6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ON A HYBRID TIME DOMAIN-LPC TECHNIQUE FOR PROSODY SUPERIMPOSING USED FOR SPEECH SYNTHESIS

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

Software ITC Cluj-Napoca, 109 Gheorghe Bilaşcu street, 3400 Cluj-Napoca, Romania
Tel.: +40-64-197681, Fax: +40-64-196787
e-mail: {Attila.Ferencz, Istvan.Nagy, TCs.Kovacs, Teodora.Ratiu, Maria.Ferencz}@sitc1.dntcj.ro

10.21437/Eurospeech.1999-399

ABSTRACT

Wishing to obtain a more natural quality of the synthesized speech and to eliminate the disadvantages of the previous text-to-speech (TTS) systems evolved in our institute (Software ITC Cluj-Napoca, Romania), we experimented and developed a new synthesis method that combines the advantages of time domain signal processing with the requirement of pitch and duration modification (required by intonation). This paper presents some theoretical considerations, signal processing and implementation aspects of this pitch alteration technique that was adapted for the new version of the ROMVOX TTS system.

1. INTRODUCTION

For analysis and synthesis purposes, speech production is often modeled with a source-filter model, presented in [3]. For voiced sounds this model consist of a source, producing a signal $g(t)$ which models the air flow passing through the vocal cords, a filter with transfer function $H(j\omega)$ which models the spectral shaping of the vocal tract (which is an acoustical resonator) 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, presented also in [3].

2. FOUNDAMENT OF THE APROACH

Figure 1 waveform **a** presents three pitch periods of the time domain waveforms belonging to the Romanian vowel **o**, respectively waveform **b** to the corespondent $\dot{g}(t)$ source signal. As it can be seen, during the open phase of the glottis in which the $\dot{g}(t)$ source signal

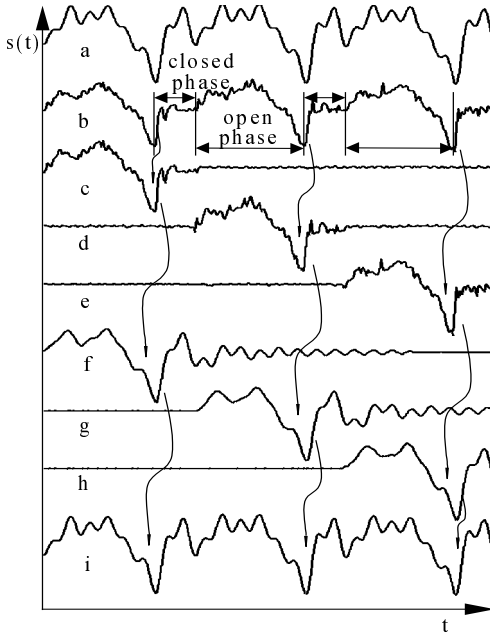

Figure 1: Waveforms of the Romanian vowel o, uttered by a male (3 pitch periods)

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 its transfer function $H(j\omega)$.

During the closed phase of the glottis (when there is no pressure wave on the level of the glottis) the vocal tract (filter) doesn't get energy anymore, so the generated waveform on the level of the lips results in this phase as combination of (convoluted) damped oscillations, due only to the energy accumulated in the filter during the previous open phase. If the source signal would consist only of a single open phase of the glottis followed by a long closed phase, the generated waveform (the output signal of the filter) would be damped (passed over in a state without oscillations). Because in reality the next open phase follows immediately after a relatively short previous closed phase, the generated waveform will contain both the effects of a limited number of previous

excitation cycles 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 (overlapping). This is equivalent to consider that the source signal consists of several individual signals (waveforms **c**, **d**, and **e**) corresponding each to an individual open-closed phase of the glottis, and each such individual source signal excites the filter generating also some individual signals which we called **individual pitch-synchronous output signals** or more simply **individual output signals** (waveforms **f**, **g**, and **h**). In concordance, the initial output signal (waveform **a**) can be considered as the superposition of all these individual output signals (waveform **i**).

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 (time intervals) between two consecutive open-closed cycles of the glottis, in which the effects of the previous cycles will be combined with the effect of the new excitation exactly in concordance with the principle of superposition. This means that it is necessary (in a previous analysis phase) to decompose the original speech signal in the above presented individual pitch-synchronous signals, respectively as those presented in Figure 1, signals **f**, **g**, **h**. In the synthesis phase we have to superimpose this individual output signals at new time intervals in concordance with the desired new pitch of the generated signal.

The main problem is the decomposition of the initial signal into individual pith-synchronous output signals. In this order 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 (pitch synchronous frames) in Figure 2.

