Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 278–294

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis

Gilles Degottex *,[1], Pierre Lanchantin [2], Axel Roebel, Xavier Rodet

Ircam – CNRS-UMR9912-STMS, Analysis-Synthesis Team, 1 Place Igor Stravinsky, 75004 Paris, France

Received 12 January 2012; received in revised form 15 August 2012; accepted 22 August 2012
Available online 30 August 2012

Abstract

In current methods for voice transformation and speech synthesis, the vocal tract filter is usually assumed to be excited by a flat amplitude spectrum. In this article, we present a method using a mixed source model defined as a mixture of the Liljencrants–Fant (LF) model and Gaussian noise. Using the LF model, the base approach used in this presented work is therefore close to a vocoder using exogenous input like ARX-based methods or the Glottal Spectral Separation (GSS) method. Such approaches are therefore dedicated to voice processing promising an improved naturalness compared to generic signal models. To estimate the Vocal Tract Filter (VTF), using spectral division like in GSS, we show that a glottal source model can be used with any envelope estimation method conversely to ARX approach where a least square AR solution is used. We therefore derive a VTF estimate which takes into account the amplitude spectra of both deterministic and random components of the glottal source. The proposed mixed source model is controlled by a small set of intuitive and independent parameters. The relevance of this voice production model is evaluated, through listening tests, in the context of resynthesis, HMM-based speech synthesis, breathiness modification and pitch transposition.

Keywords: Mixed source; Glottal model; Vocal tract filter; Voice quality; Voice transformation; Speech synthesis

1. Introduction

For voice transformation as well as for speech synthesis, it is preferable to manipulate the perceived elements of the voice rather than to model all of the details of its production. For this purpose, the source-filter model offers an interesting decomposition scheme (Miller, 1959). Basically, this model represents the acoustic source coming from the glottis by a signal which is then filtered by the resonances and anti-resonances of the vocal tract structures, namely the Vocal Tract Filter (VTF) (see Fig. 2). In order to manipulate the elements of this model, their separation from an observed acoustic signal (i.e. the inversion of the model) is a necessary preliminary step. Using spectral division, the simplicity of the inversion of the source-filter model is also attractive. Indeed, to recover the source or the filter, the speech spectrum can be divided in the frequency domain by estimates of the VTF or the source spectrum respectively instead of using deconvolution of time series. Using this model, there are currently mainly two different approaches to transform a voice recording: On the one hand, a part of the original signal can be reused in the transformed signal. For example, combined with a smooth envelope estimate (e.g. Linear Prediction (Markel and Gray, 1976), "True-Envelope" (Roebel et al., 2007; Imai and Abe, 1979)), the phase vocoder preserves a part of the original phase spectrum in the transformed waveform (Flanagan and Golden, 1966). Additionally, the methods based on Pitch-Synchronous-OverLap-Add (PSOLA) assume that the signal inside a single window can be used without being modeled (Valbret et al., 1992; Hamon

* Corresponding author. Tel.: +30 2810 391580.
E-mail address: gilles.degottex@ircam.fr (G. Degottex).

[1] Present address: Computer Science Department, University of Crete, 71409 Heraklion, Crete, Greece.

[2] Present address: Engineering Department, Cambridge University, Cambridge, CB2 1PZ, UK.

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.08.010

et al., 1989). In the following, these methods will be termed *modification methods*. On the other hand, in *encoding/decoding methods*, the speech waveform is fully encoded into a small set of parameters. The model is built so as this set is optimal in terms of information compression, optimal in terms of reconstruction of the perceived elements or meaningful in the control of its elements. For example, a speech segment can be parametrized using a set of sinusoids (McAulay and Quatieri, 1986) which can also be harmonics or quasi-harmonics in the case of monophonic signals (Pantazis et al., 2010; Stylianou, 1996). This segment can be also represented using a wideband spectrum where smooth envelopes of the amplitude and phase spectra have to be estimated (e.g. WBVPM (Bonada, 2008), STRAIGHT (Banno et al., 1998; Kawahara et al., 1999)) or modeled using a formant representation (Rodet et al., 1984). Finally, many encoding/decoding methods using a glottal model, an analytical formulation of the glottal pulse (see Fig. 1), have been proposed to represent the deterministic component of the glottal source (e.g. AutoRegressive eXogenous input (ARX) methods) (Agiomyrgiannakis and Rosec, 2008; Vincent et al., 2007; Hedelin, 1984); Glottal Spectral Separation (GSS) (Cabral et al., 2008, 2011, 2010). In addition to the deterministic component, the vocal tract is also excited by aspiration noise which appears mainly in high frequencies. In both sinusoidal methods and methods based on glottal model, the noise component can be modeled using an amplitude modulated Gaussian noise convolved by an AR envelope (Agiomyrgiannakis and Rosec, 2008; Laroche et al., 1993). For wideband spectrum models, the noise can be also segmented in multiple frequency bands using a measure of aperiodicity (Kawahara et al., 2001; Banno et al., 1998; Griffin and Lim, 1988). To circumvent the lack of precision of glottal models and also to model the noise at the same time, hybrid models using for example ARX and harmonic models have been proposed (Agiomyrgiannakis and Rosec, 2009; Vincent et al., 2007).

Current modification methods achieve excellent results in voice transformation, especially for time stretching. However, in the case of important transformation (e.g. one octave pitch transposition), artifacts often appear showing underlying limitations of the models. Indeed, one can expect that modification methods are less sensitive to modeling errors by keeping part of the original signal unchanged. This unmodelled part limits however the flexibility of the modification methods. For example, using PSOLA the VTF is not explicitly modeled, since the impulse response of the VTF is forced to decay by the windows which are especially short (2 periods length). The drawback of this method is therefore the lack of resonances in downward pitch transpositions. Conversely, although encoding/decoding methods can be more sensitive to estimation error of their parameters, they should be more flexible. Indeed, only a full modeling of the speech signal can allow a full control of its perceived elements. Although the modification and encoding/decoding methods cited above are applied to voice processing, most of these methods could be applied to any pseudo-periodic signal. One can therefore expect that a model which is more dedicated to voice production better respects some physiological or acoustic constraints. For example, it is interesting to take into account the amplitude spectrum of the glottal source for the estimation of the VTF contrary to most of the current methods which assume that the voice source is made of a flat amplitude spectrum. Accordingly, the methods using glottal models have been proposed (e.g. ARX and GSS methods). However, ARX methods are far from straightforward to implement and depend on a reliable estimation of the glottal model parameters which are high-level

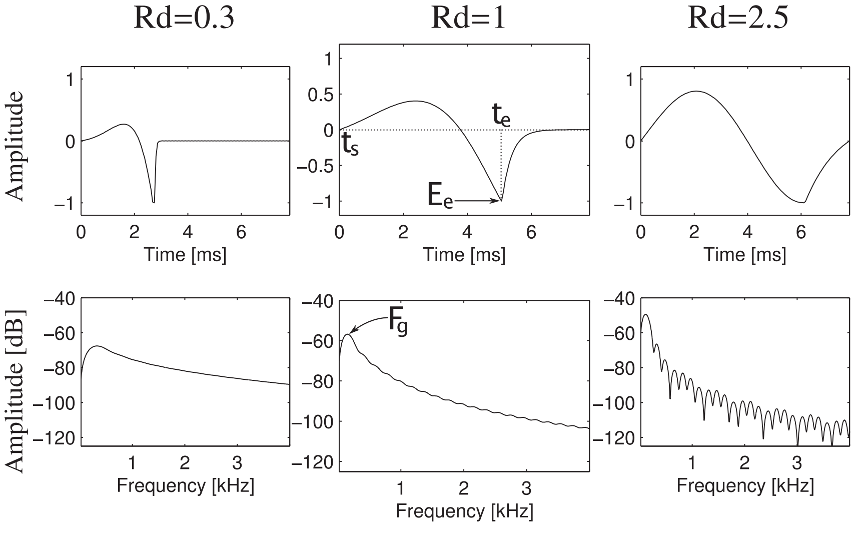

Fig. 1. Examples of the time-derivative of the glottal pulse represented by the transformed Liljencrants–Fant glottal model. Top plots show the temporal shapes and bottom plots show the corresponding amplitudes spectra with three *Rd* parameters corresponding to tense (Rd = 0.3), normal (Rd = 1) and lax sources (Rd = 2.5). The glottal closure instant is shown by t_e, the opening instant by t_s and the frequency of the glottal formant by F_g.

descriptors of the voice source and thus sensitive to inversion errors (Cabral et al., 2011; Degottex, 2010; Agiomyrgiannakis and Rosec, 2008). Therefore, the transformation and synthesis of the voiced signal using a glottal model is still a challenging question.

According to the above arguments, we developed and present in this article an encoding/decoding method using a glottal model. The Transformed Liljencrants–Fant glottal has been used to represent the deterministic component of the source whose shape is parametrized by a single shape parameter *Rd* conversely to the original version which uses 3 shape parameters (Fant, 1995), see Fig. 1. Our work was focused on using a glottal model and not on glottal models themselves. We therefore chose the widest used and studied glottal model, the Liljencrants–Fant (LF) model (Fant et al., 1985). Additionally, according to the difficulties encountered in parameters estimation of glottal models, we also decided to work with a meaningful reduced shape space using the *Rd* parametrization which is, according to Fant (1995), the most effective parameter to describe voice qualities into a single value. Testing various models and parameter configurations should be investigated in a dedicated study. To represent the random component of the source, zero-mean Gaussian noise is used. During synthesis, this noise is also amplitude modulated to improve its naturalness. Since deterministic and random components have different spectral properties, we also adapted the estimation of the VTF by taking into account this mixed source model. The whole procedure is called *Separation of the Vocal tract with the Liljencrants–fant model plus Noise* (SVLN).

Compared to the state of the art, the following points can be noticed. ARX methods jointly estimate the glottal parameters together with the VTF model parameters (e.g. an all-pole model). Conversely, in the proposed SVLN method, the glottal model parameters are first estimated in order to obtain an estimate of the glottal source spectrum, then the VTF is estimated by means of spectral division. Basically, the chosen approach is therefore very similar to the GSS method, where the spectral envelope of the signal is first obtained using STRAIGHT and the VTF estimate is then retrieved by means of spectral division using a glottal model. Compared to the ARX approach, spectral division is particularly promising since it allows to use any spectral envelope method independently of the method to estimate the parameters of the glottal model, and thus better separates the problems related to the estimation of the source and that of the VTF. Two main differences also exist between GSS and SVLN. Concerning the deterministic component, even though the LF glottal model is used in both methods, GSS uses the full parameter set (open-quotient, asymmetry and return phase) whereas SVLN uses the reduced version parametrized by *Rd*. The noise component in voiced segments is also modeled differently. In GSS, the aperiodicity measurement provided by STRAIGHT (Kawahara et al., 2001) is first used to generate a weighting function across frequency which then balances the deterministic and random components. Conversely, SVLN splits the spectrum in only two frequency bands using a Voiced/Unvoiced Frequency (VUF) (Drugman et al., 2009b). The lower band contains mainly the LF model and the upper band contains mainly Gaussian noise. Compared to GSS, SVLN simplifies therefore both representations of the deterministic and random components. In this article, we will investigate if this reduction plays an important role in the quality provided by these methods. Additionally, using SVLN, we will show results regarding the possibility to modify the breathiness of a voice as well as to transpose the pitch of an utterance.

