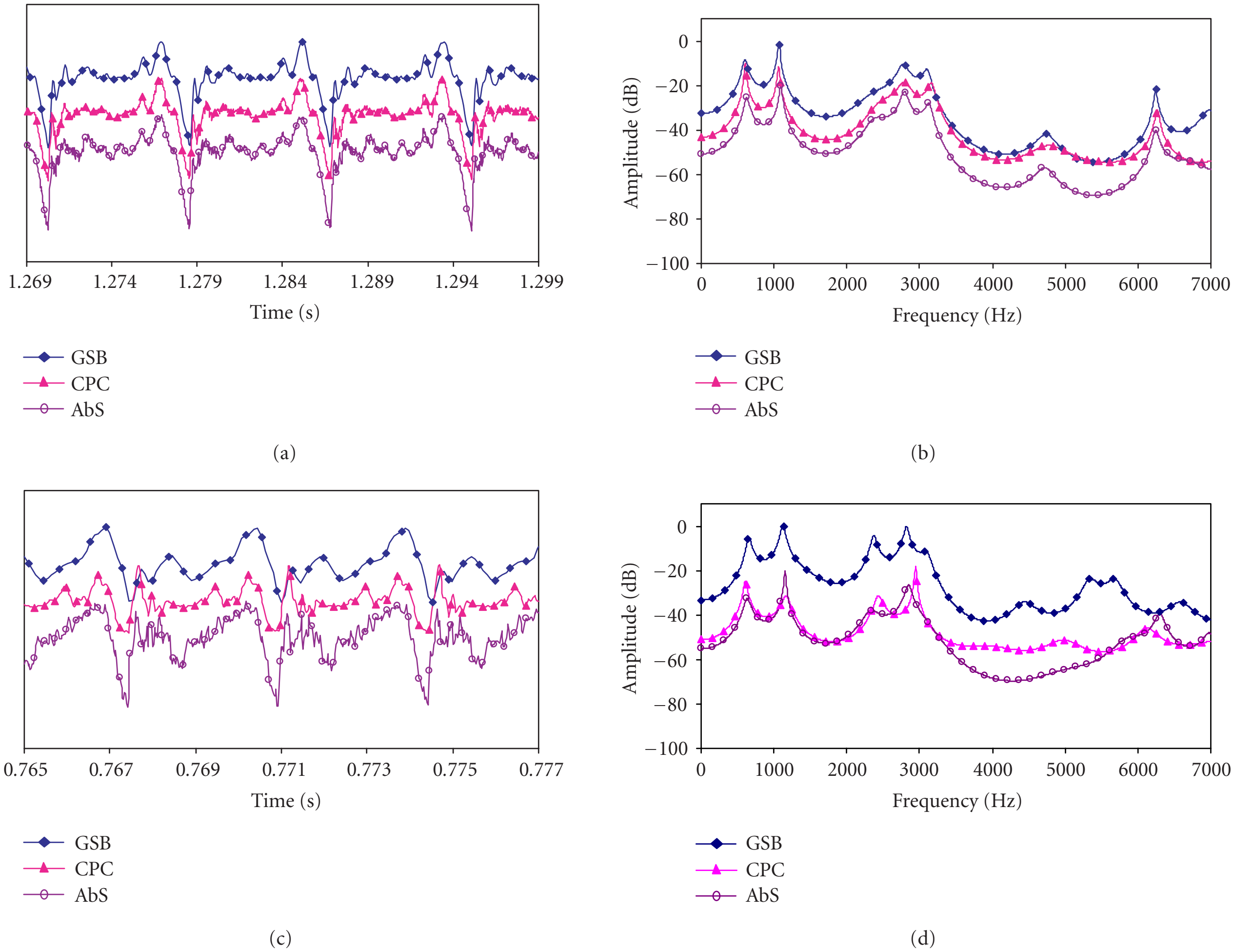
Comparing the results obtained by the three inverse filtering approaches, it is shown that as fundamental frequency increases the error in both GSD and VTR increases. Recalling the implementation of the algorithms, the CPC uses only the time interval where the GSD is zero. When the fundamental frequency is low, it is possible to see that the result of this technique is the closest one to the original one. In the case of the other two techniques, both have slight variations in the closed phase, because in both cases the glottal source effect is removed from the speech signal in an approximated manner. Otherwise, when the fundamental frequency is high, the AbS approach leads comparatively to the best result. However, it provides neither the right GSD, nor the right VTR.

In Figure 8, the relative error in the first formant central frequency and the error in the GSD are represented for the three methods, calculated according to the following expressions:

$$\text{Error}_{\text{F1}} = \frac{|F_1 - \hat{F}_1|}{F_1}, \qquad \text{Error}_{\text{GSD}} = \frac{\sum_{n=0}^{N-1} |g(n) - \hat{g}(n)|^2}{N}, \tag{4}$$

where F_1 represents the first formant central frequency and $g(n)$ and $\hat{g}(n)$ are the original and estimated GSD waveforms, respectively.

Although the simulation model does not take into account source-tract interactions, Figure 8 shows that inverse filtering results are dependent on the first formant position, being worse as it moves to lower frequencies. Also, it is possible to see that both errors increase as fundamental frequency increases. Therefore, the main conclusion of this simulation-based study is that the inverse filtering results have fundamental frequency dependence even when applied to a non interactive source-filter model.



(a)

(b)

(c)

(d)

Figure 9: (a) Estimated GSD. $F_0 = 123$ Hz, vowel "a." (b) Estimated VTR. $F_0 = 123$ Hz, vowel "a." (c) Estimated GSD. $F_0 = 295$ Hz, vowel "a." (d) Estimated VTR. $F_0 = 295$ Hz, vowel "a."

