ROMVOX – EXPERIMENTS REGARDING UNRESTRICTED TEXT-TO-SPEECH SYNTHESIS FOR THE ROMANIAN LANGUAGE

ATTILA FERENCZ*, TEODORA RAŢIU*, MARIA FERENCZ*,
TÜNDE-CSILLA KOVÁCS*, ISTVÁN NAGY*, DIANA ZAIU**

* Software ITC, 109 Republicii street, 3400 Cluj-Napoca, Romania,
tel: +40-64-197681, fax: +40-64-196787, e-mail: Attila.Ferencz@sitc1.dntcj.ro
** Technical University of Cluj-Napoca, 26 Gh. Bariţiu street, 3400 Cluj-Napoca, Romania

Abstract. The ROMVOX Text-to-Speech synthesis system developed by our team is the first one that allowed the synthesis of any unrestricted Romanian text with intonation facilities on IBM-PC compatible computers. During the last years of research several version of text-to-speech systems were achieved, trying to enhance their facilities. Our paper describes the present stage of our experiments performed in order to improve the naturalness of the generated voice.

1. Introduction

Speech synthesis systems are expected to play important roles in advanced user-friendly human-machine interfaces. Wishing to realize an as good as possible text-to-speech system for the Romanian language the research started with the development of the software for monotonous speech synthesis, which simply concatenated the elements of the speech database. Prosodic aspects need to impose a correspondent modification of the synthesized speech signal, modification performed in the second version based on LPC. The experimental results using the classical LPC synthesis method proved that the quality of the synthesized signal is limited and it cannot be considerably improved by rising the prediction order, the sampling frequency or the parameters' refreshing frequency. The following chapters present the language specific aspects of the ROMVOX system and our last approach regarding the used synthesis technique.

2. The Building Elements of the ROMVOX Text-to-Speech System

2.1 Text-preprocessing. We need a text-preprocessing module on the grapheme level in order to convert the incoming orthography into some linguistically reasonable standard forms. There are many phenomena encountered in normal orthography like: underlining, the occurrence of capitals, abbreviation containing periods, abbreviations containing no vowels, numbers, fractions, Roman numerals, dates, times, formulas and a wide variety of punctuation including periods, commas, question marks, parentheses, quotation marks and hyphens. In our system the abbreviations are stored in a vocabulary, which can be extended by the user, so field specific abbreviation can be built into the system.

2.2 Speech sound set. We used a set of 31 phonemes for Romanian language. As internal representation for special Romanian sounds we used the following symbols: g1 (in ge, gi), g (in ghe, ghi), c (in ce,ci), k (in che, chi), a1 (for Romanian letter ă), i1 (for î and â) s1 (for ş), t1 (for ţ).

2.3 Conversion of graphemes. In our system the grapheme-to-phoneme conversion rules are alphabetized according to the first letter of the sequence. Each letter of the alphabet represents a separate rule block in the table. One such block has the longest rule at the top and the shortest rule at the bottom; i.e. the last rule consists of only one letter.

Examples:

The e sound rule block	the c sound block rule
es1ti=_j1es1tj1_	coop=ko_op
este=_j1este_	cea=ca
exa=egza	cio=co
eio=ej1o	chi=ki
ea=_j1a_	che=ke
el=_yj1el_	ci=ci
ei=ej1_	ce=ce
e=e	c=k

Where _ means pause, j1 means special short i.

As result of the grapheme-to-phoneme conversion algorithm, the desired string of diphones is obtained. For example, the string corresponding to the word 'floare' is: _f fl lo oa ar re e_.

In future versions of ROMVOX, a second level processing of sound codes will be experimented. So, timing modifications could be made according to the rules of the prosody preparation module.

2.4 Word accent. For Romanian language the word accent is free, choosing between the last two syllables of the word, and there are many words with other place of accent. Semantically different words have the same orthography. For example:

cúrele (cure -plural) curéle (belt -plural)
vésela (gay -feminine, plural) veséla (dishes)

We are thinking of the possibility to formalize these kinds of problems.

2.5 Intonation. For obtaining acceptable intonation for unrestricted texts, a set of rules has to be formulated which produces natural sounding pitch contours for utterances that may have never been spoken.

In sentence intonation, one serious problem is to find such rules that make the monotonous speech more natural, so that listening to long texts would not be uncomfortable. We studied experimentally the pitch contour for different kinds of sentences (declaratives, questions, and exclamations). For declarative sentences, the fundamental frequency raises for the first word (from 100% to 140% of its value and slows down to 125% for the last part of this word), and slows down until the end of the sentence, except the last word. Here it falls at 70% and remains constant.

Questions can be with Q-word (specific word for interrogation) or without. For the former, the fundamental frequency raises on this word from 100% to 160% and comes down to 100%. For the last type of questions we adopted a conventional pitch contour, but very subtle intonation effects cannot be handled.

3. Signal processing

Our last experiments in order to improve the quality of the synthesized signal are based on a hybrid timedomain-LPC approach. This approach takes into consideration the behavior of the glottal pulse (for voiced sounds) which can be described using the Liljencrants-Fant (LF) model, [Veldhuis 96].

