Modelling the Excitation Function to Improve Quality in LPC's Resynthesis.

CELSO AGUIAR
CCRMA - Center for Computer Research in Music and Acoustics.
Stanford University, Stanford, CA 94305-8180 USA.
aguiar@ccrma.stanford.edu

ABSTRACT

LPC (Linear Predictive Coding) is a well known technic for speech analysis-synthesis. The analysis consists in finding a time-based series of n-pole IIR filters whose coefficients better adapt to the formants of a speech signal. These computations produce a residual signal which is the exact complement to the information kept in the coefficients, if we wish to recover the original. The model of the human vocal tract mechanism assumed by LPC presupposes that speech can be reduced to a succession of voiced or unvoiced sounds. Thus, an excitation function composed of pulse or noise substitutes the residual in the resynthesis. This assumption is adequate for some tasks but, in a musical context, the artifacts introduced can yield unsatisfactory results. This article proposes a simple alteration in the model to improve the quality of LPC's resynthesis. Some of the conventional problems like "buzzy" quality and loss of coloration are partially corrected.

INTRODUCTION

The problem of quality in the resynthesis with linear prediction technics has already been noticed in the Computer Music literature (Lansky, 1989) but not sufficiently addressed. Some of the critical problems impeding the accomplishment of higher quality results are the general "buzzy" quality of the sounds (due to the use of a band-limited pulse signal to model the voiced components in the speech), alterations in the coloration of the original speech and loss of energy in the region of the fundamental.

These problems altogether are enough to remind us that the technic was not initially created to generate high quality results as may be needed in Computer Music. Although quality and intelligibility were also an issue in its first applications, LPC was basically created as an analysis technic for data reduction in speech transmission. Under these conditions we must conclude that it would be necessary for the technic to undergo some kind of adaptation or improvement if it is to be applied in its full potential for our purposes.

In this paper, the model employed by LPC to reproduce the human vocal-tract mechanism is described as well as are discussed the reasons for the improvements here introduced. Due to the sufficient literature in the field, only a brief description of the LPC technic is done. Finally, we present some sound examples that demonstrate the ideas discussed.

THE LPC MODEL OF SPEECH REPRODUCTION

The main idea behind linear predictive coding is that a sample of speech can be approximated as a linear combination of past speech samples. An inverse filter must be produced so that it matches the formant regions of the speech under analysis. The coefficients for that filter are determined by minimizing the square difference between the speech samples and the linearly predicted ones. The process of LPC analysis is nothing less than the tentative application of this filter to the speech samples. These computations produce what is known as the

residual sound in this field of research.

The model used by linear prediction is the same as used conventionally in most of the models for the digital representation of speech (Schaffer & Rabiner, 1975). It is fundamentally based on the human vocal-tract mechanism, which can be thought as an acoustic tube terminated at one end by the vocal cords and at the other by the buccal cavity and lips. The shape of this tube is determined by the position of its components like lips, tongue, jaw and vellum.

The sound sources generated by the air pressure coming from the lungs can excite the vocal tract in several ways. These can be divided in two basic types of excitation: quasi-periodic pulses which generate the voiced sounds, and a noisy source which produces the unvoiced sounds of speech. Both types of excitation sources are modelled by LPC and are represented by the two boxes in the diagram below (Fig. 1).

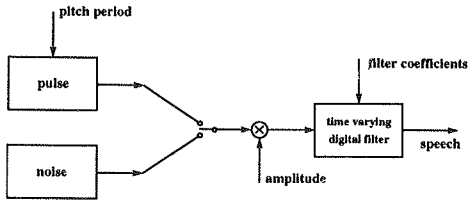

Fig. 1 - LPC's digital processing model.

The main characteristic of that residual sound from LPC analysis described above is to be a close approximation to the vocal tract input signal (Cann, 1985). If the formants are being extracted efficiently by the filtering performed during the analysis, the residual should look like a combination of noise and a series of pulses. LPC does not provide a way to gradually mix these two sources when modelling the excitation. The boundary between voiced and unvoiced speech is precisely one of the places where artifacts are bound to appear in LPC resynthesis.

The two sources cited above create a wide band excitation of the vocal tract that acts as a linear time-varying filter (Fig. 1). This filter imposes its characteristics on the frequency spectra of the excitation sources. The vocal tract is thus characterized by its natural resonances and is represented in the LPC model by the time varying digital filter whose coefficients should match the formants present in the analyzed signal.