2.4.2. *Natural singing voice results*

For this analysis, three male professional singers were recorded: two tenors and one baritone. They were asked to sing notes of different fundamental frequency values, in order to register samples of all of their tessitura. Besides, different vocal tract configurations are considered, and thus, this exercise was repeated for the five Spanish vowels "a," "e," "i," "o," "u." The singing material was recorded in a professional studio, in such a way that reverberation was reduced as much as possible. Acoustic and electroglottographic signals were synchronously recorded, with a bandwidth of 20 KHz, and stored in . wav format. In order to remove low frequency ambient noise, the signals were filtered out by a high pass linear phase FIR filter whose cut-off frequency was set to a 75% of the fundamental frequency. In the case of electroglottographic signals, this filtering was also applied because of low frequency artifacts typical of this kind of signals due to larynx movements.

In Figures 9a to 9c, the results obtained for different fundamental frequencies and vowel "a," for the same singer, are shown. These results are also representative of the other singers' recordings and of the different vowels.

By comparing Figures 9a and 9c, it is possible to conclude that in the case of a low fundamental frequency, the three algorithms provide very close results. In the case of CPC, the GSD presents less formant ripple in the closed phase interval. Regarding the VTR, the central frequencies of the formants and the frequency responses are very similar. Nevertheless, in the case of a high fundamental frequency, the resulting GSD of the three analyses are very different from those of Figure 9a, and also from the waveform model provided by the LF model. Also, the calculated VTR is very different for the three methods. Thus, conclusions with natural recorded voices are similar to those obtained with synthetic signals.

3. VIBRATO IN SINGING VOICE

3.1. *Definition*

In Section 2, inverse filtering techniques, successfully employed in speech processing, have been used for singing voice processing. It has been shown that as fundamental frequency increases, they reach a limit and thus an alternative technique should be used. As we will show in this section, the introduction of vibrato in singing voice provides more information about what can be happening.

Vibrato in singing voice could be defined as a small quasiperiodic variation of the fundamental frequency of the note. As a result of this variation, all of the harmonics of the voice will also present an amplitude variation, because of the filtering effect of the VTR. Due to these nonstationary characteristics of the signal, singing voice has been modeled by the modified sinusoidal model [25, 26]:

$$s(t) = \sum_{i=0}^{N-1} a_i(t)\cos\theta_i(t) + r(t), \tag{5}$$

where

$$\theta_i(t) = 2\pi \int_{-\infty}^{t} f_i(\tau)d\tau \tag{6}$$

and $a_i(t)$ is the *instantaneous amplitude of the partial*, $f_i(t)$ the *instantaneous frequency of the partial*, and $r(t)$ the *stochastic residual*.

The acoustic signal is composed by a set of components, (partials), whose amplitude and frequency change with time, plus a stochastic residual, which is modeled by a spectral density time-varying function. Also in [25, 26], detailed information is given on how these time-varying characteristics can be measured.

Of the two features of a vibrato signal, frequency and amplitude variations, frequency is the most widely studied and characterized. In [32, 33], the instantaneous frequency is characterized and decomposed into three main components which account for three musically meaningful characteristics, respectively. Namely,

$$f(t) = i(t) + e(t)\cos\varphi(t), \tag{7}$$

where

$$\varphi(t) = 2\pi \int_{-\infty}^{t} r(\tau)d\tau \tag{8}$$

$f(t)$ being the instantaneous frequency, $i(t)$ the *intonation* of the note, which corresponds to slow variations of pitch; $e(t)$ represents the *extent* or amplitude of pitch variations, and $r(t)$ represents the *rate* or frequency of pitch variations.

All of them are time-dependent magnitudes and rely on the musical context and singer's talent and training. In the case of intonation, its value depends on the sung note, and thus, on the context. But extent and rate are mostly singer-dependent features, typical values being a 10% of the intonation value and 5 Hz, respectively.

Regarding the amplitude variation of the harmonics during vibrato, a well-established parameterization is not accepted, and probably it does not exist, because this variation is different for all of the harmonics. It is therefore not strange that amplitude variation has been the topic of interest of some few papers. The first work on this topic is [34], where the perceptual relevance on spectral envelope discrimination of the instantaneous amplitude is proven. In [22], the relevance of this feature is experimentally demonstrated in the case of synthesis of singing voice. Also, its physical cause is tackled and a representation in terms of the instantaneous amplitude versus instantaneous frequency of the harmonics is introduced for the first time. This representation is proposed as a means of obtaining a local information of the VTR in limited frequency ranges. Something similar is done in [35], where the singing voice is synthesized using this local information of the VTR. We have also contributed in this direction, for instance in [23], where the instantaneous amplitude is decomposed in two parts. The first one represents the sound intensity variation and the other one represents the amplitude variation determined by the local VTR, in an attempt to split the contribution of the source and the vocal tract. Moreover, in [24], different time-frequency processing tools have been used and compared in order to identify the relationship between instantaneous amplitude and instantaneous frequency.

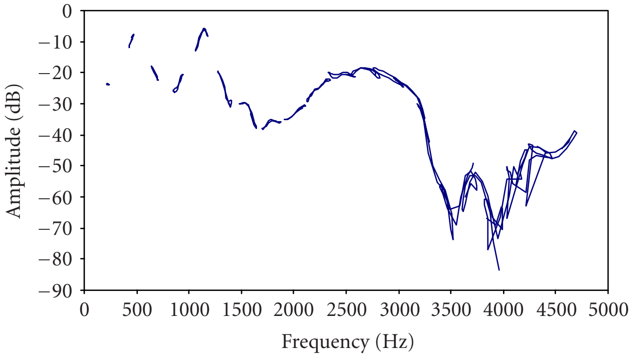


Figure 10: AM-FM representation for the first 20 harmonics. Anechoic tenor recording $F_0 = 220$ Hz, vowel "a."

