PITCH TRANSPOSITION AND BREATHINESS MODIFICATION USING A GLOTTAL SOURCE MODEL AND ITS ADAPTED VOCAL-TRACT FILTER

Gilles Degottex, Axel Roebel, Xavier Rodet

IRCAM - CNRS-UMR9912-STMS, Analysis-Synthesis Team
1, place Igor Stravinsky, 75004 Paris

ABSTRACT

The transformation of the voiced segments of a speech recording has many applications such as expressivity synthesis or voice conversion. This paper addresses the pitch transposition and the modification of breathiness by means of an analytic description of the deterministic component of the voice source, a glottal model. Whereas this model is dedicated to voice production, most of the current methods can be applied to any pseudo-periodic signals. Using the described method, the synthesized voice is thus expected to better preserve some naturalness compared to a more generic method. Using preference tests, it is shown that this method is preferred for important pitch transposition (e.g. one octave) compared to two state of the art methods. Additionally, it is shown that the breathiness of two male utterances can be controlled.

Index Terms— Voice transformation, pitch transposition, breathiness, glottal model, vocal-tract filter.

1. INTRODUCTION

Using the source-filter model, there are 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. True-Envelope (TE) [1], linear prediction), the phase vocoder preserves a part of the original phase spectrum in the transformed waveform [1]. Additionally, the methods based on Pitch-Synchronous-OverLap-Add (PSOLA) assume that the signal inside a single window can be used without being modeled [2]. On the other hand, in analysis/synthesis methods, the speech waveform can be fully encoded into a small set of parameters. For example, a short speech segment can be parametrized using a set of sinusoids [3] which can be also harmonics [4]. The same segment can be also represented using a wide-band spectrum where smooth envelopes of the amplitude and phase spectra have to be estimated (e.g. WBVPM [5], STRAIGHT [6]).

Most of these mentioned methods achieve excellent results, either for voice transformation or speech synthesis. However, in case of important modification in voice transformation (e.g. one octave pitch transposition), artifacts often appear showing underlying limitations of the chosen model. For example, when transposing pitch downward with a phase vocoder, the noise naturally produced in high frequencies arises at low frequencies where such a noise is naturally not present. This drawback increases the hoarseness of the transformed voice. The PSOLA method forces the impulse response of the vocal-tract filter to decay using a window of two periods duration. Therefore, a lack of resonances in downward pitch transposition can be percieved due to this smooth truncation of the impulse response. Moreover, whereas most of the current analysis/synthesis methods can be applied to any pseudo-periodic signals (e.g. sinusoidal models, WBVPM, STRAIGHT), one can expect that a model which is more dedicated to voice production better respect 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 Vocal-Tract Filter (VTF) contrary to most of the current methods which assume that the voice source is made of a flat amplitude spectrum. Accordingly, ARX/ARMAX methods have been proposed which use a glottal model (e.g the Liljencrants-Fant (LF) model [7]) to represent the deterministic component of the glottal source [8, 9]. However, it has been reported that these methods are sensitive to inversion errors [8]. The transformation of the voiced signal using a glottal model is thus still an open and challenging question and it is interesting to investigate other means to use such a model.

This paper addresses two applications of an analysis/synthesis method which has been previously proposed for HMM-based synthesis [10]. The source model of this method uses Gaussian noise and the LF glottal model parametrized by the single Rd shape parameter [7]. Then, we estimate the VTF by taking into account the amplitude spectrum of this source model to fit an observed speech spectrum. This method is called *Separation of the Vocal-tract with the Liljencrants-fant model plus Noise* (SVLN). Conversely to ARX/ARMAX methods, the SVLN method uses a source filter separation in frequency domain which takes benefits of the True-Envelope (TE) estimation method. We also take advantage of recent results on estimation of glottal parameters [11]. Whereas the reliability of SVLN was comparable to state of the art methods but not better than STRAIGHT for HMM-based synthesis [10], this paper shows that this method can be successfull for important pitch transposition and for modification of breathiness.

The next sections present the SVLN method, its analysis and synthesis steps. Then, its evaluation follows.

