Received 12 August 1968

Speech Analysis-Synthesis System Based on Homomorphic Filtering

ALAN V. OPPENHEIM

Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Massachusetts 02173*

A digital speech analysis–synthesis system based on a recently proposed approach to the deconvolution of speech is presented. The analyzer is based on a computation of the cepstrum considered as the inverse Fourier transform of the log magnitude of the Fourier transform. The transmitted parameters represent pitch and voiced–unvoiced information and the low-time portion of the cepstrum representing an approximation to the cepstrum of the vocal-tract impulse response. In the synthesis, the low-time cepstral information is transformed to an impulse response function, which is then convolved with a train of impulses during voiced portions or a noise waveform during unvoiced portions to reconstruct the speech. Since no phase information is retained in the analysis, phase must be regenerated during synthesis. Either a zero-phase or minimum-phase characteristic can be obtained by simple weighting of the cepstrum before transformation.

INTRODUCTION

MANY speech analysis–synthesis systems are directed toward a separation of the speech excitation function and the vocal-tract impulse response. Typically, the excitation function is characterized by a measurement of the pitch period, and the vocal-tract impulse response is characterized by samples of the spectral envelope, taken as the outputs of a filter bank with filter bandwidths on the order of several hundred hertz. This characterization of the spectral envelope is convenient, since it is obtained from a measurement of the coarse spectrum without the intermediate determination of a fine-grain spectrum.

The trend of present technology indicates the possibility of future speech-system realizations with digital components. This trend coupled with the use of recently discovered fast and efficient means for computing the spectrum digitally, suggests the possibility that speech systems in which an intermediate step involves the determination of a high-resolution spectrum may be practical. In this paper, a configuration for a speech analysis–synthesis system that requires the determination of a high-resolution spectrum is proposed and discussed. The system bases the measurement of the spectral envelope on a linear smoothing of the logarithm of the fine-grain spectrum, rather than on a linear smoothing of the complex spectrum, as in a conventional channel vocoder, or a linear smoothing of the spectral energy as proposed by Freudberg *et al.*[1] In addition to providing what appears to be a good measurement of the spectral envelope, as judged by the quality of the synthesized speech obtained, this approach provides a simple mechanism for introducing into the synthesis a speechlike phase characteristic.

The basis for the system stems from a recently proposed approach to the deconvolution of speech.[2] In this approach, we consider the speech waveform to be modeled on a short-time basis as a convolution of the components representing the excitation function and the vocal-tract impulse response. Since we consider the processing to be digital, we represent the input speech in sampled form. Letting T denote the sampling period and $s(nT)$ the input speech weighted by a window $w(nT)$,

$$s(nT)=[p(nT)\otimes v(nT)]w(nT). \tag{1}$$

where $p(nT)$ represents the excitation function and $v(nT)$ represents the vocal tract impulse response. If $w(nT)$ is a relatively smooth window, then Eq. 1 can be approximated as

$$s(nT)=p_1(nT)\otimes v(nT). \tag{2}$$

* Operated with support from the U. S. Air Force.

[1] R. Freudberg, J. DeLellis, C. Howard, and H. Shaffer, "An All Digital Pitch Excited Vocoder Technique Using the FFT Algorithm," *1967 Conference on Speech Communication and Processing* (conference preprints) (November 1967), pp. 297–310.

[2] A. V. Oppenheim and R. W. Schafer, "Homomorphic Analysis of Speech," IEEE Trans. Audio Electroacoust. **AU-16**, No. 2, pp. 221–226 (1968).

