A NEW METHOD FOR SPEECH SYNTHESIS AND TRANSFORMATION BASED ON AN ARX-LF SOURCE-FILTER DECOMPOSITION AND HNM MODELING

Damien Vincent, Olivier Rosec

France Telecom
R&D Division
2, Av. Pierre Marzin - 22307 Lannion, France
{damien.vincent, olivier.rosec}@orange-ft.com

Thierry Chonavel

ENST Bretagne
Signal & Communication Department
CS 83818 - 29238 Brest Cedex 3, France
thierry.chonavel@enst-bretagne.fr

ABSTRACT

In this paper a new method for speech synthesis is proposed. It relies on a source-filter decomposition of the speech signal by means of an ARX-LF model. This model allows the representation of the glottal signal as the sum of an LF waveform and a residual signal. The residual information is then analyzed by HNM. This signal representation enables high quality speech modification such as pitch, duration or even voice quality transformation. Experiments performed on a real speech database show the relevance of the proposed method as compared to other existing approaches.

Index Terms— Speech synthesis, speech analysis

1. INTRODUCTION

High quality speech modification is a subject of considerable importance in the speech processing area. Applications include text-to-speech synthesis by unit concatenation, for which it is often desirable to adapt the prosody, or even the spectral characteristics of the selected speech segments. Speech modification is also of great interest in the area of voice transformation and conversion, be it for speech synthesis purposes or for other applications such as movie dubbing, foreign language learning, *etc.*. . . Numerous methods have been proposed for the purpose of speech modification including non parametric techniques such as TD-PSOLA [1] and methods based on parametric models like HNM [2]. These approaches can achieve interesting speech modification methods but are limited in practice as they often degrade the quality of the resulting speech. It can be hypothesized that the problems encountered in speech modification are partly due to the fact that the signal representation does not really fit with the speech production mechanisms. More specifically, current approaches do not explicitly separate the glottal source and the vocal tract information from the speech signal. In this paper, we propose a speech analysis and synthesis scheme based on a source-filter separation using an ARX-LF model. In this model, the glottal signal is decomposed into a LF waveform [3] and a residual signal which is further analysed using an HNM model. The paper is organized as follows: section 2 presents the ARX-LF model. The analysis procedure is described in section 3 while the synthesis and modification schemes are detailed in section 4. Section 5 presents the application of the proposed method in in the context of speech synthesis and modification and section 6 concludes our work.

2. THE ARX-LF MODEL

A common approach in speech processing is to represent the speech production mechanisms by means of a source-filter model. In such representation, the source component is refered to as the glottal flaw derivative (GFD), which incorporates the derivative effect due to the lips radiation to the signal observed at the glottis. Moreover a reasonable approximation of the GFD can be obtained through the LF model [3] which enables the characterization of the glottal source signal with 5 parameters: one for the location of the glottal source (the reference is usually the glottal closure instant t_c), one for the amplitude and three to define the shape of the glottal flow. A typical LF waveform is depicted on figure 1. Among the possible parameter sets to define the shape, the vector $\theta = (O_q, \alpha_m, Q_a)$ has been chosen: O_q corre-

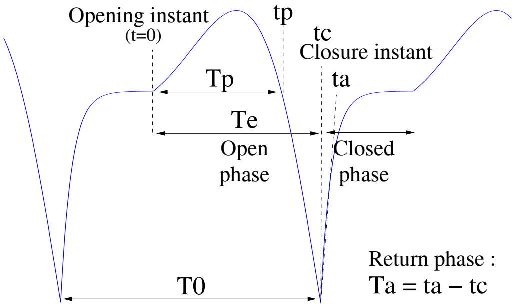

Fig. 1. The LF model

1-4244-0728-1/07/$20.00 ©2007 IEEE

sponds to the open quotient ($O_q = \frac{T_e}{T_0}$), α_m to the asymmetry coefficient ($\alpha_m = \frac{T_z}{T_e}$) and Q_a to the return phase quotient ($Q_a = \frac{T_a}{(1-O_q)T_0}$). Θ denotes the space of shape parameters. The explicit expression of the model for one fundamental period is given by:

$$u_{LF}(t) = \begin{cases} E_1 e^{at}\sin(wt) & 0 \leq t \leq T_e \\ -E_2\left[e^{-b(t-T_e)} - e^{-b(T_0-T_e)}\right] & T_e \leq t \leq T_0 \end{cases}$$

where the parameters a, b and w are implicitly related to θ and T_0.

