ON REDUCING THE BUZZ IN LPC SYNTHESIS

M. R. Sambur, A. E. Rosenberg,
L. R. Rabiner and C. A. McGonegal

Bell Laboratories
Murray Hill, New Jersey 07974

Abstract A method for reducing the characteristic buzz from LPC synthetic speech is presented. The method consists of the use of an non-impulse source for exciting the LPC synthesizer during voiced sounds. One novel feature is that the temporal parameters of the source are kept in fixed proportion to the pitch period. An extensive perceptual experiment has shown that the resulting quality of the synthesis is significantly preferred over the quality of the standard LPC synthesis.

I. Introduction

The technique of linear prediction (LPC)[1,2] has rightfully enjoyed a great deal of popularity for the analysis and synthesis of speech. Although the intelligibility of LPC speech is reasonably good, the quality of the synthesis is diminished by a discernible "buzziness". A great deal of effort has been expended in attempting to enhance the naturalness of the LPC synthesis by removing this "buzziness". Most of this effort has been directed at improving the quality by expanding the linear prediction model to account for the presence of both poles and zeros.[3,4] The inclusion of zeros in the analysis model has led to a slightly crisper sounding synthesis but has not eliminated the unnatural "buzziness".[3] In this paper we describe a method for significantly reducing the "buzziness" without changing the LPC analysis model, but instead by modifying the LPC synthesis structure.

II. Why the Buzz?

In a recent perceptual evaluation of the merits of various pitch detectors,[5] an interesting phenomena was noted about the quality of the LPC synthesis. This observation was simply that the speech synthesized from high pitched speakers was practically free of any discernible buzziness, while the quality of synthetic speech from low pitched talkers was often quite buzzy. In addition, for the low pitched speakers it was also noted that the more monotonic the pitch, the more noticeable the buzziness. Following comparisons of the various synthesized waveforms and additional informal listening, two hypotheses for the absence of buzziness in high pitched speakers were formulated. These explanations were as follows:

(a) Pitch Synchronous Interpolation

The first explanation concerns itself with the fact that the LPC parameters used in the synthesis were interpolated to allow pitch synchronous resetting of the synthesizer. Thus the higher the pitch, the more frequently the LPC parameters were updated. Conversely the lower the pitch, the less frequently the LPC parameters are changed and the longer the structure of the synthesizer is left fixed. Since speech is really a nonstationary process, it may be conjectured that the ear somehow "senses" that the LPC parameters are being held too long, and interprets this excessive stationarity as a buzzy overtone. Reduction of the buzz can then be accomplished by simply resetting the synthesizer parameters at a faster rate.

To test the above conjecture, an LPC synthesis system was implemented in which the LPC parameters were interpolated to the sampling rate of the input speech waveform - i.e., 10 KHz. The output speech sequence is then given by

$$s_n = \sum_{k=1}^{p} a_{kn} s_{n-k} + G u_n ,$$

where a_{kn} denotes the interpolated k-th LPC coefficient at the n-th sample, u_n is the standard input and G is the gain.[1,2] The interpolation scheme employed involves first transforming the LPC coefficients to the log-area ratio coefficients g_i, and then linearly interpolating the g_i's and converting back to the a_i's. The reason for this transformation is that the stability of the LPC synthesizer is quite sensitive to small perturbations in the a_i's but is relatively insensitive to shifts in the log area ratio parameters.

Using this high rate implementation, the resulting synthesis was slightly less buzzy than that of the original synthesis, but not sufficiently so as to justify the additional complexity of the structure. When quadratic or cubic interpolation of the log area ratio parameters was employed, there was no noticeable improvement over that of linear interpolation.

Before concluding that a high rate interpolation of the LPC parameters was not adequate to eliminate the buzziness, a small amount of dither noise was added to the pitch periods used by the synthesizer. This dither was introduced so as to break up any monotonicity in pitch which was believed to enhance the buzziness in the synthesis for the low pitched talkers. By randomly perturbing the pitch, it was hoped that any remaining stationarity in the synthesis parameters would be eliminated, thereby reducing the buzzy quality of the speech. However, the dithering of the pitch did not appear to reduce the buzziness, and increasing the dither above 1% or 2% of a pitch period only made the synthesis worse.