Some parts of this work have been already presented to conferences and also in the first author's Ph.D thesis (Degottex et al., 2011b; Degottex, 2010; Lanchantin et al., 2010). In this article, we encapsulate the innovative technical content of these works, we show results of listening test carried out especially for this article to evaluate the proposed method and we finally share our conclusions about voice processing using a glottal model. The next section presents the voice production model used in SVLN and its separation process, the estimation of its parameters. Follows the description of the overlap-add technique for the synthesis step. Finally, the SVLN method is evaluated by means of listening test with comparison to state of the art methods. Four different evaluation contexts are presented: resynthesis (a simple encoding/decoding procedure), HMM-based synthesis, breathiness transformation and pitch transposition.

2. The voice production model

The segments of the speech signal are assumed to be stationary in a short analysis window $w_a[t]$ (of 3.5 periods in voiced parts with a minimum of 10 ms and a fixed length of 15 ms in unvoiced segments (fricatives, plosives, silence, etc.)). A Blackman window will be used during the analysis step. Moreover, the signal is assumed to be periodic in voiced segments, where the vocal-folds vibrate. Using the source-filter model in the frequency domain, we therefore model an observed speech spectrum $S(\omega)$ computed by the Fourier transform of the windowed signal as follows (see also Fig. 2):

$$S(\omega) = \left[H^{f_0}(\omega) \cdot G^{Rd}(\omega) + N^{\sigma_g}(\omega)\right] \cdot C^{\bar{c}}(\omega) \cdot L(\omega) \qquad (1)$$

where:

$H^{f_0}(\omega)$ is the harmonic structure modeling a periodic impulse train of fundamental frequency $f_0 : H^{f_0}(\omega) = \sum_{k\in\mathbb{Z}} e^{j\omega k/f_0}$.

$G^{Rd}(\omega)$ represents the shape of the deterministic component of the glottal source in a single period, the Transformed Liljencrants–Fant glottal model. This shape is parametrized by *Rd* (Fant, 1995) and its amplitude is

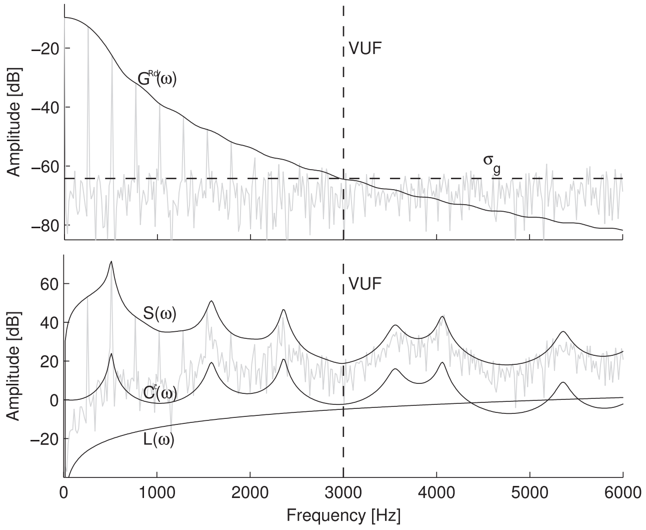

Fig. 2. Schematic representation of the used model using synthetic spectra: the glottal source model above and the full voice production model below. The spectra of one period and multiple periods are shown in black and gray lines, respectively.

parametrized by E_e at the instant t_e (see Fig. 1). Note that the original definition of this model includes a time-derivative representing the lips radiation. In the following, *glottal model* or *LF model* stands therefore for the integral of the LF formulas.

$N^{\sigma_g}(\omega)$ is the random component of the glottal source generated by aspiration noise at the glottis level. This noise is assumed to obey a Gaussian distribution of standard-deviation σ_g in the time domain. $|G^{Rd}(\omega)|$ decreasing monotonically and the noise level being assumed to be constant, $|G^{Rd}(\omega)|$ and σ_g cross at a point called Voiced/Unvoiced Frequency (VUF) in the following. Even though this frequency limit appears at the source level (Fig. 2 top), the VUF appears at the same frequency at the speech signal level (Fig. 2 bottom) since the VTF and the radiation effects are linear filters. The glottal noise has been shown to be amplitude modulated (Mehta and Quatieri, 2005; Hermes, 1991). However, we will see that this modulation does not play an important role in the estimation of the SVLN parameters (and thus not shown in Fig. 2). Nevertheless, during the synthesis step, this noise will also be modulated and colored in order to improve its naturalness.

$C^{\bar{c}}(\omega)$ is the Vocal Tract Filter (VTF) representing the resonances and anti-resonances of the vocal tract. This filter is assumed to be minimum-phase and parametrized by a vector of cepstral coefficients $\bar{c}$.

$L(\omega)$ is the filter corresponding to the radiation at the lips and nostrils level. We assume that this radiation can be modeled using a simple time derivative and therefore $L(\omega) = j\omega$ (Markel and Gray, 1976).

Consequently, the speech signal is parametrized by $\{f_0, Rd, E_e, \sigma_g, \bar{c}\}$ and can be fully encoded using this parameter set.

3. The analysis step: speech signal encoding

For a given speech utterance, the parameters of the voice production model are estimated at regular intervals of 2.5 ms. First, we assume that the spectrum of the glottal source can be split into a deterministic frequency band and a random frequency band using a Voiced/Unvoiced Frequency (VUF) (see Fig. 2) (also known as maximum voiced frequency). This VUF is also assumed to be known *a priori* thanks to existing methods (Kim and Hahn, 2007; Stylianou, 2001). In the presented study, this value is estimated by determination of Voiced/Unvoiced Frequency bands (Stylianou, 2001 [p. 3]) by means of peak classification of the speech spectrum (Zivanovic et al., 2008). Compared to a multi-band source model (Griffin and Lim, 1988) or a Harmonic+Noise Model (HNM) (Stylianou, 1996), this decomposition in only two separated frequency bands is obviously an important simplification of the voice source. Keeping in mind this reduction, we will see that such a simplification leads to a convenient estimation of the noise level of the glottal source in the next sections.

3.1. Deterministic source parameters: f_0, Rd, E_e

Numerous methods exist to compute f_0 from the speech signal. In the presented experiments, the YIN method is used (de Cheveigne and Kawahara, 2002).

To estimate the shape parameter Rd of the LF model, the recently proposed method based on Minimum Squared Phase with 2nd order Difference operator (MSPD2) is used (Degottex et al., 2011a). Basically, this method first represents the speech signal using a harmonic model. Then, both glottal and speech spectra are divided by their minimum-phase version to retrieve their minimum-phase residuals. Finally, a local search algorithm finds the best Rd value which minimizes the difference between the minimum-phase residuals (Degottex et al., 2011a; Degottex, 2010). Obviously, other methods can be used to estimate Rd like those estimating the glottal source based on maximum-phase and minimum-phase separation (through complex cepstrum or ZZT (Drugman et al., 2009a; Oppenheim et al., 1968)) or using the IAIF method (Alku et al., 1999). The estimation of glottal parameters is far from straightforward and many questions remain about the parameters range where they can be estimated in a reliable way (Degottex, 2010). In SVLN, we therefore used $Rd \in [0.3; 2.5]$ according to our previous studies (Degottex et al., 2011a; Degottex, 2010).

Concerning the amplitude of the glottal model E_e, when the VUF estimate is smaller than the f_0 estimate, E_e is set to zero, defining therefore the voiced and unvoiced segments of the analyzed signal. When the VUF is higher than f_0, the definition of E_e is actually not straightforward. Indeed, three gains co-exist in the voice production model: E_e, σ_g and the mean log amplitude of the VTF. These gains are completely dependent on each other. If E_e and σ_g are multiplied by some arbitrary value α, the VTF mean log amplitude may compensate α leading to the same gain of the observed spectrum (with $-log(\alpha)$). Consequently, a constraint is necessary. Here, the mean log amplitude of the VTF is fixed to zero. The energy variation of the speech signal is thus only modeled by the energy of the glottal source model (given by σ_g and E_e). In SVLN, E_e is therefore defined from a convention which implies the following two points. Firstly, no method for Glottal Closure Instant (GCI) detection is necessary. Conversely to the GSS method and ARX methods where the time synchronization between the LF model and the underlying glottal source (i.e. the GCIs) have to be estimated, the proposed SVLN method needs only an estimate of the shape parameter Rd. (which can be estimated without GCI detection as shown in (Degottex et al., 2011a)). Secondly, the resulting computation of E_e cannot be considered as an estimation of the actual amplitude of the glottal pulse. The ratio between E_e and σ_g represents only the ratio between noise and deterministic component.

3.2. Random source parameter: σ_g

Using the hypothesis of separability of the speech spectrum in two different frequency bands, the amplitude spectrum $|G^{Rd}(\omega)|$ crosses the expected amplitude of the noise at the VUF (see Fig. 2). Since $|G^{Rd}(\omega)|$ is known when the f_0 and Rd estimates are known, the noise level σ_g can be deduced from the VUF:

$$\sigma_g = |G^{Rd}(\text{VUF})| \cdot \frac{\sqrt{2}}{\sqrt{\pi/2} \cdot \sqrt{\sum_t w_a[t]^2}} \qquad (2)$$

where $|G^{Rd}(VUF)|$ is the expected amplitude of the LF model at the VUF which has to be converted to the Gaussian parameter σ_g: spectral amplitudes of Gaussian noise obey a Rayleigh distribution. $|G^{Rd}(VUF)|$ is thus first converted to the Rayleigh mode $(1/\sqrt{\pi/2})$, then the standard deviation of the Gaussian distribution in the time domain is retrieved from the Rayleigh mode $(\sqrt{2})$ (Yeh, 2008). Additionally, in the spectral domain, the noise level is proportional to the energy of the analysis window $w_a[t]$ used to compute $S(\omega)$. The normalization by $\sqrt{\sum_t w_a[t]^2}$ is therefore necessary. Fig. 3 illustrates estimates of source parameters.

