ARX-LF-BASED SOURCE-FILTER METHODS FOR VOICE MODIFICATION AND TRANSFORMATION

Yannis Agiomyrgiannakis and Olivier Rosec

Orange Labs, TECH-SSTP-VMI, Lannion, France
{yannis.agiomyrgiannakis, olivier.rosec}@orange-ftgroup.com

ABSTRACT

Two ARX-LF-based source/filters models for speech signals are presented. A robust glottal inversion technique is used to deconvolve the signal into an excitation component and a filter component. The excitation component is further decomposed into an LF part and a residual part. The first model, referred to as the LF-vocoder, is a high quality vocoder that replaces the residual part with modulated noise. The second model uses a sinusoidal harmonic representation of the residual signal. The latter does not degrade the signal during analysis/synthesis and provides higher quality for small modification factors, while the former has the advantage of being a compact, fully parametric representation that is suitable for low-bit-rate speech coding as well as parametric speech synthesis applications.

Index Terms— speech coding, LF, LPC vocoder, embedded speech synthesis, text-to-speech, modulated noise, pitch/time scaling, speech transformation/modification

1. INTRODUCTION

Besides linguistic information, voice conveys rich paralinguistic information regarding the expressive, organic and perspectival aspects of communication. Although the corresponding information layers seem to be interplexed in oral communication, it is desirable to develop the ability of an artificial manipulation of the corresponding qualities. The interest is significant; for example, expressiveness can increase the naturalness of speech synthesis and thus render it more desirable while high-quality transformation facilitates inexpensive creation of new voices from a single corpus. Most of the expressivity of speech can be captured via *prosodic modifications* like pitch and time scaling, usually addressed as *speech modifications*. The term *speech transformation*, on the other hand, refers mainly to the modification of the *organic nature* of the speech production system.

Most of the work in speech modification/transformation is made using sinusoidal models [1],[2], phase vocoders [3],[4] and non-parametric techniques like PSOLA [5]. The advantages and limitations of these approaches are well studied in the literature. State-of-the-art phase vocoders are able to robustly handle a wide range of speech and audio signals, but seem to be restricted in the following ways: 1) the synthesized speech doesn't sound natural for high modification factors, 2) there is reverberation when pitch is significantly lowered (i.e. during female to male conversion), and 3) they face difficulties when providing sophisticated voice qualities like a relaxed, a harsh or a breathy voice, etc. A significant portion of these deffiCiencies can be attributed to the weak connection between the signal model and the production mechanism of speech.

On the other hand, a stronger and more accurate speech model enables a wider range of transformations and increased naturalness. Unfortunately, though, the estimation of the state of the speech production system solely from the speech signal is a difficult inverse problem that may be solved only under simplifying assumptions. In previous work, it was shown that the ARX-LF model can be used to inverse filter the speech signal with increased robustness [6]. The ARX-LF model assumes that the speech production system is well approximated by an autoregressive (AR) filter excited by an LF (Liljencrant-Fant) glottal waveform. Traditionally, the reduced stability of glottal inverse filtering hindered its application to speech synthesis but, currently, the stability of ARX-LF-based inverse filtering has reached a level where it is possible to robustly analyze whole databases of certain (mostly male) speakers recorded under high-quality conditions. Hence, it can find application in speech synthesis.

A drawback of the ARX-LF model is that, although most of the energy of the AR residual is captured by the LF waveform, a significant portion of the excitation signal that is neither captured nor modeled remains. In [6], the residual signal (hereafter referred to as the *LF residual signal*) is described by a Harmonic-plus-Noise Model (HNM) similar to [1]. The HNM representation seems to be suitable for high-quality time/pitch scaling modifications but it is not evident how to modify the LF residual when the LF source is also modified. A solution to this problem was given by the LF-vocoder [7], an ARX-LF-based speech model that replaces the LF residual with modulated wideband noise. Any modification made to the LF source, directly transforms the non-deterministic LF residual in a perceptually pleasant manner.