2. THE SVLN METHOD

2.1. The voice production model

The voiced segments of the speech signal are assumed to be stationary and periodic in a short analysis window win$[t]$ (of 3.5 periods and a minimum of 10 ms). Therefore, using the source-filter model in the frequency domain, the voice production model of an observed speech spectrum $S(\omega)$ computed by the Fourier transform of the windowed signal can be described as follows (see also fig. 1):

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

where:

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

$G^{Rd}(\omega)$ is the shape of the deterministic component of the glottal source in a single period, the LF model parametrized by Rd and E_e, its shape and amplitude parameters respectively [7].

$N^{\sigma_g}(\omega)$ is the random component of the glottal source generated by turbulence at the glottal level. This noise is assumed to obey a Gaussian distribution of standard-deviation σ_g.

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

$L(\omega)$ is the filter corresponding to the radiation at the lips and nostrils level. Here, we assume $L(\omega) = j\omega$ [12].

The voiced signal can thus be parametrized by $\{f_0, Rd, E_e, \sigma_g, \bar{c}\}$

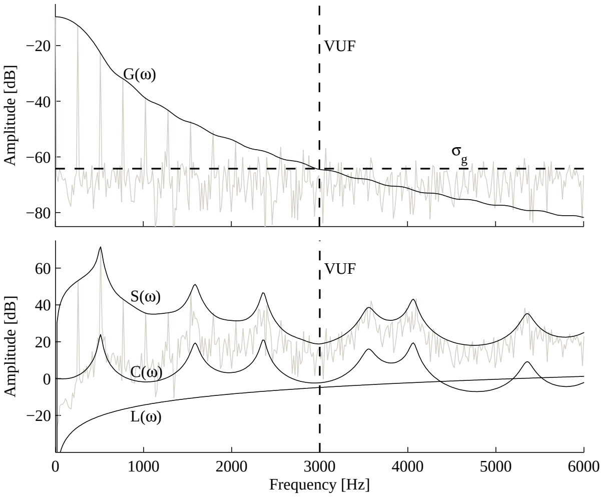

Fig. 1. The glottal source model above and the voice production model below. The spectra of one period and multiple periods is shown in black and gray lines respectively.

2.2. The analysis step: estimation of the SVLN parameters

2.2.1. The parameters of the deterministic source: f_0, Rd, E_e

The fundamental frequency f_0 can be estimated from numerous methods. For this presentation, the YIN method [13] is used.

To estimate the LF shape parameter Rd, the recently proposed method MSPD^2 is used which is based on phase minimization [11]. In order to ensure a stable estimation of the VTF in the following, the time evolution of the Rd parameter is smoothed using a 100 ms median filter followed by a zero-phase filtering whose window is of same duration. In doing so, we assume that the voice quality is almost constant inside a single phoneme.

Three gains co-exist in the voice production model: E_e, σ_g and the mean log amplitude of the VTF. These gains are 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. In this presentation, E_e is set to the mean log amplitude of the VTF and this latter is fixed to zero (see sec. 2.2.3 for the estimation of the VTF).

2.2.2. The parameter of the random source: σ_g

According to figure 1, a Voiced/Unvoiced Frequency (VUF) can be estimated to split $S(\omega)$ into a deterministic source below the VUF and Gaussian noise above. Like f_0, this frequency is assumed to be known *a priori* thanks to existing methods. In this presentation, this value is estimated by determination of voiced/unvoiced frequency bands [14, p.3] by means of peak classification of the speech spectrum [15]. $|G^{Rd}(\omega)|$ is assumed to cross the expected amplitude of the noise at the VUF (see top plot of fig. 1). Consequently, since the amplitude spectrum $|G^{Rd}(\omega)|$ is known from the f_0 and Rd estimates, the noise level σ_g can be deduced from the VUF estimate:

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

where $|G^{Rd}(\text{VUF})|$ is an expected amplitude which has to be converted to the Gaussian parameter σ_g: Spectral amplitudes of Gaussian noise obey a Rayleigh distribution. Therefore, $|G^{Rd}(\text{VUF})|$ is first converted to the Rayleigh mode $(1/\sqrt{\pi/2})$ from which the standard deviation of the Gaussian distribution is retrieved $(\sqrt{2})$ [16]. Additionally, in the spectral domain, the noise level is proportional to the energy of the analysis window $\text{win}[t]$ used to compute $S(\omega)$. The normalization by $\sqrt{\sum_t \text{win}[t]^2}$ is therefore necessary.