IMPROVING THE LPC MODEL

The experiments and conclusions reached in this research stem from a *Voice Class* course in the Spring of '93 at CCRMA. Our initial project was just to improve the quality of some programs developed by Professor Perry R. Cook to perform LPC analysis and resynthesis according to the autocorrelation method. The programs were improved, a new pitch tracking routine was included but, when experimenting with sounds sampled at 44 kHz, we noticed that the quality of the resulting sounds would not improve as normally happens when dealing with sampled sounds. On the contrary, the coefficients extracted from these 44 kHz sounds would do a worse job than the ones extracted from the same files handled at 22 kHz. The same phenomena would manifest in other LPC programs like CMix's routines for performing the covariance method. A good explanation for this fact is that the formants present in the speech will not change their position with the change in sampling rate. When analysing at a higher sampling rate, we would find ourselves trying to do the job of extracting the same formants, only that now we had more samples to handle and compute, as if trying to do the job in a less efficient manner.

Another interesting observation that can be made about the LPC process is that if the residual is the precise complement to the information kept by the coefficients then, the more our excitation function looks like the residual the closer we will be to recovering completely the original signal. The problem with the direct use of the residual in LPC is that it can only be used as it is, and only to recover the original. In other words, if we wish to stretch our sound in time and try to do this by stretching the residual and feeding it back as the excitation

function in the LPC algorithm, the artifacts introduced in the process of stretching the residual will propagate into the resulting sound and the result is the same as stretching the original sound with the same method applied to the stretching of the residual.

So, if the difference between excitation function and residual is what keeps us from recovering completely the original, or at least coming close to do that, the residual would probably be able to contribute with an extra amount of formant information that is not present in the conventional way of extracting the coefficients. With this in mind, we included another stage in the LPC model in order to diminish the discrepancies between the excitation function and the residual. The idea is to perform in the residual the same analysis done to the original sound. The extra information extracted from the residual is then stored in the coefficients that result from the second LPC analysis (Fig. 2).

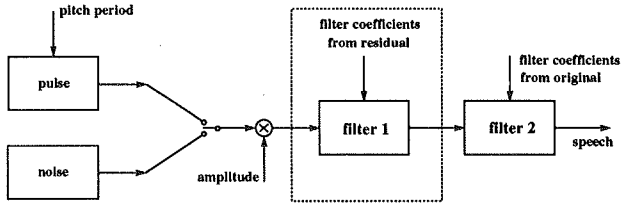

Fig. 2 - Improved LPC model including another filtering stage (residual coefficients) in cascade.

Other parameters used in the resynthesis process, like pitch, power and the information on the voiced or unvoiced content of the speech are extracted in the same way as before, from the original. This change in the algorithm together with the adjustments performed in the frequency function resulting from pitch tracking (see section below) have proved to be capable of recovering almost completely the characteristic of the original sound. As we are not directly using the residual in the resynthesis step, but rather its formant information expressed by its coefficients, the same possibilities that are present in the conventional LPC model are still present in this modified model like time stretching or changing pitch independently from duration. It can be argued that this change in the model will consume an extra amount of computing time (the double time to be more exact). It is clear that the application of this improvement can be delayed until the end of the process of obtaining the sounds in a composition, after all the choices have been made and as a means of enhancing the result.

In our experiments we have also found out that increasing the number of stages and keeping constant the number of poles (extracting the coefficients of the residual of the residual) will not yield any better results. This should be happening probably because all the useful formant information has already been completely extracted when we performed the analysis of the residual. Another possibility that we have not directly investigated in our research would be to do a increase in the number of stages while analysing the sound with a smaller number of poles.

IMPROVING THE QUALITY OF PITCH TRACKING

Another very important aspect of LPC analysis-resynthesis procedures that should also be stressed, is the quality of the pitch tracking realized in the original sound. The quality of the sounds resulting from the resynthesis depends entirely on the quality of the pitch tracking initially done on the sound.

The pitch detector developed by Professor Perry Cook and used in our programs is a lag-domain type of pitch detector. It uses AMDF technic and refinements to determine the pitch of a quasi periodic sound. In this pitch detector a delay line size is determined according to lower and higher pitch limits specified. For the female voice used in our experiments the limits of 90 Hz and 525 Hz have been sufficient. In case of a male speaker the lower limit can be adjusted to 50 Hz. The pitch tracking is also realized frame by frame. In our tests a frame and hop size of 600 and 200 samples respectively (the same values for the equivalent parameters used in the LPC analysis) for the female speech sampled at 22050 Hz, have proven to be the most effective.

Once the pitch tracking is done, a Frequency X Time function like the one in Fig. 3 emerges as a result of the algorithm. At this point it is very important to be able to adjust this curve for resynthesis. This adjustment is mostly necessary due to imperfections in the result emerging from the pitch tracking algorithm. These imperfections come in the form of glitches that can be easily detected by a small deglitching routine or avoided for different sections of the Frequency X Time curve by a redefinition of the frame and hop sizes in the pitch tracking input.