In that work, the AM-FM representation is defined as the instantaneous amplitude versus instantaneous frequency representation, with time being an implicit parameter. This representation is compared to the magnitude response of an all-pole filter, which is typically used for VTR modeling. Two main conclusions are derived, the first one is that only when anechoic recordings are considered, these two representations can be compared. Otherwise, the instantaneous magnitudes will be affected by reverberation. The second one is that, as a frequency modulated input is considered, and frequency modulation is not a linear operation, the phase of the all-pole system will affect the AM-FM representation, leading to a different representation than the vocal tract magnitude response. However the relevance of this effect depends on the formant bandwidth and vibrato characteristics, vibrato rate in this case. It was also shown that in natural vibrato the phase effect of VTR is not noticeable, because vibrato rate is slow comparing to formant bandwidths.

Figure 10 constitutes a good example of the kind of AM-FM representations we are talking about. In it, each harmonic's instantaneous amplitude is represented versus its instantaneous frequency. For this case, only two vibrato cycles, where the vocal intensity does not change significantly, have been considered. As the number of harmonic increases, the frequency range swept by each harmonic widens. Comparing Figure 10 and Figure 9b, the AM-FM representation of the former one is very similar to the VTR of Figure 9b. However, in the case of the AM-FM representation, no source-filter separation has been made, and thus both elements are melted in that representation. The results obtained by other authors [22, 35] are quite similar regarding the instantaneous amplitude versus instantaneous frequency representation, however, in those works no comment is made about the conditions of recordings.

3.2. Simplified noninteractive source-tract model with vibrato

The main conclusion from the results presented above could be that vibrato might be used in order to extract more information about glottal source and VTR in singing voice. Therefore, we will propose here a simplified noninteractive source-filter model with vibrato that will be a signal model of vibrato production and will explain the results provided by sinusoidal modeling. We will first make some basic assumptions regarding what is happening with GSD and VTR during vibrato. These assumptions are based on perceptual aspects of vibrato, and on the AM-FM representation for natural singing voice.

(1) The GSD characteristics remain constant during vibrato, and only the fundamental frequency of the voice changes. This assumption is justified by the fact that perceptually there is no phonation change during a single note.

(2) The intensity of the sound is constant, at least during one or two vibrato cycles.

(3) The VTR remains invariant during vibrato. This assumption relies on the fact that vocalization does not change along the note.

(4) The three vibrato characteristics remain constant. This assumption is not strictly true, but their time constants are considerably larger than the signal fundamental period.

Taking into account these four assumptions, the simplified noninteractive source-filter model with vibrato could be represented by the block diagram in Figure 11.

Based on this model, we will simulate the production of vibrato. The GSD characteristics are the same as in Section 2.4, and the VTR has been implemented as an all-pole filter whose frequency response represents Spanish vowel "a." A frequency variation, typical of vibrato, has been applied to the GSD with a 120 Hz intonation, an extent of 10% of the intonation value, and a rate of 5,5 Hz. All of them are kept constant in the complete register.

We have applied to the resulting signal both inverse filtering (where the presence or absence of vibrato does not influence the algorithm), and sinusoidal modeling, where instantaneous amplitude and instantaneous frequency of each harmonic need to be measured. Results obtained for this simulation are shown in Figures 12, 13, 14, and 15. In Figure 12a inverse filtering results are shown for a short window analysis. When fundamental frequency is low, GSD and VTR are well separated. In Figures 12a, 13a , sinusoidal modeling results are shown. The frequency variations of the harmonics of the signal are clearly observed and, as a result, the amplitude variation. On the other hand, in Figure 14, the AM-FM representation of the partials is shown. Taking into account the AM-FM representation of every partial, and comparing this to the VTR shown in Figure 12a, it is possible to conclude that a local information of the VTR is provided by this method. However, as no source-filter decomposition has been developed, each AM-FM representation is shifted in amplitude depending on the GSD spectral features. This effect is a result of keeping GSD parameters constant during vibrato. Comparing Figures 14 and 15, it can be noticed that if the GSD magnitude spectrum is removed from the AM-FM representation of the harmonics, the resulting AM-FM representation would provide only VTR information. The result of this operation is shown in Figure 16.

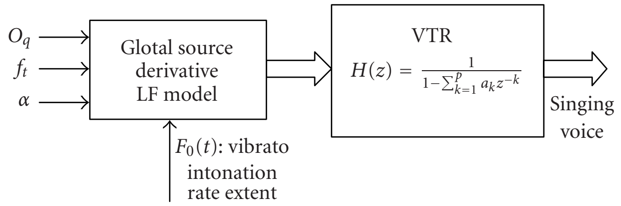

O_q
f_t
α
Glotal source derivative LF model
$F_0(t)$: vibrato intonation rate extent
VTR
$H(z) = \frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}}$
Singing voice