The LF-vocoder is a fully parametric speech representation that yields significantly higher quality than typical LP-based vocoders and provides a wide range of high-quality glottal source modifications while retaining the ability of low bit-rate coding. The first version of the LF-vocoder (presented in [7]) suffered a quality degradation from the treatment of transient and unvoiced frames. This paper presents an improved version suitable for low-rate speech coding at approximately 5 kbps. However, the LF-vocoder model is lossy and suffers a slight quality degradation (i.e. a loss of presence), undesirable for high-quality text-to-speech (TTS) systems. Thus, a new ARX-LF-based model is proposed in this paper. It provides transparent analysis-synthesis, improved time/pitch scaling and the ability of modifying the LF glottal source while preserving most of the quality of the original signal. The new model (LF+HM) describes the LF residual using a harmonic sinusoidal representation (HM). A series of modifications of the harmonic LF residual is then proposed for the case where pitch, Vocal Tract Length (VTL) and/or LF source are altered. Finally, the two models (LF vocoder and LF+HM) are evaluated in analysis-synthesis, time/pitch scaling modifications as well as interesting transformations like gender/age transphonation, conversion to dwarf/giant, voice relaxation, whispering voice synthesis etc.

The LF model and ARX-LF-based inverse filtering is described in Section 2. Section 3 presents the new improved version of LF-vocoder. Section 4 proposes a novel ARX-LF-based model that uses a harmonic representation for the LF residual and presents a methodology that incorporates LF, pitch and VTL modifications into that model. Informal evaluation results and voice transformation examples are discussed in Section 5. Section 6 concludes the paper.

2. GLOTTAL INVERSION

The glottal flow signal is usually addressed through its derivative (GFD), which incorporates the effect of the lip's radiation to the signal observed at the glottis. The LF model [8] represents the glottal source signal with 5 parameters: one for the location of the glottal source (the reference is usually the GCI; the Glottal Closure Instant), one for the amplitude and three to define the shape of the glottal flow. Among the possible parameter sets to define the shape, the vector $\theta = (O_q, \alpha_m, Q_a)$ has been chosen: O_q corresponds to the open quotient ($O_q = \frac{T_e}{T_0}$), α_m to the asymmetry coefficient ($\alpha_m = \frac{T_p}{T_e}$) and Q_a to the return phase quotient ($Q_a = \frac{t_a - t_c}{(1-O_q)T_0}$). The explicit expression of the glottal flow derivative for one glottal cycle is given by:

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

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

Given the above assumptions, the speech signal $s(n)$ can be represented by means of an ARX model [9]:

$$s(n) = -\sum_{k=1}^{p} a_k(n)s(n-k) + b_0 g(n) + r(n), \qquad (1)$$

where $g(n)$ denotes the LF glottal flow derivative and $a_k(n)$ are the time-varying coefficients of the order p AR model characterizing the vocal tract. Coefficient b_0 is related to the LF waveform amplitude while $r(n)$ will be referred to as the *LF residual signal*.

The estimation of the parameters of the ARX-LF model is a hard optimization problem that in [6] is solved in time-domain using a codebook of LF glottal waveforms, a Viterbi algorithm to search for the optimal sequence of LF waveforms and GCI using some continuity constraints, a simplex algorithm to compensate the discrete nature of the LF codebook, warped linear prediction for the estimation of the vocal tract and a suitable AR order selection.

The LF residual $r(n)$ contains a variety of information that is neither captured by the glottal flow derivative $g(n)$ nor by the AR filter: 1) fine spectral information from the misfit of the AR spectral envelope, 2) LF parameter estimation errors, 3) nonlinear effects (i.e. ripples) from the unmodeled source/vocal tract interactions [10], 4) non-deterministic components of the excitation (i.e. friction noise, high-band noise, etc), and 5) potential modeling mismatches associated with the accuracy of the linear source/filter speech production model and the LF excitation model (i.e. at higher frequencies). Since, modeling mismatches and analysis errors may always exist in practice, it is interesting to develop methods that are natively robust to such conditions. Accordingly, we propose and evaluate three representations of the LF residual $r(n)$: a HNM [6], a modulated noise model [7] and a harmonic model (Section 4).