Given the above assumptions, the speech signal $s(n)$ can be represented by means of an ARX model [4]:

$$s(n) = -\sum_{k=1}^{p} a_k(n)s(n-k) + b_0 u_{LF}(n) + e(n) \quad , \quad (1)$$

where $u_{LF}(n)$ denotes the LF glottal flow derivative and where $a_k(n)$ are the time-varying coefficients of the order p AR model characterizing the vocal tract. Coefficient b_0 is related to the LF waveform ampliftude while $e(n)$ is a residual signal. $e(n)$ contains the information that is not explicitly captured by the linear ARX-LF model, including: i) the mismatch between the deterministic part of glottal waveform and the LF model, ii) nonlinear effects such as ripple which results derom source-vocal tract interactions and iii) the noise components of the glottal waveform.

3. ANALYSIS

The analysis procedure has two main purposes: i) estimating the LF-ARX model parameters and ii) decomposing the residual signal by means of an HNM model. These steps are detailed in the next two sections.

3.1. LF-ARX parameter estimation

The estimation of the LF-ARX model parameters can be fomulated as the minimization of the energy of the residual signal. However, it can be seen from equation 1 that this optimization is highly nonlinear and thus rather intricate. A first solution to this estimation problem has been proposed in [5] and [6]. The method proposed in this paper capitalizes on these previous works and can be decomposed in the following steps:

1. Estimation of f_0 using a modified Yin algorithm including f_0 tracking [6];
2. Estimation of the glottal closure instants by using f_0 continuity constraints and an appropriateness measure to the ARX-LF model [6];
3. Regularization of the LF source sequence by means of a Viterbi algorithm;
4. Estimation of the AR parameters characterizing the vocal tract.

The regularization step is essential in order to alleviate the well known ill-conditionned character of semi-blind deconvolution problems. Its principle is to use a set of quantized LF waveform for which the appropriateness cost introduced in [5] is computed. Given this treillis, the join cost between two nodes is defined as $1 - r_{ij}$ where r_{ij} denotes the correlation between the quantized LF waveforms indexed by i and j. Then a Viterbi algorithm gives the optimal LF waveform sequence. In order to reduce the complexity of the algorithm it is worth noting that all the join cost are computed off-line and tabulated.

Once the position and the shape of the LF waveform has been obtained, the b_0 coefficient and the vocal tract parameters can be estimated by classical linear prediction techniques. However, in order to get a better resolution in the lower part of the spectrum, it is preferable to use a warped frequency scale. In this case, the AR coefficient can be estimate using a warped linear prediction (WLP) algorithm presented in [7]. A classical shortcoming of linear prediction is that it tends to associate a formant to prominent harmonics, especially in the case of high pitched speech signal. In order to prevent such estimation errors, we propose a regularization procedure. For that purpose we hypothesize that in the lower part of the spectrum ($f < 1.5$ kHz) the spectral enveloppe can be approximated by sampling the amplitude spectrum of the AR model at each harmonic frequency. Based on this assumption, we estimate a new spectral enveloppe consisting of i) a discrete cepstrum representation of the spectrum for $f < 1.5$ kHz [8] and ii) the obtained AR frequency response for $f > 1.5$ kHz. In the vicinity of 1.5kHz (more precisely, for 1.2kHz $< f <$ 1.8kHz, the amplitude spectrum is obtained by linear interpolation of both spectral representations. Then, given this amplitude spectrum, the final AR model can be determined from the corresponding autocorrelation coefficients.

3.2. HNM residual decomposition

The residue is analyzed by a version of HNM similar to the one described in [2]. It is worth noting that the fundamental frequency refinement procedure and the maximum voicing frequency estimation are done on the speech signal rather than on the residue itself. This enables more robust estimates of these two parameters.

The harmonic part is determined by minimizing the least square criterion as in [2]. Note that the residual signal is supposed not to contain the vocal tract information. Its spectral enveloppe can be considered as smooth, but includes the effect of the WLP. For that reason, it is not necessary to estimate an AR model for the noise part. The only information needed to generate the noise component is its variance corrected in order to include the frequency warping effect.

3.3. Analysis of unvoiced frames

The analysis scheme presented above suppose that the signal is voiced. Of course for an unvoiced frame, the ARX-LF model is inappropriate. So our analysis scheme needs a voicing decision mechanism. In this paper, the algorithm proposed in [2] was used. When a frame is declared unvoiced, the above analysis scheme is simply replaced by a WLP based analysis.