3.3. The estimation of the Vocal Tract Filter (VTF)

In SVLN, according to the difference of the underlying source properties, the frequency bands below and above the VUF are modeled using two different envelopes (see Fig. 2). For the sake of simplicity, to estimate the VTF, we therefore assume that the deterministic and random components of Eq. (1) can be represented separately:

$$S(\omega) = \begin{cases} H^{f_0}(\omega) \cdot G^{Rd}(\omega) \cdot C^{\bar{c}}(\omega) \cdot L(\omega) & \text{for } \omega < \text{VUF} \\ N^{\sigma_g}(\omega) \cdot C^{\bar{c}}(\omega) \cdot L(\omega) & \text{for } \omega > \text{VUF} \end{cases} \qquad (3)$$

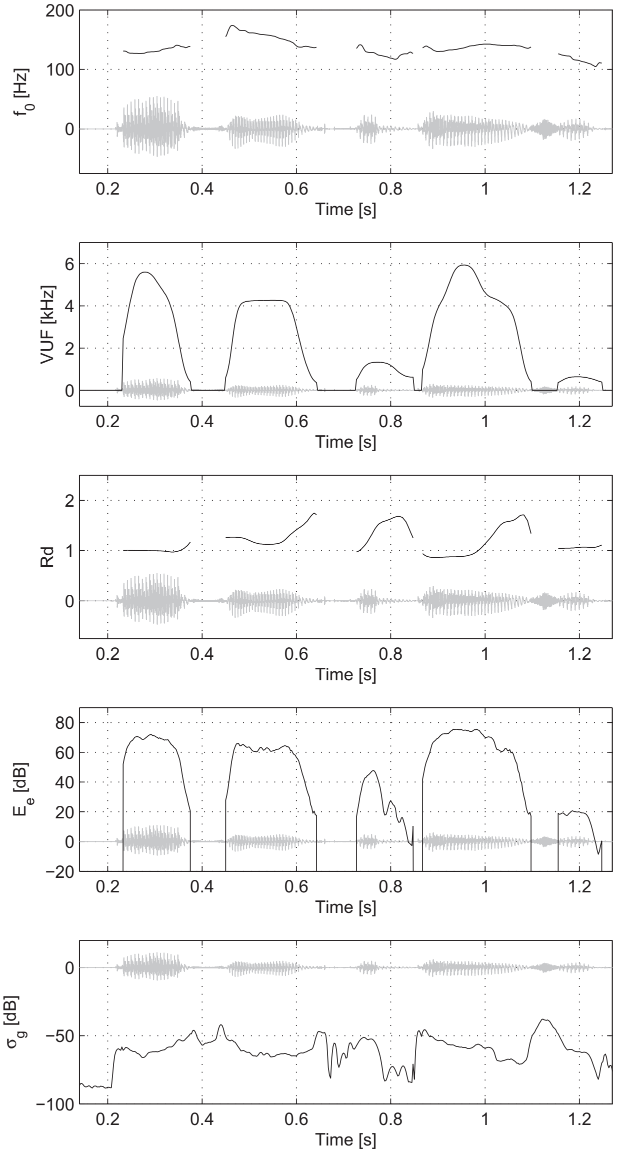

Fig. 3. An example of parameter trajectories of an American male utterance: "Author of the danger". The parameter is shown in black line and the waveform in gray (which is scaled for the sake of clarity).

The envelopes estimated on each part of Eq. (3) are then aligned to ensure a VTF estimate which is independent of the nature of the source. The envelopes estimation and the alignment is described here below.

In the deterministic band, where $\omega <$VUF, the contribution of the radiation $L(\omega)$ and the deterministic source $G^{Rd}(\omega)$ are removed from $S(\omega)$ by spectral division (see Eq. (4)). An iterative cepstral envelope $\mathcal{T}(.)$ (called *true-envelope*, (Roebel et al., 2007; Imai and Abe, 1979)) is then used to fit the top of the harmonics of the division result. Note that this envelope corresponds to the expected amplitude of the VTF frequency response since the top of a harmonic is the expected amplitude of its corresponding sinusoidal component.

$$T(\omega) = \mathcal{T}\left(\frac{S(\omega)}{L(\omega) \cdot G^{Rd}(\omega)}\right) \cdot \frac{1}{\gamma} \qquad (4)$$

where $\gamma = \sum_t w_a[t]/(f_s/f_0)$ stands for the number of periods in the analysis window. This normalization is necessary regarding to the synthesis step where the VTF is convolved with each period of the source. The gain of the estimated VTF has to be normalized according to the shape and the duration of the analysis window.

In the random band, where $\omega >$VUF, $S(\omega)$ is divided by $L(\omega)$ and by the crossing value $|G^{Rd}(\mathrm{VUF})|$ to ensure a continuity between the two frequency bands. The result of this division is modeled by computing its real cepstrum $\mathcal{P}(.)$ truncated to a given order (discussed below). According to the Rayleigh distribution of the spectral amplitudes of this band, the mean log amplitude measured by $\mathcal{P}(.)$ has to be converted to the Rayleigh mode on a linear scale (factor $e^{0.058}$ in Eq. (5) below) (Yeh, 2008). Then, the expected amplitude is retrieved from the Rayleigh mean value ($\sqrt{\pi/2}$).

$$P(\omega) = \mathcal{P}\left(\frac{S(\omega)}{L(\omega) \cdot G^{Rd}(\mathrm{VUF})}\right) \cdot \frac{\sqrt{\pi/2}}{\gamma \cdot e^{0.058}} \tag{5}$$

To obtain the final VTF estimate $C(\omega)$, the two envelopes $T(\omega)$ and $P(\omega)$ have to be aligned. $T(\mathrm{VUF})$ and $P(\mathrm{VUF})$ cannot be perfectly equal due to their different estimation methods. Therefore, a smooth transition has to be ensured to avoid artifacts in the synthesis. For this reason, the envelopes are cross-faded in the frequency domain using a weighting function:

$$C(\omega) = T(\omega) \cdot (1 - W(\omega)) + P(\omega) \cdot W(\omega) \tag{6}$$

where $W(\omega)$ is a sigmoid function whose inflection point is centered on VUF and the slope in the transition band is of 140 dB/kHz which has been chosen empirically. Finally, the cepstral coefficients $\bar{c}$ of the VTF are retrieved from the minimum-phase cepstrum of $C(\omega)$ to represent the VTF with a small and meaningful set of parameters.

It is worth mentioning the three following technical details. Firstly, concerning the order of the envelopes, it is necessary that $\mathcal{T}(.)$ and $\mathcal{P}(.)$ do not fit the harmonic structure of the observed spectrum $S(\omega)$. For $\mathcal{T}(.)$, the optimal order $0.5 \cdot f_s/f_0$ is used (Roebel et al., 2007). The same order is also used for the cepstral envelope $\mathcal{P}(.)$. Indeed, although no harmonic partial appears in the frequency band of the random source, sinusoidal peaks with distance of f_0 (but not multiples of f_0) arise in this band because the glottal noise is amplitude modulated by the glottal area (Mehta and Quatieri, 2005; Hermes, 1991) (such peaks are visible in Fig. 4 around 9 kHz). Secondly, the division by $L(0) = 0$ has to be avoided in Eqs. (4) and (5). $L(0)$ can be either extrapolated from $L(\omega)$ or $j\omega$ can be replaced by $1 - \mu e^{j\omega}$ with μ close to unity. In this work $L(0)$ has been extrapolated. Finally, the amplitude spectrum of the observed speech signal $|S(\omega)|$ is almost perfectly represented by the SVLN method. Indeed, the estimation of the VTF always completes the source and radiation models in order to obtain $|S(\omega)|$. The phase spectrum can however be modeled only by the LF model, Gaussian noise and the minimum-phase property of the VTF. Therefore, in the context of a simple encoding and decoding of the voice (without transformation), a bias of the Rd value implies an error of resynthesis of the phase spectrum only. In terms of stability, this robustness related to the shape parameter is interesting regarding the risk incurred by the estimation of high-level descriptors like Rd.

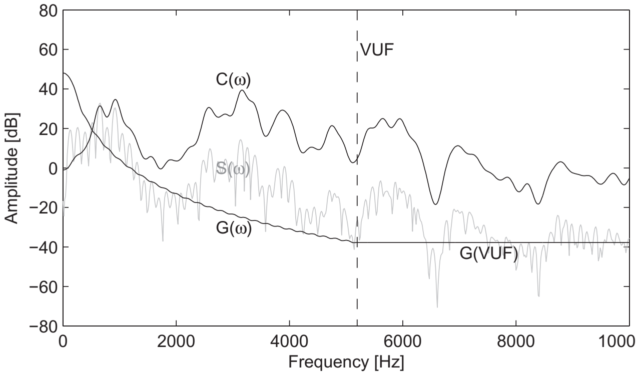

Fig. 4. An example of VTF estimate. The glottal model and the VTF estimate are in black lines and the speech spectrum is in gray.

4. The synthesis step: speech signal decoding

This section describes the synthesis of a speech utterance given a parameters set. Small segments of stationary signals are first synthesized and these segments are then overlap-added to construct the whole signal. Follows, the definition of a segment, the synthesis of its content and the final concatenation.

4.1. Segment position and duration

In voiced parts, temporal marks m_k are placed at intervals according to the fundamental period $1/f_0$ (see Fig. 5), one mark for each segment. The maximum excitation instant t_e (see Fig. 1) of each LF pulse is placed at m_k. Then the starting time t_k of the kth-segment is defined as the opening instant t_s of the LF model and the ending time of this segment is the starting time of the next. In unvoiced parts, a segment has a 5 ms duration, and its mark m_k is placed in the center, as illustrated in Fig. 5.

4.2. The noise component: filtering, modulation and windowing

For all segments, noise is generated. To improve its naturalness, the following post-processing steps are used. Firstly, the lowest frequencies of the aspiration noise are weaker than higher frequencies (Stevens, 1971). If the noise is white in the voiced segments, the synthesized voice sounds hoarse because the noise randomizes the lowest harmonics of the deterministic component. The noise is therefore filtered with a high-pass filter $F_{hp}^{\mathrm{VUF}}(\omega)$ defined by a cutoff frequency equal to the VUF and a slope of

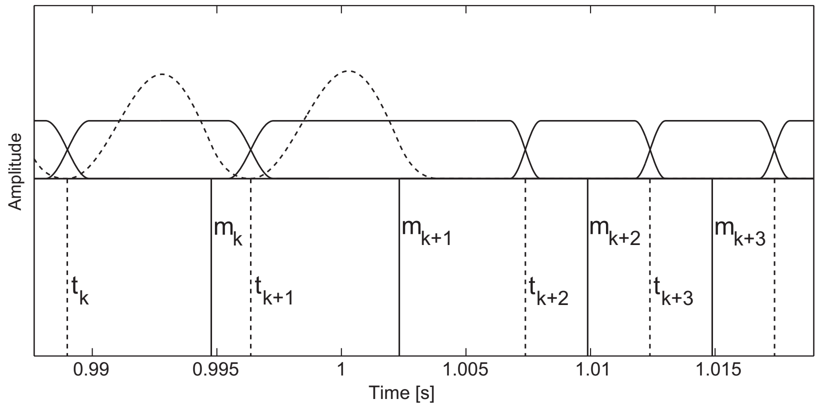

Fig. 5. Example of two voiced segments followed by two unvoiced segments during the synthesis step: marks m_k and starting times t_k are shown with vertical lines. Synthesized LF pulses are in dashed lines, and windows $w_k[t]$ are in solid lines.