where

$$p_1(nT) = p(nT)w(nT).$$

[The assumption that Eq. 1 can be replaced by Eq. 2 corresponds to assuming that $w(nT)$ is approximately constant over the duration of $v(nT)$. Results based on this assumption presented in Ref. 2 and in the present paper indicate that for the case of a 40-msec Hanning window, this assumption is justified.] Letting $P_1(\omega)$ represent the spectrum of $p_1(nT)$ and $V(\omega)$ represent the spectrum of the vocal-tract impulse response,

$$S(\omega) = P_1(\omega)V(\omega). \tag{3}$$

The approach to recovering the vocal-tract impulse response by means of homomorphic filtering is based on the observation that in the logarithm of $S(\omega)$, the contributions of excitation and vocal tract are added. Furthermore, the contribution from the vocal tract tends to vary slowly with frequency, while the contribution from the excitation tends to vary more rapidly and periodically with frequency. Consequently, we may expect that to some approximation, each of these contributions can be separated by means of linear filtering. Specifically, if we consider the inverse transform of $\log S(\omega)$, we may expect the contribution due to the excitation to occur at multiples of the pitch period, while the contribution from the vocal-tract impulse response tends to occur near the origin. Thus, to recover the component $v(nT)$ we would retain those values in the inverse transform near the origin, corresponding to smoothing the log spectrum and then transform, exponentiate and inverse transform to obtain an impulse response function. The operations described above can be carried out either retaining phase information or discarding it. In the first of these the complex logarithm of the spectrum is obtained with the real part corresponding to the logarithm of the magnitude and the imaginary part corresponding to the phase. In the latter case, the imaginary part of the logarithm is taken to be zero.

The inverse transform of the logarithm of the transform has been termed the cepstrum when phase information is discarded, and the complex cepstrum when phase information is retained. Thus, the strategy for obtaining an impulse response function is to compute the cepstrum or complex cepstrum of a segment of the input speech, retain only that portion near the origin and transform the result by the inverse set of operations.

In considering the relation between the impulse response functions obtained using the cepstrum or the complex cepstrum, we note that they both have the same spectral magnitude and differ only in their phase characteristics. In particular, truncating the cepstrum in such a way that it remains an even function results in an impulse response function with zero phase. This can be seen by noting that if the cepstrum is an even function, its transform, and hence also the exponential of its transform, is a real function. As an alternative, we may discard phase in the initial computation and generate a minimum phase characteristic from the spectral magnitude information. (We define a minimum phase sequence as one for which the phase is the Hilbert transform of the log magnitude. The properties of the cepstrum of minimum phase sequences and the use of the cepstrum in realizing the Hilbert transform are discussed in Ref. 3.) It has been argued previously that a minimum phase characteristic can be obtained by simple weighting of the cepstrum. We note that an impulse response function obtained in any of the three ways described will have identical spectral magnitudes and differ only in the phase associated with their Fourier transforms.

With a measurement of pitch and a voiced–unvoiced decision, an excitation function can be generated, which, when convolved with the impulse response function, will result in synthesized speech. During voicing, the excitation function consists of a train of unit impulses or unit samples with individual spacing corresponding to the pitch periods. During unvoiced intervals, a noiselike waveform with a flat spectrum is used, for example, a train of impulses with random polarity.

The above discussion outlines the basic strategy for the analysis and synthesis. In the next Sections, the analyzer and synthesizer configurations are described in more detail. The analysis is considered as the determination of the cepstrum and a measurement of the parameters of the excitation. The synthesis is considered as the conversion of this to an impulse response function and the generation of the synthesized speech.

I. ANALYZER CONFIGURATION

The analysis consists of a measurement of the cepstrum and a characterization of the excitation function by means of a voiced–unvoiced decision and a measurement of the pitch period during voicing. The parameters used to characterize the spectral envelope are samples of the cepstrum. Since the excitation function introduces into the cepstrum sharp peaks at multiples of a pitch period, we would generally choose the cutoff time to be less than the smallest expected pitch period.

Since we are considering digital processing, the input waveform and the cepstrum will be sampled functions. They will have the same sampling rate. This follows from the fact that the spectrum of a sampled function is periodic in frequency with a period equal to the reciprocal of the sampling rate. If the spectrum is periodic in frequency, then the logarithm of the spectrum is also, with the same period. Hence, the cepstrum is a sampled function with the same sampling rate as the original waveform. The parameters used to characterize the spectral envelope or vocal-tract impulse

[3] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, "Non-linear Filtering of Multiplied and Convolved Signals," IEEE Proc. **56**, No. 8, 1264–1291 (1968).