3. THE LF VOCODER

The LF vocoder replaces the LF residual $r(n)$ with a non-deterministic component that is a mixture of modulated and unmodulated wide-band noise. A schematic representation of the LF vocoder speech synthesis procedure is presented in Figure 2. The LF glottal flow derivative $g(n)$ is integrated with a 2-tap IIR filter to produce the glottal flow signal. A DC component is then added to the glottal flow and the modified glottal flow signal is used to modulate white noise. The idea of augmenting the LF excitation with LF modulated noise is not new [11],[12]. Similarly, it has been observed that noise is well incorporated into the signal when it exhibits a pitch synchronous time-domain distribution [13]. The DC component accounts for a percentage of friction noise that passes through the vocal chords during both the open and the closed phase [12], and it has been fixed to be 20% of the maximum value of the glottal flow. Since the modulated noise signal exhibits a quasi-harmonic structure in the spectral domain, the introduction of the DC offset rises the interharmonic noise level.

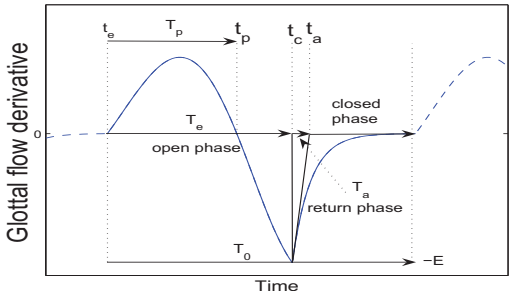

Fig. 1. The Liljencrant-Fant model

This type of modulated noise is suitable for most voiced phonemes but not for partially devoiced phonemes, voiced fricatives and transient frames between voiced and unvoiced speech segments. For the latter cases, it is better to bury the modulated noise into unmodulated noise in order to compensate the strong unvoiced character of the corresponding frame. In Figure 2 this is made by first decorrelating the noise using a delay element with a random delay of k samples and then mixing it with the modulated noise according to a proportion $\nu \in [0, 1]$. Again, in the spectral domain, this procedure rises the interharmonic noise level. The non-deterministic part $e(n)$ is then mixed with the LF glottal flow $g(n)$ according to a proportion $\gamma \in [0, 1]$. Both proportions ν and γ are directly estimated from the *energy ratio* between the glottal flow derivative $g(n)$ and the LF residual $r(n)$ using some simple weighting functions [7]. The synthesis of the voiced frames ends by derivating the excitation subsequently by applying the vocal tract AR filter.

The first version of the LF vocoder [7] suffered a performance degradation from the treatment of unvoiced frames and transients (between voiced and unvoiced segments). In [7], transient frames were synthesized by mixing a voiced component that was extrapolated from the nearest voiced frame with a locally computed unvoiced component. In this paper we propose to synthesize transient frames using only the voiced component. The parameters of each transient frame are now re-estimated by extrapolating the AR coefficients and the LF parameters from the nearest voiced frame, estimating the corresponding GCI locations using pitch information and, finally, performing a local search to refine the estimated GCI locations and the corresponding fundamental period. A Signal-To-Noise (SNR)-based criterion ensures that unvoiced frames will not be characterized as voiced. Thus, in the new version of the LF vocoder, there are only voiced and unvoiced frames.