6 dB/kHz in the transition band. Since the VUF is only used to estimate σ_g and is not part of the model parameters, this value is retrieved from the intersection of σ_g and $G^{Rd}(\omega)$ in the synthesis step. Secondly, the time amplitude of the aspiration noise depends on the glottal flow and the glottal area. If the glottal noise is not amplitude modulated synchronously with the fundamental period, a second source is perceived separately from the deterministic source (Agiomyrgiannakis and Rosec, 2009; Mehta and Quatieri, 2005; Hermes, 1991). Accordingly, a modulation $v^{Rd}[t]$ is built from the LF pulse as proposed by del Pozo and Young (2008):

$$v^{Rd}[t] = \beta \cdot g^{Rd}[t] + (1-\beta) \tag{7}$$

where $g^{Rd}[t]$ is the LF pulse with voicing amplitude $A_v = 1$ and β a constant balancing the quantity of modulated and constant noise. Here, the Rd parameter is set to the same value as the one of the deterministic source. Then, from informal listening of 10 different voices and their corresponding resynthesis, we fixed the value $\beta = 0.75$ according to the naturalness of the resynthesis. Obviously, if these two values were properly estimated from the observed signal, the naturalness of the synthesized noise could be improved (Mehta and Quatieri, 2005).

No window is necessary to cross fade the glottal pulses since they start and end at zero amplitude. However, the noise is generated continuously across the signal and this noise can have different color and amplitude between segments. For each k-segment, a window $w_k[t]$ is therefore built with a fade-in center on t_k and a fade-out center on t_{k+1} (see Fig. 5). The fade-in/out function is a Hanning half window of duration $0.25 \cdot \min(t_{k+1} - t_k, t_k - t_{k-1})$. Additionally, the fade-out of w_k is the complementary of the fade-in of w_{k+1} and the sum of all windows is 1 at any time of the synthesized utterance. Once the synthesized segments of speech are overlap-added at the end of the synthesis step, it is therefore not necessary to normalize the result by the sum of the windows. According to the discussion above, the noise spectrum of the kth segment in voiced segments is synthesized by:

$$N_k(\omega) = \mathcal{F}\Big(w_k[t] \cdot v^{Rd_k}[t] \cdot \mathcal{F}^{-1}\Big(F_{hp}^{\mathrm{VUF}_k}(\omega) \cdot N^{\sigma_{gk}}(\omega)\Big)\Big) \tag{8}$$

where $N^{\sigma_{gk}}(\omega)$ is the spectrum of zero-mean Gaussian random signal $n^{\sigma_{gk}}[t]$ and $\mathcal{F}(.)$ is the Fourier transform. In unvoiced segments, the noise source reduces to:

$$N_k(\omega) = \mathcal{F}(w_k[t] \cdot n^{\sigma_{gk}}[t]) \tag{9}$$

4.3. The glottal pulse and the filtering components

In this last step, the deterministic source $G^{Rd_k}(\omega)$ is added to the noise and the VTF and radiation filters are applied to the source:

$$S_k(\omega) = \left[e^{-j\omega m_k} \cdot G^{Rd_k}(\omega) + N_k(\omega)\right] \cdot C^{\bar{c}_k}(\omega) \cdot j\omega \tag{10}$$

where $e^{-j\omega m_k}$ is a delay placing the instant t_e of the LF pulse at m_k and $C^{\bar{c}_k}(\omega)$ is the minimum-phase frequency response of the VTF corresponding to the cepstral coefficients $\bar{c}_k$. The entire signal is finally constructed by successively overlap-adding the time segments which are retrieved through the inverse Fourier transform of $S_k(\omega)$. Note that the sum of the windows $w_k[t]$ being always equal to one, it is not necessary to use any other extra windows in this overlap-add process.

As a last technical detail, the analysis and synthesis steps are not perfectly symmetric. Indeed, according to the estimation of the VTF (Eqs. (4)–(6)), one may expect that $G^{Rd_k}(\omega)$ is low-pass filtered like $N^{\sigma_{gk}}(\omega)$ is high-pass filtered. However, according to informal listening, we were not able to notice any difference with or without low-pass filtering of $G^{Rd_k}(\omega)$. For the sake of simplicity, this filtering has been therefore discarded. A reason might be that the bias introduced by $G^{Rd_k}(\omega)$ is smooth across frequency ($G^{Rd_k}(\omega)$ decreasing monotonically) so as the frequencies above VUF are only slightly increased.

5. Evaluation

Listening tests have been carried out to evaluate various properties of SVLN compared to state of the art methods.

Before discussing the results, for the sake of precision, we first discuss the influence of the parameters set on the quality of SVLN and describe globally the used listening tests.

5.1. Used features

According to our experiments, irregularities of the *Rd* estimate are observed whatever the used estimation method (e.g. using MSPD2, ZZT or IAIF). Additionally, the stability of the separation process of SVLN across adjacent frames is linked to the stability of the *Rd* parameter. To ensure a stable estimate of the VTF and avoid audible artifacts, it is therefore necessary to remove possible erratic values in the estimated *Rd* curve. We therefore filtered this curve using a median filter. Then, using a Hanning window, a zero-phase filter is used to smooth the steps made by the median filtering. However, over-smoothed or erratic *Rd* values as well as a lack of flexibility of the glottal model to represent the actual shape of the glottal pulse have some consequences. Indeed, in any method using a glottal model, the used glottal model may not filter out properly the amplitude spectrum of the actual glottal pulse during the estimation of the VTF (Eq. (4) for SVLN). The spectral difference between the actual pulse and its model therefore remains in the estimated VTF. For example, a remaining glottal formant tends to generate an additional erroneous low-frequency resonance in transformed voices. It is therefore important to avoid erratic behaviors and over-smoothing of the *Rd* curve at the same time. Consequently, we used a window length of 100 ms according to informal listening. In doing so, we implicitly assume that the voice quality is almost constant inside the duration of a single phoneme.

The VUF has also an impact on the synthesized voice. If the VUF is underestimated, noise is generated at low frequencies, and the synthesized voice sounds hoarse. Conversely, if the VUF is overestimated, the voice may sound buzzy. The voicing decision in the time domain is equally critical for a proper reconstruction of the transients. If a plosive is classified voiced during the analysis step, the source at low frequencies will be generated by the LF model which will create a bubble-like artifact.

Finally, in the following tests, some methods have common features (e.g. f_0 for STRAIGHT, GSS and SVLN). To ensure that the estimation of these features do not influence the results, the same data have been used across all methods. The octave errors of the f_0 estimate were also corrected manually. The VUF estimator is used initially to determine the voicing decision in the time domain. A given time is voiced if VUF$> f_0$. To avoid that errors of the voicing decision influence the results, the VUF values have been manually corrected based on the inspection of the speech waveform. The start and end of voiced segments are set to the first and last glottal closure respectively (voiced fricatives being considered as voiced). If a VUF value was initially zero in a voiced segment, it has been set to $4 \cdot f_0$. According to informal tests, this default value provides a satisfactory resynthesis quality.

5.2. Design of the listening tests

The listening tests have been conducted according to crowd-sourcing using web pages. Basically, listeners are invited to visit a web page where audio files have to be evaluated following basic recommendations. For this evaluation, we sent the tests to two mailing lists (`AUDITORY@lists.mcgill.ca` and `parole@ml.univ-avignon.fr`) and to a personal contact list of musicians and researchers also in audio or speech community. The first language of the listeners are therefore mainly English or French. However, people of German, Greek and Spanish language have also answered the tests. Web-based tests have advantages as well as drawbacks. Compared to a local test carried in a single place where the population is mainly made of native speakers of one language, they allow to cover a wider population of listeners. Nevertheless, the listening condition of web-based tests can not be fully controlled. A controlled context in an anechoic chamber would be mandatory to carry out evaluation of human perception. However, the presented study targets applications used in natural environment. We therefore consider that some variability in the listening conditions is interesting to, at least, avoid bias due to the listening material (e.g. due to the headphones). However, to ensure minimum conditions, it was also recommended to use absolutely headphones or earphones. At the end of the test, the listeners were asked if they used headphones, earphones or loudspeakers and all answers made using loudspeakers were discarded. Also, if any technical problem arises with an audio file, the listener had the possibility to indicate the problematic file and the corresponding answer was discarded.

For any listening test (web-based or not), some constraints have to be respected. First, the focus of the listeners degrades quickly after 15 min which limits the length of the tests and thus the number of tested utterances. The duration of the utterances were between 3 s and 5 s with a sampling rate of 44.1 kHz and each utterance was produced by a different speaker in American English, French, Japanese and Greek to ensure some speaker variability. In each test, there were always the same number of female and male voices. Note that the audio files used in these tests can be found at `gillesdegottex.eu/ExDegottexG2012svln`.

5.3. Evaluation of resynthesis

This first test evaluates the quality of the resynthesis, the reconstruction of the speech signal from the model parameters, without transformation or further modeling. Two state of the art methods are included in this test: the Glottal Spectral Separation (GSS) (Cabral et al., 2011, 2008; Cabral, 2010) (provided by the author) and STRAIGHT (Kawahara et al., 2001, 1999) (version `V40pcode`). Basically, STRAIGHT uses the standard source-filter model where the filter is a minimum-phase spectral envelope and the source is a weighted sum in the frequency domain between a Dirac impulse and noise. GSS can be seen as an intermediate model between SVLN and STRAIGHT because GSS uses the spectral envelope of STRAIGHT and replace the source by the LF model, the phase spectrum being randomize by noise as in STRAIGHT. SVLN also uses the LF model but, conversely to GSS, it adapts the spectral envelope estimation to the underlying nature of the two frequency bands above and below the VUF.

In this test, for each recording (among a total of 8 utterances made of 4 languages with female and male voices), listeners were asked to grade the quality of resynthesis according to the recommendation ITU-R BS (Assembly, 2003): Excellent (5), Good (4), Fair (3), Poor (2), Bad (1). Additionally, in breathiness and pitch transposition, only the voiced segments have to be modified. The original recording can be therefore kept unchanged in unvoiced segments. In this resynthesis test, two resynthesized audio files were therefore proposed for each method: one version with only the voiced segments resynthesized and another version with the whole utterance resynthesized. Finally, to check the consistency of the answers, the original recordings were also added in the audio files set.

