THE NEW VERSION OF THE ROMVOX TEXT-TO-SPEECH SYNTHESIS SYSTEM BASED ON A HYBRID TIME DOMAIN-LPC SYNTHESIS TECHNIQUE

Attila Ferencz, István Nagy, Tünde-Csilla Kovács,
Maria Ferencz, Teodora Rațiu

Software ITC, 109 Republicii street, 3400 Cluj-Napoca, Romania,
tel: +40-64-197681, fax: +40-64-196787,
e-mail: Attila.Ferencz@sitc1.dntcj.ro

ABSTRACT

Through the years we developed several TTS systems for the Romanian language, each of them presenting some advantages and disadvantages [2]. Taking into account that waveform coding (time domain) methods assures a maximum level of intelligibility and naturalness of the synthesized speech, and that prosodic effects superimposing requires the alteration of pitch (frequency domain), we developed a hybrid time domain-LPC method, obtaining a better quality of the synthesized voice than those obtained with our former systems. This paper presents some theoretical considerations, signal processing and implementation aspects of this new synthesis method developed for the ROMVOX TTS system.

1. INTRODUCTION

Researches concerning with TTS synthesis started in 1991 at Software ITC Cluj-Napoca. Besides language specific aspects (diphone database construction, text-preprocessing, grapheme to phoneme conversion, prosodic effects superimposing for the Romanian language) [1] our researches dealt with signal processing, resulting several architectures of the correspondent TTS systems like those presented in [2]. Wishing to obtain a better quality of the synthesized sound and to eliminate the disadvantages of the previous systems ([2]), we developed a new synthesis method that combines the advantages of time domain signal processing with easy pitch modification (required by intonation).

2. THEORETICAL ASPECTS

For analysis and synthesis purposes, speech production is often modeled with a source-filter model, presented in [3]. This model consist of a source, producing a signal $g(t)$ which models the air flow passing the vocal cords, a filter with transfer function $H(j\omega)$ which models the spectral shaping of the vocal tract and a differential operator R which models the conversion of the air flow to a pressure wave $s(t)$ as it takes place at the lips and which is called lip radiation. It is possible to combine the differential operator with the source that now produces the time derivative $\dot{g}(t)$ of the airflow passing the vocal cords, resulting a simplified model. The time derivative $\dot{g}(t)$ can be considered a pressure wave at the level of the glottis. Our approach takes into consideration the behavior of the glottal pulse (for voiced sounds) which can be described using the Liljencrants-Fant (LF) model, [3].

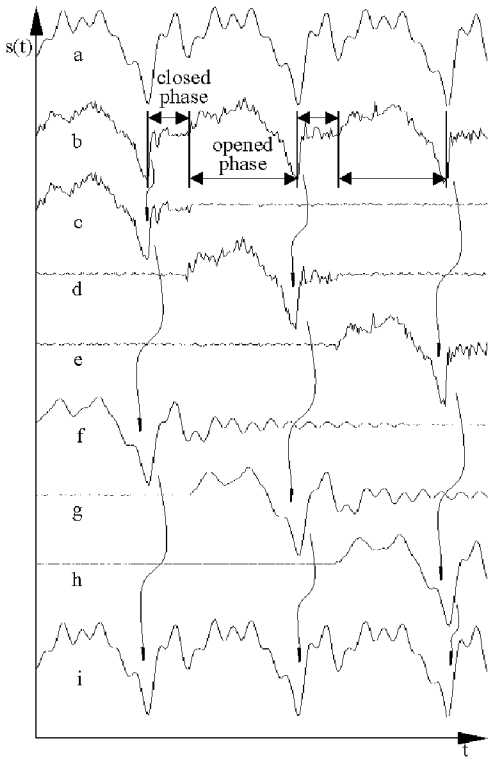

Figure 1: Waveforms of the Romanian vowel o, uttered by a male (3 pitch periods)

Figure 1.**a** presents the time domain waveforms of the Romanian vowel **o**, the corespondent $\dot{g}(t)$ source signal (Figure 1.**b**.). As it can be seen, during the opened phase of the glottis in which the $\dot{g}(t)$ source signal contains values which are different from zero (also positive and negative values), the source signal assures the excitation of the filter, resulting a generated waveform which depends on the resonance characteristics of the vocal tract, respectively on the transfer function $H(j\omega)$.