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

Performing LPC analysis on waveform **b** the obtained prediction coefficients will encapsulate information which is related to the resonance characteristics of the vocal tract. The next step is the extension of waveform **b** with its dumped regime, extrapolation that is based on

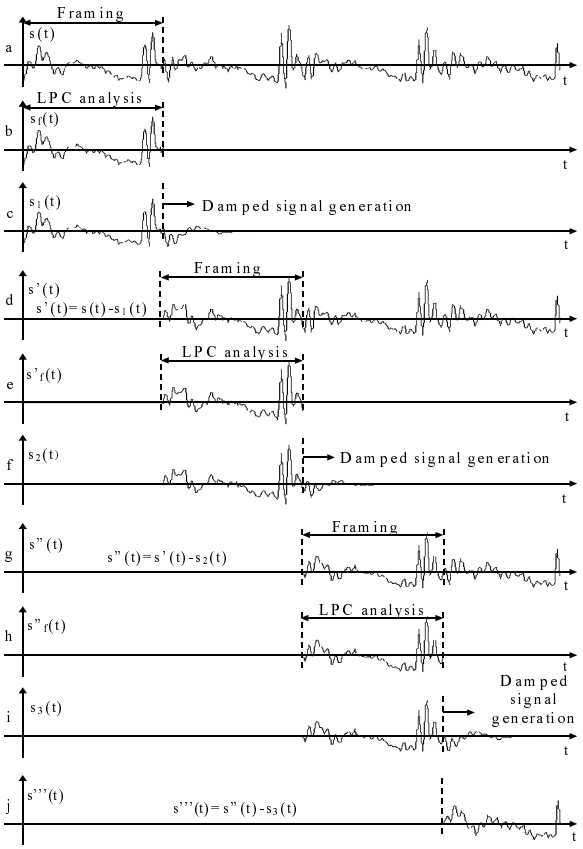

Figure 2: Operations performed during the second, pitch-synchronous analysis phase

the technique of linear prediction and can be computed using the time-domain recurrence formula of a digital IIR filter (in the absence of excitation), using the above LPC coefficients as filter 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 this 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 $s_3(t)$.

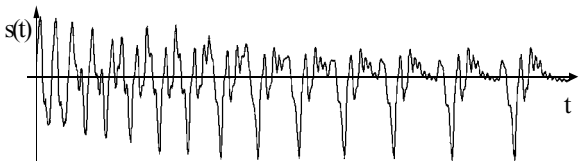

Figure 3: Waveform of vowel o, re-synthesized with modified pitch

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 are the expected 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 converted for this new version of the system, using the decomposition method described above. It was obtained a new sound database in which each diphone was decomposed in its individual output signals. Although the proposed technique was initially dedicated only for the pitch-modification of voiced sounds, it can successfully applied for limited lengthening or shortening of unvoiced sounds, too. Therefor in the case of 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.

4. SPEECH SYNTHESIS

In the synthesis phase the individual pitch-synchronous signals are overlapped (in concordance with the principle of superposition) with the disparity, calculated upon the desired pitch modification. Some of these signals can be repeated or omitted if we desire a higher or lower speech rate, being applied the same technique as that presented in [4]. If amplitude (energy) variation is desired, before overlapping the individual segments, they can be multiplied with the desired factor.

Figure 3 presents a case in which one individual pitch-synchronous signal was 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 and continues to decrease to one octave lower values.

For prosody effects superimposing we considered that the average fundamental frequency for monotonous male speech is 130Hz, a value very close to the fundamental frequency of the voice of the actor we collaborated to record and to implement the primary diphone database. Using this synthesis technique the fundamental frequency can be altered between 50% and 200% relatively to the initial value without considerable voice quality loss. In the present version of ROMVOX this interval is not exploited completely because such a large

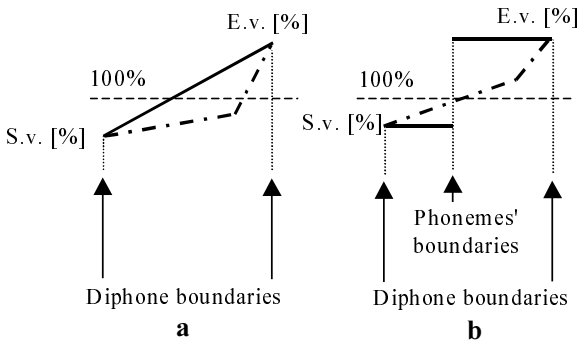

Figure 4: Linear interpolation inside the diphones of the contour of the relative fundamental frequency and of the relative amplitude variation (a) and of the relative duration variation (b).

Note: 1. The dash-dot line represents the variation of the parameters before the interpolation, the continuous line represents the new variation of the parameters after interpolation
2. S.v. = start value, E.v. = end value

variation of the pitch is not frequent in the common speech, but this interval will be extended in future versions which will have the aim of emotional speech synthesis.