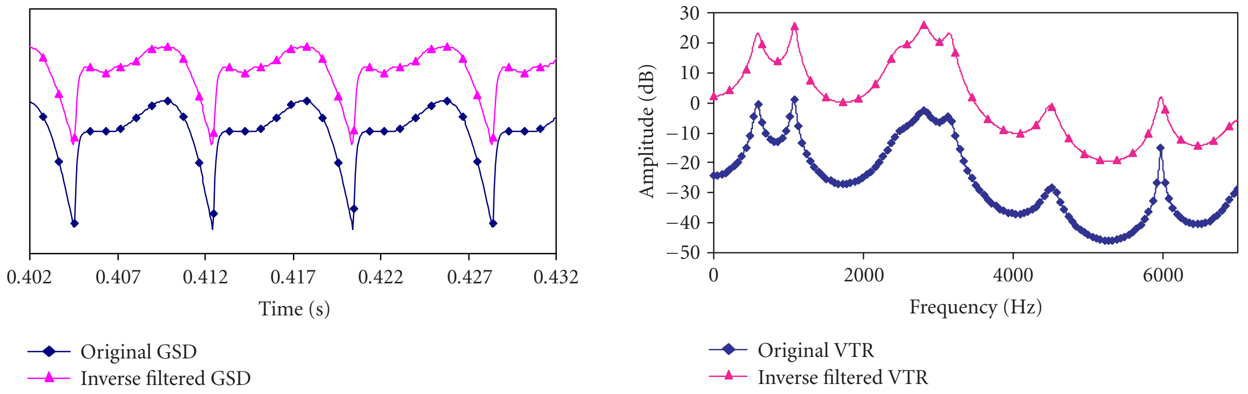
FIGURE 11: Noninteractive source-filter model with vibrato.

For this simplified noninteractive source-filter model with vibrato, instantaneous parameters of sinusoidal modeling provide a complementary information about both GSD and VTR. When inverse filtering works, the GSD effect can be removed from the AM-FM representation provided by sinusoidal modeling and only the information of the VTR remains.

3.3. Natural singing voice

The relationship between these two signal models, noninteractive source-filter model and sinusoidal model, has been established for a synthetic signal where vibrato has been included under the four assumptions stated at the beginning of the section. Now, the question is whether this relationship holds in natural singing voice too. Therefore, both kinds of signal analysis will be now applied to natural singing voice. In order to get close to simulation conditions, some precautions have been taken in the recording process.

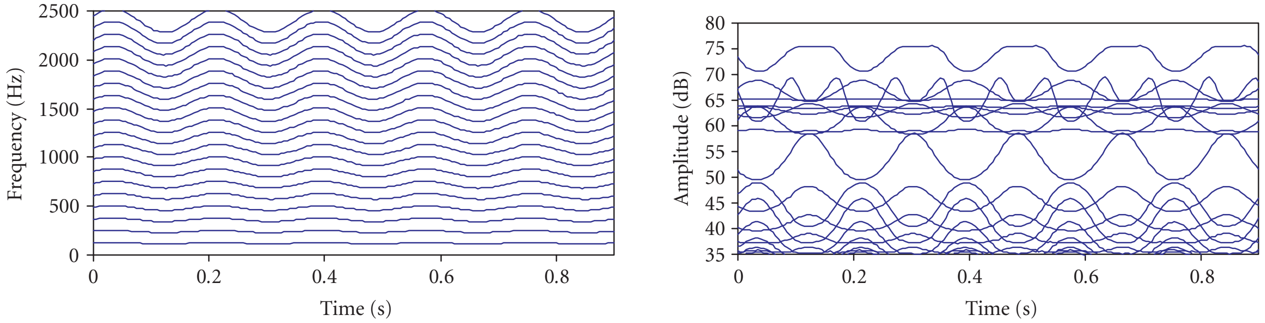
(1) The musical context has been selected in order to control intensity variations of the sound. Singers were asked to sing a word of three notes, where the first and the last one simply provide a musical support and the note in between is a long sustained note. This note is two semitones higher than the two accompanying ones.



(a)

(b)

FIGURE 12: Inverse filtering results. GSB inverse filtering algorithm. (a) GSD. (b) VTR.



(a)

(b)

FIGURE 13: Sinusoidal modeling results. (a) Instantaneous frequency. (b) Instantaneous amplitude.

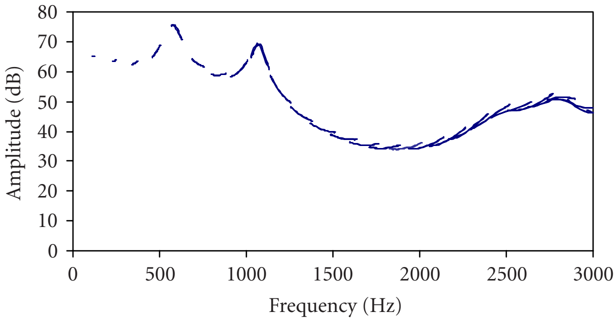


FIGURE 14: AM-FM representation.

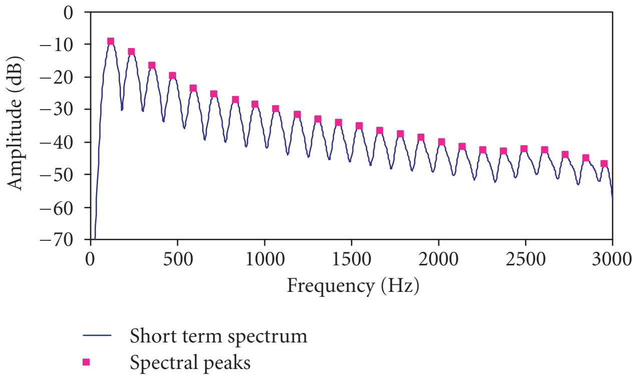


FIGURE 15: GSD short term spectrum. Blackman-Harris window.

(2) Recordings have been done in a studio where reverberations are reduced but not completely eliminated as in an anechoic room. In this situation, the AM-FM representation will present slight variations from the actual VTR, but it is still possible to develop a qualitative study.