During the closed phase of the glottis (no pressure wave) the vocal tract respectively the filter doesn't get energy anymore, so the generated waveform results in this phase as combination of damped oscillations, due to the energy accumulated in the filter during the open phase. If the source signal would consist of a single opened phase of the glottis followed by a long closed phase the generated waveform would be damped, ending with no oscillations. Because in reality the next opened phase follows immediately after a relatively short previous closed phase, the generated waveform will contain the effects of both the effects of the previous state and the effect of the new excitation. Taking into account that the above model is a linear model, the two effects are combined by simple addition, in concordance with the theorem of superposition. This is equivalent to consider that the source signal consists of a few individual signals (waveforms **c**, **d**, and **e**) corresponding each to an individual opened-closed phase of the glottis, and each such individual source signal will excite the filter resulting also individual output signals (waveforms **f**, **g**, and **h**). From the superposition of these output signals results the initial, whole output signal. The waveforms in Figure 1 presents such a case for three pitch periods.

3. SIGNAL PROCESSING ASPECTS AND IMPLEMENTATION

Our synthesis approach that assures the re-synthesis of the initial signal with modifiable pitch is based on this principle of superposition. Pitch modification means the modification of the distances between two consecutive opened-closed cycles, in which the effect of the previous cycle will be combined with the effect of the new excitation in a different manner but exactly in concordance with the principle of superposition. This means that it is necessary (in a previous analysis phase) to decompose the original signal in pitch-synchronous individual signals as those presented in Figure 1, signals **f**, **g**, **h**. In the synthesis phase we have to superimpose this individual signals at new distances in concordance with the desired new pitch.

The main problem is the decomposition of the initial signal into individual, pith-synchronous signals. A first (asynchronous) analysis phase has the goal to generate pitch-synchronous markers that will be used during the second, pitch-synchronous analysis phase. The operations performed during the second phase are illustrated for three pitch-periods in Figure 2.

Waveform **a** presents the initial signal $s(t)$ in concordance with the pitch-synchronous markers determined in the first phase, the first pitch-synchronous window is applied, resulting the waveform **b**. Because the pitch-synchronous markers are placed at the beginning of the opened phase of the glottis each frame will contain one opened phase followed by the closed phase of the glottis.

Using LPC analysis for waveform **b** the obtained prediction coefficients will encapsulate information which reflects the resonance characteristics of the vocal tract. The next step is the extension of waveform **b** by generating its dumped regime based on the time-domain recurrence formula of a digital IIR filter using the LPC coefficients:

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + ... + a_p s(n-p)$$

where p is the prediction order and a_1 , a_2 , ..., a_p are the prediction coefficients. As presented before this damped signal is due to the accumulated energy in the filter, and is determined by the resonance characteristics of the filter. The obtained signal $s_1(t)$ (waveform **c**) represents the first individual pitch-synchronous signal that was extracted from the initial $s(t)$ signal. In order to extract the next individual pitch-synchronous signal it is necessary to annul the effect of the first (previous) individual pitch-synchronous signal on the second (next) one, on the third one, a. s. o. This is performed by subtracting signal $s_1(t)$ from signal $s(t)$, respectively:

$$s'(t) = s(t) - s_1(t)$$

The resulted $s'(t)$ signal is presented in waveform **d**. So the $s(t)$ signal was prepared to be extracted the second individual pitch-synchronous signal $s_2(t)$. In order to realize this goal, the previous presented steps are repeated, respectively:

- framing,
- LPC analysis,
- damped signal extension,
- annulment of the actual (second) individual pitch-synchronous signal on the next ones ($s''(t) = s'(t) - s_2(t)$).

In the same way, repeating these steps for the third frame results $s_3(t)$.

As it can be observed, the initial $s(t)$ signal is decomposed step by step in $s_1(t)$, $s_2(t)$ and $s_3(t)$, which we called **individual pitch-synchronous signals**. At the end of all these steps $s''(t)$ (waveform **j**) is equal to zero for all these three pitch periods, demonstrating that the decomposition was performed correctly.

The ROMVOX text-to-speech synthesis system is based on diphone concatenation. So the diphone sound inventory was transformed for this new system, using the decomposition method described above. It was obtained a new sound database in which each diphone was decomposed in pitch periods. Although the proposed technique was initially dedicated only for the pitch-modification of voiced sounds, it was successfully applied for unvoiced sounds, too. For unvoiced sounds there were placed some "artificial pitch markers" in the first, asynchronous analysis phase, the second phase being the same as for voiced sounds. Pitch modification can be easily obtained reconstructing the signal from the elementary pitch periods of each diphone. For speech rate variations (lengthening or shortening) it was applied the same technique as that presented in [4] and [5].