Another improvement is made to the synthesis of unvoiced speech segments. In [7], the unvoiced segments were synthesized using a parametric modulated noise model suitable for low bit-rate

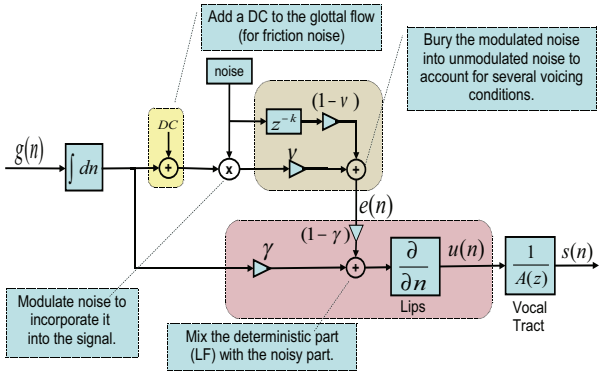

Fig. 2. Voiced synthesis using the LF Vocoder

quantization and modification but lacking high-quality. In this paper we retain the original unvoiced speech waveform and we use a PSOLA-like technique to concatenate it with the voiced (potentially modified) speech signal. Initially, during analysis, we assign a *virtual GCI* to the unvoiced region. For unvoiced to voiced transitions, for example, the virtual GCI is at $t_c' = t_c - T_0$ where t_c is the nearest GCI of the voiced region and T_0 the corresponding fundamental period. Similarly, for voiced to unvoiced transitions, $t_c' = t_c + T_0$. During synthesis, the virtual GCI t_c' is used to concatenate the unvoiced signal with the voiced signal as if both were voiced using a PSOLA-like approach [5]. During pitch/source modifications, the proposed scheme reduces the concatenation distortion between unvoiced and voiced segments, especially when the unvoiced segment (undesirably) contains a voiced glottal period or a vowel onset. In that case, the *virtual GCI* is an approximation of the location of the GCI of the glottal cycle that is incorrectly classified into the unvoiced region, and as such, it can be efficiently used for concatenation with PSOLA-type approaches. Furthermore, the quality of the concatenation is not effected when the unvoiced segment contains only unvoiced information. Finally, note that the unvoiced segments can be efficiently quantized using CELP techniques [14]. Therefore, the new LF vocoder can still be used as a low bit-rate representation suitable for embedded text-to-speech synthesis.

A significant advantage of the LF vocoder is that its fully parametric nature allows a wide range of modifications to be made natively in the parametric space. Thus, time/pitch modifications are made by simply changing the fundamental period of voiced frames and using a frame insertion/deletion mechanism that ensures the desirable tempo. A glottal source modification is made directly on the LF parameters, while vocal tract modifications are made directly on the AR spectral envelope.

4. THE LF + HARMONICS MODEL

The LF+HM model represents the LF residual $r(n)$ as a sum of harmonically related sinusoids:

$$r(n) \approx \sum_{k=1}^{K} A_k \cos(k\omega_0 n + \phi_k), \qquad (2)$$

where ω_0 is the fundamental frequency (in rad/s), A_k the amplitudes and ϕ_k the phases of the harmonically related sinusoids. The analysis of $r(n)$ is made up to a cutoff frequency of 7800 Hz using typical least squares methods [1]. In analysis-synthesis, LF+HM closely approximates the original waveform because the harmonic residual $r(n)$ increases the reconstruction SNR higher than 20 dB. Upon pitch scaling, LF source or vocal tract modifications, though, the LF residual $r(n)$ must be altered accordingly.

Let us consider vocal tract length (VTL) modifications. An increase of the length of the vocal tract corresponds to a linear shift of the formants towards lower frequencies and vice versa [15]. Since pitch and VTL are correlated with speaker size, gender and age [15], it is possible to perform the corresponding transformations using simple frequency scaling $\omega' = \lambda\omega$, or more generally using a frequency warping function $\omega' = f_w(\omega)$. Such a vocal tract transformation may however lead to a potential misalignment between the spectral details contained in the residual $R(z)$ and the modified vocal tract. In order to compensate this, it is necessary to resample $R(e^{j\omega})$ at the warped frequency axis ω': $\tilde{R}(e^{jk\omega_0}) = R(e^{jf_w^{-1}(k\omega_0')})$, where $\tilde{R}(z)$ is the new (modified) LF residual. Since we have access only to the harmonics $R(e^{jk\omega_0})$ of the LF residual, we must resort to some type of interpolation to obtain $\tilde{R}(e^{jk\omega_0})$. Satisfying results are obtained if amplitudes are linearly interpolated in the logarithmic domain and phases are computed using *nearest neighbor* interpolation.