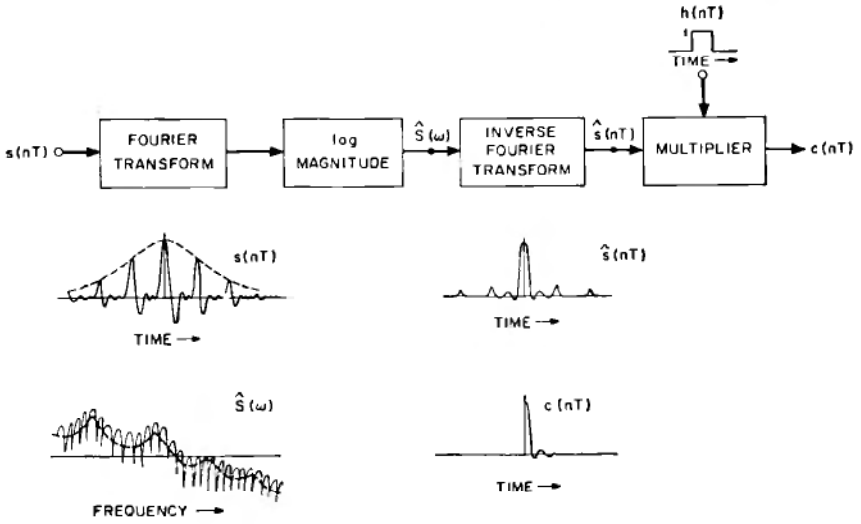


FIG. 1. Representation of the analyzer configuration.

response are taken as the first M samples of the cepstrum, where M represents the number of channels allotted to specifying spectral envelope information.

The cepstrum is obtained by weighting the input speech with a suitable window. Because the synthesis strategy is to excite an impulse response function with a train of constant amplitude impulses, then to within the accuracy of Eq. 2 the specific choice of the analysis window is not critical. For the specific system simulated, a Hanning window with a duration of 40 msec was used. With the speech waveform sampled at 10 kHz, the discrete Fourier transform (DFT) was computed for 512 points, followed by a computation of the log magnitude and the inverse DFT, resulting in samples of the cepstrum. A new cepstrum is obtained every T_c msec, where T_c is typically taken to be 10 or 20 msec.

Parameters characterizing the excitation function may be extracted by detecting the peak in the cepstrum which occurs at the pitch for voiced intervals and basing a voiced–unvoiced decision on its presence or absence. An alternative approach is to base a characterization of the excitation function on measurements on the speech waveform that are independent of those used to characterize the spectral envelope, using any one of a variety of pitch and buzz–hiss detectors. Measurement of the excitation function from the cepstrum is perhaps the most reasonable if the excitation parameters are to be sampled at the same rate as the vocal-tract parameters. If the excitation parameters are to be sampled more rapidly, which is typically the case, then the use of other pitch-detection algorithms is very likely to be more practical. The analyzer configuration is summarized in Fig. 1.

II. SYNTHESIZER CONFIGURATION

Speech synthesis was accomplished by converting the results of the analysis to an impulse response function, which, when convolved with an excitation function, produced the synthetic speech. The excitation function was generated from a knowledge of the pitch period and a voiced–unvoiced decision. During voiced portions, the excitation function consists of a train of