Figure 3 presents a case in which one individual pitch-synchronous signal is used to generate a longer output signal with modified pitch. The signal starts with a higher fundamental frequency (one octave higher), decreases to the initial value of the pitch (at the middle of the signal), continuing to decrease to lower values (one octave lower).

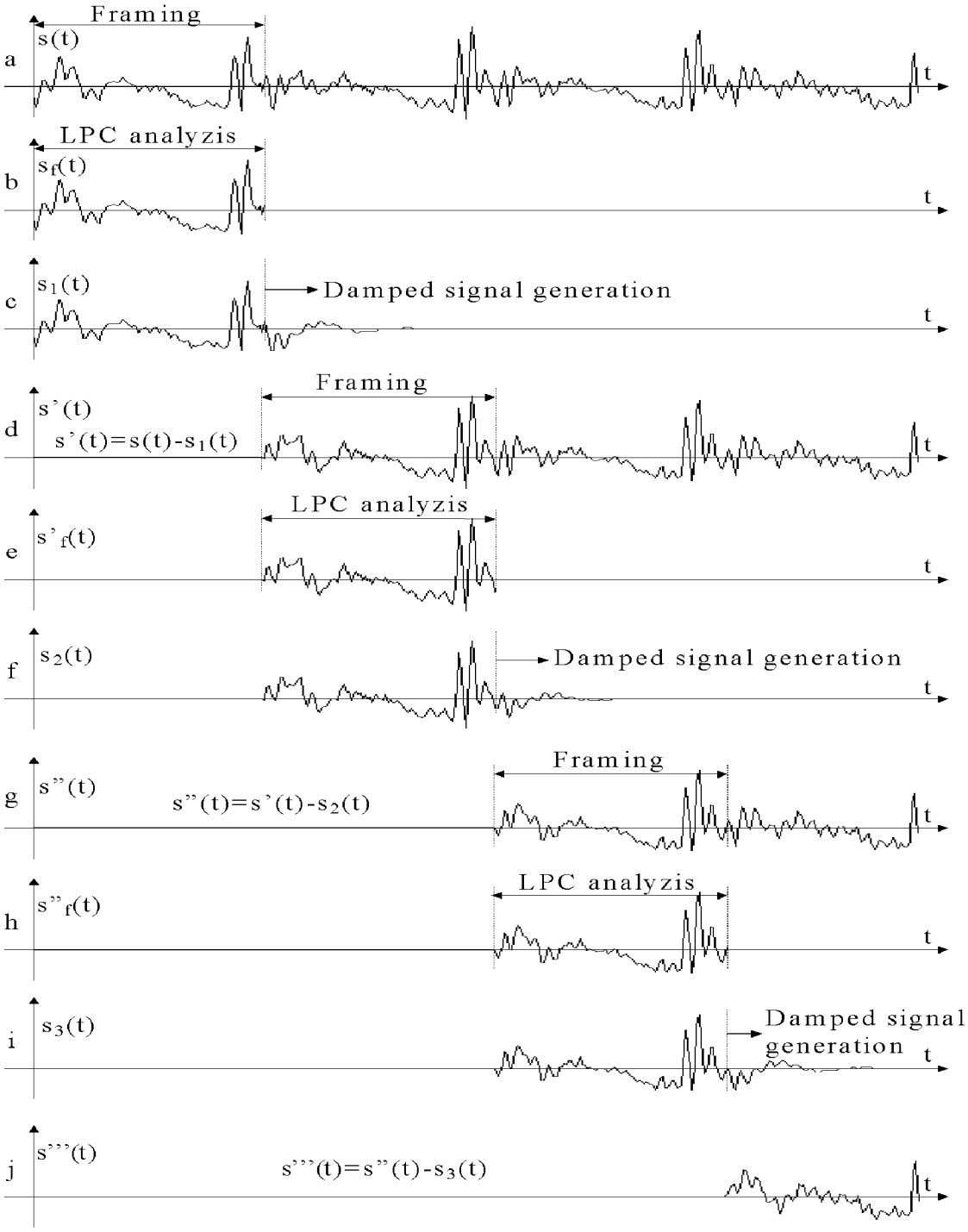

Figure 2: Operations performed during the second, pitch-synchronous, phase of analysis

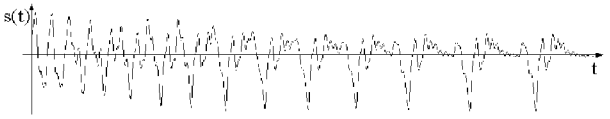

Figure 3: Waveform of vowel o, re-synthesized with modified pitch

4. CONCLUSIONS

As presented before, the aim of our research was to develop an improved synthesis technique that should assure a better quality of the generated signal. The improvement concerns the signal processing part and it presents the following aspects and advantages with respect to our previous developments [2], respectively to other synthesis techniques.