In Figures 17, 18, 19, and 20 the results of these analyses are shown for a low-pitched baritone recording, $F_0 = 128$ Hz, *vowel "a"*. Contrarily to Figures 12, 13, 14, and 15, here there is no reference for the original GSD and VTR. Comparing Figures 12b, 13b and 17b, 18b, instantaneous frequency variation is similar in simulation and natural singing voice.

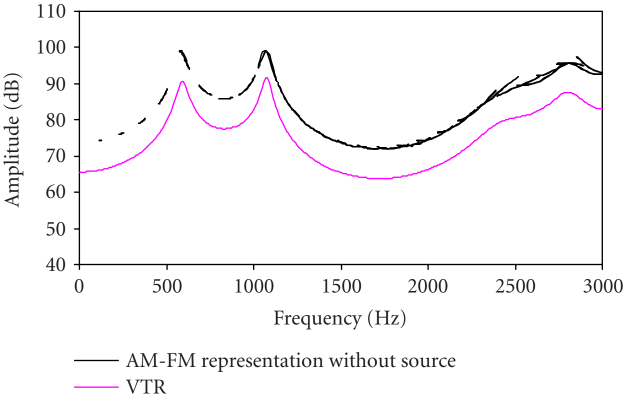


Figure 16: AM-FM representation without source.

However, the extent of vibrato in this baritone recording is lower than in synthetic signal. In the case of instantaneous amplitude, natural singing voice results are not as regular as synthetic ones. This is because of reverberation and irregularities of natural voice. Regarding intensity of the sound, there are not large variations in instantaneous amplitude, and so, for one or two vibrato cycles it could be considered constant. In this situation, the AM-FM representation of the harmonics, shown in Figure 19, is very similar to synthetic signal's AM-FM representation, though the already mentioned irregularities are present. In Figure 20, the GSD spectrum is shown for the signal of Figures 17a, 18a. It is very similar to the synthetic GSD spectrum, both are low frequency periodic signals, although it has slight variations in its harmonic amplitudes that will be explained later.

Now, the so-obtained GSD spectrum will be used to extract from the AM-FM the information of the VTR. The result of this operation is shown in Figure 21.

As in the case of synthetic signal, the compensated AM-FM representation is very close to the VTR obtained by inverse filtering. However, the matching is not as perfect as for the synthetic signal.

From this two-signal model comparison, it is possible to conclude that the simplified noninteractive source-filter model with vibrato can explain, in an approximated way, what is happening in singing voice when vibrato is present. Now, it is possible to say that GSD and VTR have not large variations during a few vibrato cycles. In this way, the instantaneous amplitude and frequency obtained by sinusoidal modeling provide more, and complementary, information about GSD and VTR during vibrato than known analysis methods.

It is important to note that the AM-FM representation by itself does not provide information of GSD and VTR separately, but it represents, in the vicinity of each harmonic, a small section of the VTR. In order to know what is exactly happening with GSD and VTR during vibrato, precautions have to be taken with recording conditions. Even in nonoptimum conditions, AM-FM representation of vibrato provides complementary information to that of inverse filtering methods.

4. DISCUSSION OF RESULTS AND CONCLUSIONS

In Section 2, inverse filtering techniques have been reviewed, and their dependence on the fundamental frequency has been shown. It seems to be obvious that, regardless of the particular technique, inverse filtering in speech fails as frequency increases. In natural singing voice, where pitch is inherently high, there are no references in order to make sure whether this is the only cause of this failure. In Section 3, and with the aim to give an answer to this question, a novel noninteractive source-filter model has been introduced for singing voice modeling, including vibrato as an additional feature. It has been shown that this model can represent the vibrato production in singing voice. In addition, this model has allowed a relationship between sinusoidal modeling and source-filter model, through which authors have coined as AM-FM representation.

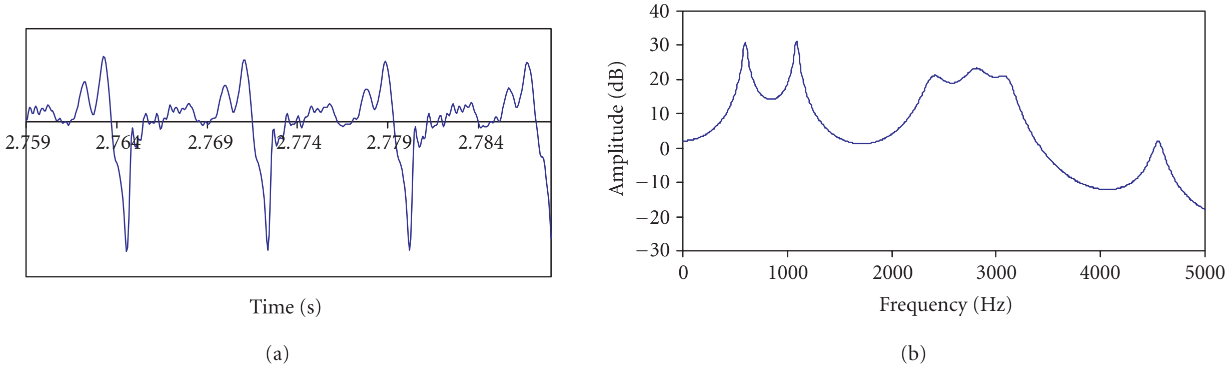
In this last section, AM-FM representation will be used again in singing voice analysis, in order to determine whether there are other effects in singing voice when fundamental frequency increases. To this end, the same analysis of Section 3 has been applied to the signal database of Section 2 corresponding to three male singers' recordings. On the one hand, inverse filtering is applied and GSD and VTR are estimated. On the other hand, sinusoidal modeling is considered and the two instantaneous magnitudes (frequency and amplitude for each harmonic) are measured. Then, the AM-FM representation is obtained for each (frequency modulated) harmonic, and the GSD is removed from this representation using the GSD obtained by the inverse filtering.