(b) High Peak Factor

A second explanation for the apparent absence of buzziness in the synthesis of high pitched speakers is that the resulting synthesis is less "peaked" than that of low pitch speakers. Figure 1b shows an example of the synthesized waveform for a low pitched speaker and the waveform of the original speech (Fig. 1a) from which the LPC parameters were obtained. The high degree of peakedness of the LPC synthesis is apparent. For high pitched speakers, however, the synthesized waveform (Fig. 1d) has about the same peak factor as the original (Fig. 1c). It should be noted that the differences in peak factor between the synthesis of the low pitched talkers and the high pitched speaker is not due to any significant differences in bandwidths of the formants. Instead, the waveform of the high pitched talker has not decayed nearly as much as for the low pitched speaker, due to the significantly shorter pitch period.

It has been argued that the ear preceives the high peak factor in the synthesized waveform as a buzzy overtone.[6,7] Unfortunately, in experiments using a variety of all-pass filters to spread out the waveform and thus reduce the peak factor, we have observed no reduction in buzziness. However, as we shall see in the next two sections, the buzz can be significantly reduced by a scheme that both reduces the peak factor and helps to destroy any preceived regularity in the synthesis.

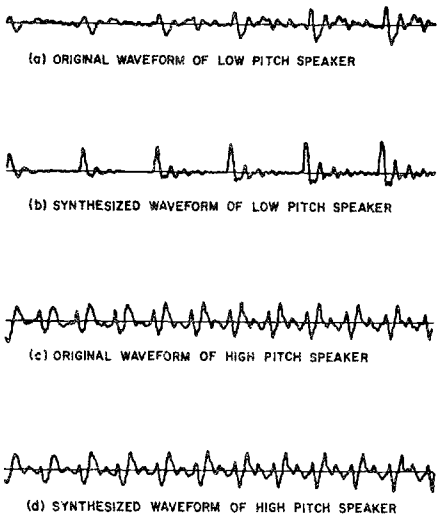

Figure 1

III. Non-Impulse Source Excitations

The high peak factor of LPC speech can be attributed to the use of an impulse source for exciting the synthesizer during voiced sounds. If an excitation could be constructed that was wider in duration than an impulse and still possessed a nearly flat spectral characteristic, then the amplitude spectrum of the output speech would be left unchanged but the speech would be less peaked. In addition, if the excitation was altered in a pitch synchronous fashion, there would be a nonstationary component added to the speech that would help remove any regularity that the ear could preceive as buzziness. It should be noted that the glottal excitation in real speech is known to be subject to wide variations across an utterance, and the modelling of this variation is important in preserving the naturalness of synthetic speech.[8]

In using a time varying non-impulse source, we are trying to simultaneously change the LPC synthesis formulation in two ways. Namely, we are trying to reduce the peak factor of the signal, and at the same time introduce a certain amount of controlled irregularity in the excitation source. In our experimental study, we shall show that this combination helps greatly to reduce the buzzy quality of the resulting synthetic speech.

In our experimental study, we have used the basic LPC structure with a non-impulse (finite duration) source to excite the synthesizer for voiced sounds. For a typical input sound, the shape of the source is varied in accordance with the parameters defined in Fig. 2. Source pulses are separated by the pitch period T. The parameter T_p (opening time) is the portion of the pulse with positive slope; the parameter T_N (closing time) is the portion of the pulse with negative slope. For a given perceptual evaluation, the relative opening and closing times (T_p/T and T_N/T) are specified and held fixed throughout the utterance. The object of the evaluation is to determine whether any particular combination of opening and closing time and pulse shape would help eliminate the buzziness and improve the naturalness of the synthetic speech.

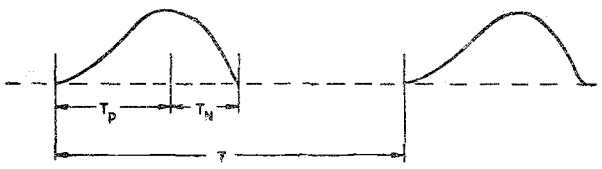

Figure 2

The pulse shapes considered in our study are depicted in Fig. 3. (The mathematical definitions of these pulses are given in Ref. 6.) These shapes have been studied previously because of their close resemblance to the actual glottal waveform.[6] In order to approximate the flat spectrum of an impulse, the duty cycle (ratio of $(T_p + T_N)/T$) employed in our investigation was less than 12%. A higher duty cycle results in a decidely lowpass sounding synthesis. However, as illustrated in Fig. 4, even for lower duty cycles, the spectra of the pulses are attenuated somewhat for high frequencies.

