.

Effect of the glottal source and the vocal tract on the partials amplitude of vibrato in male voices

Ixone Arroabarren, Alfonso Carlosena, and NHF

Citation: J. Acoust. Soc. Am. **119**, 2483 (2006); doi: 10.1121/1.2177584

View online: http://dx.doi.org/10.1121/1.2177584

View Table of Contents: http://asa.scitation.org/toc/jas/119/4

Published by the Acoustical Society of America

Effect of the glottal source and the vocal tract on the partials amplitude of vibrato in male voices

Ixone Arroabarren[a] and Alfonso Carlosena[b]
*Department of Electrical and Electronic Engineering, Universidad Pública de Navarra,
E-31006 Pamplona Spain*

(Received 7 October 2004; revised 22 January 2006; accepted 23 January 2006)

In this paper the production of vocal vibrato is investigated. The most relevant features of the acoustical vibrato signal, frequency and amplitude variations of the partials, will be related to the voice production features, glottal source (GS) and vocal tract response (VTR). Unlike previous related works, in this approach, the effect on the amplitude variations of the partials of each one of the above-mentioned voice production features will be identified in recordings of natural singing voice. Moreover, we will take special care of the reliability of the measurements, and, to this aim, a noninteractive vibrato production model will be also proposed in order to describe the vibrato production process and, more importantly, validate the measurements carried out in natural vibrato. Based on this study, it will be shown that during a few vibrato cycles, the glottal pulse characteristics, as well as the VTR, do not significantly change, and only the fundamental frequency of the GS varies. As a result, the pitch variations can be attributed to the GS, and these variations, along with the vocal tract filtering effect, will result in frequency and amplitude variations of the acoustic signal partials. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2177584]

PACS number(s): 43.75.Rs, 43.75.Bc, 43.75.−z, 43.70.−h [NHF] Pages: 2483–2497

I. INTRODUCTION

Singing voice constitutes an interesting challenge to the study of voice quality because of its differences from every day speech. Among these, the most interesting one is perhaps the vocal vibrato. This is a specific musical feature not present in speech and has been by itself the topic of interest of many researchers in distinct areas such as physiology and musicology.

It is quite easy to describe vocal vibrato from the acoustical point of view. Borrowing Sundberg's definition:[1] "vibrato is a regular fluctuation in pitch, timbre and/or loudness;" however, some of its basic aspects remain hidden still today.

From this definition, it is clear that vibrato is a regular pitch variation and, in fact, this is the most widely studied aspect from the pioneering work of Seashore[2] to our day. Several features related to the pitch variation, or, in parallel, fundamental frequency variation, have been analyzed, and the most relevant works are those dealing with the parametrization of this variation.[3–5] This is usually characterized by three parameters: the intonation, which represents slow variations around the average frequency value of the note; the vibrato extent, which represents the amplitude of fundamental frequency variations; and the vibrato rate, which represents the frequency of the fundamental frequency variations. Making use of these parameters, differences among singers have been measured and studied,[6,7] and even links between vibrato rate and physiological features have been explored based on a reflex resonance model.[8] In particular, this latter model is proposed to describe the relationship between the muscular activation and the typical values of vibrato rate.

Going back to Sundberg's definition, it is obvious that it does not specify what happens with timbre and loudness during vibrato. Both concepts are not unambiguously defined since they depend on the acoustic signal as well as the listener. Focusing on the acoustical signal, it can be seen as a particular combination of the amplitude of the different partials (a chord of sine tones of different frequencies) composing the voiced sound. The position of these partials is harmonically related to the fundamental frequency, and in the case of vibrato, it is expected that harmonicity is preserved. On the other hand, the amplitude of the partials also shows temporal variations during vibrato, but its origin and characterization still remains unclear, as will be shown later in more detail. These variations also reflect in the timbre and loudness temporal evolution, and, since the amplitude variations are not well understood, neither are the timbre and loudness variation.

Regarding the amplitude variation of the partials, several works have dealt with this topic during the last decades, but apart from minor differences among them, almost all simply measure these variations and argue about their perceptual relevance. In particular, the first step on this direction was taken by McAdams and Rodet,[9] who demonstrated that the amplitude behavior of a partial was coupled with its frequency behavior according to a given spectral envelope, which could be used by the auditory system to discriminate the timbre of a complex sound. Later, Horii[10] proposed a resonance-harmonic interaction to explain the phase relationship between fundamental frequency and amplitude modulation of vibrato. A parallel work was carried out by Maher and Beauchamp,[11] who concluded that the amplitude fluctuation

[a] Electronic mail: ixone.arroabarren@unavarra.es
[b] Electronic mail: carlosen@unavarra.es

of a partial during vibrato varies in form, amplitude, and phase according to the position of the partial within the vocal tract resonances. However, in neither case was it demonstrated how the vocal tract, as well as the glottal source (GS), behaves such that those variations are generated. In the same way, Imazumi *et al.*[12] investigated the sources of vibrato by considering the correlation between the amplitude and frequency variations of the partials, and argued about the interrelationships between the fundamental frequency, the level of the harmonics, the formant frequencies, and the overall amplitude of the singing voice. The main limitation of their approach is that they did not identify the individual contribution of the voice production elements, GS and vocal tract response (VTR), on the amplitude and frequency of the partials, making it difficult to demonstrate the precise behavior of these elements in vocal vibrato production.

The correlation between the amplitude and frequency variations of the partials associated with vocal vibrato has been subsequently analyzed by other authors,[13,14] and the main conclusion in all cases is that, while this correlation is determined by the vibrato production mechanisms, the exact role of each voice production element cannot be identified based only on the direct observation of the amplitude versus frequency representation, since this represents a global information where the individual contributions of the voice production elements are combined.

By analyzing the different efforts devoted to the study of amplitude variation of the partials, it can be seen that all the works have in common the use of the sinusoidal model[15,16] to represent the acoustic signal, since the parameters of this model are the amplitude and frequency variations of the partials. However, in the mentioned works no voice production description is considered, which precludes any statement concerning the vibrato production mechanisms.

Additionally, regarding the voice production process, different mathematical models have been proposed during the last decades in order to describe it. For instance, physical models are closer to the voice production system,[17,18] since the GS and VTR are obtained based on some physical features of the voice organ elements. Additionally, the source-filter model proposed by Fant[19] (also known as the noninteractive source-filter model) can be qualified as a signal voice production model, since it does not contain any physical parameters, but it directly depends on a waveform representing the GS and a transfer function corresponding to the VTR. Each model represents a tradeoff in the voice production descriptions, for instance physical models provide a more realistic description, but is not easy to extract the values of the model parameters from a given acoustic signal. On the contrary the source-filter model constitutes a more simplified description, but it makes it easier to extract the GS and VTR for a given recorded sound.

With this in mind, in this paper the vibrato production will be investigated, but, unlike earlier works, the individual contribution of the GS and VTR on the amplitude variation of the partials will be identified. For this purpose, the acoustical features of the vibrato signal, amplitude, and frequency variations, as well as the voice production process, will be jointly considered by combining two different signal descriptions: the sinusoidal model and the noninteractive source-filter model. At this point, it is important to emphasize how different the two approaches are, which makes it very important to assess the reliability of the measurements that will help to combine such different approaches.

For this purpose, we will propose a vibrato production signal model, the noninteractive vibrato production model, that will allow us to validate the measurements made in natural singing voice recordings and will describe, and thus help to understand, how the voice production elements, GS and VTR, behave such that the specific correlation between the amplitude and frequency of the partials is produced. Finally, a model evaluation process will allow us to examine if the behavior of the proposed model fits well with natural vocal vibrato production. It is important to note that the voice material considered for this study corresponds to male recordings and, thus, the conclusions derived from it are of application strictly to this kind of voice.

The organization of the work is as follows. In Sec. II the acoustical properties of vibrato are described, where the sinusoidal model is used to describe the acoustic signal, and, based on its parameters, the amplitude variations of the partials, as well as the intensity of the sound, is analyzed. In Sec. III the voice production process is briefly described by the noninteractive source-filter model, defining its main features (GS and VTR) and briefly reviewing the analysis techniques associated to this particular model.

In Sec. IV the vibrato production is examined, and its noninteractive vibrato production model will be proposed. Finally, in Sec. V, this model is evaluated, and the conclusions that follow are discussed in Sec. VI.

II. ACOUSTICAL CHARACTERIZATION OF VIBRATO

Before starting with the acoustical analysis, there are some details that must be clarified: First, it has to be recalled that vibrato is a time-dependent musical effect and, thus, its features can be affected by its musical context (pitch transition,[20] crescendo, decrescendo,[21] etc.). For this particular study, a simple musical context has been selected in such a way that other musical effects are avoided or at least minimized. Thus, the specifications of the voice material used for this study will be briefly described.

Second, as already mentioned, the sinusoidal model has been described as the most suitable signal model for vocal vibrato description. However, there are different possibilities for the calculation of its parameters from a given recorded signal,[11,13,14] which might affect the final result. Therefore, it is necessary to provide a brief description of the particular analysis procedure adopted for this study.

Third, from Sundberg's definition of vocal vibrato mentioned above, the main conclusion is that the less-understood acoustical aspects are the timbre and loudness. In particular, regarding the pitch variations characterization, the authors refer the readers to a recent work of Arroabarren *et al.* concerning pitch variations.[5] On the other hand, timbre and loudness are also dependent on the perceptual system of the listener,[22] thus it is important to note that in this particular study we will only focus on the objective acoustic signal features, frequency and amplitude variations, as well as the intensity variations of the sound.

A. Voice material

During a short musical excerpt (a few seconds) of natural singing voice performance many acoustical changes may take place, such as, for instance, pitch and intensity variations. Therefore, for a sensible acoustical characterization of vocal vibrato a very simple musical context has been selected, in order to make easier the isolation of vibrato out of other musical effects, such as those mentioned.

For this study three semi-professional male singers were enlisted, two tenors and one baritone. Each singer was asked to sing an exercise of three notes, the second one corresponding to a long sustained note separated three semitones from the others (Do-Mi-Do). The first and third notes behave as a musical support for the second one, so that the singer can sing a long sustained vowel, and the sung vowel of these notes was the vowel [i]. This exercise was repeated by each singer, increasing the frequency of each note in one semitone, so that their whole frequency range was covered. Regarding the sung vowel of the central note, the whole exercise was repeated for the five Spanish vowels, [a], [e], [i], [o], [u], in order to have available several vocal tract configurations.