The database of ROMVOX being a diphone-based one, in concordance with desired prosody there has to be established the values of the relative fundamental frequency variation, relative duration variation and relative amplitude variation of the synthesized signal for the onset and the offset of each diphone belonging to the phonetic transcription of a given text. The synthesis program performs a linear interpolation between the two values of these parameters corresponding to the diphone's boundaries as it is shown in Figure 4. Maintaining during synthesis the initial values (100%) for all these parameters would result a monotonous synthesized voice.

The starting values (S.v.) and ending values (E.v.) of these parameters result from the analysis performed upon the stylized intonation curve of the entire sentence or phrase. The stylized intonation contour is obtained by superimposing the one corresponding to the sentence type (e.g. declarative, interrogative or imperative, commas' position) - as it was described in [2] - with the one which is related to the word accent. An example is shown in Figure 5. The sentence from the example is: "Citeşte piesa, nu-i aşa?" (He is reading the play, doesn't he?).

Because the word accent in Romanian is mainly free, there can not be developed rule based algorithms, so the accent's position is determined using a database which contains the words including special accent markers. For those words, which were not included so far in the database, it is possible to mark the stressed syllable/syllables during typing the text.

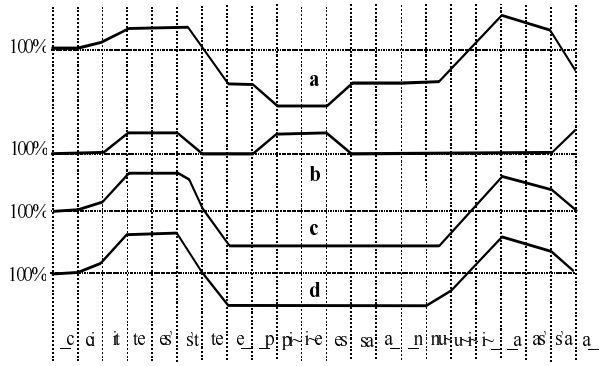

Figure 5: The determination of the stylized pitch contour
a - stylized intonation contour for the sentence type
b - stylized intonation contour for word accent
c - stylized intonation contour resulted superimposing contours in **a** and **b** (**a**+**b**-100%)
d - stylized intonation contour resulted after the linear interpolation between S.v. and E.v. for each diphone

To be able to experiment and establish different prosodic patterns it was developed a version which allows the modification of the fundamental frequency, of the duration and of the energy of the signal by editing their contours on the screen (shown in Figure 6) using the mouse as input device. These contours were drawn for the same sentence.

5. CONCLUSION

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 the previous versions of ROMVOX [2], respectively to other synthesis techniques.

The classical Linear Prediction Coding (LPC) synthesis method presented inconveniences because on the one hand it required an additional homemade DSP-board to assure real time operation, and on 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. It could not handle special cases in which the formant tracks suffered discontinuities, as it happens in reality.

Our approach doesn't use any windowing technique, so at least this source of spectral distortion is eliminated [5]. 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.

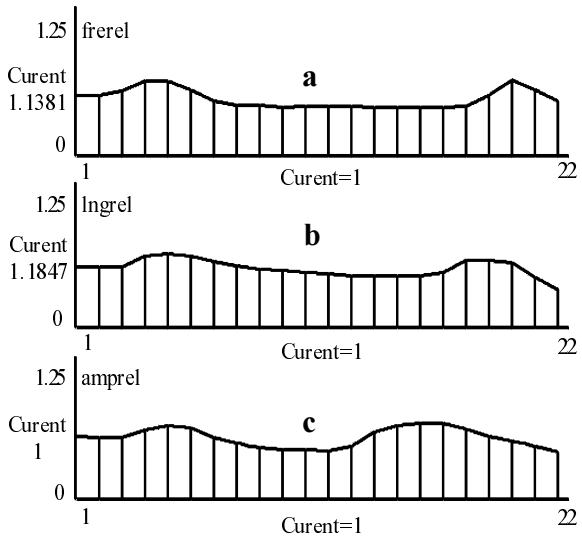

Figure 6: The desired parameter variation performed by drawing the contour of the parameters' relative variation (**a -** fundamental frequency, **b -** duration, **c -** energy)

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.

6. REFERENCES

[1] Ferencz, A., et al., *ROMVOX - Experiments regarding Unrestricted Text-to-Speech Synthesis for the Romanian Language*, Proceedings of the 9th 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 Pitch-Synchronous Synthesis Methods for the ROMVOX Text-to-Speech System*, Proceedings of the 5th European Conference on Speech Communication and Technology, EUROSPEECH'97, Rhodes, vol. 5, page 2439-2442, 1997

[5] 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