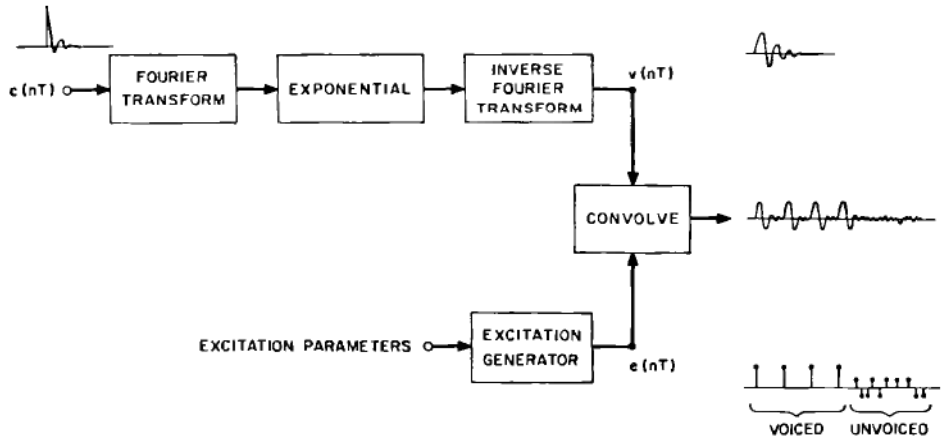
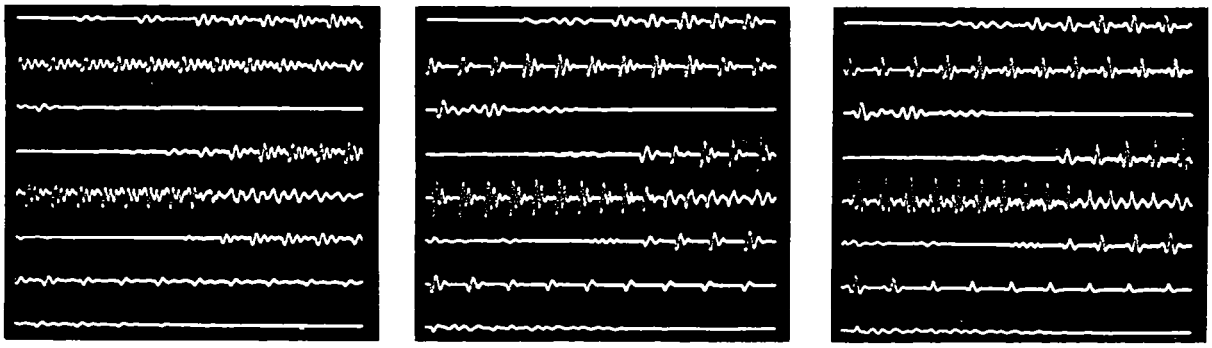


FIG. 2. Representation of the synthesizer configuration.



(a) (b) (c)

FIG. 3. Example of original and processed speech. (a) Sample of original speech. (b) Corresponding sample of processed speech with minimum-phase synthesis. (c) Corresponding sample of processed speech with zero-phase synthesis. Each sample corresponds to 819.2 msec with 102.4 msec per line. The sentence is "Your jumping thrilled him."

pulses with a spacing equal to the measured pitch period. During unvoiced portions, the excitation function consists of a train of equally spaced pulses with random polarity. The amplitudes of the pulses during hiss are adjusted to achieve a proper relationship in the relative energies during voicing and unvoicing.

The impulse response function is obtained from the cepstrum by computing the DFT of the cepstral values, followed by an exponential transformation and the inverse DFT. Thus, if C_n represents the nth channel, then the cepstrum $\hat{v}(nT)$ may be taken to be

$$\hat{v}(nT)=\hat{v}(-nT)=C_n, \qquad 0\leq n\leq M-1,$$

and

$$\hat{v}(nT)=\hat{v}(-nT)=0, \qquad M<n, \tag{4}$$

so that the cepstrum is taken to be an even function. The resulting impulse response function $v(nT)$ is then an even function.