20 listeners answered the test and Fig. 6 shows the results for each method averaging the scores among the 8 utterances. These results suggest the following two points. Firstly, for voiced segments only, the quality of STRAIGHT cannot be distinguished from the two others methods according to the confidence intervals. Since STRAIGHT will be compared to SVLN in pitch transposition in the last test of this evaluation, it ensures that the naturalness of the transpositions will be evaluated and the influence of the overall quality on the comparison will be minimized. Secondly, and most importantly, the sound quality provided by SVLN and GSS is clearly degraded when both voiced and unvoiced segments are resynthesized. Contrarily, we can not infer the same conclusion for STRAIGHT. When resynthesizing a full utterance, the quality provided by STRAIGHT is therefore clearly more stable than SVLN and GSS. Compared to STRAIGHT, the methods using a glottal model (i.e. ARX, GSS and SVLN) introduce indeed a new problem. In STRAIGHT, the voice source has always a flat and unity amplitude spectrum, in both voiced and unvoiced segments. In transients, the STRAIGHT envelope moves therefore between voiced and unvoiced frames without any specific adaptation of the underlying source properties. However, using a glottal model, the amplitude spectrum of the glottal source can change quickly inside a single analysis window since a glottal pulse has a non-flat amplitude spectrum. Ideally, the VTF estimate should be therefore adapted within the analysis window using a non-stationary analysis, which is not the case in SVLN or GSS. This difference between SVLN/GSS and STRAIGHT could therefore explain the quality difference between the full resynthesis and the resynthesis of the voiced segments only.

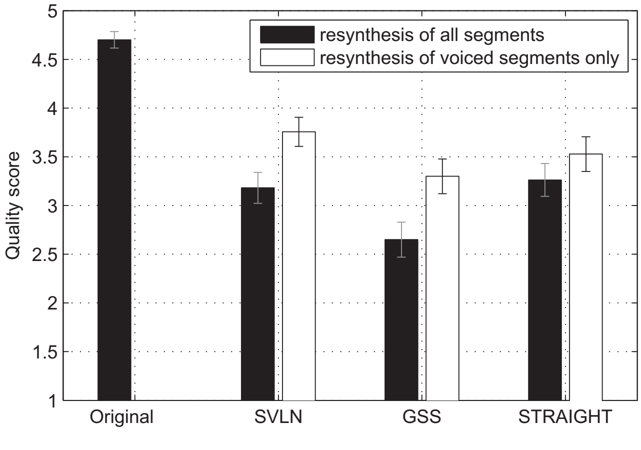

Fig. 6. Quality scores of the resynthesis according to a listening test with the 95% confidence intervals.

Looking at the results in more details, three additional elements can be noticed in this test. Firstly, the variance results across the utterances is substantial. The top plot of Fig. 7 illustrates this variability qualitatively using different bars for each utterance (resynthesizing the voiced segments only). Quantitatively, the bottom left plot shows quantitatively the estimated standard-deviation of this variance. The voice is made of elements of different nature: periodicity, creakiness, noise, resonances, etc. which are not balanced the same way in each voice. Since the methods do not represent each element of the voice with the same accuracy, the quality of resynthesis can not indeed be the same. Consequently, it is important to remember that no systematic improvement can be inferred from the listening tests while no method provides the same quality for any utterance in a resynthesis test. By showing differences of quality or preference, we show only trends,

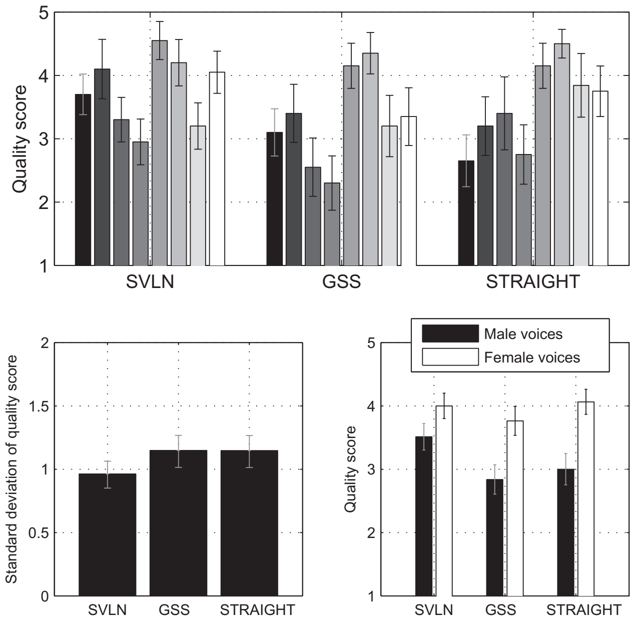

Fig. 7. Details of the quality scores. Top plot illustrates qualitatively the variance of the quality scores with respect to the used utterances and bottom left plot shows this variance quantitatively. Bottom right plot shows the detailed scores related to gender.

an average improvement for the used recordings. Secondly, according to the bottom right plot, the variance is smaller with SVLN across gender compared to GSS and STRAIGHT. Whereas the three methods provide the same quality for female voices, SVLN better reconstructs the male utterances than the two other methods. At last but not least, concerning the listening conditions of the test, in top plot of Fig. 7, some confidence intervals do not overlap (e.g. 4th and 5th bars). A score difference therefore exists between utterances whereas the listening conditions can be different. Although a local test may reduce the variance for each utterance, the variance across utterances will therefore remain.

5.4. Preference test for speech synthesis based on Hidden Markov Models (HMM)

Even though results on voice transformation are mainly presented in this study, a preliminary study has been also conducted about French speech synthesis using an HMM-based synthesis system (Zen et al., 2007) (HTS version 2.1.1). The main goal of this preliminary step is to provide useful information about the advantages and drawbacks of SVLN in the context of statistical modeling for any other future applications. We present the results of a preference test which evaluated and compared the efficiency of three encoding/decoding methods: SVLN, STRAIGHT and a basic method using impulses train for the source in voiced segments, Gaussian noise in unvoiced segments and amplitude spectral envelope for the VTF (Zen et al., 2007) (termed *impulse-source* method in the following).

For all compared methods the STRAIGHT method was used to estimate the f_0 curves and to compute the time domain voicing. For the baseline systems (STRAIGHT and impulse-source), the STRAIGHT method was used to extract the mel-cepstrum and to estimate aperiodicity. f_0 and aperiodicity parameters were used to generate the mixed-excitation and the mel-cepstral coefficients using a Mel Log Spectrum Approximation (MLSA) filter. Both orders of cepstral and aperiodicity coefficients were 30. For the SVLN method, in order to reduce the number of parameters in the learning procedure, the amplitude E_e was merged into the first cepstral coefficient of the VTF. To keep the relative level between the deterministic and random sources, the gain of the random source σ_g was therefore normalized by E_e. Finally, the cepstral coefficients were encoded using a mel scale like in the baseline methods.

The set of parameters were split into several independent streams and different configurations were tested. f_0 was modeled by a single Gaussian distribution for voiced parts, and the voiced/unvoiced decision was taken into account by a specific weight applied on each space of a Multi-Space Distribution (MSD) (Tokuda et al., 2002a). Knowing that Rd is only defined in voice segment of speech, Rd was first included in the same MSD stream as f_0, with full covariance matrix in order to also take into account the correlation between both parameters. Despite the fact that this configuration is conceptually better, using a configuration with the Rd value in the same stream as σ_g and $\bar{c}$ provides a slightly better quality according to informal listening. Although Rd is only meaningful in voiced segments, this parameter can technically be calculated for both voiced and unvoiced segments. During the analysis step of SVLN, σ_g is directly expressed from Rd and VUF. Moreover, $\bar{c}$ is also highly dependent on Rd and σ_g. Taking into account these dependencies in the statistical model can therefore play a significant role in the robustness of the synthesis. In the formal listening test, we therefore adopted the following configuration:

- One single Gaussian distribution with semi-tied covariance (Gales, 1999) for $\{Rd, \sigma_g, \bar{c}\}$;
- One multi-space distribution (Tokuda et al., 2002a) for f_0,

where both streams include first and second time derivatives of their parameters. Note that among the different tested configurations, we also tried to model the VUF instead of the noise level. However, in both cases the same artifacts were audible according to informal listening. To be consistent with the model, we therefore preferred to modeled the noise level.

In order to avoid unnatural discontinuities in the prosody and obtain co-articulation in a synthesized utterance, it is necessary to take into account the context of each phoneme. Therefore, contextual features are used to describe the phonetic, lexical and syntactic context of the phonemes. These contextual features, detailed in Table 1, have been automatically extracted from the speech recordings and their text transcriptions using *ircamAlign* (Lanchantin et al., 2008), an HMM-based segmentation system relying on the HTK toolkit (Young, 1994) and the French phonetizer Lia_phon (Bechet, 2001). For each utterance of the training set, the text was first converted into a phonetic graph with multiple pronunciation possibilities. Then, the best phonetic sequence was chosen according to the corresponding audio file and aligned temporally with it. The context features were finally extracted according to the aligned text and the extracted phonetic sequence. A 5-states left-to-right HSMM was finally used to model each contextual phoneme (Zen et al., 2004).

The training procedure was similar to the one described in (Tokuda et al., 2002b): monophones models were first trained and then converted to context-dependent models. Moreover, decision-tree clustering was performed according to the extracted context features in order to obtain reliable model parameters. During the synthesis step of each compared method, a parameter sequence was first generated using HTS with a constrained maximum likelihood algorithm (Tokuda et al., 1995). The same procedure was used for STRAIGHT and the impulse-source method using their respective parameters.

The compared synthesis systems have been trained on a database containing 1995 sentences (approximately 1h30 of

Table 1
Context features extracted by ircamAlign for the HMM-based speech synthesis.

Phonetic features:

- **Phoneme identity (SAMPA code)**, and the following phonological features: vowel (length, height, fronting, rounding) consonant (type, place, voicing) for the central phoneme and for its neighbors (2 before and 2 after)

Lexical and syntactic features

- **Phoneme and syllable structure**: position of the phoneme in its syllable; number of phonemes in the current, previous and next syllable; position of the phoneme in the word; position of the phoneme in the phrase; nucleus of the syllable
- **Word related**: Part Of Speech (POS) of the word and its neighbors (1 before and 1 after); number of syllables in the current, previous and next word; number of content words from the start and from the end of the phrase, number of non-content words up to the previous and next content word
- **Phrase related**: number of syllables in the phrase; number of words in the phrase; position of the phrase in the utterance
- **Utterance related**: number of syllables, words and phrases in the utterance
- **Punctuation related**: punctuation of the last phrase

speech) spoken by a French non-professional male speaker and recorded at 16 kHz in an anechoic room. 5 utterances were finally synthesized by each system and used as test samples. In the preference test, the listeners were asked to give a grade between −3 and 3 for a pair of audio files using two different systems by answering the question "which sound do you prefer". Comparing each method with each other, a total of 15 comparison pairs were evaluated by each listener.