The classical LPC synthesis method presented inconveniences because on the one hand it required an additional homemade DSP-board to assure real time operation, and the other hand the quality of the generated voice was less natural. An other version of the system based on the Philips PCF8200 formant synthesizer presented the disadvantage that the transformation of the sound inventory could not be fully automated and required always continuous formant tracks, could not handle special cases in which the formant tracks suffered discontinuities, as it happens in reality.

The TD-PSOLA (Time Domain Pitch Synchronous Overlap-add) developed by CNET is a very simple but ingenious method which assures high voice quality, the only disadvantage is that it is based on a time-domain windowing technique which can introduce some spectral distortions during the pitch modification. The result of these spectrum distortions can be interpreted as a reverberation of the desired pitch-modified signal. TD-PSOLA requires at the same time a very exact pitch synchronous framing; any framing error may cause the unpleasant increase of this reverberation effect. The first disadvantage was solved by CNET through adopting the LP-PSOLA technique.

Our approach doesn't use any windowing technique, so this source of spectral distortion is eliminated. Figure 4 presents the spectral behavior of a generated signal with k=0.66 fundamental frequency modification (decreasing fundamental frequency) respectively with k=1.5 (increasing fundamental frequency), both cases in comparison with the spectrum of the initial signal. As both figures show, the peaks of the modified harmonics are situated almost on the ideal imaginary spectrum envelope.

Although the proposed technique was initially dedicated only for pitch-modification of voiced sounds, the experimental results demonstrated that the same approach can be applied for limited lengthening/shortening of unvoiced sounds. Other advantage of this technique is that it runs in real time without any additional hardware.

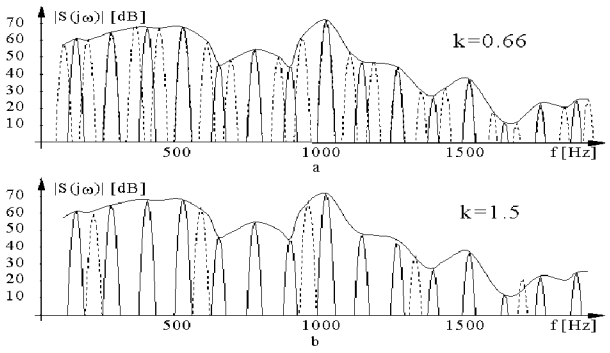

Figure 4: Power spectrum of vowel o:
- continuous line: initial pitch
- dotted line: modified pitch

The disadvantage of the method is that pitch modification is based only on the modification of the duration of the closed phase of the glottis, so the correspondent virtual open/close ratio suffers an undesired modification, but which is not perceptually very relevant.

5. REFERENCES

1. Ferencz, A., et al., *ROMVOX - Experiments regarding Unrestricted Text-to-Speech Synthesis for the Romanian Language*, Proceedings of the Ninth International Workshop on Natural Language Generation, Niagara-on-the-Lake, Ontario, Canada, page 304-307, 1998

2. Ferencz, A., et al., *The Evolution of the ROMVOX Text-to Speech Synthesis System from Monotonous to Enhanced, DSP-based Version*, Proceedings of SPECOM'97 International Workshop, Cluj-Napoca, page 179-184, 1997

3. Veldhuis, R.N.J., *An alternative for the LF model*, IPO Annual Progress Report 31, Eindhoven, page 100-108, 1996

4. Ferencz A. et al., *Experimental Implementation of the LPC-MPE (Multi-Pulse Excitation) Synthesis Method for the ROMVOX Text-to-Speech System*, Proceedings of the International Workshop Speech and Computer, SPECOM'96, St. Petersburg, Russia, page 159-164, 1996

5. Ferencz A. et al., *Experimental Implementation of Pitch-Synchronous Synthesis Methods for the ROMVOX Text-to-Speech System*, Proceedings of the 5th European Conference on Speech Communication and Technology, EUROSPEECH'97, vol. 5, page 2439-2442, 1997

6. Charpentier, F., Moulines, E., *Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones*, Proceedings of the 1st European Conference on Speech Communication and Technology, EUROSPEECH'89, Paris, vol. 2, page 13-19, 1989