Recordings were made in a professional recording studio, in such a way that reverberations were reduced considerably though not completely eliminated as in an anechoic chamber.[14] The signals were sampled at a standard rate of 44.1 kHz, with 16-bit linear encoding on a single channel, and the recording levels were adjusted on the DAT machine during the recording process to achieve a (roughly) constant output level. These changes were necessary to account for the large dynamic range produced by the singer across his tessitura. In this way, the full 16 bits of resolution were used for each note produced. This means that the relative power of the notes sung is not preserved in our recordings; however, we deemed it more important to avoid quantization error than to preserve relative power.

B. Sinusoidal modeling

Concerning the acoustic signal analysis, the additive analysis-synthesis approach has been selected. It was developed during the 1980's by J. O. Smith III[23] in the context of computer music, and, in parallel but independently, McAulay and Quatieri proposed a similar approach for speech applications.[15] Later, a more complete sinusoidal model was proposed by Serra and Smith,[16] avoiding the problems of the preliminary versions in nonharmonic sounds. In particular, in the preliminary versions of the model, the whole sound was modeled by a set of sinusoids. However, in the case of nonharmonic sounds the number of sinusoids required to represent the sound was too high. Therefore, in the sinusoidal model proposed by Serra and Smith, an additional term was included to represent the nonharmonic part of the sound, and only the harmonic part of the sound was represented by a set of sinusoids.

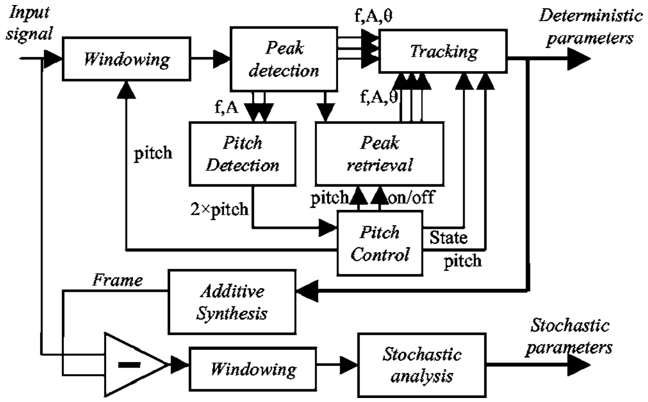

FIG. 1. Block diagram of the analysis part of the additive analysis-synthesis approach.

The continuous time sinusoidal model, which this analysis-synthesis scheme is based on, assumes the following expression for the acoustic signal $s(t)$:

$$s(t) = \sum_{i=1}^{M} a_i(t)\cos\,\theta_i(t) + e(t), \tag{1}$$

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

$$e(t) = \int_{0}^{t} h(t,\tau)u(\tau)d\tau, \tag{3}$$

where $a_i(t)$ and $f_i(t)$ are the instantaneous amplitude (IA) and instantaneous frequency (IF) of partial i, respectively, which characterize the deterministic component of the signal, and $e(t)$ is the stochastic one. The stochastic component $e(t)$ is usually described as a filtered white noise $u(\tau)$ by a time-varying filter characterized by its impulse response $h(t,\tau)$. In natural sound analyses it is seldom clear which spectral components are associated with each part, hence, the estimation algorithm has to be flexible enough.

To carry out our particular analysis we have implemented a tool based on the sinusoidal model that is described by the block diagram of Fig. 1.

According to the block diagram of Fig. 1, the input signal is analyzed frame by frame, with a typical frame duration of 6 ms. Then, in the first step a short time window length of the signal (three or four fundamental periods) is selected. The short-term spectrum of the signal is calculated and the spectral peaks detected, estimating their frequencies, amplitudes, and phases. This information is conveyed to a pitch detection block, where the fundamental frequency (F_0) of the frame is estimated. As shown, this is a pitch synchronous analysis, because the pitch information is used to select the most appropriate window length. Once the spectral peaks are detected, and the pitch estimated, a peak retrieval routine is applied, in order to recover nondetected peaks. Considering the spectral peaks of the frame, they are linked to earlier frame analyses in the peak tracking step, generating the trajectories of each partial, and obtaining the sinusoidal parameters of (1), namely, $a_i(t)$ and $f_i(t)$, the IA and IF of the signal

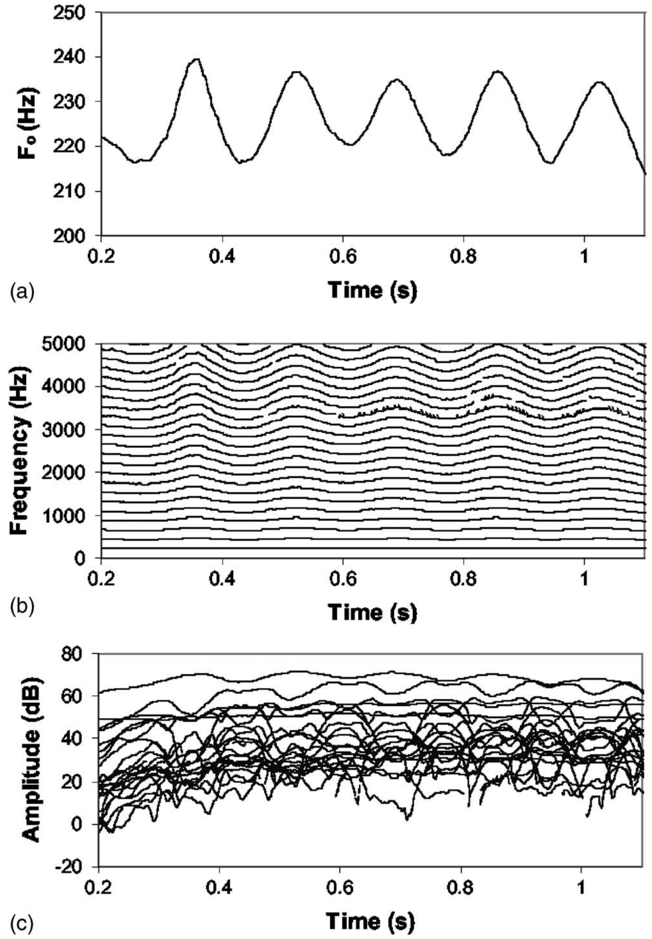

FIG. 2. Sinusoidal model features. The signal corresponds to a tenor recording, singing the vowel [a] with a F_0 of 220 Hz. The selected analysis frame size is 6 ms. (a) F_0 (b) IF (c) IA.

harmonics, respectively. Finally, the stochastic part is estimated by removing the deterministic part from the original signal. For more details of this analysis procedure readers are referred to the works of McAulay and Quatieri[15] or Serra and Smith.[16]

Going back to the recorded samples, the vocal vibrato recordings correspond to clean solo recordings and only voiced sounds, since there is no consonant in the musical exercise. Moreover, they correspond to the normal phonation, where the aspiration noise is minimum,[24] so that the whole signal can be modeled deterministically. In Fig. 2 results obtained for a representative tenor recording are shown.

From Fig. 2 it is apparent that all the signal partials are harmonically related to the fundamental frequency, and all of them follow the vibrato pattern, which is more evident as the harmonic order increases. Regarding the amplitude of the partials, in some cases it is possible to see a semi-periodic pattern but, in contrast to the frequency variation of the partials, no correlation among their IA is obvious.

C. Frequency and amplitude of the partials

In this subsection, the frequency and amplitude variations of the partials associated with vibrato will be analyzed, as well as the intensity of the sound, and some conclusions about their behavior during vibrato will be drawn.

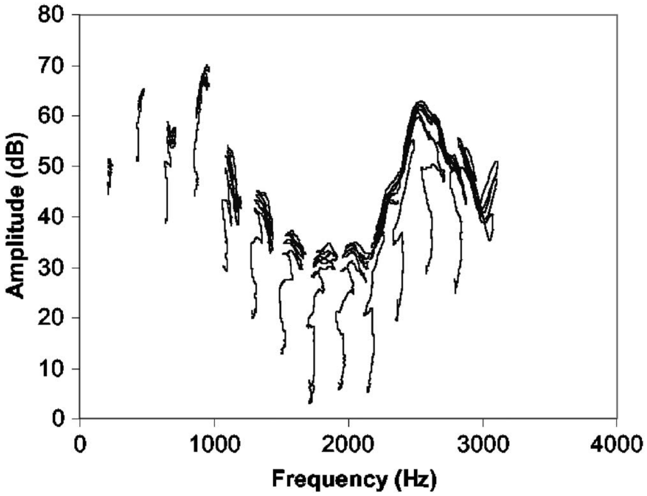

FIG. 3. AM-FM representation. The signal corresponds to a tenor recording, singing the vowel [a] with a F_0 of 220 Hz.

In Fig. 2 the frequency and amplitude variations of a set of partials of a tenor recording are shown. They are representative of a typical behavior during vocal vibrato. From Fig. 2(c) it is clear that, unlike the frequency variation of the partials, the amplitude variation does not follow a regular pattern, as all the harmonics show a different amplitude variation; however, in most of the partials it is still possible to see a semi-periodic pattern in their amplitude variation synchronous with the frequency variation.

In order to show the implicit correlation between the amplitude and frequency variations, the amplitude variation of the partials shown in Fig. 2(c) are now represented in Fig. 3 versus their frequency variation appearing in Fig. 2(b). This representation has been used by several authors, as already explained, and has received different names, *average frequency characteristics*,[12] *composite transfer function*,[13] or *AM-FM representation*.[14] All these names imply that this information reflects the dependence on both vocal tract and the spectral variation in the glottal source.