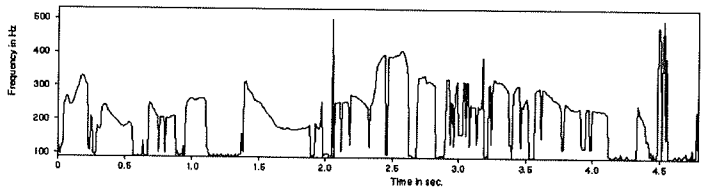

Fig. 3 - Frequency function obtained from the pitch tracking algorithm.

The worst artifacts introduced in the resynthesis by these imperfections occur in the boundaries between voiced and unvoiced frames. In case the pitch tracker fails to produce a good estimate of the pitch for a voiced frame, which is very conceivable due to the proximity to a noisy section of the speech, the resynthesis will yield a false frequency for the pulsed excitation in that frame, producing a very annoying effect. The best option in cases like this is to just eliminate these points and interpolate between the remaining ones. A smother curve like the one displayed in Fig. 4 should be sufficient to produce a more faithful result.

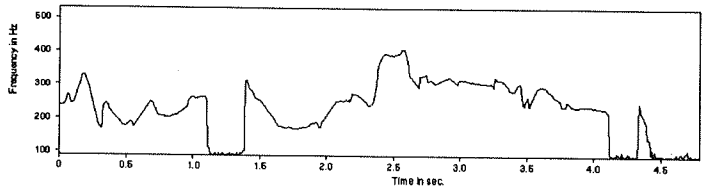

Fig. 4 - Adjusted version of frequency function used to improve resynthesis

Apparently, any compositional elaboration to be introduced in the frequency dimension of a speech phrase with LPC, should probably be based on this adjusted version of the curve rather than in the inaccurate version coming out of the pitch tracker.

SOUND EXAMPLES

With the purpose of demonstrating the ideas discussed in this paper, an example phrase of female speech was analyzed and resynthesized in several different conditions. An audio tape containing these sound examples can be ordered from the author.

1. Original speech phrase used: "I burst out in scorn, at the reprehensible poverty of our sex". Female voice sampled at 22050 Hz. Duration of 4.79 sec.

2. Residual produced by the LPC analysis of sound example 1. Autocorrelation method using the standard LPC model was employed. Number of poles was 32, block size 600 and hop size 200. Amplitude has been boosted up for demonstration purposes only.

3. Resynthesis of the phrase using the residual as excitation.

4. Resynthesis of the phrase using a pulse-noise combination as excitation. No adjustments were made in the frequency function obtained from pitch tracking (Fig. 3).

5. Resynthesis of the phrase using a pulse-noise combination as excitation. Adjustments were made in the frequency function according to what is displayed in Fig. 4.

6. Improved resynthesis of the original according to our refined model proposed. The LPC analysis program was modified so that it also extracts the LPC coefficients from the residual sound. Number of poles for the analysis of the original was 32, number of poles for the analysis of the residual was also 32, block size 600 and hop size 200. Resynthesis is further improved by including the adjustments made in the frequency function obtained from pitch tracking.

7. Excitation function used to produce sound example 5.

8. Excitation function used to produce sound example 6.

9. Resynthesis via covariance analysis method using CMix's LPC programs. Parameters were set according to Lansky's advice: number of poles of 24, block size of 200, hop size of 100. Adjusted version of frequency function was employed.

10. Time stretched version of sound example 9 (3 to 1 stretch).

11. Time stretched version of sound example 6 (3 to 1 stretch).

REFERENCES

Cann, R. (1985). "An Analysis/Synthesis Tutorial." In *Foundations of Computer Music,* ed. Curtis Roads and John Strawn. Cambridge, Mass.: MIT Press, 114-44.

Lansky, P. (1989). "Compositional Applications of Linear Predictive Coding." In *Current Directions in Computer Music Research,* ed. Max V. Mathews and JohnPierce. Cambridge, Mass.: MIT Press, 5-8.

Schaffer, R. & Rabiner, L. R. (1975). "Digital Representations of Speech Signals." *Proceedings of the IEEE,* 63(4), 662-77.

ACKNOWLEDGMENTS

My acknowledgements go to Professor Perry R. Cook for his challenge and support during the *Voice Class* course in the Spring of '93 at CCRMA, Stanford. I'd also like to thank Mr. Paul Lansky for confirming that I should write this paper, and composer Marco Trevisani for providing the speech samples used during this research. The paper was written while under a Master of Arts fellowship from the Brazilian Agency CAPES, to which I am most indebted.