For pitch scaling modifications, the modified LF residual $\tilde{R}(e^{j\omega})$ must be resampled at the new harmonic frequencies $k\omega_0'$, where ω_0' is the new fundamental frequency. For that purpose, the same interpolation scheme as described above is used.

Glottal source modifications are less straightforward. Let $U(z) = b_0 G(z) + R(z)$ be the z-tranform of the AR excitation from equation (1) and $\tilde{G}(z)$ be the desired glottal flow derivative according to the LF model. We propose the following transformation of the glottal source $U(z)$:

$$\tilde{U}(z) = \mu \frac{\tilde{G}(z)}{G(z)} U(z) = \mu b_0 \tilde{G}(z) + \mu \frac{\tilde{G}(z)}{G(z)} R(z), \qquad (3)$$

where μ is a scale factor that preserves the energy of the excitation. The latter transformation performs two tasks; first, it changes the LF source $G(z)$ with the new one $\tilde{G}(z)$, second, it shapes the spectrum of the residual in a way that is *consistent* with the modification of the source. Therefore, the new residual is

$$\hat{R}(k\omega_0') = \mu \frac{\tilde{G}(k\omega_0')}{G(k\omega_0')} \tilde{R}(k\omega_0'). \qquad (4)$$

The latter operation can be computed directly on the amplitudes and the phases of the harmonics of $\tilde{R}(z)$,$G(z)$ and $\tilde{G}(z)$. Furthermore, all the modifications of the LF residual can be computed efficiently in one step. The synthesis is made by mixing the new LF source and the LF residual $\hat{R}(k\omega_0')$ while ensuring that the energy ratio between the two components is not altered.

Analysis, synthesis and modification of voiced frames is made in a pitch-synchronous manner using a window of two fundamental periods. Unvoiced speech and voiced/unvoiced concatenation are treated as described in Section 3. In fact, LF+HM and LF-vocoder share a considerable amount of code and differ only on the synthesis of the (voiced) glottal cycle. In addition, LF+HM uses the same information (pitch, energy, energy ratio, LF, vocal tract) with the LF vocoder, augmented only by a second layer of information that holds the harmonic representation of the residual.

5. EXPERIMENTS AND RESULTS

The voice transformation abilities of the two presented models are demonstrated by making compound modifications that alter pitch, vocal source and VTL simultaneously. With an appropriate pitch

scaling and VTL modification we can alter the perceived age, gender and size of the speaker [15]. For example, starting from a *male* voice, a *female* voice can be obtained by shifting formants 20% higher and raising pitch 60-100% (depending on the average pitch of the utterance). Reverse modifications convert female voices to male voices. A *dwarf* voice can be synthesized by shifting formants higher (reducing VTL) and lowering pitch by 0-25%. Raising pitch significantly (100%-300%) and scaling formants by 33-50% gives the impression of a child. Lowering pitch significantly while scaling formants downwards generates a toy voice that sounds like a *giant* character. All these modifications can be made together with glottal source modifications. For example, the LF-vocoder can synthesize *whispering voice* by ignoring the LF source (i.e. setting energy ratio to zero) and applying a dynamic compression scheme on the energy of each frame. It can be used to synthesize a *whispering child*. A *relaxed voice* can be synthesized by increasing the asymmetry coefficient α_m and the return phase of the LF glottal flow derivative: Thus, we are able to generate the toy voice of a *relaxed giant* character. An elaborated evaluation of the quality, the naturalness or the pertinence of the transformations is beyond the scope of this paper. Furthermore, the aim of the paper is to demonstrate the ability of changing voice properties and not to propose systematic modifications for voice generation. For this purpose, we provide a range of proof-of-concept transformation examples to the multimedia database of the conference.