However, by observing Fig. 3 it is not possible to see how the GS and the VTR behave during vibrato, since the effect of the two elements is combined. However, the AM-FM representation shows some common features concerning the IA of the partials, which were mathematically described recently.[25] The IA of a partial, $A_i(n)$, was modeled as the product of three elements:

$$A_i(n) = \alpha_i \beta_i(n) \gamma_i(n) \tag{4}$$

where α_i is the relative weight of the harmonic compared to the fundamental, $\beta_i(n)$ is a time-varying function representing the intensity change of the partial, which models the amplitude variation independently of the spectral envelope tracing, and $\gamma_i(n)$ is a time-varying function representing those variations related to the spectral envelope traced by the AM-FM representation. These three elements can be identified by observing the AM-FM representation of Fig. 3. From Fig. 3 it is apparent that when the IF of the partials shows its periodic pattern, their IA traces a local spectral envelope, which is related to the $\gamma_i(n)$ parameter of the IA, representing the local spectral envelope traced by each partial.

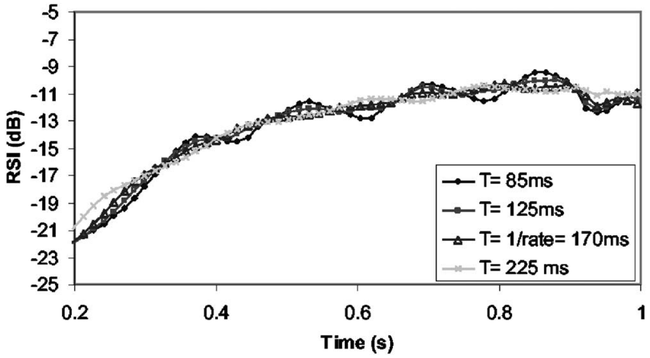

FIG. 4. Relative sound intensity. The signal corresponds to a tenor recording, singing the vowel [a] with a F_0 of 220 Hz.

In spite of this mathematical description, the main conclusion about the amplitude variations during vibrato is that a direct relationship with the voice production features cannot be extracted by a simple observation of the AM-FM representation, since the amplitude of the partials will be affected by all the voice production elements. However, the AM-FM representation gives at least a clue that the amplitude behavior of the partials is not arbitrary, but rather involves the specifics of vocal vibrato production.

On the other hand, regarding the intensity variations, they can be easily characterized based on the former analysis. The intensity of the sound is defined as the sound energy transmitted per unit time through a unit area, and it is proportional to the square of the sound pressure.

We have already mentioned that, during the recording process, no calibration was made for determining the exact value of vocal intensity, because the main concern of the measurement was to avoid the quantization error. Therefore, in order to quantify the sound intensity, the *relative sound intensity* (RSI) will be calculated from the recorded signal samples as

$$\mathrm{RSI}(n)[\mathrm{dB}] = 10\log_{10}\left(\frac{1}{N}\sum_{k=n-(N/2)}^{n+(N/2)} s^2(k)\right), \tag{5}$$

where N is the window length and $s(k)$ are the samples of the acoustical signal. According to this definition, the RSI will be used for a relative quantitative description of the sound intensity evolution during vibrato.

Considering the RSI definition of (5), the IA model of (4), and depending on N values, the resulting intensity may or may not show a periodic variation pattern. If the window length were selected such that the $\gamma_i(n)$ component of the partials is compensated for, the resulting RSI would not show periodic variations, as it is shown in Fig. 4, where the N values are represented by the T values in seconds, for a sampling frequency, f_s, of 44.1 kHz, and $T{=}N/f_s$. In Fig. 4 four RSI calculations are shown, corresponding to different N values. Comparing these results the RSI does not show periodic variations when the window length is fitted to one vibrato period.

To conclude, we can say that sound intensity variations are possible depending on the averaging window selected to calculate this magnitude. These variations will be a consequence of the frequency and amplitude variations of the partials.

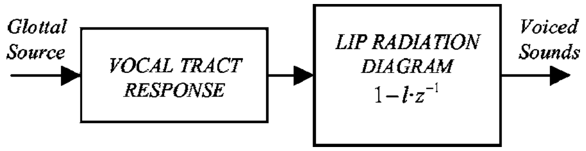

FIG. 5. Block diagram of the noninteractive source-filter model.

III. NONINTERACTIVE SOURCE-FILTER MODEL

Once the main features of the acoustical analysis of vibrato have been determined, the voice production process will be described. In this way, we will determine which voice production elements must be considered in order to understand the production of vocal vibrato. For this purpose, the noninteractive source-filter model will be considered as the voice production model,[19] because it is simpler than physical models and will make it easier to link the sinusoidal model to the voice production mechanisms. Thus, a brief review of this voice production description will be provided, as well as some important key points about the signal analysis associated to this model.

A. Model definition

The noninteractive source-filter model can be represented by the block diagram in Fig. 5. There, the voice production is modeled by the glottal source (GS) that is linearly modified by the vocal tract response (VTR) and the lip radiation impedance, which is approximated by a derivative system. Typically, the lip radiation system is usually combined with the GS, in such a way that the glottal source derivative (GSD) is considered as the vocal tract excitation. It is important to note that this model is denominated as the noninteractive source-filter model since both voice production elements are considered independent. This assumption does not exactly hold in natural voice production.[26–28]

Additionally, and in voiced sounds, the GS excitation is a periodic pulsed signal, determining the vocal texture. The most extended glottal pulse characterization defines five independent parameters: fundamental frequency, F_0, or fundamental period T_0, amplitude of voicing, Av, open quotient, O_q, asymmetry coefficient, α, and return phase interval T_a, which is related to the spectral tilt, f_t. All of the proposed GS waveform mathematical models try to include these parameters, such as for instance the LF model[29] or the KLGLOTT88.[24]

In Fig. 6, an idealized GS pulse and its derivative are shown, illustrating the five parameters defined. From this figure, it is clear that O_q controls the time interval where the GS is not null; α, defined as $\alpha{=}T_P/O_qT_0$, determines the symmetry degree of the glottal pulse; and T_a parametrizes the closure abruptness of the waveform.

Regarding the VTR, it is normally modeled by an all-pole autoregressive (AR) filter

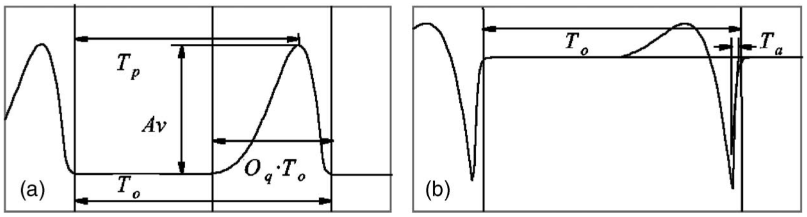

FIG. 6. Glottal pulse parametrization. (a) GS. (b) GSD.

$$H(z) = \frac{G}{1 - \Sigma_{k=1}^{p} a_k z^{-k}}, \tag{6}$$

where a_k are the prediction coefficients and G is the gain of the all-pole filter. This model corresponds to nonnasal sounds, as it is the case in our particular recordings.

B. Analysis procedure

As illustrated in Fig. 5, the noninteractive source-filter model is a very simple voice production one, but the analysis techniques associated to it allow extracting from the acoustic signal the information of the GS and the VTR, which is not the case in other voice production models. However, when this model is considered, a thorough analysis should be carried out, particularly in singing voice analysis, for a correct interpretation of the results. Therefore, some key issues pertaining to this model and its associated analysis procedures will be highlighted.

In natural voice production, the physical system is composed by two resonant cavities (subglottal and supraglottal cavities), which are connected by a valve, the glottis, where the vocal folds are located. In voiced sounds, the vocal folds are opened and closed providing the harmonic nature to the air flow, but, additionally, the vocal tract response varies during a single fundamental period because both cavities are connected and disconnected along every fundamental period, resulting in a nonlinear element. This effect is one of the effects associated to the source-tract interaction,[26,30] and it is not included in the noninteractive source-filter model.

On the other hand, the complete analysis procedure can be decomposed into two different steps: the first one is represented by the inverse filtering techniques, which allow us to separate both the GS and the VTR based on the acoustic signal, and the second one corresponds to a parametrization step, where the GS and VTR parameters are obtained in order to reduce to a few numerical values both voice production elements.

Regarding the inverse filtering techniques, there are many different approaches available in the literature,[31–35] however, since all of them assume a simplified voice production description, the resulting GS waveform will be affected in a different way by the so-called formant ripple resulting from the source-tract interaction and incomplete cancellation of the formants. Additionally, all of them have a fundamental frequency dependence, which limits their application as the fundamental frequency of the signal increases, and thus some singing voice signals.[36]

On the other hand, the GS parametrization is an important analysis step, particularly when a large number of signals are to be analyzed, and there are available several possible methods:[37] We have, for instance, direct estimation methods, where the source parameters are obtained by measuring time domain landmarks. However, they are not very robust because those landmarks are not easy to determine in inverse filtered signals, particularly the opening instant. Additionally, there are fitting estimation methods, where a mathematical model is fitted to the inverse filtered GSD. However, all of these approaches have a high computational load. Alternatively, the normalized amplitude quotient (NAQ) has been recently proposed[38] as a suitable parametrization method. It is amplitude and fundamental frequency normalized, avoids time domain landmarks measurement, and has a minimum computational cost. It is defined as the quotient of the maximum of the GS and the minimum of the GSD. Moreover, it has been shown that it is also a global parameter as it depends on the three above-mentioned GS parameters.[39]

To summarize these paragraphs, it is important to note that the noninteractive source-filter model allows for a simple voice production description, since it is a signal model. However, there are different factors that must be taken into consideration for a correct interpretation of the obtained results.

In order to illustrate the application of this kind of technique to the voice material described in Sec. II A, the GSD and VTR corresponding to a representative baritone recording are shown in Fig. 7. In this particular case three representative inverse filtering techniques have been selected: the analysis by synthesis (AbS) approach,[31] the glottal spectrum based (GSB) inverse filtering,[34] and the closed phase covariance (CPC).[35] From Fig. 7 we can conclude that the three techniques provide similar results. Regarding the GSD, three fundamental periods are represented, illustrating a similar pulse shape to that in Fig. 6. Concerning the VTR, it can be seen that, as this recording corresponds to a male singing voice recording, the highest formants are concentrated in one frequency region, ranging from 2000 to 3000 Hz, which is known as the singer's formant.[22]

It is interesting to note that the inverse filtering analysis is a short time analysis (the window length is three or four fundamental periods), compared to the slow fundamental frequency variations of vibrato (a typical vibrato rate is about 5 Hz). Then, it can be said that these techniques are insensitive to the presence or absence of vibrato, since they are based on short time window analysis.