An impulse response function obtained in this way is a zero phase function, that is, it has a Fourier transform which is real. An alternative is to modify the cepstrum in such a way that a nonzero phase characteristic is introduced without altering the spectral magnitude. One specific means for accomplishing this is by forming the cepstrum from the channel signals as

$$\hat{v}(nT)=\begin{cases} 2C_n & 0<n\leq M-1 \\ C_n & n=0 \\ 0 & n<0 \end{cases}. \tag{5}$$

It is easily verified that the even part of the cepstrum defined either through Eq. 4 or 5 is the same, and consequently the magnitude of the spectrum of the impulse response function obtained from either is identical. However, the impulse response function obtained through Eq. 5 is not zero phase, since its Fourier transform has a nonzero imaginary part. In particular, reconstruction of the cepstrum by means of Eq. 5 corresponds to generating a minimum phase characteristic as defined through the Hilbert transform. Thus, the impulse response function obtained in this way can be referred to as a minimum-phase impulse response function.

Synthesis of the speech is carried out by means of an explicit convolution of the impulse response function and the excitation function as depicted in Fig. 2.

III. COMPUTATIONAL CONSIDERATIONS

The system described above was simulated on an 18-bit digital computer with fixed-point arithmetic. The input speech was pre-emphasized, low-pass filtered at 5 kHz, and sampled at 10 kHz. The speech was digitized to 9 bits.

The analysis was carried out by weighting the input speech with a Hanning window having a duration of 40 msec. The spectral analysis consisted of a 512 point DFT corresponding to a spectral resolution of approximately 20 Hz. Similarly, the cepstrum was computed with a 512 point inverse transform. The channel signals consisted of the first 32 points of the cepstrum. A new cepstrum was computed at 20-msec intervals along the speech waveform, thus providing samples of the channel signals at 20-msec intervals.

In conventional channel vocoder systems in which transmission is digital, but for which the analyzer and synthesizer processing is analog, the channel signals are low-pass filtered and sampled before digitization and desampled by means of low-pass filtering at the receiver. In the system under discussion in this paper, we likewise interpolated the pitch information and the impulse response function information between sampling instants to avoid roughness that would be introduced into the synthesized speech due to sudden changes in pitch or spectral envelope information. While several means for implementing this interpolation suggest themselves, the most straightforward and the one incorporated in the system that was simulated, is a linear interpolation. Thus, in carrying out the synthesis, with a new impulse response computed at 20-msec intervals along the speech waveform, an im-

