AUTOMATIC GENERATION OF MIXED EXCITATION IN A LINEAR PREDICTIVE SPEECH SYNTHESIZER

Sverre Holm

The Norwegian Institute of Technology
Division of Telecommunications
N-7034 Trondheim-NTH, Norway

ABSTRACT

The main objective of this work has been to add the model for fricative excitation to the LPC synthesis model. From the LPC model one finds the acoustic tube section with the greatest constriction and adds a modulated noise signal. The results from this model demonstrate that one is able to produce noise bursts at the right time instants that are shorter than the frame length. This gives a more natural sound for certain phonemes, but adds a quantization type of background noise.

INTRODUCTION

Linear Predictive Coding is one of the most efficient and best understood methods for speech compression (1). Its simple speech production model is capable of reproducing speech of high quality, but with a loss of naturalness. Much of the data reduction in these coders is due to the assumed separation between the excitation signal and the vocal tract transfer function. Another factor causing great data reduction is the classification of the excitation as either voiced or unvoiced. An important question is how these two assumptions influence the quality of the synthesized speech.

Fricative sounds are generated by a turbulent flow at a constriction in the vocal tract. Thus the excitation and the transfer function are not spatially separable. Still the LPC model works quite well for the unvoiced fricatives, because the transfer function for these sounds is so simple. Often only 4 reflection coefficients are sufficient to reproduce these sounds. However for the voiced fricatives the problem is more severe. There should be at least two excitation sources, one for the vocal cords and the other where the turbulence occurs. But the LPC coder must decide on one excitation source and the sounds usually turn out to be too buzzy.

Even more difficult to model are the stop consonants which are generated by releasing a built-up pressure. They may be both voiced and unvoiced too. For these sounds the short-term stationarity assumed by the division into 15-30 ms frames is not always valid. Therefore the critical startpoints are not faithfully reproduced, but an improvement can be heard when the frame length is decreased.

Linear prediction can be interpreted as a way of finding a whitening filter, but it can also be regarded as an estimation of the cross-sectional areas in an acoustic tube model of the vocal tract. A natural question is whether these area estimates can be used to find the constriction where fricative sounds are generated. And then the next question is whether we can insert a proper noise signal at that point in the model and get an improvement. This would be a help in reproducing the voiced fricatives and it might also improve the reproduction of the stop consonants as the inserted signal dependent on the voiced excitation and thus adds some dynamics to the model. This mixed source model has many similarities to the one used by Flanagan and Ishizaka (2) and was originally described by Meyer-Eppler (3) (partly referenced by Flanagan (4)).

SYNTHESIS MODEL

Physical formulation

In order to use the proposed expanded synthesis model one have to reformulate the LPC synthesizer so the signals in the model actually describe the volume flows in the acoustic tube. First of all the common one-pole deemphasis filter must be split up into a pitch-pulse filter and a lip radiation filter. The pitch-pulse filter is a two-pole filter which shapes the impulse excitation into a more triangular form. The lip-radiation filter has one zero. The whole model is shown in fig. 1.

The structure used in the synthesis filter is a four multiplier form shown in fig. 2. The signals in this structure are the positive-going and negative-going volume flows in the acoustic tube.

Compared to the common two-multiplier structure (1) this realization has an extra branch in the forward direction. The branch consists of a multiplication with $1 + r_1$. To get the right overall gain the excitation standard deviation must be scaled accordingly. The four-multiplier exci-

CH1610-5/81/0000-0118 $00.75 © 1981 IEEE

Reprinted from PROCEEDINGS OF IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, March 30, 31 & April 1, 1981

tation standard deviation given in terms of the M reflection coefficients and the two-multiplier gain is

$$\sigma_{4m} = \sigma_{2m} / \prod_{i=1}^{M} (1 + r_i)$$

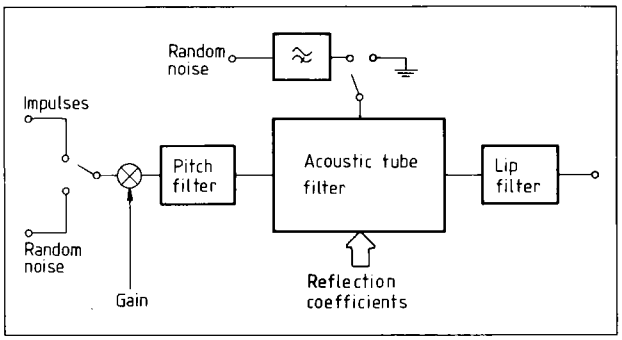

Figure 1. Block diagram of an LPC synthesizer with automatic generation of mixed excitation. The switches are controlled by the voiced/unvoiced descision and are shown in the voiced position.