Therefore, one question arising is how VTR and GSD parameters evolve during vibrato. Additionally, by comparing Fig. 3 to Fig. 7(b), the AM-FM representation is very similar to the VTR of Fig. 7(b). However, in the case of the

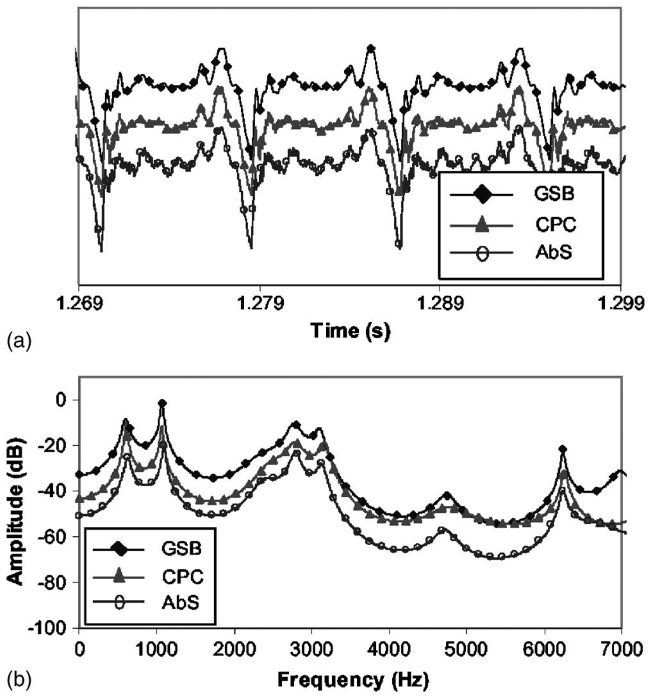

FIG. 7. Inverse filtering results. The signal corresponds to a baritone recording, singing the vowel [a] with a F_0 of 123 Hz. (a) GSD. (b) VTR.

AM-FM representation, no source-filter separation has been made, and thus both elements are melted in this representation.

It can be concluded that in natural vocal vibrato nothing can be said, *a priori*, about what is happening with the GSD and VTR during vibrato, but, whatever happens, the resulting global AM-FM representation provides a spectral envelope very similar to the VTR obtained by inverse filtering.

IV. NONINTERACTIVE VIBRATO PRODUCTION MODEL

After analyzing the acoustic signal corresponding to vibrato, assuming the sinusoidal model, and reviewing the voice production process based on the noninteractive source-filter model, it might seem difficult to establish a relationship between these two different mathematical descriptions of the acoustic signal and, even more, to identify the individual contribution of the voice production elements to the amplitude variation of the partials. With this in mind, a noninteractive vibrato production model will be proposed in order to describe, and help to understand, how the GSD and the VTR behave during vibrato such that the specific amplitude variation of the partials comes out. This model will allow us to assure the goodness of the analysis carried out in natural singing voice recordings. In particular, this signal model will permit us to generate synthetic signals that will be subsequently analyzed in the same way as natural singing voice recordings, which will help to infer what is happening in vocal vibrato production.

Before proposing the vibrato production model, some basic assumptions will be made regarding the behavior of GSD and VTR during vibrato. Assuming that the RSI, calculated by using a window length equal to one vibrato period, does not change during a few vibrato cycles, we have the following.

(1) The GSD characteristics, or glottal pulse shape features, O_q, α, and f_t, remain also 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 VTR does not appreciably change along with vibrato. This assumption supports the fact that vocalization does not change along the note.

Taking into account these assumptions, the proposed noninteractive vibrato production model is represented by the block diagram of Fig. 8.

As illustrated in Fig. 8, for a given vowel where the RSI does not significantly change, the GSD parameters, controlling the glottal pulse shape, along with the VTR features, remain constant while the fundamental frequency of the excitation varies.

This model anticipates that a long-term relationship can be established between the GS and VTR and the AM-FM representation: Taking into account that the GSD features remain constant during vibrato, the AM-FM representation of each harmonic should represent a local section of the VTR, and each representation will be shifted (linearly distorted) in amplitude depending on the GSD spectral shape. Therefore, by removing the GSD effect from the AM-FM representation, only the VTR will be left.

This relationship is graphically shown in Fig. 9 for a synthetic signal described by the proposed model. For this particular example, the GSD has been modeled according to the LF model, and its parameters correspond to a normal phonation glottal pulse shape. The VTR filter corresponds to vowel [a] of a baritone, and the three vibrato parameters remain constant during the note: the intonation 100 Hz, the extent value of 10 Hz, and a rate value of 5 Hz. The resulting signal has been analyzed by both inverse filtering (GSB inverse filtering algorithm), where the presence or absence of vibrato has no influence on the algorithm, and by a sinusoidal model, where the IA and IF of each harmonic need to be measured. Results obtained for this simulation are shown in Fig. 9.

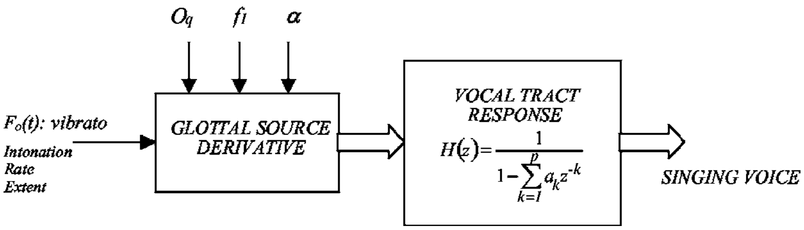

FIG. 8. Noninteractive vibrato production model.

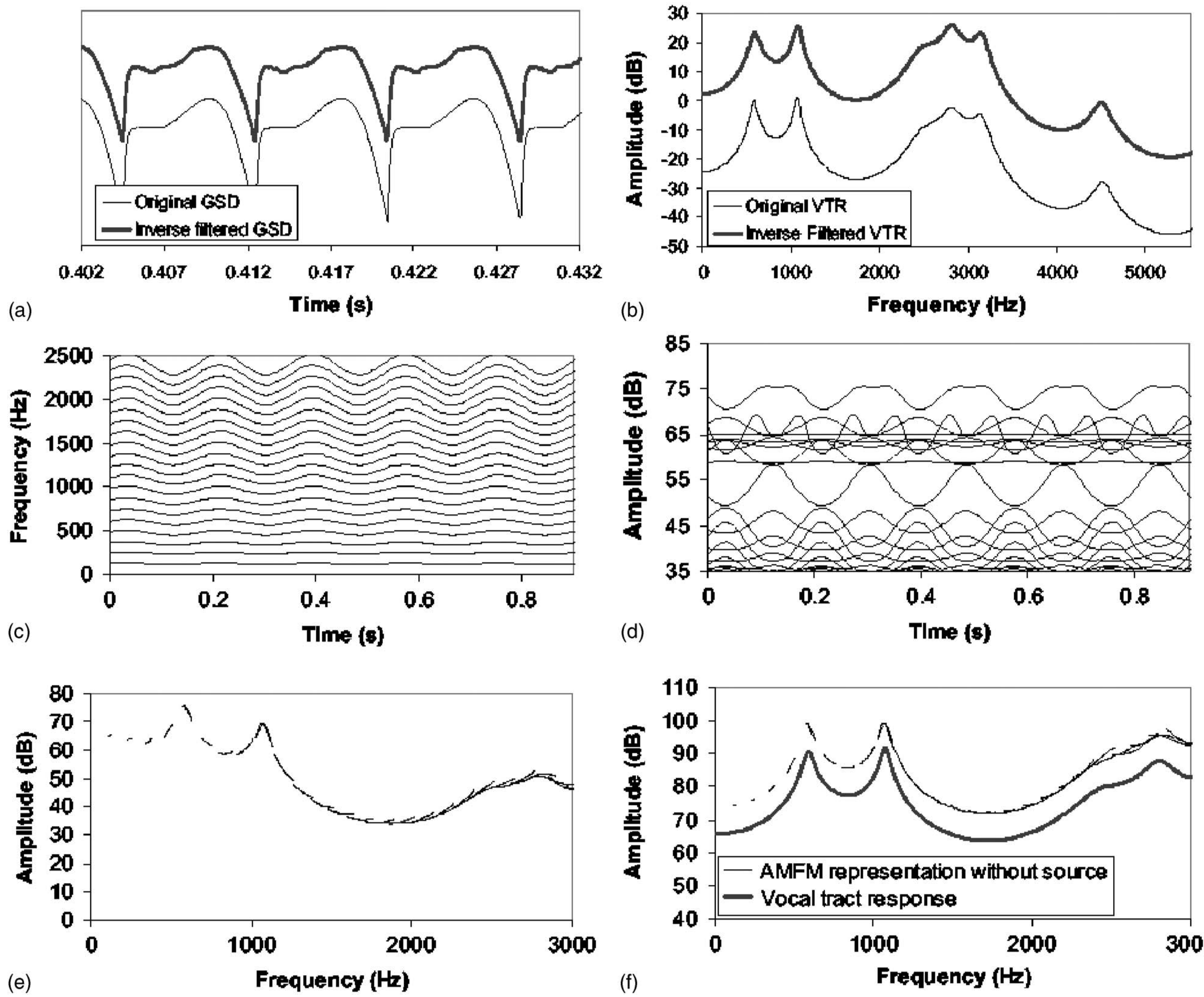

FIG. 9. Relationship between voice production and sinusoidal model for the synthetic signal. Inverse filtering results. (a) GSD. (b) VTR. Sinusoidal modeling results. (c) IF. (d) IA. (e) AMFM representation. (f) AMFM representation without source effect.

In Figs. 9(a) and 9(b) inverse filtering results are shown for a short window analysis. Since fundamental frequency is low, GSD and VTR are well separated. In Figs. 9(c) and 9(d), sinusoidal modeling results are shown. The frequency variations of the signal harmonics are clearly observed and the resulting amplitude variation, too. On the other hand, in Fig. 9(e) the AM-FM representation of the partials is shown. Taking into account the AM-FM representation of every partial, and comparing it to the VTR shown in Fig. 9(b), it is possible to conclude that local information of the VTR is provided by this representation. However, since no source-filter decomposition has been carried out, each AM-FM representation is offset in amplitude depending on the GSD spectral features. This effect results from keeping GSD parameters constant during vibrato. Comparing Figs. 9(e) and 9(b), it can be guessed that if the GSD magnitude spectrum were removed from the AM-FM representation of the harmonics, the resulting AM-FM representation would be very similar to the VTR. The result of this operation is shown in Fig. 9(f), and it can be seen that the compensated AM-FM representation is very close to the VTR.