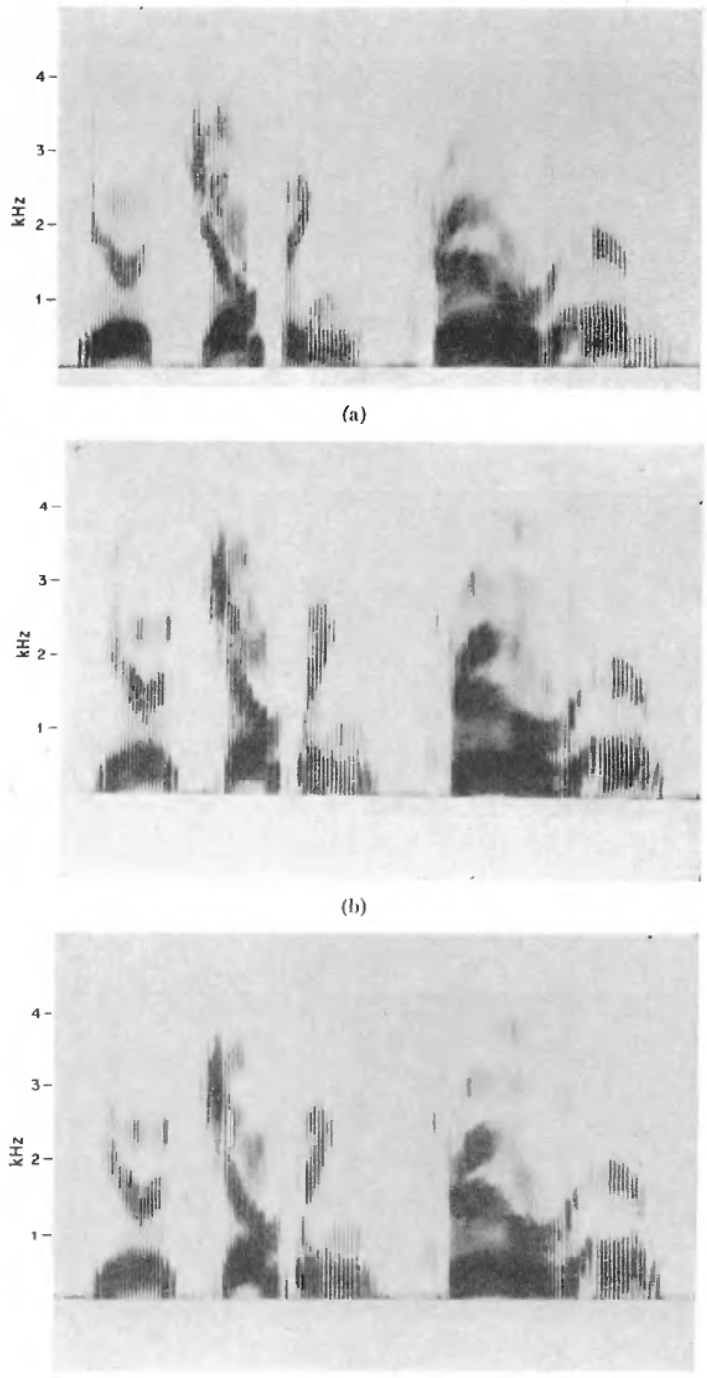


FIG. 4. Spectrograms of original and processed speech, for the sentence illustrated in Fig. 3. (a) Original speech. (b) Processed speech with minimum-phase synthesis. (c) Processed speech with zero-phase synthesis.

pulse response beginning at a time T into the interval would be computed as

$$f(nT)=f_1(nT)+[f_2(nT)-f_1(nT)]T/(20\times10^{-3}).$$

This is mathematically equivalent to carrying out a linear interpolation (in time) of the complex spectrum of the impulse response function.

The convolution of the excitation function and impulse-response functions requires that at the beginning of each pitch period the impulse response be added to the tail from the previous impulse response. An alternative that is considerably simpler is to truncate the previous impulse response at the beginning of the new pitch period. This can either be done for voiced portions

only or for both voiced and unvoiced. Comparison of this method of synthesis indicates that if it is only carried out during voiced portions, there is no detectable effect as judged over a cross section of speakers and sentences. If it is also used during unvoiced synthesis, the effect is to reduce somewhat the quality of the noise excitation. This degradation is mild and might very well be justified in light of potential hardware simplicity.

The excitation parameters were obtained by means of cepstral pitch detection as described by Noll.[4] The long-time portion of the cepstrum was digitally interpolated to a sampling rate of 20 kHz and low-pass filtered. The interpolation was carried out by first generating a sequence consisting of the sequence to be interpolated alternating with samples of zero value. This new sequence was then low-pass filtered. Filtering was carried out using a recursive digital fourth-order Butterworth filter with a cutoff frequency of 2.5 kHz. The filtering was carried out forward and backward on the sequence to achieve an effective filter characteristic that has zero phase.

It was found that a particularly strong indication of the absence of voicing was a lack of correspondence between the location of a cepstral peak before and after interpolation and low-pass filtering. This coupled with a measurement of the input energy formed the basis for a voiced–unvoiced decision. The location of the peak during voicing was used as a measurement of pitch.

Measurement of the excitation parameters with the system described above led to errors in pitch consisting of pitch doubling and pitch halving. Errors in voiced–unvoiced decisions tended to occur primarily at boundaries, i.e., in transition from voiced to unvoiced. For the specific experiments described below, these errors in excitation were corrected by visual inspection and hand editing. The only changes made in pitch values were to correct errors of doubling or halving. The occurrence of errors of this type and algorithms for correcting them have been discussed by Noll.