14 French native listeners answered the test. Left plot of Fig. 8 shows the preference scores of each method compared to each other and the right plot shows the mean preference scores which are computed by averaging all grades among all comparisons (+N for each grade advantaging the method, and −N for each grade penalizing the method). Note that the differences of the parameters between the methods resulted in different clustering of the context features. This may generate slight prosodic differences between the methods and may alter the evaluation. The mean preferences (right plot) show that the speech synthesized by SVLN has a preference between that of STRAIGHT and that of the impulse-source method. Detailed preferences (left plot) show that SVLN is preferred compared to the impulse-source method. In a context of simple resynthesis, without HMM modeling, Cabral et al. (2008) have also shown that GSS is preferred against a synthesis without using noise in voiced segments. The detailed preferences show also that STRAIGHT is clearly preferred compared to SVLN. As seen in the resynthesis test, STRAIGHT provides similar quality scores between voiced and unvoiced segments conversely to SVLN and GSS (Fig. 6). Even though a resynthesis on a frame by frame basis may not reveal an overall instability of the separation method, a statistical modeling is sensitive to this stability. It seems therefore consistent that the STRAIGHT method provides indeed a better quality in HMM-based synthesis than SVLN. Additionally, Cabral et al. (2011) have shown that GSS is slightly preferred to STRAIGHT when mixing the LF model with noise in the context of HMM synthesis. Even though the resynthesis test shows that SVLN provides a better quality than GSS on a frame by frame basis, GSS can be more stable in statistical modeling and thus provides a better quality in HMM-based synthesis. Finally, Raitio et al. (2011) have shown interesting results using a glottal separation procedure which does not require any glottal model. Using their method the preference compared to STRAIGHT can be clearly increased. Using a more flexible model of the glottal

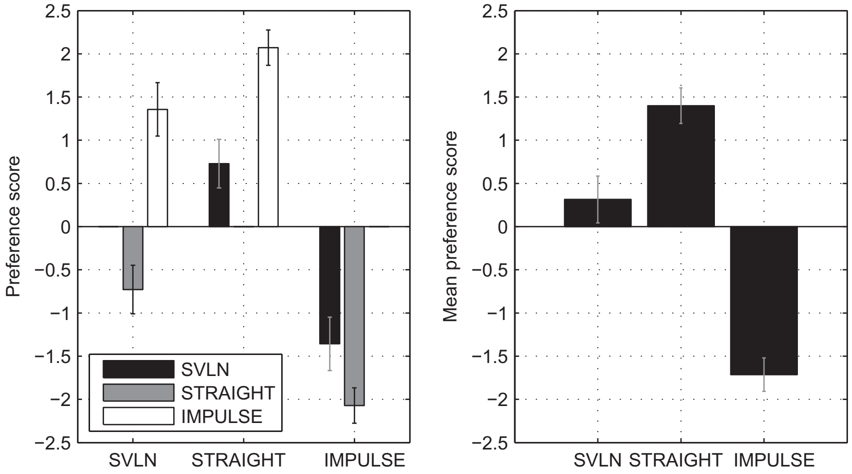

Fig. 8. Preference scores and their 95% confidence interval. Detailed preference scores of each method compared to each other to the left and mean preference scores to the right.

source than a glottal model, the stability of the separation method can be indeed better. Cabral et al. (2011) remarked instability of the separation of GSS due to the estimation of the glottal parameters. From our experiments we also noticed the same sensitivity of SVLN as discussed in the introduction of this evaluation section.

5.5. Evaluation of breathiness transformation

According to Fant (1995), the *Rd* parameter of the LF model is linked to the breathiness and tenseness of the voice. A test evaluating the capability of SVLN to modify this voice quality has been therefore conducted. Even tough the voice quality is linked to f_0 (Tooher and McKenna, 2003), we modified only *Rd* in this test in order to evaluate its impact on the breathiness independently of f_0. Through the voice production model of SVLN, a modification of the *Rd* parameter changes the perception of both the deterministic and random components. For example, by decreasing *Rd*, the VUF is expected to increase since *Rd* controls also the spectral tilt of the glottal pulse. A frequency band, previously excited by noise, can so be made of harmonics (see Fig. 9).

In this test, the listeners were asked to compare transformed recordings by modifying *Rd* to different extents. For example, *Rd* was first multiplied by 2 for one transformation and divided by 2 in another one. Then, each listener evaluated to which extent the first is breathier than the second. Only the voiced segments were transformed in this test and the original signal was kept unchanged in the unvoiced segments. 4 different transformations were compared The transformations were obtained by multiplying the *Rd* parameter by four different powers of 2: $2^{-1} = 0.5; 2^{-1/2} \approx 0.71; 2^{1/2} \approx 1.41$ and 2. The original recordings were also present in the test set and each listener were therefore asked to compared 10 pairs of audio files. After listening to the two audio files of each comparison pairs, a grade was then selected by the listener: "+3 if the left sound is much breathier than the right one; +2 if the left sound is breathier than the right one; +1 if the left sound is slightly breathier than the right one; 0 if the two sounds are about the same or if a difference exists which is, from the point of view of the listener, not related to breathiness; and the same on the other side of the comparison grid". The "Mean breathiness score" of each audio file is then computed like a mean preference score. The test was proposed on two different web pages, one English and one French, where two voices were used on each page, one female and one male. 4 different recordings were therefore used for the whole test. Also before the test, the listeners had the possibility to listen to recorded utterances of real speakers imitating normal and breathy voices in order to illustrate the target effect.

To avoid one language having more weight than the other one, we kept the results of the first 10 participants who conducted the test for each English and French pages. Left plot of Fig. 10 shows the mean breathiness scores averaging the 4 voices. Globally, the breathiness of the used voices can be clearly modified by the SVLN method. However, it is interesting to see that the score of the original recording is not aligned with the other scores. It is by far evaluated as being less breathy than expected since the original recording should have a score around 0. In a previous publication (Degottex et al., 2011b), we have shown that the resynthesis is fairly well aligned with the transformed sounds on this breathiness axis. Therefore, it seems that the SVLN method adds breathiness in the resynthesis. According to Fig. 11, one can see that this effect is mainly present in the male voices. One can note also that the score corresponding to the factor 2 is almost 50% more important than the score corresponding to 0.5. A simple linear coefficient on *Rd* does not imply therefore a linear

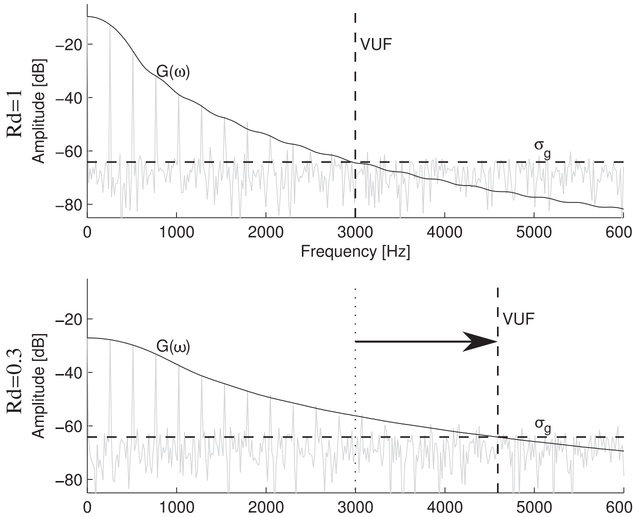

Fig. 9. Example of modification of *Rd* using synthetic spectra. When *Rd* is explicitly reduced, the VUF is implicitly raised.

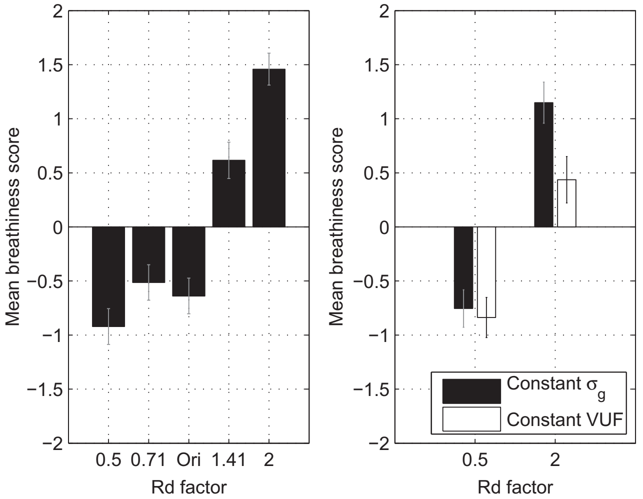

Fig. 10. Evaluation of breathiness according to a listening test. To the left, mean breathiness scores. To the right, mean breathiness scores while keeping a constant σ_g or a constant VUF.

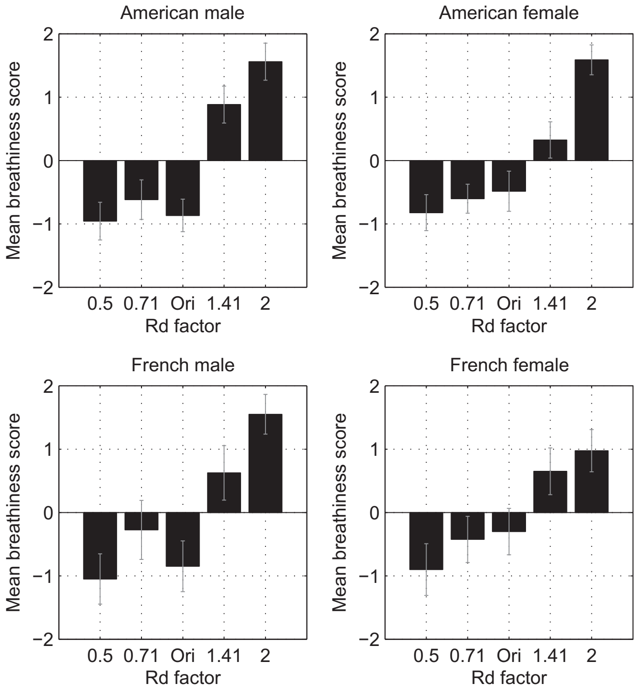

Fig. 11. Details of the breathiness evaluation for each voice.

modification of the breathiness. Cabral et al., 2008 have also shown that the breathiness of an utterance can be modified using GSS. Additionally, they also found that, by transforming a model voice, GSS increases more easily the breathiness than tenseness.

In this test, as illustrated by Fig. 9 the VUF is modified implicitly by the modification of *Rd* through the spectral tilt. It is therefore interesting to evaluate the impact of this implicit modification on the breathiness perception. Therefore, within the same tests (i.e. with the same listeners), it was asked to compare complementary pairs of transformations where the noise level σ_g was kept constant when modifying *Rd* (as assumed by the voice production model) with transformations where the VUF was kept constant. Right plot of Fig. 10 shows the resulting scores for the same 4 utterances. With a factor 2, which increases the breathiness, the noise entering the low frequencies by keeping σ_g constant plays an important role in the perceived breathiness conversely to the case where the VUF is kept constant. The presence of noise in low frequencies is therefore very important in this voice quality. However, the same effect does not appear towards tenseness. Compared to breathiness, the perception of tenseness can be more related to the glottal pulse shape. Therefore, keeping the VUF or σ_g constant may not have a significant impact for transformations towards tenseness since the glottal pulse is always modified the same way in this comparison (i.e. $Rd' = 0.5 \cdot Rd$). Finally, Fig. 11 shows that the scores depend clearly on the transformed voices (e.g. see the scores corresponding to the female voices and *Rd* factor 2). Indeed, each voice having different extent of breathiness, it might be more difficult to add breathiness in a voice which already breathy compared to another less breathy voice.