For this noninteractive vibrato production model, the individual contribution of the GS and the VTR on the amplitude variation of the partials has been identified. In particular, when inverse filtering works, the GSD effect can be removed from the AM-FM representation provided by the sinusoidal model and the VTR information is isolated.

At this point, the relationship between the two signal models, noninteractive source-filter model and sinusoidal model, has been established for a synthetic signal where vibrato has been included under two assumptions stated at the beginning of the section. Now the question is if this relationship also holds in natural singing voice, where many other effects are present. Therefore, both kinds of signal analyses will now be applied to natural singing voice recordings. In order to set up conditions similar to the simulated signal, 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, as detailed in Sec. II A.
(2) Recordings have been made in a studio, where reverberations are reduced. Under these conditions the

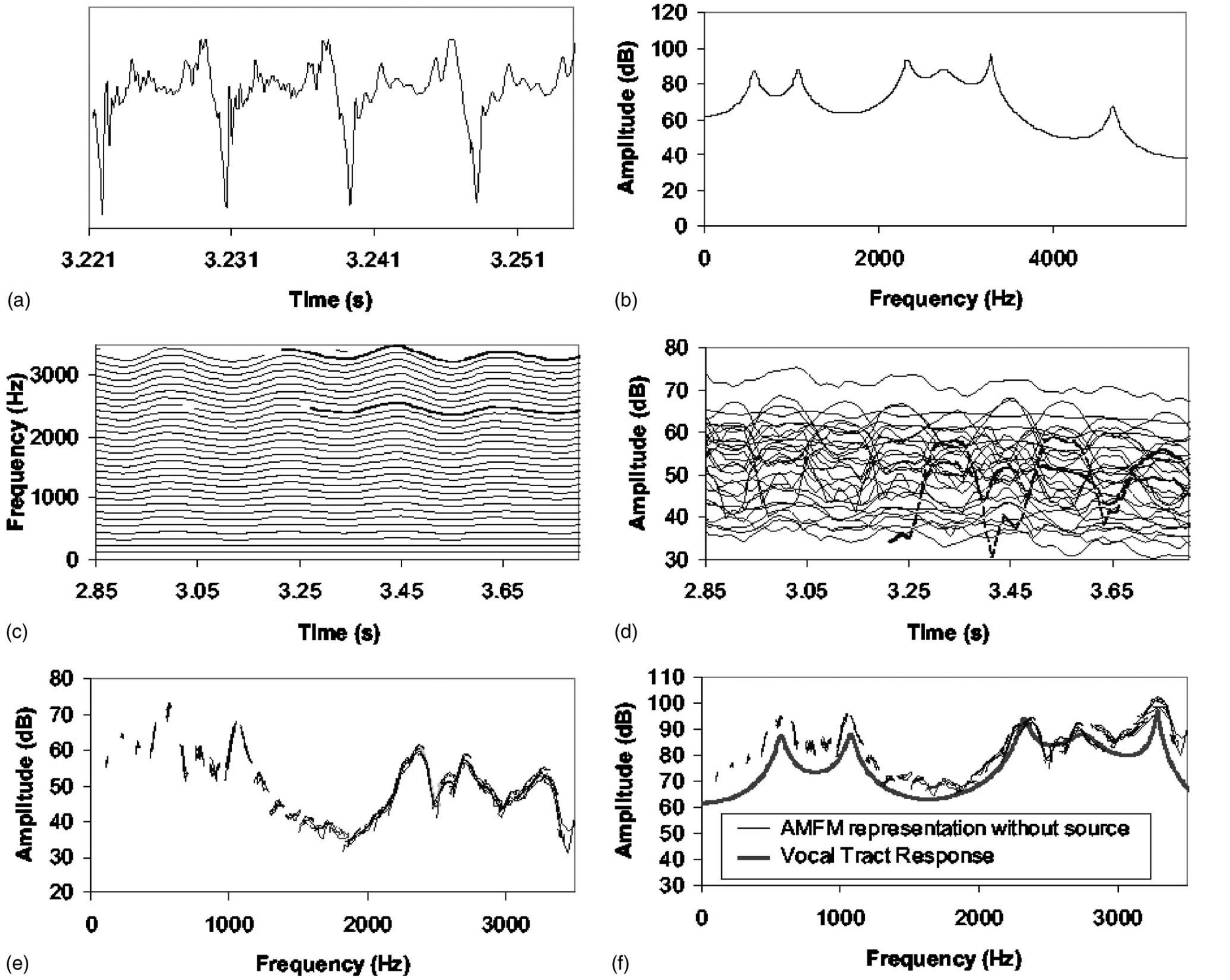

FIG. 10. Relationship between voice production and sinusoidal model for natural singing voice. Inverse filtering results. (a) GSD. (b) VTR. Sinusoidal modeling results. (c) IF. (d) IA. (e) AMFM representation. (f) AMFM representation without source effect.

> AM-FM representation will present slight variations from the actual VTR, but it is still possible to carry out the same study.[14]

In Fig. 10 these analysis results are shown for a baritone recording, characterized by a F_0 of 107 Hz and a vocal tract configuration corresponding to vowel [a]. Unlike the measurements shown in Figs. 9(a) and 9(b), there is no reference for the original GSD and VTR. By comparing Figs. 9(c) and 10(c), the IF variation is similar in simulation and natural singing voice. However, the vibrato extent in this baritone recording is lower than in synthetic signal. In the case of the IA, natural singing voice results are obviously not as regular as synthetic ones. This is because of reverberation and irregularities of natural voice. Regarding the RSI of the sound, there are not large variations on the IA and, so, for one or two vibrato cycles, it can be considered constant.

In this situation, the AM-FM representation of the harmonics, shown in Fig. 10(e), is very similar to the synthetic signal AM-FM representation of Fig. 9(e), though the already mentioned irregularities are present. Now, the so-obtained GSD spectrum will be used to extract from the AM-FM representation the VTR information. The result of this operation is shown in Fig. 10(f) and, as in the case of the synthetic signal, the compensated AM-FM representation is very close to the VTR.

In this way, it can be concluded that, as in the case of the synthetic signal, the individual contribution of the voice production elements has been identified from the AM-FM representation corresponding to natural vocal vibrato, since the compensated AM-FM representation is similar to the VTR obtained by inverse filtering. However, the matching is not as close as for the synthetic signal.

For not limiting ourselves to a unique signal analysis, the same analysis procedure has been applied to a representative set of signals of the database described in Sec. II A, and their corresponding results are shown in Fig. 11. For this particular set of signals, recordings corresponding to the three male singers have been selected (tenor 1, tenor 2, and baritone), considering also different vocal tract configurations ([a], [e], [i]). Additionally, in order to analyze the behavior of the model in different situations, three synthetic signals have been added to the comparison, representing also

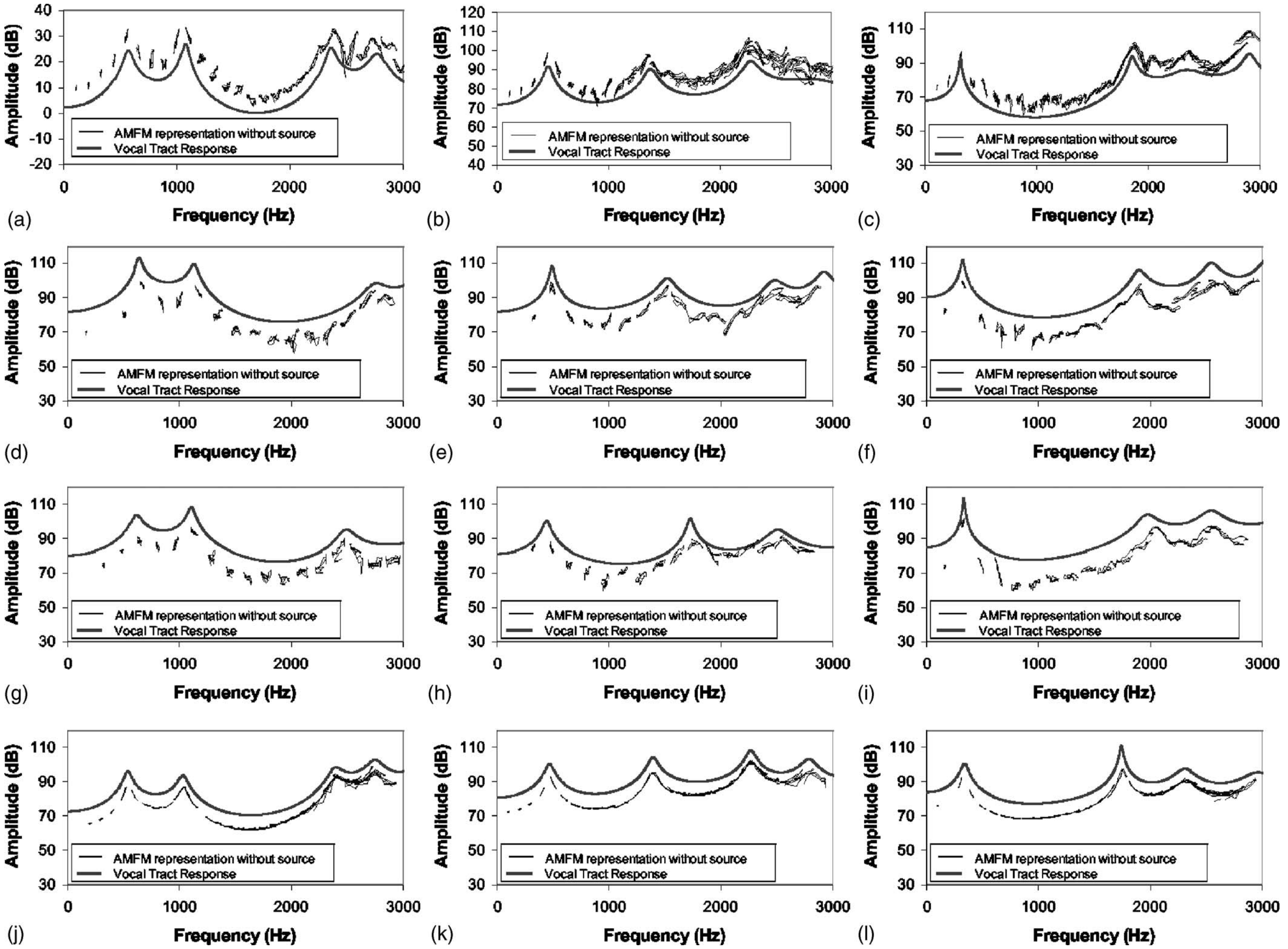

FIG. 11. Compensated AM-FM representation for a representative set of recordings. For each singer three different vocal tract configurations have been selected, as well as for the noninteractive vibrato production model. The results corresponding to different voices have been distributed in different rows, and each column corresponds to the same vocal tract configuration. Baritone recordings: (a) vowel [a], (b) vowel [e], and (c) vowel [i]. Tenor 1 recordings: (d) vowel [a], (e) vowel [e], and (f) vowel [i]. Tenor 2 recordings: (g) vowel [a], (h) vowel [e], and (i) vowel [i]. Synthetic signals: (j) vowel [a], (k) vowel [e], and (l) vowel [i].