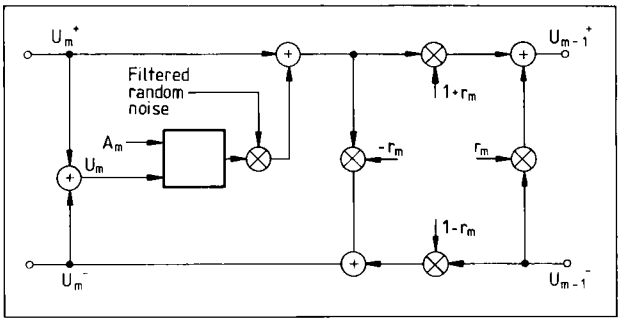

Figure 2. One section of the synthesis filter showing the modulated noise source.

Adding fricative excitation

The fricative excitation is produced when the air flow becomes turbulent. This gives rise to a random noise source with a spectrum which is flat in the mid-audio range. The source is located at or in front of the constriction. The Reynold's number in a section shows whether the flow is turbulent or not (3, 4) and its square is given by the volume flow and the cross-sectional area. In section m this number is:

$$R^2 = K_1 \cdot \frac{U_m^2}{A_m}$$

where K_1 is a constant. The volume flow in section m is the positive-going wave minus the negative-going wave, see fig. 2. When the squared Reynolds number exceeds the threshold K_0 the magnitude of the pressure source is:

$$P = (R^2 - K_0)$$

To avoid positive feedback the random numbers are high-pass filtered at 500 Hz, and the volume flow is low-pass filtered at 500 Hz. The low-pass filter is a three section recursive one. It is important that the noise source signal is not delayed too much compared to the signal to which it is added. This is ensured by having a filter with little time-delay.

As a short-cut to trying this model we have just added the pressure source to the positive going volume flow, as fig. 2 shows. The proportionality constant and the threshold have been adjusted by listening to the synthesizer output and by examining the synthesized time-series.

Another simplification is that only noise source is used. It is always placed in the section with the greatest constriction.

EXPERIMENTAL RESULTS

Our result so far have been encouraging. The buzziness heard on male utteranced when reproduced by the standard LPC synthesizer is greatly reduced. Instead there appears a signal-correlated type of background noise.

Figures 3 to 6 give an example consisting of the stop-part of a "d" followed by an "o". Figure 3 shows the original male utterance sampled at 8 kHz. The plosive part is about 6 ms long. Figure 4 is the output of an LPC synthesizer using Burg's estimation algorithm with 22.5 ms frame-length, 10 reflection coefficients and quantization to 2.4 kbit/s. The plosive part is not reproduced but the pitch-periods close to it have a fast oscillation "simulating" the noise part in the original. The synthesized "d" is more buzzy than the original because the sound is reproduced as all voiced. The output of the expanded synthesizer without quantization in figure 5 has the same general looks as the ordinary synthesizer. But in one of the pitch-periods a noise excitation has been automatically added. The noise burst is about 9 ms long which is quite similar to the original noise burst. In listening the synthesized "d" is more natural sounding than the ordinary LPC synthesizer, as the buzziness is masked by the noise burst. Figure 6 shows the same expanded synthesizer output when the LPC parameters have been quantized to 2.4 kbit/s. We have now two noise bursts, but in listening the perceived improvement is still the same.

The new model is expected to be quite dependent on the way the parameters are quantized. We have used the "log-area" ratio method (1) for all 10 reflection coefficients. Another critical factor influencing the area estimates is the way the pitch-synchronous interpolation in the synthesizer is done. In our examples the reflection coefficients have been linearly interpolated between two frames.

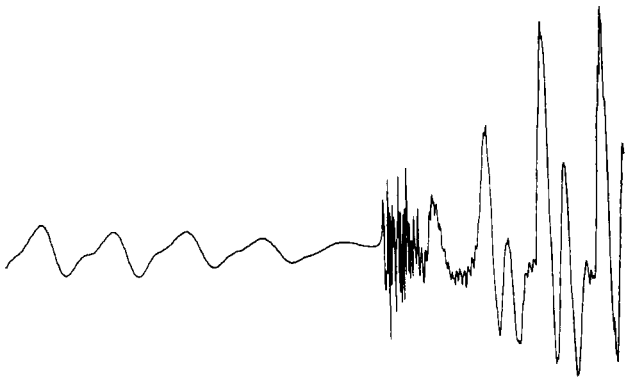

Figure 3. 100 ms of male utterance "d", sampled at 8 kHz.