5.6. Evaluation of pitch transposition

Using a glottal model, we assumed that the voice quality can be better preserved in voice transformation. A last test is therefore presented in order to evaluate the quality provided by SVLN in pitch transposition. Also, since the breathiness can be modified using SVLN, the influence of the modification of *Rd* in pitch transposition is first evaluated in a preliminary test.

5.6.1. Preference of breathiness in pitch transposition

In the time domain, the glottal pulse shape is always stretched by the fundamental period. Consequently, in the frequency domain the glottal spectrum always follows the variation of f_0. For example, if f_0 is increased by 100 cents, the glottal formant is equally increased. Taking into account the spectral shape of the glottal source in the estimation of the VTF, this property will be respected using the proposed separation procedure of SVLN. Additionally, the voice quality is known to be correlated to f_0 (Tooher and McKenna, 2003; Henrich, 2001). The higher the pitch, the more lax the source and thus the bigger the *Rd* value. More specifically in the context of pitch transposition, a relation between the transposition factor and *Rd* exists. Accordingly, we propose to modify *Rd* following this simple formula:

$$Rd' = 2^{\alpha \cdot T/1200} \cdot Rd \tag{11}$$

where *T* is the transposition factor given in cents and α is a constant which controls the modification of *Rd* according to the transposition factor.

The choice of α is obviously not straightforward. To avoid a choice based only on informal listening, the following preliminary test has been carried out. A preference test was used to compare pairs of transformed utterances between α values $\{0, 0.5, 1, 1.5\}$ using transpositions of ± 900 cents. According to informal listening, differences with a higher resolution than the proposed values are hardly noticeable (e.g. between 0.25 and 0.5). Also, for $\alpha \geqslant 1.5$, the synthesized voice sounds either over-stressed or over-lax. The test was proposed on two different web pages for only two languages, one English and one French, where two voices were used on each page, one female and one male. 4 different recordings were therefore used for the whole test. The listeners had to give their preferences about “the naturalness of the first sound compared to the second”, using a grade between −3 and 3. Again, only voiced segments were transformed. For each page the results of the first 16 listeners who conducted the test were kept and the mean preference scores are shown in Fig. 12 for each voice (averaging both preferences for downward and upward transpositions). Globally, as expected from the results of the resynthesis test, the results of this test also vary with respect to the utterance. This simple test does not allow obviously to conclude that transforming the *Rd* parameter leads to better transpositions. The expression (11) is too simplistic and could be the subject of a dedicated study. As a preliminary test for pitch transposition, it supports our informal listening showing that an α value between 0 and 1.5 should be convenient. More specifically, we used the following expression to obtain an optimal α value from the listening test:

$$\alpha^\star = \sum_i \frac{p_i}{\sum_j p_j} \cdot \alpha_i \qquad (12)$$

where p_i is the preference related to the factor $\alpha_i = \{0, 0.5, 1, 1.5\}$. To ensure an improvement, only α_i values providing an improvement are considered in (12) (i.e. we consider only the indices i such as $p_i > 0$). Finally, according to the data of the listening test and expression (12), $\alpha^\star = 0.4$

5.6.2. Preference in pitch transposition

In this last listening test, we compared different transposition methods using a preference test. Three methods are compared: PSOLA (Hamon et al., 1989) (implementation from Peeters (2001) using randomization of the frequency band above the VUF), STRAIGHT (Kawahara et al., 1999) (version `V40pcode`) and the proposed SVLN method. PSOLA is not an encoding/decoding method like STRAIGHT and SVLN. It modifies the original signal assuming that windows of two periods length placed on local maxima of energy can be kept unchanged. According to the transposition factor, the windows are then duplicated (upward transposition) or decimated (downward transposition) and placed at new time positions. In the case of an upward transposition, the phase spectrum is randomized above the VUF such as the duplication of the original periods does not create an artificial correlation and a buziness effect.

Four different transposition factors were evaluated, ±600 cents (half an octave) and ±1200 cents (one octave). In order to keep a moderate number of pairs to evaluate by the listeners, the test was split into two web pages, one for the downward transpositions and one for the upward transpositions and we kept only the answers of listeners who answered both pages. In each page, pairs of transpositions made by two different methods were proposed to the listeners who were asked to give a grade between -3 and $+3$, according to the first audio file compared to the second, based on "their preference about the naturalness of the sounds". 8 utterances were used (both female and

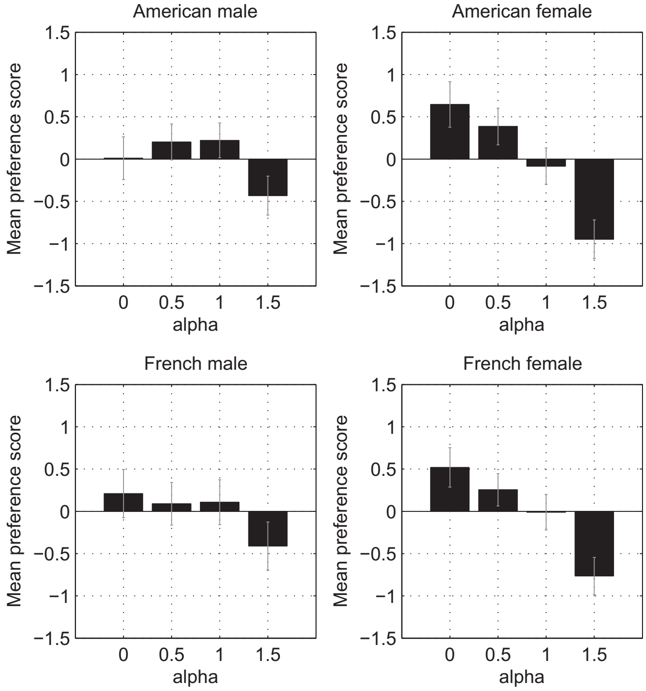

Fig. 12. Mean preference scores across breathiness modification in pitch transposition of ±900 cents with the 95% confidence intervals.

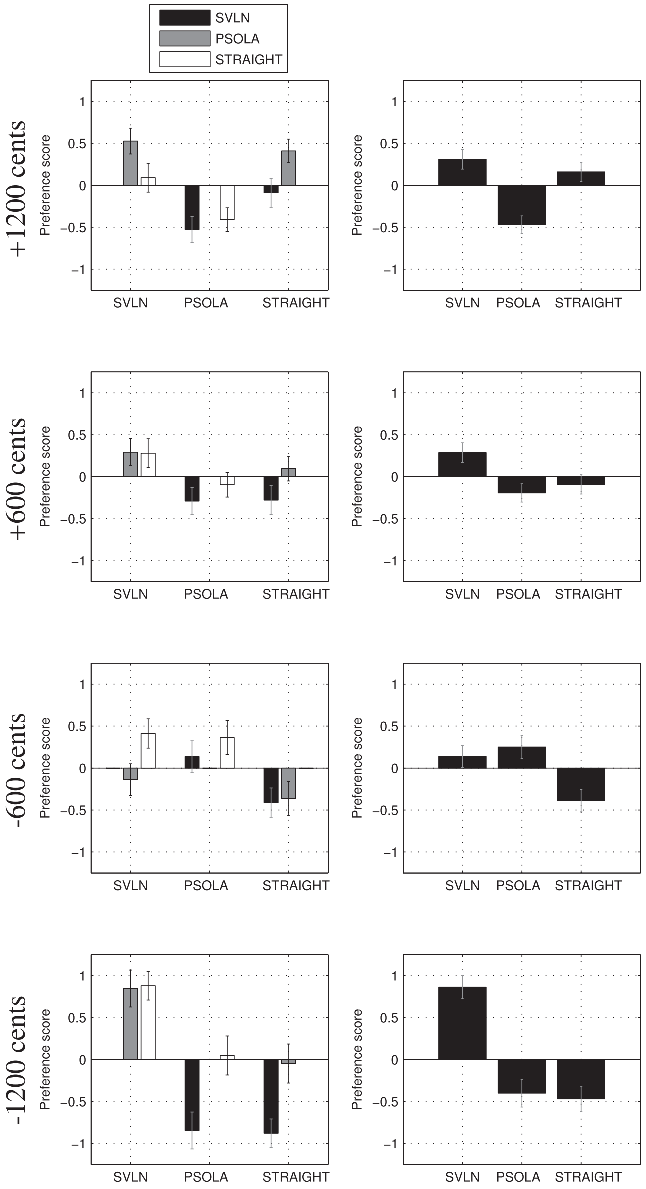

Fig. 13. Preference scores for pitch transposition with the 95% confidence intervals. Left plots shows the detailed preference scores for one method compared to each other whereas right plots show the mean preference scores averaging all scores related to each method.

male voices in 4 languages). Only voiced segments were transformed and the unvoiced segments were taken from the original recordings. 21 listeners answered the two pages of the test and the corresponding results are shown in Fig. 13.

Globally, SVLN clearly improves the transpositions of −1200 cents. According to informal listening, distortions in PSOLA and STRAIGHT increase when the transposition factor decreases whereas the quality given by SVLN seems more constant. On the one hand, the glottal model is a constrained representation of the glottal pulse in terms of temporal shape and spectral characteristics which are related to know acoustic properties that a glottal source should have. SVLN always respects these constraints whatever the transposition factor and may thus ensure some basic naturalness. On the other hand, methods which does not consider these constraints may generate a source signal which does not have the basic properties that a glottal source should have. About upward transpositions, the same behavior does not appear. Globally, differences between upward transpositions are hardly noticeable (which has been also reported spontaneously by many listeners). The only clear difference comes from the PSOLA method whose preference is reduced for +1200 transpositions. Despite the post-processing randomizing the phase spectrum in the PSOLA method, the noise component seems to suffer of a lack of naturalness according to informal listening. Moreover, higher the transposition, more important this effect. For upward transpositions, it seems therefore more important to ensure the quality of the noise component whereas the naturalness of the glottal pulse seems to play a more important role in the downward transpositions.