different vowels. In order to give more details of these recordings, in Table I the values of the three fundamental frequency parameters (intonation, vibrato rate, and vibrato extent) are collected. It can be seen that all the signals have a low fundamental frequency, so that the main problems associated to the inverse filtering techniques are avoided, and in particular the fundamental frequency of the baritone recordings is slightly lower than the one corresponding to the tenor recordings. In the case of the synthetic signals, the fundamental frequency was chosen to be 100 Hz, the vibrato rate 5.5 Hz, and the vibrato extent 84.46 cents.

TABLE I. Fundamental frequency parameters for the representative data set. This data set covers different vocal tract configurations for the three male singers.

		Rate (Hz)	Intonation (Hz)	Extent (cent)
Baritone	[a]	4.47	110.86	48.20
	[e]	4.60	109.92	60.51
	[i]	4.46	108.21	51.69
Tenor #1	[a]	5.98	166.54	43.62
	[e]	5.16	161.60	70.02
	[i]	5.28	166.48	74.99
Tenor #2	[a]	5.89	160.25	27.65
	[e]	5.60	162.27	26.57
	[i]	5.74	161.44	34.50

By observing Fig. 11 the different vocal tract configurations are evident, because of the different formant positions. Additionally, since all the recordings correspond to singing male voices, the singers' formant is in the high-frequency region (above 2000 Hz). Also, the differences between the fundamental frequencies of the signals can be observed on the AM-FM representation, since for the same frequency region (0–3000 Hz) there are more harmonics in the case of the synthetic signal or in the baritone recordings. On the other hand, it can be seen that in all cases, different singers and vocal tract configurations as well as in the synthetic case, the compensated AM-FM representation is very close to the VTR, as was pointed for Fig. 10, which allows us to conclude that the noninteractive vibrato production model can explain, in an approximated way, what is happening in singing voice when vibrato is present.

V. MODEL EVALUATION

In the last section, a noninteractive vibrato production model has been proposed in order to relate the amplitude and frequency variation of the partials to the voice production elements. The key point of the model, as should be clear from our analysis, is the fact that the GSD and VTR features, O_q, α, and f_t, and formants central frequencies and bandwidths, respectively, remain almost constant during vibrato. In order to confirm these assumptions, the proposed model will be evaluated and further analyses will be carried out. To this purpose, the GSD and VTR features will be measured and their time evolution evaluated.

The evaluation procedure will be as follows: First, inverse filtering techniques will be applied for the GSD and VTR estimation in such a way that the analysis window will be moved over time. Later, and using the appropriate VTR and GSD parametrization, the behavior of these elements during vibrato will be characterized in order to check if their parameters remain constant during vibrato or, on the contrary, if they vary.

Moreover, it is important to note that because of the limitation of the inverse filtering techniques for high fundamental frequencies, the evaluation procedure will be applied in recordings corresponding with low fundamental frequencies. However, after this evaluation process the AM-FM representation of natural vibrato recordings will be compared to the one corresponding to the model, in order to analyze what this model predicts for higher fundamental frequencies.

A. Evaluation procedure

In Sec. III B the analysis procedure associated to the noninteractive source-filter model was shortly reviewed, and some relevant notes were provided for a correct interpretation of the obtained results. In this section, this analysis procedure will be used for the evaluation procedure of the noninteractive vibrato production, and, taking into account the above-mentioned details, some decisions have been made. Concerning the inverse filtering techniques, three representative approaches have been selected (the three were addressed in Sec. III B) for not limiting ourselves to a unique, and maybe biased, calculation. Additionally, as has been highlighted before, there are different factors that affect the inverse filtering results (differences among the selected techniques, fundamental frequency dependence,…), which has led us to apply the same analysis procedure to the synthetic signals, in order to determine the accuracy of the estimation procedure. Finally, regarding the VTR and GS parametrization, the normalized amplitude quotient (NAQ) has been chosen as a representative GS parameter, since its measurement is simple and robust and condenses in a single numerical value the whole glottal pulse information.[39] In parallel, and as the VTR representative parameters, the central frequencies of the first two formants, F1 and F2, have been selected.

This analysis procedure has been applied to the representative data set specified in Table I, as well as to the synthetic signals of Sec. IV, and the results corresponding to the vowel [a] are shown in Fig. 12. It is important to note that, in order to make easier the comparison among different singers, vocal tract configurations, and synthetic and natural signals, the absolute values of NAQ, F1, and F2 have been replaced by their relative values in percentage. To that purpose their mean value has been removed and they have been normalized by their average value in the considered interval.

In Fig. 12, different magnitudes for the four representative signals are shown: the acoustic signal, the RSI, the fundamental frequency variation, and the relative values of the NAQ, F1, and F2 and the plots corresponding to the different signals have been grouped in different columns, (a), (b), (c), and (d).

In all cases it can be seen that the considered time interval is long in the sense that it contains several vibrato cycles (at least four), and in all cases the RSI does not change significantly (perfectly constant in the synthetic signal). By observing the NAQ, F1, and F2 variations, it can be seen that they are not perfectly constant in all cases, either natural or synthetic signals. Additionally, by comparing the results corresponding to the three natural vibrato signals and the synthetic one, it can be seen that the maximum value of the NAQ relative value is very similar, around 10% for all signals. Something similar can be said concerning the F1 and F2, where the maximum value is around 5%. By considering that the variations appreciated in Fig. 12 for the NAQ, F1, and F2 have very similar values, it can be concluded that these variations can be mainly attributed to measurement errors.

In order to analyze other vocal tract configurations, the measurements of the NAQ, F1, and F2 obtained along the considered time interval have been reduced to a single numerical value: their standard deviation along the considered time interval. To this purpose, the results corresponding to the representative data set of Sec. IV are shown in Table II for the three singers and the noninteractive vibrato production model.

By analyzing Table II, it can be seen that the standard deviation of the normalized parameters is quite similar in all cases, and less than 10%, regardless of the conditions (different singers and proposed model, vocal tract configurations, different inverse filtering techniques). Additionally, by considering the three different vocal tract configurations for the three male singers, it can be seen that the standard deviation of the normalized NAQ is higher in the case of vowel [i], which corresponds to the lowest first formant central frequency, as can be seen in Fig. 11. This difference might be a consequence of the source-tract interaction, since it increases as the first formant central frequency decreases,[40] and it will be reflected on the inverse filtered GSD waveform. Also, it can be noticed that the standard deviation of the three parameters obtained for the baritone recordings are lower than those obtained for the tenor's recordings, which corresponds to differences in the intonation value, as is the case of the synthetic signals.

To summarize, and according to the results shown in Fig. 12 and Table II, it can be concluded that the GS and VTR parameters do not significantly change during vibrato, which validates the assumptions made when we proposed the