2.2.3. The estimation of the vocal-tract filter $C^{\bar{c}}(\omega)$

According to the properties of the two frequency bands below and above the VUF, two different envelopes are used to model the VTF. Then, these envelopes have to be properly normalized to ensure a VTF estimate which is independent of the excitations properties.

In the deterministic frequency band (top of eq. 3), the contribution of $L(\omega)$ and $G^{Rd}(\omega)$ are removed from $S(\omega)$ by division in the frequency domain. The True-Envelope (TE) $\mathcal{T}(.)$ [1] is then used to fit the top of the harmonics of the division result. Note that this envelope fits the expected amplitude of the VTF frequency response since the top of a harmonic is its expected amplitude.

In the random frequency band (bottom of eq. 3), $S(\omega)$ is divided by $L(\omega)$ and $|G^{Rd}(\text{VUF})|$ to ensure a continuity between the two frequency bands. The division result is modeled by computing its real cepstrum $\mathcal{P}(.)$ truncated to a given order (discussed below). According to the Rayleigh distribution, the mean log amplitude measured by $\mathcal{P}(.)$ has first to be converted to the Rayleigh mode on a linear scale (factor $e^{0.058}$ in eq. 3) [16]. Then, the expected amplitude is retrieved from the Rayleigh mean value $(\sqrt{\pi/2})$.

$$C(\omega) = \begin{cases} \mathcal{T}\left(\frac{S(\omega)}{L(\omega)G^{Rd}(\omega)}\right) \cdot \gamma^{-1} & \text{if } \omega < \text{VUF} \\ \\ \mathcal{P}\left(\frac{S(\omega)}{L(\omega)G^{Rd}(\text{VUF})}\right) \cdot \frac{\sqrt{\pi/2}}{\gamma \cdot e^{0.058}} & \text{if } \omega \geq \text{VUF} \end{cases} \tag{3}$$

where $\gamma = \sum_t \text{win}[t]/(f_s/f_0)$ stands for the number of periods in the analysis window (f_s is the sampling frequency). This normalization is necessary regarding to the synthesis step where the VTF is convolved with each period of the source. It is also necessary that the two envelopes $\mathcal{T}(.)$ and $\mathcal{P}(.)$ do not fit the harmonic structure $H^{f_0}(\omega)$ of the observed spectrum. For the envelope $\mathcal{T}(.)$, the optimal order $0.5 \cdot f_s/f_0$ is used [17]. The same order is used for the cepstral envelope. Indeed, although no harmonic partial appears in the frequency band of the random source, sinusoidal peaks with distance of f_0 (but not of multiples of f_0) arise in this band because the glottal noise is amplitude modulated by the glottal area [9, 18]. Finally, the cepstral coefficients $\bar{c}$ of the VTF are retrieved from the minimum-phase realization of $C(\omega)$.

Note that this separation method is always able to model the observed amplitude spectrum $|S(\omega)|$. Indeed, the estimation of the VTF always completes the source and radiation models in order to obtain $|S(\omega)|$. Conversely, the phase of the model is imposed either by the LF model or Gaussian noise.

2.3. The synthesis step using SVLN parameters

A speech utterance is synthesized from the estimated parameters using the procedure described in the sections below. Short segments of stationary signals are first synthesized and then overlap-added.

2.3.1. The definition of the segments, their position and duration

Temporal marks m_k of the k^{th}-segment are first placed at intervals according to f_0 (see fig. 2). The maximum excitation instant t_e [7] of each LF model will be placed on this mark. Then the starting time t_k of the k^{th}-segment is defined as the opening instant of the LF model and the ending time of this segment is the starting time of the next.

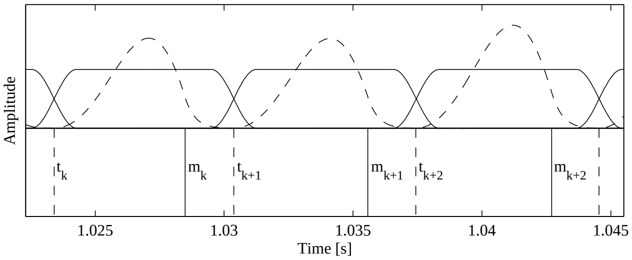

Fig. 2. Three segments: LF models are in dashed lines, and synthesis windows $\text{win}_k[t]$ are in solid lines.