4. SYNTHESIS AND MODIFICATION

4.1. Synthesis algorithm

The synthesis is done pitch-synchronously by passing the reconstructed glottal source signal through a time-varying filter. For the filter determination, a classical interpolation of the line spectral pairs (LSP) with a Hanning window is used. From section 3, the glottal source $u(n)$ is the sum of three components: the LF glottal waveform $u_{LF}(n)$, as well as the harmonic part $e_h(n)$ and the noise part $e_n(n)$ of the HNM decomposition of the residue. Note that noise part is synthesized by high-pass filtering (with a cut-off frequency equal to the maximum voicing frequency F_c a white gaussian noise. Of course for unvoiced frames, only the noise component is to be considered. Each component is generated using a classical OLA procedure with a Hanning window whose length is twice the local fundamental period. Thus the reconstructed glottal source signal is

$$u(n) = w(n)u^l(n) + (1 - w(n))u^{l+1}(n) \qquad (2)$$

where $u^l(n)$ and $u^{l+1}(n)$ denote the short term glottal signals respectively obtained from the l^{th} and $l+1^{\text{th}}$ analysis instants.

4.2. Synthesis with modifications

This section gives a brief overview of the speech modification algorithm based on the ARX-LF model. The overall process can be split into two steps: i) the determination of the sequence of analysis frame indices together with their corresponding synthesis instants given a stream of f_0 and time modification coefficients and ii) the modification and synthesis of the selected speech frames. The first step is classical in prosodic modification and is detailed in [1]. For the generation step itself, we concentrate here on f_0 modification as time scale modification essentially implies at most duplication of some of the selected speech frames. We describe the way the LF component and the residual signal are modified.

Considering the LF waveform model, it is clear that the shape and amplitude parameters can be controled regardless of the fundamental frequency. Thus, a simple way of handling f_0 modification is to use a shape invariant processing. In that case, the modified LF glottal waveform $\tilde{u}_{\text{LF}}(t)$ is related to the original glottal waveform $u_{\text{LF}}(t)$ by :

$$\tilde{u}_{\text{LF}}\left(\frac{t}{\tilde{T}_0}\right) = u_{\text{LF}}\left(\frac{t}{T_0}\right)$$

which means that the spectrum of the modified glottal source is a stretched version of the original one with a duration equal to $\tilde{T}_0$. This modification process scales the absolute duration of the return and open phases by $\frac{\tilde{T}_0}{T_0}$.

Another way of modifying the fundamental frequency of the LF waveform consists in keeping the durations of the open and return phases unchanged. In this alternative method the fundamental frequency is modified along with the shape parameters :

$$\begin{array}{lcl} \tilde{T}_e = T_e & \Leftrightarrow & \tilde{O}_q = O_q \frac{T_0}{\tilde{T}_0} \\ \tilde{T}_a = T_a & \Leftrightarrow & \tilde{Q}_a = Q_a \frac{1-O_q}{1-\tilde{O}_q}\frac{T_0}{\tilde{T}_0} \end{array} ,$$

where $\tilde{O}_q$ and $\tilde{Q}_a$ respectively denote the modified open and return phase qiotients. Note that this method leads to the modification of the LF shape parameters in such a way that the spectral envelope of the LF waveform is preserved. By contrast, the first method stretches the spectral content of the LF waveform. For instance, for a fundamental frequency decrease, it increases the open phase and return phase durations and thus lower the high frequency content of the speech signal, in such a way that the resulting voice sounds more lax.

Moreover, the second method is expected to give results closer to the TD-PSOLA method which performs overlapping between frames and thus tends to preserve the open and return phase durations. For all these reasons, the second method will be used in the experiments.

5. EXPERIMENTS AND RESULTS

The experiments focus on time and fundamental frequency modifications using one of the three following methods : TD-PSOLA, HNM and the proposed ARX-LF method. These algorithms are applied on 9 sentences : 5 from a French male speaker and 4 from a female English speaker. Each sentence is transformed according to five schemes : 1) a time scaling factor of 2, 2) two fundamental frequency scalings (using factors 0.7 and 1.4), 3) two combinations of time and frequency scalings (using a time scaling of 2.0 and frequency scaling of 0.7 or 1.4).

To evaluate the proposed algorithm, two preference tests (TD-PSOLA vs ARX-LF and HNM vs ARX-LF) have been performed. For each sentence and transformation pattern, listeners select the best sounding transformed stimulus if they have a preference. As the stimuli do not sound natural especially when a time streching is applied, listeners have been asked to focus on voice quality and on possible artefacts.