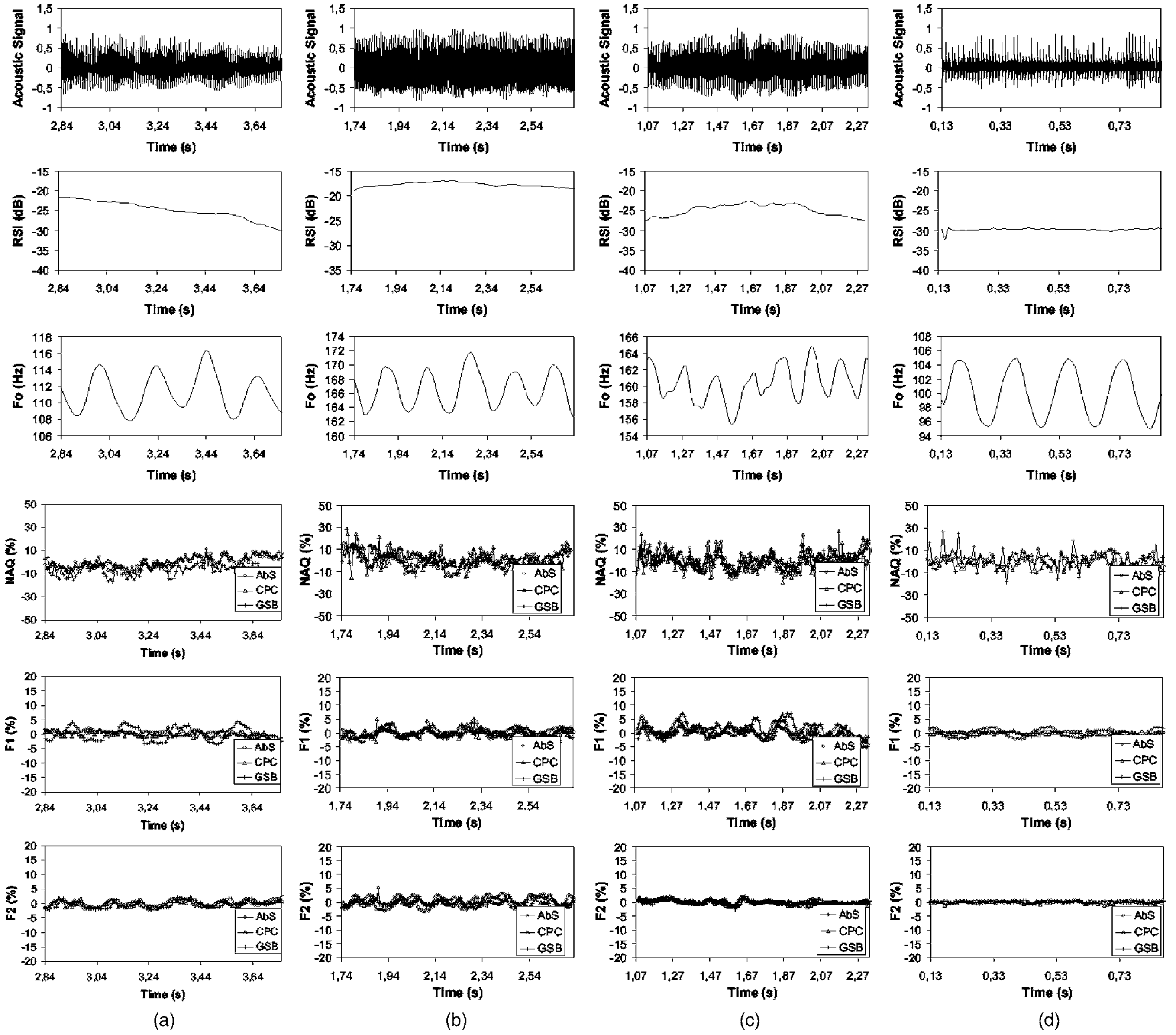

FIG. 12. Model evaluation. The signals associated to the model evaluation process are shown for four representative signals, corresponding to the three singers and the synthetic signal, all of them corresponding to the vowel [a]. For each one of these recordings the represented signals are the acoustic signal, the relative sound intensity, the fundamental frequency, the normalized NAQ, F1, and F2. The plots corresponding to the same voice are grouped by columns: (a) baritone, (b) tenor 1, (c) tenor 2, and (d) synthetic signal.

vibrato production model. This allows us to state that the noninteractive vibrato production model closely represents natural vibrato production.

B. Vibrato in higher pitched signals

As has been mentioned, in Sec. V A an evaluation procedure has been applied in order to validate the assumptions made along the model proposition. This evaluation procedure has been applied in low F_0 recordings such that the possible inverse filtering errors are avoided. However, it would be important to see how the same model performs under different conditions, and in particular for higher F_0 signals. This will allow us to determine if the above extracted conclusions concerning the GS and VTR behavior also hold under different pitch conditions.

Therefore, in this subsection, and in order to compare the noninteractive vibrato production model and natural vibrato signals, the AM-FM representation of the signal will be considered, since the analysis associated to the sinusoidal model has no limitation with the fundamental frequency value.

In this case a representative set of recordings for the vowel [a] has been selected, corresponding to three different F_0 values for the three male singers, as well as for the noninteractive vibrato production model: 166, 216, and 360 Hz, representing low, medium, and high fundamental frequency values (for male voices). In the case of synthetic signals, the VTR has been the same in all cases, corresponding to the vowel [a], and only the fundamental frequency of the signal has been varied, the vibrato rate being 5.5 Hz and the extent 84.46 cents.

TABLE II. Standard deviation of the normalized voice production parameters NAQ, F1, and F2 for the recordings corresponding to the baritone, tenor 1, tenor 2, and the noninteractive vibrato production model. The results correspond to different vocal tract configurations and different inverse filtering techniques results.

		Baritone			Tenor 1		
		sd_{NAQ}	sd_{F1}	sd_{F2}	sd_{NAQ}	sd_{F1}	sd_{F2}
	AbS	4.88	1.17	1.11	6.62	1.42	2.06
[a]	CPC	4.71	1.14	0.97	7.72	1.42	0.90
	GSB	5.65	2.28	1.30	5.84	1.24	3.14
	Abs	4.77	2.73	1.05	6.10	3.20	1.63
[e]	CPC	6.68	1.90	0.99	9.87	2.90	0.94
	GSB	5.02	3.32	1.54	6.39	2.59	1.60
	AbS	8.71	1.13	0.50	6.66	1.37	1.23
[i]	CPC	7.85	1.20	0.39	7.50	1.84	2.07
	GSB	9.32	1.57	0.60	9.20	2.18	1.36
		Tenor 2			Model		
	AbS	7.23	3.06	0.67	3.51	1.43	0.35
[a]	CPC	8.80	1.49	0.79	5.64	0.50	0.47
	GSB	6.12	2.33	0.87	4.37	1.69	0.59
	AbS	5.16	6.57	3.53	3.78	0.92	0.30
[e]	CPC	5.14	7.39	3.41	3.47	0.12	0.07
	GSB	7.38	5.32	3.56	6.41	0.87	0.31
	AbS	9.08	1.23	1.81	3.83	1.51	0.39
[i]	CPC	8.88	3.86	1.23	2.15	0.71	0.09
	GSB	9.79	2.56	1.35	3.76	1.90	0.55

The AM-FM representation of the above-mentioned signals are collected in Fig. 13. In this figure rows represent different singers (two tenors and one baritone) and the synthetic signals, and columns correspond to different fundamental frequency values.

In this case, as was detailed in Sec. II C, the AM-FM representation provides global information, since both voice production elements are included. However, the proposed vibrato production model allows us to understand how this representation is related to the voice production process. The noninteractive vibrato production model predicts that as the fundamental frequency increases, the harmonics of the GSD signal will be located in higher frequencies, and they will be located in different relative positions with respect to the vocal tract resonances. As a result, the AM-FM representation for each partial will correspond to a different local section of the VTR, shifted in amplitude depending on the spectral content of the GSD.

With this in mind, it can be observed in Fig. 13, for both natural and synthetic signals, that as the fundamental frequency increases, the AM-FM representations of the partials are obviously more separated from each other. Additionally, the spectral envelope foreseen in all cases is quite similar, since all correspond to the same vocal tract configuration, vowel [a], and all are of male singing voices. By considering one row of figures, for instance, (a), (b), and (c), it can be seen that, for a given singer, the only difference among the different F_0 values is reflected on the harmonic position, and the location of the AM-FM representation of each partial, which is predicted by the proposed model [panels (j), (k), and (l)]. This points out that there are no additional changes, concerning vibrato production, as the fundamental frequency increases. This behavior can also be observed in the results corresponding to other singers (different rows of figures) and, as a consequence, the conclusions can be extended to other male voices.

To conclude, it can be said that the noninteractive vibrato production model can be used to describe the behavior of the GS and VTR during the vocal vibrato production, at least from the signal point of view. It is clear that for a physiological description of the process other kind of models should be considered.

VI. CONCLUSIONS

In this work an acoustical characterization of vocal vibrato has been carried out. Such characterization has been based on the sinusoidal model parameters, IF and IA of the partials, focused on the amplitude variations and their relationship with the frequency variations. It has been shown that these variations are correlated, which is illustrated by the so called AM-FM representation of the partials. Additionally, we have shown that the amplitude and frequency variation of the partials will be translated into intensity variations depending on the window length imposed for the estimation. In particular, it has been shown that the RSI will not show pitch-related variations when the averaging window length is adjusted to one vibrato period. The main conclusion derived from this acoustical analysis is that the origin of the amplitude variations must be pursued taking into consideration the voice production mechanisms, since the sinusoidal model is a pure signal model.

In order to identify the contribution of the voice production elements, GS and VTR, on the amplitude variation of the partials, a noninteractive vibrato production model has been proposed. Consequently, this model has described the behavior of both voice production features during vibrato. In particular, considering a RSI constant interval, this model assumed that both voice production mechanisms remain almost constant during vibrato, and only the fundamental frequency of the GS changes. As a result, this model predicted that the AM-FM representation of each harmonic corresponds to a local section of the VTR but shifted (i.e., linearly distorted) in amplitude according to the GS spectrum. By comparing both synthetic and natural singing voice signals it was concluded that the above-mentioned prediction holds quite well in natural vocal vibrato.

Next, in order to validate the assumptions that the noninteractive vibrato production model is based on, the time evolution of the GSD and VTR features was evaluated in natural and synthetic signals making use of inverse filtering techniques. From this evaluation process, it was demonstrated that the GSD and VTR do not significantly vary during vibrato, which agrees well with the assumptions made during the model proposition. Additionally, the proposed model was compared to natural vibrato signals for different pitch values, making use of the AM-FM representation, and it was shown that the proposed model predicted similar representations to those obtained for natural vocal vibrato. This