2.3.2. The noise component: filtering, modulation and windowing

For all segments, noise is generated. However, if this noise is white, the synthesized voice sounds hoarse because the lowest harmonics of the deterministic source are disturbed by the noise. The synthesized noise is thus high-pass filtered in spectral domain by a multiplicative term $F_{hp}^{\text{VUF}}(\omega)$ defined by a cutoff frequency equal to the VUF and a slope of 6 dB/kHz in the transition band (the VUF is retrieved from the noise level and $|G^{Rd}(\omega)|$). Additionally, if the glottal noise is not amplitude modulated synchronously with f_0, a second source is perceived separately from the deterministic source [18]. Therefore, a modulation function $v^{Rd}[t]$ is built as follows:

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

where β is the magnitude of the modulation and $g^{Rd}[t]$ is the LF model normalized by its maximum amplitude. In this presentation, the Rd parameter is set to the same value as that of the deterministic component. Then, from informal listening, we fixed the value $\beta = 0.75$ according to the naturalness of the synthesis. The estimation of these two parameters should be addressed in a future work. Conversely to the glottal pulses, the noise does not stop at zero amplitude at the end of each segment. Therefore, a cross fade is necessary between noise segments of different color and amplitude. For each k^{th}-segment, a synthesis window $\text{win}_k[t]$ is built with a fade-in center on t_k and a fade-out center on t_{k+1} (see figure 2). The fade-in/out function is a hanning half window of duration $0.25 \cdot \min(t_{k+1} - t_k, t_k - t_{k-1})$ where the fade-out of win_k is the complementary of the fade-in of win_{k+1} such as the sum of all windows is 1 at any time of the synthesis. According to these descriptions, the noise spectrum of the k^{th}-segment is synthesized by:

$$N_k(\omega) = F_{hp}^{\text{VUF}_k}(\omega) \cdot \mathcal{F}\big(v^{Rd_k}[t] \cdot \text{win}_k[t] \cdot n^{\sigma_{gk}}[t]\big) \qquad (4)$$

where $n^{\sigma_{gk}}[t]$ is a zero-mean Gaussian random signal of standard-deviation σ_{gk} and $\mathcal{F}(.)$ is the Discrete Time Fourier Transform.

2.3.3. Glottal pulse and filtering elements

Finally, the deterministic glottal pulse $G^{Rd_k}(\omega)$ is added to the noise segment and the VTF and radiation filters are applied in order to synthesize the speech segment $S_k(\omega)$:

$$S_k(\omega) = \left(e^{-j\omega m_k} \cdot G^{Rd_k}(\omega) + N_k(\omega)\right) \cdot C^{\bar{c}_k}(\omega) \cdot j\omega \qquad (5)$$

where $e^{-j\omega m_k}$ is a delay placing the instant t_e of the LF model at the mark m_k and $C^{\bar{c}_k}(\omega)$ is the minimum-phase VTF corresponding to the cepstral coefficients $\bar{c}_k$.

Finally, the time domain sequence of each segment is retrieved through the inverse Fourier transform of $S_k(\omega)$. Then, the entire signal is constructed by overlap-adding the time segments.

3. EVALUATION

3.1. Preference tests for pitch transposition

Two on-line preference tests have been carried out to compare the SVLN method to two other methods: the Shape-Invariant Phase vocoder (SHIP) [1] and the STRAIGHT method [6]. These preference tests were dedicated to transpositions of ±900 cents and ±1200 cents respectively. The f_0 and VUF estimates were common to all compared methods and if necessary, we corrected manually the errors of f_0 estimation related to octaves errors. Additionally, the voiced segments have been manually annotated.