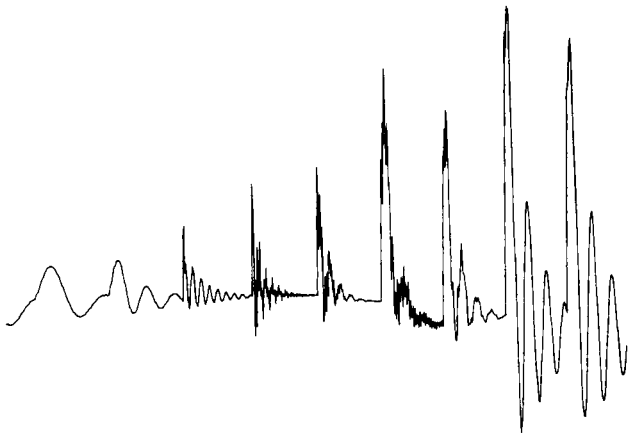

Figure 4. Output from 2.4 kbit/s linear predictive coder.

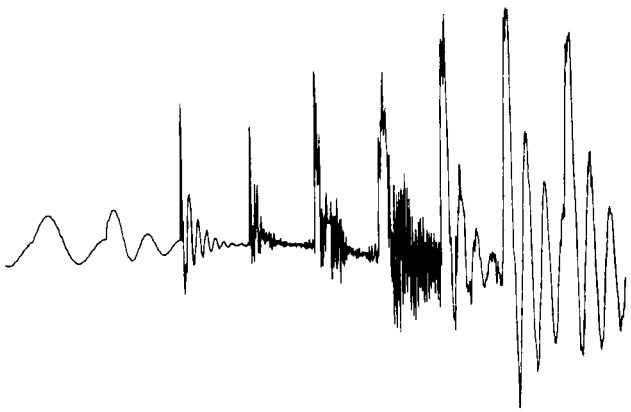

Figure 5. Output from the expanded synthesis model, no quantization of parameters.

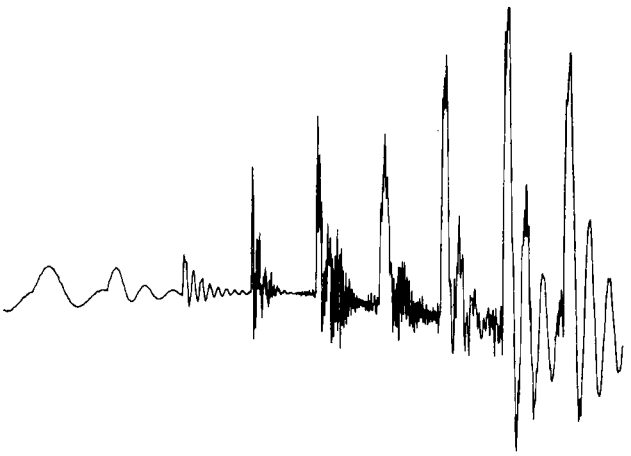

Figure 6. Output from the expanded synthesis model, the parameters are quantized to 2.4 kbit/s.

CONCLUSION

The results from automatically adding fricative noise in the LPC synthesis model are encouraging so far. It is a way of improving the LPC synthesizer output without a need for more parameters transmitted. Thus it is a way of drawing out unused information from the area parameters. A disadvantage is that the speech production model is less robust than the standard LPC, as it depends on receiving good area estimates.

Still more work will have to be done on the model. Adding a pressure source to a volume flow is obviously wrong, and the best way of finding where to insert the noise source is not yet found. A solution to these problems might help to reduce some of the background noise.

References

(1) J.D. Markel and A.H. Gray, jr., "Linear Prediction of Speech", Springer Verlag, New York 1976.

(2) J.L. Flanagan and K. Ishizaka, "Automatic Generation of Voiceless Excitation in a Vocal Cord-Vocal Tract Speech Synthesizer", IEEE Trans. on Acoust., Speech, and Sign. Proc., Vol ASSP-24, no 2, April 1976.

(3) W. Meyer-Eppler, "Zum Erzeugungsmechanismus der Geräuschlaute", Z. Phonetik 7, 1953.

(4) J.L. Flanagan, "Speech Analysis, Synthesis and Perception", 2nd ed. Springer Verlag, 1972.