IV. RESULTS AND DISCUSSION

An illustration of the synthesized speech is shown in Fig. 3. Figure 3(a) represents approximately 0.8 sec of original speech and Figs. 3(b) and (c) represent the same portion of the sentence after processing. Figure 3(b) corresponds to a minimum-phase synthesis and Fig. 3(c) corresponds to a zero-phase synthesis. Spectrograms of these sentences are shown in Figs. 4(a), (b), and (c).

The effect of introducing into synthesized speech a nonzero phase characteristic has been discussed by

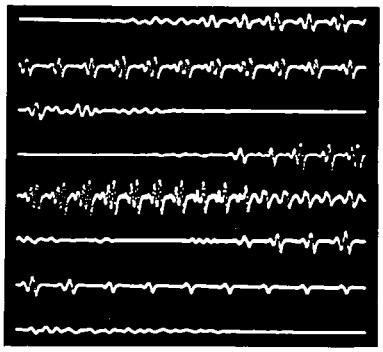


FIG. 5. Speech sample of Fig. 3(a) processed using maximum-phase synthesis.

David, Miller, and Mathews[5] and by Gold.[6] In the experiments reported in Ref. 5, a pitch synchronous analysis of the speech waveform was carried out and the individual pitch periods processed in an all-pass manner followed by resynthesis of the speech. Their results were phrased in terms of the peak factor of the synthesized speech. They reported that speech samples with a higher peak factor tended to have "raucous" quality, while a lower peak factor sounded more tonal. They comment, however, that the effect seemed highly variable to the observers with the magnitude of the effect varying considerably between speakers and sentences.

Gold describes a set of experiments that are directed toward introducing into speech generated by means of a spectrally flattened channel vocoder a speechlike phase characteristic. In his system, the channel vocoder synthesizer was preceded by a formant synthesizer based on the notion that after spectral-flattening the effect of the formant synthesizer is to introduce a phase characteristic into the synthesized speech without affecting the spectral magnitude characteristics. Since a set of formant networks is a minimum-phase network, this system can be considered to generate an approximation to a minimum phase characteristic. Gold hypothesizes that ". . . this improvement occurs because a formant-tracking vocoder more faithfully reproduces the phase of the actual speech than does a channel vocoder."

On the basis of these two sets of experiments, we may expect that within the context of the present system the minimum-phase synthesis is preferable to the zero-phase synthesis both because it has a lower peak factor and because it is closer to the phase of the original speech. Informal listening tests were conducted using an AB forced-choice preference test with experienced listeners. With careful listening over headphones a preference for minimum-phase speech was

[4] A. M. Noll, "Cepstrum Pitch Determination," J. Acoust. Soc. Amer. **41**, 293–309 (1967).

[5] E. E. David, J. E. Miller, and M. V. Mathews, "Monaural Phase Effects in Speech Perception," Proc. Third Int. Congr. Acoust., 3rd, Stuttgart, 1959, I. 227 (1961).

[6] B. Gold, "Experiment with Speechlike Phase in a Spectrally Flattened Pitch-Excited Channel Vocoder," J. Acoust. Soc. Amer. **36**, 1892–1894 (1964).

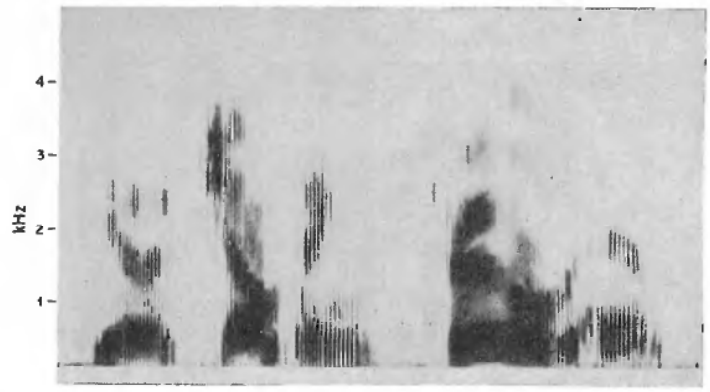


FIG. 6. Spectrogram of processed speech with maximum-phase synthesis.

given in roughly 75% of the pairs. Consistent with the results of David *et al.*, the effect was subtle, being evident for some sentences and speakers and not all evident for others. It was generally agreed among the listeners that the effect was not first order.