The estimation of the VTF of the SVLN method (sec. 2.2.3) takes into account the amplitude spectrum of the source. Consequently, if the parameters controlling the source are left unmodified in transposition, the glottal formant (the main peak of the amplitude spectrum of the source [7]) will be shifted proportionally to the transposition factor. However, the voice quality is correlated to f_0 [19]. The higher the pitch, the more lax the source and thus the bigger the Rd value. In order to obtain a natural voice in pitch transposition, the glottal formant and f_0 should not be equally shifted. In these preference tests, the Rd parameter of the transposed voices was modified using the formula $Rd' = 2^{\kappa \cdot T/1200} \cdot Rd$, where T is the transposition factor given in cents and κ is a proportionality parameter between the modified and original Rd parameters. From informal listening of transpositions with various κ values, this parameter has been fixed to $\kappa = 0.5$ for the preference tests.

Each preference test comprised two web pages dedicated to English and French recordings. For each language, two utterances (one from a female voice and another from a male voice) were transposed by $T = \pm 900$ cents ($T = \pm 1200$ cents for the second test) using the three compared methods. The participants were asked therefore to compare each method to the others through 24 comparison pairs. For each pair, the participants attributed a score depending on their preference about the overall quality of the first transposed utterance compared to the second one like with a scale of Comparison Mean Opinion Score (CMOS) (*much better (+3); better (+2); slightly better (+1); about the same (+0) and the same on the other way*). We preferred an overall quality test compared to an evaluation focusing on a particular quality of the sound (e.g. naturalness) in order to assess to which extent the artifacts influence the preference. 50 participants answered the first test and 40 answered the second. Figure 3 shows the mean preferences.

First, the SVLN method is close to the other methods. For ±900 cents, one can see that the STRAIGHT method outperforms globally the other compared methods and the quality of the SVLN method is between that of STRAIGHT and that of SHIP. In addition, one can note that the preference for the SHIP method is significantly lower than that of SVLN and STRAIGHT methods in downward

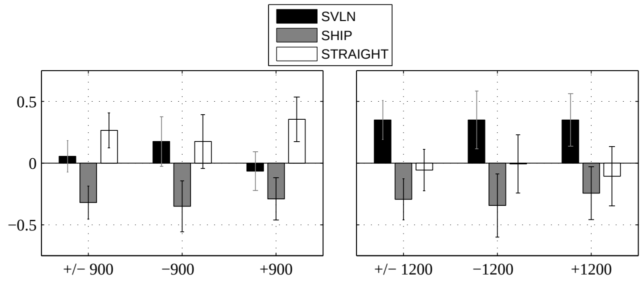

Fig. 3. Mean preferences with their 95% confidence interval of the methods for various transposition factors in cents (±T represents the overall scores regardless the direction of the transpositions).

transpositions which is consistent with the remark made in the introduction about the noise which arises in low frequencies in that case. One can see that the SVLN method performs significantly better for ±1200 cents than for ±900 cents. One the one hand, the SVLN method imposes either the phase spectrum of the deterministic LF model or that of Gaussian noise. It is possible that these two extrema cannot properly represent the real glottal source. On the other hand, using the LF model, one can expect that the synthesized source always respects some constraint imposed by this model which is important for the naturalness of the voice. These two observations can explain the disparity between the two tests regarding the preferences for the SVLN method.

3.2. Evaluation of breathiness modification

A last comparison test was carried out to evaluate the capability of the SVLN method to modify the breathiness of a given recording. Similar to the previous tests, the participants were asked to compare two utterances of different breathiness obtained by a modification of the Rd parameter between the analysis and synthesis steps of the SVLN method. The test was proposed in the same two languages, English and French, using only the male voices. The original recordings and four different modifications were compared. The latter was 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. Like in a CMOS test, the participants attributed a score to the pair comparing the second sample to the first one according to the scale: *much breathier (+3); breathier (+2); slightly breathier (+1); about the same or if a difference exists which is, from the point of view of the listener, not related to breathiness (0); and the same on the other way*. 43 participants answered the test and figure 4 shows the mean breathiness scores which are computed like a mean preference score. In conclusion, the maximum score of breathiness is close to 50% times bigger than the minimum score. The used scale is therefore not linearly related to the perception of this voice quality. However, the breathiness of the voiced segments of the two evaluated utterances can be clearly modified by the SVLN method.