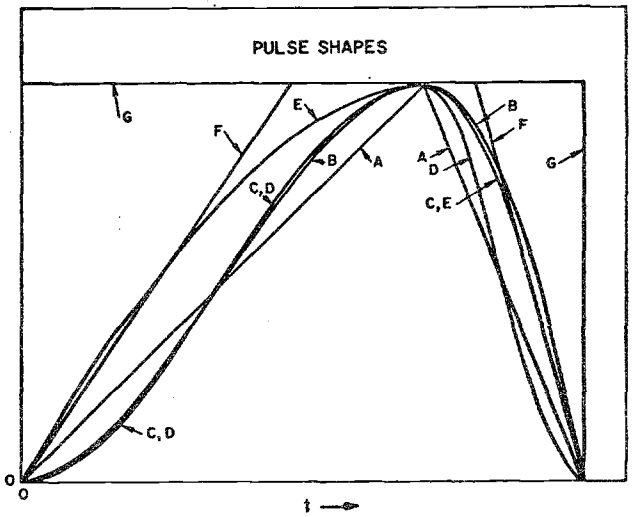

Figure 3

On the basis of preliminary listening, it was determined that the use of a non-impulsive source could indeed significantly reduce the buzz and enhance the quality of the LPC synthesis. In addition, it was found that, except for the rectangular pulse (G), no difference could be detected between synthesized sentences with equivalent values of the parameters T_P/T and T_N/T but with different pulse shapes. The rectangular pulse shape was associated with a distinctly inferior sounding synthesis. For all the pulse shapes examined, the synthesized speech changed from very "buzzy" and sharp to very "bassy" and muffled as T_P/T and T_N/T were varied. For a particular range of T_P/T and T_N/T, the buzziness and bassiness appeared to diminish and the quality of the synthesis was enhanced. To confirm our informal study, an extensive experiment was performed to determine whether a combination of T_P/T and T_N/T which minimized "buzziness" and "bassiness" was, in fact, preferred over the quality of the synthesis using an impulse source. It was also of interest to determine whether there was agreement among listeners over the proper combination of T_P/T and T_N/T, and to determine which combinations were associated with equally preferred stimuli.

IV. Experimental Evaluation

The experimental stimuli consisted of the two sentence utterances (spoken by a male speaker), "We were away a year ago" and "I was stunned by the beauty of the view". These sentences were synthesized with a triangular pulse source (shape A in Fig. 3) using sixteen combinations of T_P/T and T_N/T.* These combinations are specified in Table 1. The combination (T_P/T, T_N/T) equal to (1,0) is a special case which reverts to the conventional impulse source for LPC synthesis.

The experimental evaluation consisted of a simultaneous ranking, in order of preference, of all 16 combinations of (T_P/T, T_N/T) for each sentence. Listeners, seated in a sound booth, were able to listen and sort the stimuli by means of a computer-controlled sort board[9] until a satisfactory ranking was

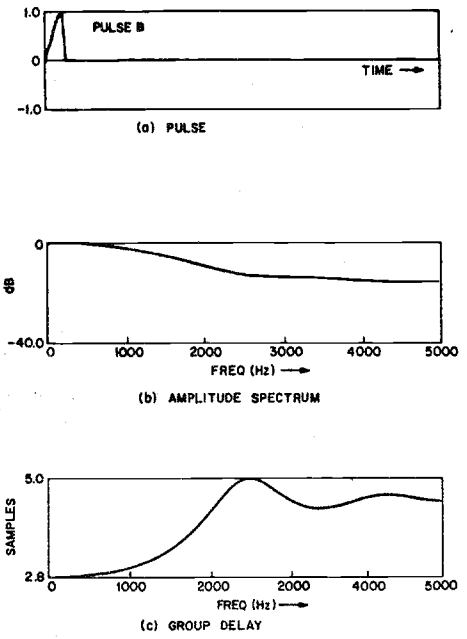

Figure 4

obtained. Sixteen column positions were available to sort the sixteen stimuli. Ties, obtained by placing two or more stimuli in a single column, were allowed. Each complete ranking of a sentence took approximately 10 minutes per listener. Ten listeners participated in the evaluation, 6 male and 4 female, with varying amounts of sophistication and experience in speech listening experiments. Three of the listeners, as co-authors of this paper, were aware of the nature of the stimuli. Each listener made four rankings of each sentence in different experimental sessions.

The data were processed by replacing the actual ranks assigned by the listeners with their rank order letting the least preferred stimulus have the rank order 1 and the most preferred stimulus have the rank order 16. Ties were handled by assigning the average rank order to each of the tied values. Thus if there were three least preferred stimuli, they would be assigned rank order 2 which is the average of the rank orders 1, 2, and 3 they occupy among the 16 ranked stimuli. This conditioning of the data has the affect of minimizing variability among the listeners.