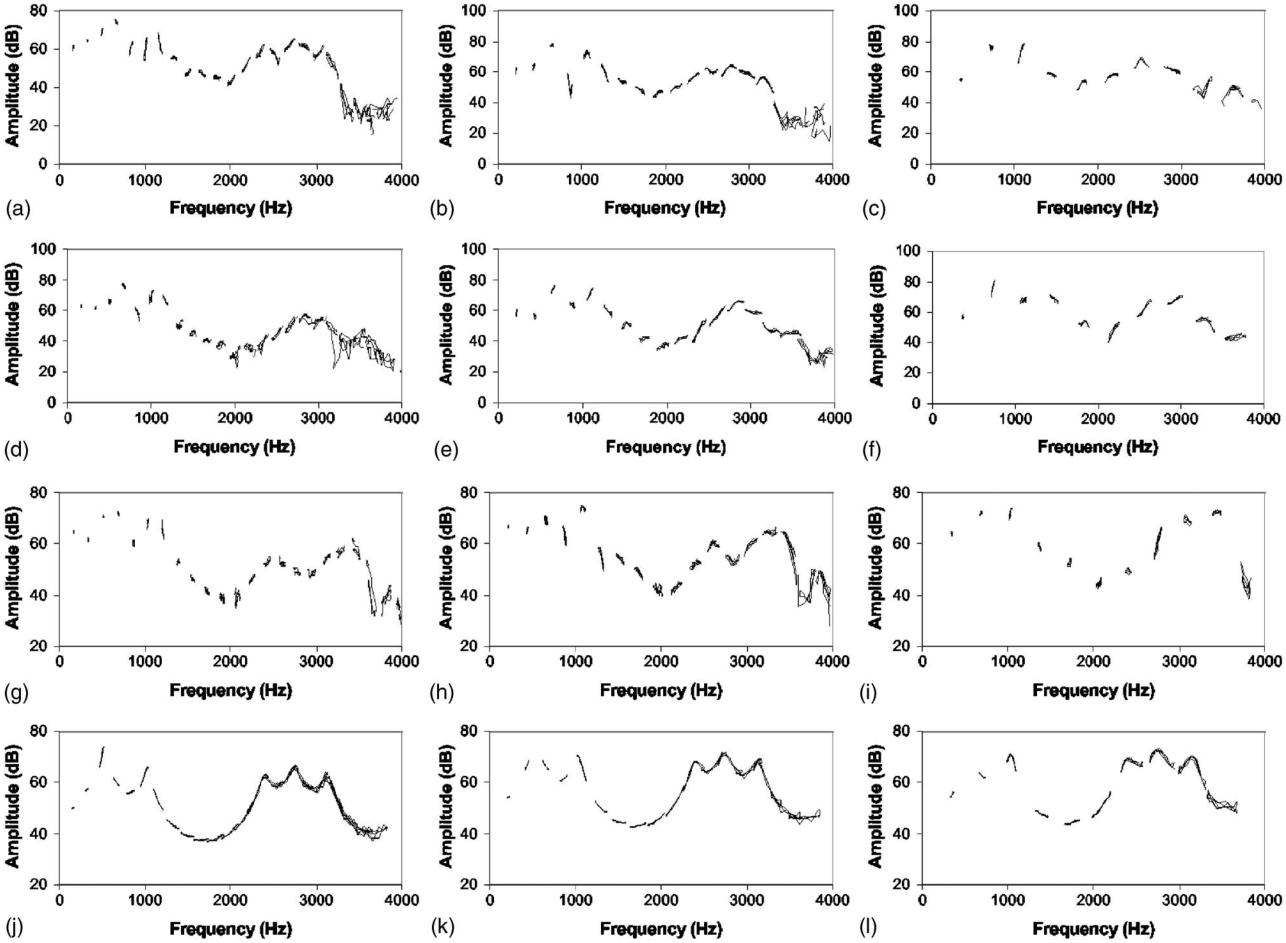

FIG. 13. Vibrato production in higher pitched signals. The proposed noninteractive vibrato production model is compared to natural vocal vibrato for different pitch values and different singers. The selected vocal tract configuration corresponds to vowel [a]. The results corresponding to the same voice are grouped by rows and each column correspond to a similar fundamental frequency value. Baritone recordings: (a) F_0=166 Hz, (b) F_0=216 Hz, and (c) F_0=350 Hz. Tenor 1 recordings: (d) F_0=166 Hz, (e) F_0=216 Hz, and (f) F_0=350 Hz. Tenor 2 recordings: (g) F_0=166 Hz, (h) F_0=216 Hz, and (i) F_0=350 Hz. Synthetic signal: (j) F_0=166 Hz, (k) F_0=216 Hz, and (l) F_0=350 Hz.

validation process has shown that the proposed vibrato production model is able to describe the production of this vocal effect.

Based on the noninteractive vibrato production model, it can be concluded that, during the vibrato production, the pitch variations are generated on the glottal source and these variations, along with the vocal tract filtering effect, reflect the frequency and amplitude variations of the acoustic signal partials.

ACKNOWLEDGMENTS

This work was supported in part by the Ministerio de Educación y Ciencia under Grant FPU, AP2000-4674. The Gobierno de Navarra and the Universidad Pública de Navarra are gratefully acknowledged for financial support. Finally, authors would also like to acknowledge support from Xavier Rodet and Axel Röbel (IRCAM, Paris) and the collaboration from Daniel Erro who implemented some of the algorithms.

[1] J. Sundberg, "Acoustic and psychoacoustics aspects of vocal vibrato," in *Vibrato*, edited by P. H. Dejonckere, M. Hirano, and J. Sundberg (Singular, San Diego, 1995), pp. 35–62.

[2] C. Sheashore, "The vibrato," in *University of Iowa Studies in the Psychology of Music*, Vol. 1 (University Press, Iowa City, 1932).

[3] E. Prame, "Measurements of the vibrato rate of ten singers," J. Acoust. Soc. Am. **96**, 1979–1984 (1994).

[4] E. Prame, "Vibrato extent and intonation in professional western lyric singing," J. Acoust. Soc. Am. **102**, 616–622 (1997).

[5] I. Arroabarren, M. Zivanovic, J. Bretos, A. Ezcurra, and A. Carlosena, "Measurement of vibrato in lyric singers," IEEE Trans. Instrum. Meas. **51**, 660–665 (2002).

[6] C. Dromey, N. Carter, and A. Hopkin, "Vibrato rate adjustment," J. Voice **17**, 168–178 (2003).

[7] J. A. Díaz and H. B. Rothman, "Acoustical comparison between samples of good and poor vibrato in singers," J. Voice **17**, 179–184 (2003).

[8] I. R. Titze, B. H. Story, M. Smith, and R. Long, "A reflex resonance model of vocal vibrato," J. Acoust. Soc. Am. **111**, 2272–2282 (2002).

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

[10] Y. Horii, "Acoustic analysis of vocal vibrato: A theoretical interpretation of data," J. Voice **3**, 36–43 (1989).

[11] R. C. Maher and J. W. Beauchamp, "An investigation of vocal vibrato for synthesis," Appl. Acoust. **30**, 219–245 (1990).

[12] S. Imaizumi, H. Saida, Y. Shimura, and H. Hirose, "Harmonic analysis of the singing voice: Acoustic characteristics of vibrato," in Proc. of the Stockholm Music Acoustics Conf., 28 July–August 1 1993, pp. 197–200.

[13] M. Mellody, F. Herseth, and G. H. Wakefield, "Modal distribution analysis and synthesis of a soprano's sung vowels," J. Voice **15**, 469–482 (2001).

[14] I. Arroabarren, M. Zivanovic, X. Rodet, and A. Carlosena, "Instantaneous frequency and amplitude of vibrato in singing voice," in Proc. of the IEEE ICASSP, 6–10 April, 2003, Hong Kong, China.

[15] R. J. McAulay and Th. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," IEEE Trans. Acoust., Speech, Signal Process. **34**(4), 744–754 (1986).

[16] X. Serra and J. Smith III, "Spectral Modeling Synthesis: A sound Analysis/Synthesis system based on deterministic plus stochastic decomposition," Comput. Music J. **14**, 12–24 (1990).

[17] P. R. Cook, "SPASM, a real-time vocal tract physical model controller; and Singer, the companion software synthesis system," Comput. Music J. **17**, 30–43 (1993).

[18] B. H. Story, "An overview of the physiology, physics and modeling of the sound source for vowels," Acoust. Sci. Technol. **23**(4), 195–206 (2002).

[19] G. Fant, *Acoustic Theory of Speech Production* (Mounton, The Hague, 1960).

[20] J. F. Michel and R. D. Myers, "Vibrato and pitch transitions," J. Voice **1**, 157–161 (1987).

[21] J. F. Michel and R. D. Myers, "The effects of crescendo on vocal vibrato," J. Voice **5**, 292–298 (1991).

[22] J. Sundberg, "Perceptual aspects of singing," J. Voice **8**, 106–122 (1994).

[23] J. O. Smith III and X. Serra, "PARSHL: An Analysis/Synthesis program for non-harmonic sounds based on a sinusoidal representation," in Proc. of the ICMC (1987).

[24] D. H. Klatt and L. C. Klatt, "Analysis, synthesis and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857 (1990).

[25] I. Arroabarren, M. Zivanovic, and A. Carlosena, "Analysis and synthesis of vibrato in lyric singers," in Proc. of the EUSIPCO, 3–6 September 2002, Toulouse, France.

[26] T. V. Ananthapadmanabha and G. Fant, "Calculation of the true glottal flow and its components," Speech Commun. **1**(3–4), 167–184 (1982).

[27] A. Barney, C. Shadle, and P. O. A. L. Davis, "Fluid flow in a dynamic mechanical model of the vocal folds and tract: I. Measurements and theory," J. Acoust. Soc. Am. **105**, 444–455 (1999).

[28] C. Shadle, A. Barney, and P. O. A. L. Davis, "Fluid flow in a dynamic mechanical model of the vocal folds and tract: II. Implications for speech production studies," J. Acoust. Soc. Am. **105**, 456–466 (1999).

[29] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," STL-QPSR **85**, 1–13 (1985).

[30] D. G. Childers and C.-F. Wong, "Measuring and modeling vocal source-tract interaction," IEEE Trans. Biomed. Eng. **41**, 663–671 (1994).

[31] H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for the glottal source waveform," in Proc. of the IEEE ICASSP, April 1986, Vol. 11, pp. 1605–1608.

[32] H.-L. Lu and J. O. Smith, "Joint estimation of vocal tract filter and glottal source waveform via convex optimization," in Proc. of the IEEE WASPAA, October 1999, pp. 79–92.

[33] P. Alku and E. Vilkman, "Estimation of the glottal pulseform based on Discrete All-Pole modeling," in Proc. of the ICSLP, September 1994, Yokohama Japan, pp. 1619–1622.

[34] I. Arroabarren and A. Carlosena, "Glottal spectrum based inverse filtering," in Proc. of the EUROSPEECH, 1–4 September 2003, Geneva, Switzerland.

[35] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Trans. Acoust., Speech, Signal Process. **27**, 350–355 (1979).

[36] 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, November 2001.

[37] H. Strik, "Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses," J. Acoust. Soc. Am. **103**, 2659–2669 (1998).

[38] P. Alku and T. Bäckström, "Normalized amplitude quotient for parametrization of the glottal flow," J. Acoust. Soc. Am. **112**, 701–710 (2002).

[39] I. Arroabarren and A. Carlosena, "Glottal source parameterization: A comparative study," in Proc. of the VOQUAL, 27–29 August 2003, Geneva, Switzerland.

[40] B. Guerin, M. Mrayati, and R. Carre, "A voice source taking account of coupling with the supraglottal cavities," in Proc. of the IEEE ICASSP, April 1976, Vol. 1, pp. 47–50.