Figure 1.a. presents the time domain waveforms of the Romanian vowel **o**, the corespondent source signal (Figure 1.b.). As it can be seen, during the opened phase of the glottis in which the 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

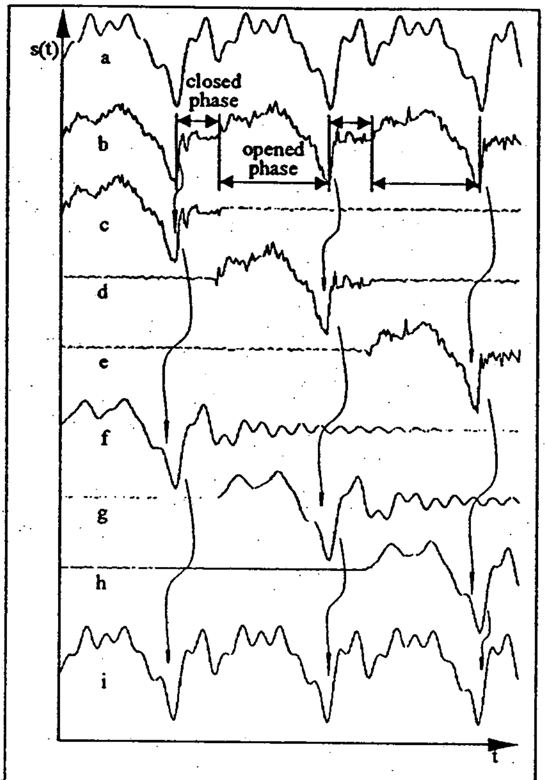

Figure 1. Waveforms of the Romanian vowel o, uttered by a male (3 pitch periods)

the vocal tract. 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. 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 considering 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 presented in Figure 1. present such a case for three pitch periods.

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 theorem 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 only to superimpose this individual signals at new distances in concordance with the desired new pitch.

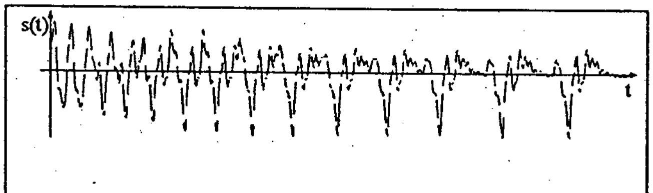

Figure 2. Waveform of vowel o, re-synthesized with modified pitch

Figure 2. presents such 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 lower fundamental frequency (one octave lower), which increases to the initial value of the pitch (at the middle of the signal), continuing to increase to higher values (one octave higher).

The main problem is the decomposition of the initial signal into individual, pith-synchronous signals. This implies two aspects. First of all it is necessary to determine the evolution of damped regime for each individual signal. 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. We used the LPC analysis method, which is one of the most used methods for the determination of the filter characteristics of the vocal tract. If the parameters of the filter are determined and if the filter is placed in the initial state from the beginning of the closed phase of the glottis, it will generate automatically the desired damped signal which can last over 2-3 other pitch periods. The other task is to eliminate the effect of this new determined damped signal upon the next pitch-synchronous individual signal (signals). This can be done by simply subtracting the current determined individual signal from the initial one.

These two operations will be performed consecutively for the whole signal, and each intermediate individual signal is saved in a database (sound inventory). Because the sound inventory contains diphones, the above procedure must be applied for each diphone.

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, respectively to other synthesis techniques.

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.

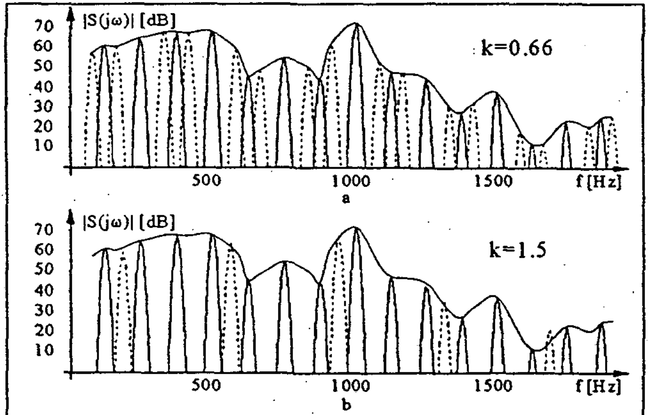

Figure 3. Power spectrum of vowel o:
- continuos line: initial pitch
- dotted line: modified pitch

Our approach doesn't use any windowing technique, so this source of spectral distortion is eliminated. Figure 3. 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.

5. References

[Ferencz et al. 96] Ferencz, A., et al. 1996. Experimental Implementation of the LPC-MPE Synthesis Method for the ROMVOX Text-to-Speech Synthesis System. Proceedings of SPECOM'96 International Workshop, St-Petersburg, 159-164.

[Ferencz et al. 97] Ferencz, A., et al. 1997. 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, 179-184.

[Olaszy et al. 91] Olaszy G., and G. Németh. 1991. Multilingual Text-to-Speech Converter. In Journal on Communications No. 2, 1991.

[Veldhuis 96] Veldhuis, R.N.J.1996. An alternative for the LF model. In IPO Annual Progress Report 31, Eindhoven, 100-108.