The results are depicted in table 1. They show a clear preference for the proposed method in the case of time modification. The superiority of our approach over TD-PSOLA

	Male voice			Female voice		
	D	f_0	D + f_0	D	f_0	D + f_0
A1	3%	29%	5%	0%	50%	3%
A2	72%	39%	65%	81%	25%	78%
N	25%	32%	30%	19%	25%	19%

	Male voice			Female voice		
	D	f_0	D + f_0	D	f_0	D + f_0
B1	0%	3%	1%	16%	39%	22%
B2	70%	86%	74%	47%	47%	41%
N	30%	11%	25%	37%	14%	37%

Table 1. Upper table : preference test between TD-PSOLA (A1) and ARX-LF (A2), N means no preference. Lower table : preference test between HNM (B1) and ARX-LF (B2). Results merged into three class : duration transformation (D), f_0 scaling (f_0), duration and f_0 modifications (D+f_0).

was expected for time stretching. The fact that it also outperforms the HNM can be explained by the ability of our method to capture a larger amount of the deterministic part of the speech signal than the HNM would do. For instance, in the presence of jitter, the estimated maximum voicing frequency can be very low even if the signal clearly exhibits a deterministic structure. This phenomenon does not appear so critical with the LF-ARX model as it gives a reasonable source filter decomposition even when the glottal closure instants are irregularly spaced.

In the case of pitch modification, the ARX-LF model compares favourably to the HNM, especially for the male voice. This can be explained by the tonal noise present in the HNM modified stimuli. Note that this problem of phase coherence is attenuated for the female voice, which is consistent with the works of Kim [9] who shows that the phase coherence problem is more prominent when dealing with low pitched speech signals. The results are more contrasted when comparing our model to TD-PSOLA pitch modifications. Indeed, on one hand the pitch modification factors remain acceptable for TD-PSOLA algorithm. On the other hand, it can be observed that our method sometimes causes a slight degradation of the quality of the speech signal. More careful inspection of the signal shows that the analysis can lead to sudden change in the source-filter decomposition, in spite of the tracking mechanism proposed in section 3.1. This irregularity in the speech signal deconvolution process is not annoying for resynthesis nor for time scale modification, but tends to be problematic in the case of pitch modification as it can locally alter the speech signal coherence.

6. CONCLUSION

In this paper, we have presented a new method for speech synthesis and modification based on the ARX-LF model. This model can handle transparent speech resynthesis and therefore can be seen as an interesting coder that can be used in corpus based speech synthesis. Experiments have shown its abilty to yield high quality prosodic modifications, as it is superior to both TD-PSOLA as well as HNM based modification schemes. We have also briefly illustrated the usefulness of the ARX-LF model in voice quality modification. Future works will be directed to the definition of more coherent modification schemes including pitch, voice quality and even timbre modification.

7. REFERENCES

[1] E. Moulines and J. Laroche, “Non-parametric techniques for pitch-scale and time-scale modification of speech,” *Speech Communication*, vol. 16, pp. 175–205, 1995.

[2] Y. Stylianou, *Harmonic plus Noise Models for speech, combined with statistical methods for speech and speaker modification*, Ph.D. thesis, ENST, 1996.

[3] G. Fant, J. Liljencrants, and Q. Lin, “A four-parameter model of glottal flow,” *STL-QPSR*, vol. 4, pp. 1–13, 1985.

[4] W. Ding, H. Kasuya, and S. Adachi, “Simultaneous estimation of vocal tract and voice source parameters based on an ARX model,” *IEICE Trans. Inf. Syst.*, vol. E78-D, no. 6, pp. 738–743, June 1995.

[5] D. Vincent, O. Rosec, and T. Chonavel, “Estimation of LF glottal source parameters based on ARX model,” *Interspeech*, pp. 333–336, 2005.

[6] D. Vincent, O. Rosec, and T. Chonavel, “Glottal closure instant estimation using an appropriateness measure of the source and continuity constraints,” *IEEE ICASSP*, pp. 381–384, 2006.

[7] A. Harma and U.K. Laine, “A comparison of warped and conventional linear predictive coding,” *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 579–588, July 2001.

[8] O. Cappé, J. Laroche, and E. Moulines, “Regularized estimation of the spectral envelope from discrete frequency points,” *IEEE ASSP Workshop on application of signal processing to audio and acoustics*, 1995.

[9] D.S. Kim, “Perceptual phase quantization of speech,” *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 4, pp. 355–364, July 2003.