The overall impression that we obtained from several informal listening tests is that LF+HM provides analysis/synthesis results that are nearly indistinguishable from the original speech. This is not the case of the LF-vocoder which provides speech slightly inferior than LF+HM, although the newer version of the LF vocoder is much improved over the previous version. The advantage of LF+HM over LF vocoder is kept when the modifications are relatively minor, i.e. when pitch scaling is within $\pm 40\%$ of the original value or when the glottal source is only slightly altered. However, the advantage is gradually lost as modifications become stronger; for example, the two models provide more-or-less the same quality when pitch modifications are higher than 100%. In high modification factors and/or compound modifications of pitch, VTL and glottal source, we noticed that LF-vocoder sometimes outperforms LF+HM. This may be attributed to its parametric nature that allows a higher degree of consistency between the deterministic and the stochastic part of the AR excitation.

On the other hand, it must be stated that the performance of ARX-LF-based source/filter models is intimately related to the success of the glottal inversion process, which is not always guaranteed. For instance, in the vicinity of transient frames, the estimated LF glottal flow occasionally fails to capture a plausible glottal waveform. The inversion is also sensitive to the recording conditions (e.g. reverberation and noise) and also to the speaker's speaking style (e.g. creaky or breathy phonations). Such failures manifest themselves upon speech modification but remain very difficult to predict by means of objective measures. Still, using the proposed models it is possible to analyze entire speech corpora dedicated to speech synthesis for which very pleasant and high quality transformations can be done in a fully parametric manner.

6. CONCLUSION

Two ARX-LF-based speech models that allow a wide range of speech modifications/transformations were presented in this paper. They are versatile representations that map speech to parameters directly associated with the speech production system. This permits us to go beyond typical time/pitch scaling and conduct *exotic* transformations like softening the voice of a speaker, synthesizing whispering, breathy, creaky, harsh voice, etc. When coupled with vocal tract transformations they can be used for gender and age modifications. Furthermore, the LF-vocoder is suitable for low-bit-rate coding.

7. REFERENCES

[1] Y. Stylianou, "Applying the Harmonic-plus-Noise Model in concatenative speech synthesis," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[2] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice Hall, 2001.

[3] P. Depalle and G. Poirrot, "SVP: A modular system for analysis, processing and synthesis of sound signals," in *Proceedings of the International Computer Music Conference*, 1991.

[4] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Audio and Speech Processing*, vol. 7, no. 3, 1999.

[5] 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.

[6] D. Vincent and O. Rosec, "A new method for speech synthesis and transformation based on a ARX-LF source-filter decomposition and HNM modeling," in *ICASSP*, 2007.

[7] Y. Agiomyrgiannakis and O. Rosec, "Towards Flexible Speech Coding for Speech Synthesis: an LF+Modulated Noise Vocoder," in *INTERSPEECH*, 2008.

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

[9] 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.

[10] T.V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its component," *Speech Communication*, vol. 1, pp. 167–184, Dec. 1982.

[11] C. d'Alessandro, N. d'Alessandro, S. Le Beux, J. Simko, F. Cetin, and H. Pirker, "The Speech Conductor: Gestural Control of Speech Synthesis," in *eNTERFACE 2005: The Summer Workshop on Multimodal Interfaces*, July 2006, http://www.isca-speech.org/archive/einterface05/eint5_52.html.

[12] D. Mehta and T. F. Quatieri, "Synthesis, Analysis and Pitch Modification of the breathy vowel," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 2005.

[13] J. Skoglund and B. Kleijn, "On time-frequency masking in voiced speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, 2000.

[14] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems*, John Wiley and Sons, 2004.

[15] D. R. Smith and R. D. Patterson, "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," *Journal of Acoustical Society America*, vol. 118, no. 5, pp. 3177–3186, November 2005.