To help focus the question of whether a minimum-phase characteristic was preferable to other phase characteristics with the same peak factor, speech was synthesized with a different phase characteristic that was chosen to maintain the same peak factor as in the minimum-phase case. Specifically, a minimum-phase impulse response function was obtained, and prior to convolution with the excitation function, was reversed in time. That is, if $v_1(nT)$ denotes a minimum-phase impulse response, and $v_2(nT)$ represents the impulse-response function used in the synthesis, then

$$v_2(nT)=v_1(-nT).$$

The phase characteristic associated with $v_2(nT)$ is the negative of the phase associated with $v_1(nT)$. The magnitude of the spectrum is the same in both cases. If $v_1(nT)$ is a minimum-phase function, that is, it has all its poles and zeros in its z transform inside the unit circle, then $v_2(nT)$ has all its poles and zeros outside the unit circle and could consequently be referred to as maximum phase. We would expect that speech based on a maximum-phase synthesis would have the same peak factor as with minimum phase. Figure 5 shows the same segment of speech displayed in Fig. 3, but using maximum phase synthesis. In Fig. 6 is shown a spectrogram for comparison with Fig. 4. Again, informal listening tests were conducted with experienced listeners using an AB forced-choice format. Maximum-phase synthesis was compared with minimum-phase synthesis and with zero-phase synthesis. Minimum-phase synthesis was preferred to maximum phase in approximately 85% of the pairs and zero phase was preferred to maximum phase in approximately 85% of the pairs. Maximum phase has a noticeably rougher quality with the effect being considerably more evident than in the comparison between minimum phase and zero phase.

V. SUMMARY AND CONCLUSIONS

A speech analysis–synthesis system has been described that requires a digital computation of a high-resolution spectrum and is based on a recently proposed method for carrying out a deconvolution of waveforms. In effect, the system analyzer computes the cepstrum and uses samples near the origin to characterize the spectral envelope or equivalently the vocal-tract impulse response. In the synthesizer, the cepstral samples are converted to an impulse-response function and an explicit convolution carried out with an excitation function. Either zero-phase or minimum-phase impulse-response functions may be obtained. The system was simulated and informal tests were carried out using a cross section of sentences and speakers. Minimum phase was slightly preferred over zero phase, although the difference was subtle and both speaker and sentence dependent. This result is consistent with those reported by David *et al.* and by Gold. In contrast, maximum-phase speech, which has the same peak factor as minimum-phase speech, had a markedly rougher quality, supporting the hypothesis that a speechlike phase is preferable in synthetic speech.

In general, comments from experienced listeners were that this system produced very high quality, natural-sounding speech. There are three primary aspects in which this system differs from conventional channel vocoder systems. One difference is in the means for obtaining spectral envelope information. In the present system, this information is obtained by linear smoothing of the log spectrum. A conventional channel vocoder is more nearly equivalent to carrying out a linear smoothing of the complex Fourier transform of a weighted sample of speech. The second difference lies in the phase characteristics of the synthesized speech. The phase associated with the equivalent impulse response in a conventional channel vocoder is introduced

by the phase characteristics of the synthesizer filter bank and is essentially constant phase. The third major difference lies in the method of synthesis.

For the experiments discussed above, the system was considered to be operating in an analog mode, that is no additional quantization was imposed on the channel signals beyond the 18-bit quantization imposed by the finite register length of the computer. In a separate experiment, it was ascertained that the number of channel signals could be reduced to 26 (that is, the upper six channels discarded) and quantized to six bits per channel corresponding to 7800 bits/sec for the channel signals without noticeable degradation in the quality of the synthesized speech. It is anticipated that a further reduction in the bit rate can be accomplished.