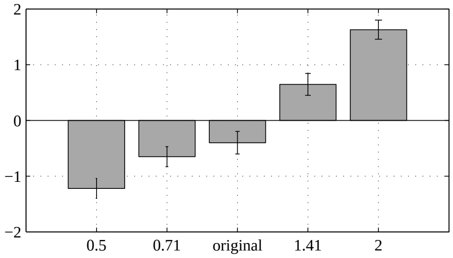

Fig. 4. Breathiness scores for various scaling of the Rd parameter.

4. CONCLUSIONS

In this paper, the transformation of voiced segments using a method called *Separation of the Vocal-tract with the Liljencrants-Fant model plus Noise* (SVLN) is addressed. Preference tests comparing the SVLN, SHIP and STRAIGHT methods have been carried out to compare the overall quality of their results in pitch transposition. For small transpositions, the quality of the SVLN method seems between those of SHIP and STRAIGHT. However, for a significant transposition (one octave below or above), the glottal model used in SVLN constraints the determinist component of the glottal source in some way that a minimal naturalness is ensured. The SVLN method therefore outperforms the two other methods in this case. Using a comparison test, we also showed that the breathiness of two neutral male utterances can be modified using the SVLN method. [1]

5. ACKNOWLEDGMENTS

This research was partly supported by the *Affective Avatar* ANR project, by the *Respoken* FEDER project and the Centre National de la Recherche Scientifique (CNRS). The presentation was also supported by the Swiss National Science Foundation (SNSF). Authors also would like to thank Chunghsin Yeh for his precious help.

6. REFERENCES

[1] Axel Roebel, “A shape-invariant phase vocoder for speech transformation,” in *DAFx*, 2010.

[2] H. Valbret, E. Moulines, and J.P. Tubach, “Voice transformation using psola technique,” in *ICASSP*, 1992, pp. 145–148.

[3] R. McAulay and T. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 4, pp. 744–754, Aug 1986.

[4] Yannis Pantazis, Olivier Rosec, and Yannis Stylianou, “On the properties of a time-varying quasi-harmonic model of speech,” in *Interspeech*, 2008.

[5] J. Bonada, *Voice Processing and Synthesis by Performance Sampling and Spectral Models*, Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, 2008.

[6] H. Kawahara, I Masuda-Katsuse, and A. Cheveigne, “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,” in *Speech Communication*, 1999, vol. 27.

[7] G. Fant, “The LF-model revisited. transformations and frequency domain analysis.,” *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 1995.

[8] Y. Agiomyrgiannakis and O. Rosec, “Towards flexible speech coding for speech synthesis: an LF + modulated noise vocoder,” in *Interspeech*, 2008.

[9] H.-L. Lu, *Toward a High-quality Singing Synthesizer with Vocal Texture Control*, Ph.D. thesis, Stanford, 2002.

[10] P. Lanchantin, G. Degottex, and X. Rodet, “A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method,” in *ICASSP*, 2010, pp. 4630–4633.

[11] G. Degottex, A. Roebel, and X. Rodet, “Phase minimization for glottal model estimation,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, no. 99, pp. 1 –1, 2010.

[12] J Markel and A Gray, *Linear Prediction of Speech*, Springer Verlag, 1976.

[13] A. de Cheveigne and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *JASA*, vol. 111, April 2002.

[14] Y. Stylianou, “Applying the harmonic plus noise model in concatenative speech synthesis,” *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 1, pp. 21 –29, jan 2001.

[15] Miroslav Zivanovic, Axel Roebel, and Xavier Rodet, “Adaptive threshold determination for spectral peak classification,” in *DAFx*, 2007.

[16] C. Yeh, *Multiple fundamental frequency estimation of polyphonic recordings*, Ph.D. thesis, UPMC, juin 2008.

[17] Axel Roebel, Fernando Villavicencio, and Xavier Rodet, “On cepstral and all-pole based spectral envelope modeling with unknown model order,” *Pattern Recognition Letters*, vol. 28, no. 11, pp. 1343–1350, 2007.

[18] D. Mehta and T. F. Quatieri, “Synthesis, analysis, and pitch modification of the breathy vowel,” in *WASPAA*, 2005.

[19] M. Tooher and J. G. McKenna, “Variation of the glottal LF parameters across f0, vowels, and phonetic environment,” in *VOQUAL*, 2003.

[1] Some demonstration audio samples are available at: http://recherche.ircam.fr/anasyn/degottex/index.php?n=Main.SVLN