The mean rank orders across the four judgments of all listeners for both sentences is shown in Table 1. An analysis of variance indicated that the interaction effects of sentence, stimulus combination, ranking session and stimulus combination were negligible but that of listener and stimulus combination was significant. For convenience the results pooled over all listeners are given in Table 1.

T_P/T (percent)

T_N/T (percent)	1	2	4	6	10
0	2.6	x	x	x	x
1	2.4	5.7	10.8	12.0	8.8
2	1.9	5.6	11.3	11.9	8.9
4	8.9	11.3	11.6	x	x
6	11.6	10.5	x	x	x

TABLE 1

Note that a stimulus condition with a mean rank order of 1 would indicate that it was consistently least preferred while a mean rank order of 16 would indicate a consistently most preferred stimulus condition. There are three conditions with mean rank orders from 2 to 3 which indicate generally good agreement as least preferred stimuli. These conditions, with low values of T_P/T and T_N/T, are associated with significant amounts of "buzziness" and include the standard impulse source condition. The most preferred stimulus conditions occupy a broader range. The mean rank orders take on values of 11 to 12 indicating less general agreement as the most preferred stimuli. These conditions are shown in the enclosed area within the Table. Inspection of the Table suggests that the simplest way to characterize the most preferred stimuli is by the sum of T_P/T and T_N/T. For these stimuli this sum assumes values of 6, 7, and 8 with a slight edge in preference to those stimuli with sum equal to 7. Stimuli with a more "buzzy" quality have values of this sum less than these preferred stimuli while a more "bassy" quality is associated with stimuli with greater sums. Thus, the most preferred stimuli pooled over all listeners, while occupying a broad range of conditions, are in the region in which both "buzziness" and "bassiness" are minimized.

An additional processing of the data using the two-sample sign test[10] was carried out to determine which stimuli were ranked significantly different than the others and which were not. A 1% level of significance was chosen for the hypothesis that a given pair of stimuli were given the same rank pooled over listeners, sentences, and sessions. The results confirm our conclusions obtained from inspection of Table 1, namely that the hypothesis of equal preference can generally not be rejected for pairs of stimuli within the enclosed area but that rankings are generally significantly different between stimuli within the enclosed area and those outside. For example, Table 2 shows the results of pairing the rankings of stimulus condition $(T_P/T,\ T_N/T) = (1,6)$ with all other stimuli.

T_P/T (percent)

T_N/T (percent)	1	2	4	6	10
0	-	x	x	x	x
1	-	-	0	0	-
2	-	-	0	0	-
4	-	0	0	x	x
6	*	0	x	x	x

TABLE 2

The 0's indicate that the hypothesis of equal rank with (1,6) could not be rejected for any of these conditions, while -'s indicate that these stimuli are ranked significantly lower. In contrast, Table 3 shows the results of pairing the rankings for stimulus condition (1,0), the standard impulse source with all other conditions.

T_P/T (percent)

T_N/T (percent)	1	2	4	6	10
0	*	x	x	x	x
1	0	+	+	+	+
2	-	+	+	+	+
4	+	+	+	x	x
6	+	+	x	x	x

TABLE 3

The +'s indicate that the associated stimuli were ranked significantly higher than (1,0).

A preliminary hypothesis before carrying out the formal experiment was that pairs of stimuli with combinations of T_P/T and T_N/T symmetric about the main diagonal $T_P/T = T_N/T$ are indis-

* Due to the lack of differences between the nonrectangular pulses, the triangular shape was selected arbitrarily as representative of the unipolar pulse source excitation.

tinguishable. For example, it was conjectured that the stimulus with combination (6,1) is ranked equally with the stimulus with combination (1,6). The two-sample sign test demonstrated that this hypothesis could not be rejected except for the pair of stimuli with combinations (2,1) and (1,2). The combination (2,1) was ranked significantly higher than (1,2). It is concluded that for these low values of T_P/T and T_N/T at the given sampling rate that the pulse shapes are degenerate and the hypothesis of symmetry not meaningful.

An illuminating method of presenting the results, particularly from the point of view of showing inter-listener variability, is by means of a multidimensional preference scaling analysis.[11] This method of analysis represents the stimuli as points in a multidimensional subjective space and the subjects as vectors in this same space. The subjects vectors are located so that the projection of the stimulus points onto each subject vectors are maximally correlated with the original rankings.