6. Conclusions

In this article, an encoding/decoding method has been presented, called Separation of the Vocal tract with the Liljencrants–fant model plus Noise (SVLN). Whereas most of the existing techniques can be applied to any pseudo-periodic signal (e.g. vocoders, PSOLA, STRAIGHT), the presented method aims to separate a given speech spectrum into four parts related to voice production: a deterministic source modeled by a glottal model, a random source, a Vocal Tract Filter (VTF) and a radiation filter. This method is thus dedicated to voice processing like ARX methods and the Glottal Spectral Separation (GSS) method. Compared to the former, the presented method makes use of the simplicity of the source-filter model in the spectral domain, using spectral division like GSS, allowing to use any VTF estimation method. In terms of speech analysis, or more specifically glottal source analysis, spectral division therefore provides also a promising mean by widening the possible techniques which can be used in inverse filtering. Compared to GSS, the presented method simplifies both representation of the deterministic and random components using the unique Rd shape parameter of the Liljencrants–Fant (LF) glottal model and a standard-deviation of Gaussian noise.

A first listening test about resynthesis on a frame by frame basis has shown that SVLN and the widely used STRAIGHT method provide similar overall qualities whereas SVLN seems to have a slightly better quality than GSS. Among the methods, the evaluated quality is different if both voiced and unvoiced segments are resynthesized or if only the voiced segments are resynthesized (keeping the original signal in the unvoiced segments). Using SVLN and GSS, the quality is clearly better when resynthesizing only the voiced segments whereas this difference can not be established for STRAIGHT. Using glottal models, as in SVLN and GSS, there is therefore a stability problem between voiced and unvoiced segments. A preliminary test using HMM-based speech synthesis led us to similar conclusions, the utterances synthesized by STRAIGHT being preferred compared to that of SVLN. Nevertheless, keeping the unvoiced segments unchanged, by transforming only the voiced segments, we carried out two other listening tests to evaluate the capacity of the proposed method to transform the breathiness and the pitch of a recording. The first test has shown that, whereas SVLN introduces breathiness in the resynthesis, this voice quality can be clearly modified using the proposed method. Finally, the last test about pitch transposition has shown that SVLN is slightly preferred or similar to STRAIGHT and PSOLA methods, except for downward transpositions of one octave where it is clearly preferred. For important downward transpositions, the LF glottal model constraints the deterministic component of the glottal source in a way that it obeys, at least, to known basic a priori on the glottal pulse provided by many studies on glottal source analysis.

Acknowledgments

This research was partly supported by the *Affective Avatar* ANR project, by the *Respoken* and *AngelStudio* FEDER projects and the Centre National de la Recherche Scientifique (CNRS) for the PhD grant. Authors also would like to thank Chunghsin Yeh and Joao Cabral for the discussions, their time and their code, the reviewers for their numerous and precise remarks and especially the listeners for their precious ears.

References

Agiomyrgiannakis, Y., Rosec, O., 2009. ARX-LF-based source-filter methods for voice modification and transformation. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 3589–3592.

Agiomyrgiannakis, Y., Rosec, O., 2008. Towards flexible speech coding for speech synthesis: an LF + Modulated Noise Vocoder. In: Proc. Interspeech, pp. 1849–1852.

Alku, P., Tiitinen, H., Naatanen, R., 1999. A method for generating natural-sounding speech stimuli for cognitive brain research. Clin. Neurophysiol. 110, 1329–1333.

Assembly, T.I.R., 2003. ITU-R BS.1284-1: EN-General methods for the subjective assessment of sound quality. Technical Report. ITU.

Banno, H., Lu, J., Nakamura, S., Shikano, K., Kawahara, H., 1998. Efficient representation of short-time phase based on group delay, in: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 861–864.

Bechet, F., 2001. Liaphon: un système complet de phonetisation de textes. Traitement Automatique des Langues 42, 47–67.

Bonada, J., 2008. Voice processing and synthesis by performance sampling and spectral models. Ph.D. thesis. Universitat Pompeu Fabra. Spain.

Cabral, J.P., 2010. HMM-based speech synthesis using an Acoustic Glottal Source Model. Ph.D. thesis. CSTR, University of Edinburgh, UK.

Cabral, J., Renals, S., Richmond, K., Yamagishi, J., 2008. Glottal spectral separation for parametric speech synthesis. In: Proc. Interspeech, Brisbane, Australia, pp. 1829–1832.

Cabral, J., Renals, S., Yamagishi, J., Richmond, K., 2011. HMM-based speech synthesiser using the LF-model of the glottal source. In: IEEE Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 4704–4707.

de Cheveigne, A., Kawahara, H., 2002. YIN, A fundamental frequency estimator for speech and music. J. Acoust. Society Amer. 111, 1917–1930.

Degottex, G., 2010. Glottal source and vocal tract separation. Ph.D. thesis. UPMC-Ircam. France.

Degottex, G., Roebel, A., Rodet, X., 2011a. Phase minimization for glottal model estimation. IEEE Trans. Audio Speech Lang. Process. 19, 1080–1090.

Degottex, G., Roebel, A., Rodet, X., 2011b. Pitch transposition and breathiness modification using a glottal source model and its adapted vocal tract filter. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 5128–5131.

del Pozo, A., Young, S., 2008. The linear transformation of LF glottal waveforms for voice conversion. In: Proc. Interspeech, pp. 1457–1460.

Drugman, T., Wilfart, G., Dutoit, T., 2009b. A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In: Interspeech.

Drugman, T., Bozkurt, B., Dutoit, T., 2009a. Complex cepstrum-based decomposition of speech for glottal source estimation. In: Proc. Interspeech, pp. 116–119.

Fant, G., 1995. The LF-model revisited. Transformations and frequency domain analysis. STL-QPSR 36, 119–156.

Fant, G., Liljencrants, J., Lin, Q.G., 1985. A four-parameter model of glottal flow. STL-QPSR 26, 1–13.

Flanagan, J.L., Golden, R.M., 1966. Phase Vocoder. Technical Report. The Bell System Technical Journal.

Gales, M.J.F., 1999. Semi-tied covariance matrices for hidden markov models. IEEE Trans. Speech Audio Process. 7, 272–281.

Griffin, D.W., Lim, J.S., 1988. Multiband excitation vocoder. IEEE Trans. Acoust. Speech Signal Process. 36, 1223–1235.

Hamon, C., Mouline, E., Charpentier, F., 1989. A diphone synthesis system based on time-domain prosodic modifications of speech. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 238–241.

Hedelin, P., 1984. A glottal LPC-vocoder. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 21–24.

Henrich, N., 2001. Etude de la source glottique en voix parlée et chantée. Ph.D. thesis. UPMC, France (In French).

Hermes, D.J., 1991. Synthesis of breathy vowels: some research methods. Speech Comm. 10, 497–502.

Imai, S., Abe, Y., 1979. Spectral envelope extraction by improved cepstral method. Electron. Comm., 10–17 (In Japanese).

Kawahara, H., Estill, J., Fujimura, O., 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In: MAVEBA.

Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A., 1999. Restructuring speech representations using a pitch-adaptative time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. Speech Comm. 27, 187–207.

Kim, S.J., Hahn, M., 2007. Two-band excitation for HMM-based speech synthesis. IEICE – Trans. Inf. Systems, 378–381.

Lanchantin, P., Morris, A.C., Rodet, X., Veaux, C., 2008. Automatic phoneme segmentation with relaxed textual constraints. In: Proc. Language Resources and Evaluation Conference, pp. 2403–2407.

Lanchantin, P., Degottex, G., Rodet, X., 2010. A HMM-based speech synthesis system using a new glottal source and vocal tract separation method. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, USA, pp. 4630–4633.

Laroche, J., Stylianou, Y., Moulines, E., 1993. HNS: Speech modification based on a harmonic+noise model. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 550–553.

Markel, J., Gray, A., 1976. Linear Prediction of Speech. Springer, Verlag.

McAulay, R., Quatieri, T., 1986. Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. Acoust. Speech Signal Process. 34, 744–754.

Mehta, D., Quatieri, T.F., 2005. Synthesis, analysis, and pitch modification of the breathy vowel. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 199–202.

Miller, R.L., 1959. Nature of the vocal cord wave. J. Acoust. Soc. Amer. 31, 667–677.

Oppenheim, A., Schafer, R., Stockham, T., 1968. Nonlinear filtering of multiplied and convolved signals. Proc. IEEE 56, 1264–1291.

Pantazis, Y., Rosec, O., Stylianou, Y., 2010. Adaptive AM–FM signal decomposition with application to speech analysis. IEEE Trans. Audio Speech Lang. Process. 19, 290–300.

Peeters, G., 2001. Modeles et modification du signal sonore adaptees a ses caracteristiques locales. Ph.D. thesis. UPMC, France (In French).

Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., Alku, P., 2011. HMM-based speech synthesis utilizing glottal inverse filtering. IEEE Trans. Audio Speech Lang. Process. 19, 153–165.

Rodet, X., Potard, Y., Barriere, J.B., 1984. The CHANT project: from synthesis of the singing voice to synthesis in general. Comput. Music J. 8, 15–31.

Roebel, A., Villavicencio, F., Rodet, X., 2007. On cepstral and all-pole based spectral envelope modeling with unknown model order. Pattern Recognition Lett. 28, 1343–1350.

Stevens, K.N., 1971. Airflow and turbulence noise for fricative and stop consonants: static considerations. J. Acoust. Soc. Amer. 50, 1180–1192.

Stylianou, Y., 1996. Harmonic plus noise models for speech combined with statistical methods, for Speech and Speaker Modification. Ph.D. thesis. TelecomParis. France.

Stylianou, Y., 2001. Applying the harmonic plus noise model in concatenative speech synthesis. IEEE Trans. Speech Audio Process. 9, 21–29.

Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T., Imai, S., 1995. An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. In: Proc. Eurospeech, pp. 757–760.

Tokuda, K., Masuko, T., Myizaki, N., Kobayashi, T., 2002a. Multi-space probability distribution HMM. IEICE Trans. Inf. Systems E85-D, 455–464.

Tokuda, K., Zen, H., Black, A., 2002b. An HMM-based speech synthesis system applied to English. In: Proc. IEEE Workshop on Speech synthesis, pp. 227–230.

Tooher, M., McKenna, J.G., 2003. Variation of the glottal LF parameters across F0, vowels, and phonetic environment. In: Proc. ISCA Voice Quality: Functions, Analysis and Synthesis (VOQUAL), pp. 41–46.

Valbret, H., Moulines, E., Tubach, J., 1992. Voice transformation using PSOLA technique. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 145–148.

Vincent, D., Rosec, O., Chonavel, T., 2007. A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and HNM modeling. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 525–528.

Yeh, C., 2008. Multiple fundamental frequency estimation of polyphonic recordings. Ph.D. thesis. UPMC-Ircam. France.

Young, S., 1994. The HTK hidden markov model toolkit: design and philosophy. Technical Report. University of Cambridge.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2004. Hidden semi-Markov model based speech synthesis. In: Proc. of ICSLP.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., Tokuda, K., 2007. The HMM-based speech synthesis system (HTS) version 2.0. In: Proc. ISCA Workshop on Speech Synthesis (SSW). <http://hts.sp.nitech.ac.jp>.

Zivanovic, M., Roebel, A., Rodet, X., 2008. Adaptive threshold determination for spectral peak classification. Comput. Music J. 32, 57–67.