In Figure 22, the results obtained for several fundamental frequencies, for the baritone singer, are shown. As in Section 2, these results are representative of other singers' recordings and other vowels.

Regarding the AM-FM representation, it is possible to say, looking at Figure 22, that as fundamental frequency increases, the frequency range swept by one harmonic is wider, because of the extent and intonation relationship. Also, as fundamental frequency increases, the AM-FM representations of two consecutive harmonics are more separated, which is a direct consequence of their harmonic relationship. In addition to these obvious effects, there is no other evident consequence of fundamental frequency increase in this analysis, and thus the simplified noninteractive source-filter model with vibrato can model high-pitched singing voice with vibrato, from the signal point of view.

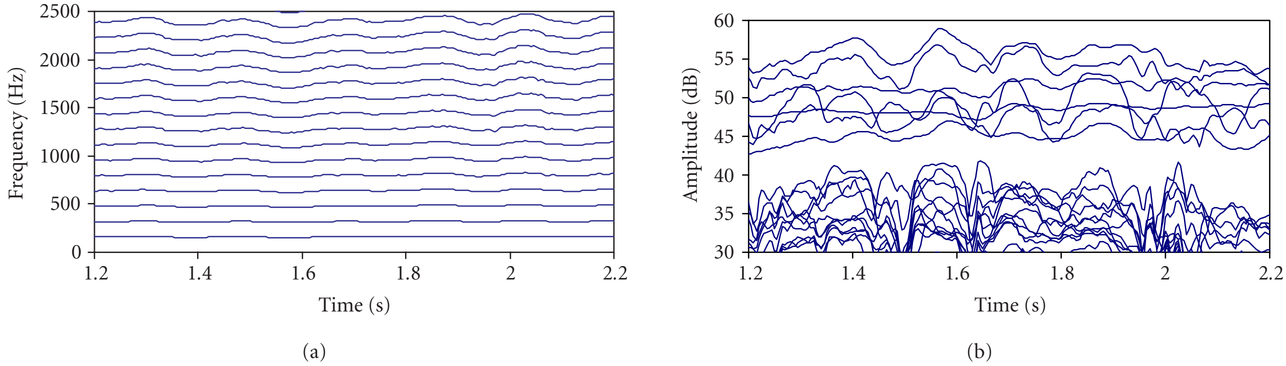
The main limitation of the plain AM-FM representation is that no source-filter separation is possible unless it is combined with other method, and thus, from here, nothing can be said about the exact shape of GSD and VTR. However, the main advantage of this representation is that it has no fundamental frequency limit, and so, it can be applied in every singing voice sample with vibrato. This conclusion brings along another evidence: the noninteractive source-filter model remains valid in singing voice.

We can summarize the main contributions and conclusions of this work as follows.



(a) (b)

Figure 17: Inverse filtering results. GSB inverse filtering algorithm. (a) GSD (b) VTR.



(a) (b)

Figure 18: Sinusoidal modeling results. (a) Instantaneous frequency. (b) Instantaneous amplitude.

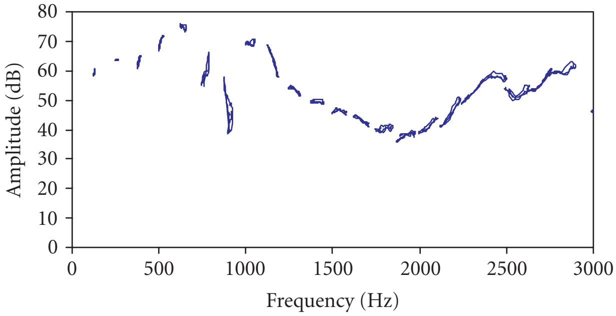


Figure 19: AM-FM representation.

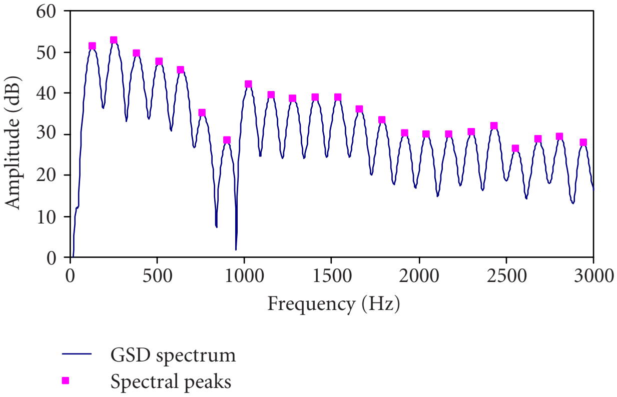


Figure 20: GSD Short term spectrum. Blackman-Harris window.