The results of this analysis are displayed in two dimensions in Fig. 5. These two dimensions account for 60% (dimension 1) and 20% (dimension 2) of the subject variance. The stimulus locations are indicated by labels in the form (T_P/T, T_N/T) while endpoints of subject vectors are scattered along an arc on the left hand side. A sample subject vector together with projections from stimulus points along the vector are shown. The most "buzzy" stimuli with the smallest values of $T_P/T + T_N/T$ are found at one extreme of dimension 1, while the most "bassy" stimuli with the largest values of $T_P/T + T_N/T$ are found at an extreme of dimension 2. If we propose to label these dimensions "buzziness" and "bassiness" respectively, we find that while the most "buzzy" stimuli are consistently least preferred there is a good deal of inter-subject variability with regard to the ranking of the most "bassy" stimuli.

Summarizing briefly, the results indicate that a range of stimuli with T_P/T and T_N/T combinations which are associated with minimization of both "buzziness" and "bassiness" are most preferred. In addition, the simplest way to characterize these percepts is by the sum $T_P/T + T_N/T$.

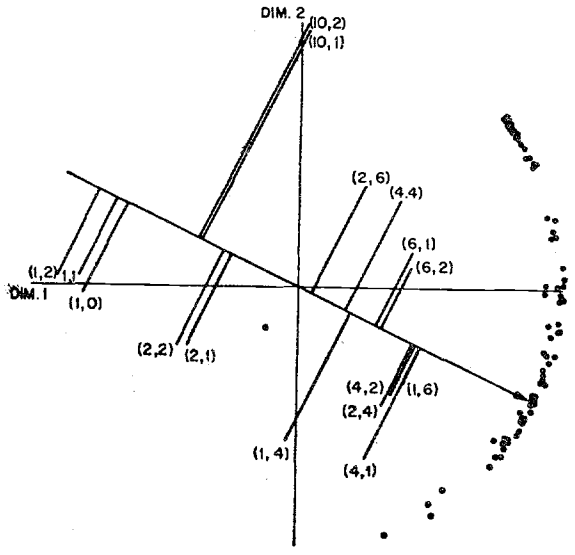

Figure 5

V. Summary

The goal of this paper was to develop an LPC synthesis scheme whose output did not possess the characteristic "buzzy" and unnatural quality of the standard LPC formulation. The scheme used to achieve this goal consisted of the use of an non-impulse source, whose temporal parameters are in fixed proportion to the pitch, for exciting the standard LPC synthesizer during voiced sounds. The success of this scheme was confirmed by an extensive perceptual experiment.

REFERENCES

1. B.S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", J. Acoust. Soc. Am., Vol. 50, pp. 637-655, 1971.
2. J. D. Markel and A. H. Gray, Jr., Linear Prediction of Speech, New York: Springer-Verlag, 1976.
3. B. S. Atal and M. R. Schroeder, "Recent Advances in Predictive Coding-Applications to Speech Synthesis", in Proceedings of the Speech Communications Seminar, Stockholm, August 1974.
4. B. S. Atal and M. R. Schroeder, "Linear Prediction Analysis of Speech Based on a Pole-Zero Model", J. Acoust. Soc. Amer., Vol. 58, Supplement No. 1, Fall 1975, Paper VV4, p. S96.
5. J. B. Allen, personal communication.
6. R. C. Mathes and R. L. Miller, "Phase Effects in Monaural Perception," J. Acoust. Soc. Am., Vol. 19, pp. 780-797, Sept. 1947.
7. C. A. McGonegal, L. R. Rabiner and A. E. Rosenberg, "A Subjective Evaluation of Pitch Detection Methods using LPC Synthesized Speech", to appear IEEE ASSP, 1977.
8. A. E. Rosenberg, "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," J. Acoust. Soc. Am., Vol. 49, No. 2 (Pt. 2), pp. 583-590, 197-.
9. C. H. Coker and S. Pruzansky, "Sort Board for Random Access of Auditory Stimuli," J. Acoust. Soc. Am., Vol. 47, p. 95(A), 1970.
10. J. E. Freund, *Modern Elementary Statistics*, Prentice-Hall, 1973.
11. J. D. Carroll, "Individual Differences and Multidimensional Scaling" in R. N. Shepard, A. K. Romney and S. Nerlove, eds., *Multidimensional Scaling: Theory and Application in the Behavioral Sciences*. Vol. I, New York: Seminar Press, Inc., 1972.