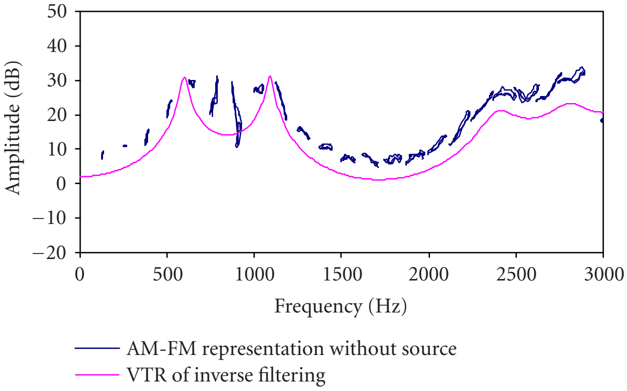
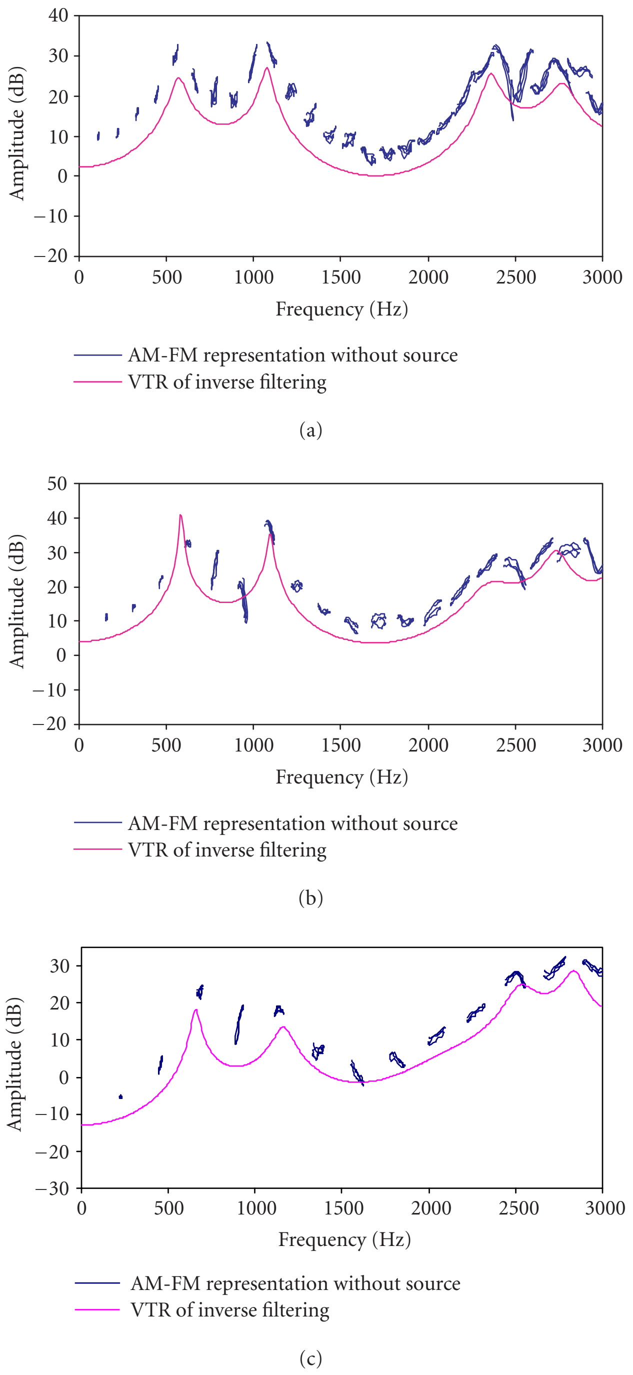


Figure 21: AM-FM representation without source.

(i) Several representative inverse filtering techniques have been critically compared when applied to speech. It has been shown how all of them fail as frequency increases, as it is the case in singing voice.

(ii) A novel noninteractive source-filter model has been proposed for singing voice, which includes vibrato as a possible feature.

(iii) The existence of vibrato and the above mentioned model has allowed to relate source-filter model (i.e., inverse filtering techniques) and the simple sinusoidal



(a)

(b)

(c)

Figure 22: AM-FM representation removing the source and VTR given by inverse filtering. (a) $F_0 = 110$ Hz, vowel "a," (b) $F_0 = 156$ Hz, vowel "a," (c) $F_0 = 227$ Hz, vowel "a."

Model. In other words, although both are signal models for singing voice, the first one is related to the voice production and the second one is a general signal model, but thanks to vibrato both can be linked.

(iv) Even though sinusoidal modeling does not allow to obtain separate information about the sound source and VTR, the AM-FM representation gives complementary information particularly in high frequency ranges, where inverse filtering does not work.

ACKNOWLEDGMENTS

The Gobierno de Navarra and the Universidad Pública de Navarra are gratefully acknowledged for financial support. Authors would also like to acknowledge the support from Xavier Rodet and Axel Roebel (IRCAM, Paris), material and medical support from Ana Martínez Arellano, and the collaboration from student Daniel Erro who implemented some of the algorithms.

REFERENCES

[1] N. Henrich, *Etude de la source glottique en voix parlée et chantée : modélisation et estimation, mesures acoustiques et électroglottographiques, perception*, Ph.D. thesis, Paris 6 University, Paris, France, 2001.

[2] B. H. Story, "An overview of the physiology, physics and modeling of the sound source for vowels," *Acoustical Science and Technology*, vol. 23, no. 4, pp. 195–206, 2002.

[3] B. Guerin, M. Mrayati, and R. Carre, "A voice source taking account of coupling with the supraglottal cavities," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '76)*, vol. 1, pp. 47–50, Philadelphia, Pa, USA, April 1976.

[4] T. V. Ananthapadmanabha and G. Fant, "Calculation of the true glottal flow and its components," *Speech Communication*, vol. 1, no. 3-4, pp. 167–184, 1982.

[5] M. Berouti, D. G. Childers, and A. Paige, "Glottal area versus glottal volume-velocity," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '77)*, vol. 2, pp. 33–36, Cambridge, Mass, USA, May 1977.

[6] G. Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, The Netherlands, 1960.

[7] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[8] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *Speech Transmission Laboratory-Quarterly Progress and Status Report*, vol. 85, no. 2, pp. 1–13, 1985.

[9] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '86)*, vol. 11, pp. 1605–1608, Tokyo, Japan, April 1986.

[10] A. K. Krishnamurthy and D. G. Childers, "Two-channel speech analysis," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 730–743, 1986.

[11] P. Alku and E. Vilkman, "Estimation of the glottal pulseform based on discrete all-pole modeling," in *Proc. 2nd International Conf. on Spoken Language Processing (ICSLP '94)*, pp. 1619–1622, Yokohama, Japan, September 1994.

[12] H.-L. Lu and J. O. Smith, "Joint estimation of vocal tract filter and glottal source waveform via convex optimization," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '99)*, pp. 79–92, New Paltz, NY, USA, October 1999.

[13] I. Arroabarren and A. Carlosena, "Glottal spectrum based inverse filtering," in *Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, Geneva, Switzerland, September 2003.

[14] E. L. Riegelsberger and A. K. Krishnamurthy, "Glottal source estimation: methods of applying the LF-model to inverse filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '93)*, vol. 2, pp. 542–545, Minneapolis, Minn, USA, April 1993.

[15] B. Doval, C. d'Alessandro, and B. Diard, "Spectral methods for voice source parameters estimation," in *Proc. 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*, vol. 1, pp. 533–536, Rhodes, Greece, September 1997.

[16] I. Arroabarren and A. Carlosena, "Glottal source parameterization: a comparative study," in *Proc. ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, Geneva, Switzerland, August 2003.

[17] H.-L. Lu, *Toward a high-quality singing synthesizer with vocal texture control*, Ph.D. thesis, Stanford University, Stanford, Calif, USA, 2002.

[18] N. Henrich, B. Doval, and C. d'Alessandro, "Glottal open quotient estimation using linear prediction," in *Proc. International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, Firenze, Italy, September 1999.

[19] N. Henrich, B. Doval, C. d'Alessandro, and M. Castellengo, "Open quotient measurements on EGG, speech and singing signals," in *Proc. 4th International Workshop on Advances in Quantitative Laryngoscopy, Voice and Speech Research*, Jena, Germany, April 2000.

[20] N. Henrich, C. d'Alessandro, and B. Doval, "Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data," in *Proc. 7th European Conference on Speech Communication and Technology (EUROSPEECH '01)*, Aalborg, Denmark, September 2001.

[21] H.-L. Lu and J. O. Smith, "Glottal source modeling for singing voice synthesis," in *Proc. International Computer Music Conference (ICMC '00)*, Berlin, Germany, August 2000.

[22] R. Maher and J. Beauchamp, "An investigation of vocal vibrato for synthesis," *Applied Acoustics*, vol. 30, no. 2-3, pp. 219–245, 1990.

[23] I. Arroabarren, M. Zivanovic, and A. Carlosena, "Analysis and synthesis of vibrato in lyric singers," in *Proc. 11th European Signal Processing Conference (EUSIPCO '02)*, Toulose, France, September 2002.

[24] I. Arroabarren, M. Zivanovic, X. Rodet, and A. Carlosena, "Instantaneous frequency and amplitude of vibrato in singing voice," in *Proc. IEEE 28th Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '03)*, Hong Kong, China, April 2003.

[25] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[26] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, C. Roads, S. Pope, A. Picialli, and G. De Poli, Eds., Swets & Zeitlinger, Lisse, The Netherlands, May 1997.

[27] C. Ma, Y. Kamp, and L. F. Willems, "A Frobenius norm approach to glottal closure detection from the speech signal," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 258–265, 1994.

[28] J. Makhoul, "Linear prediction: a tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[29] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Processing*, vol. 39, no. 2, pp. 411–423, 1991.

[30] B. Doval and C. d'Alessandro, "Spectral correlates of glottal waveform models: an analytic study," in *Proc. IEEE 22th Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '97)*, pp. 1295–1298, Munich, Germany, April 1997.

[31] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[32] E. Prame, "Vibrato extent and intonation in professional western lyric singing," *Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 616–621, 1997.

[33] I. Arroabarren, M. Zivanovic, J. Bretos, A. Ezcurra, and A. Carlosena, "Measurement of vibrato in lyric singers," *IEEE Trans. Instrumentation and Measurement*, vol. 51, no. 4, pp. 660–665, 2002.

[34] S. McAdams and X. Rodet, "The role of FM-induced AM in dynamic spectral profile analysis," in *Basic Issues in Hearing*, H. Duifhuis, J. Horst, and H. Wit, Eds., pp. 359–369, Academic Press, London, UK, 1988.

[35] M. Mellody and G. H. Wakefield, "Signal analysis of the singing voice:low-order representations of singer identity," in *Proc. International Computer Music Conference (ICMC '00)*, Berlin, Germany, August 2000.

Ixone Arroabarren was born in Arizkun, Navarra, Spain, on December 11, 1975. She received her Eng. degree in telecommunications in 1999, from the Public University of Navarra, Pamplona, Spain, where she is currently pursuing her Ph.D. degree in the area of signal processing techniques as they apply to musical signals. She has collaborated in industrial projects for the vending machine industry.

Alfonso Carlosena was born in Navarra, Spain, in 1962. He received his M.S. degree with honors and his Ph.D. in physics in 1985 and 1989, respectively, both from the University of Zaragoza, Spain. From 1986 to 1992 he was an Assistant Professor in the Department of Electrical Engineering and Computer Science at the University of Zaragoza, Spain. Since October 1992, he has been an Associate Professor with the Public University of Navarra, where he has also served as Head of the Technology Transfer Office. In March 2000, he was promoted to Full Professor at the same University. He has also been a Visiting Scholar in the Swiss Federal Institute of Technology, Zurich and New Mexico State University, Las Cruces. His current research interests are in the areas of analog circuits and signal processing, digital signal processing and instrumentation, where he has published over sixty papers in international journals and a similar number of conference presentations. He is currently leading several industrial projects for local firms.