THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 4, April 1979
Printed in U.S.A.

A Digital Simulation of the Telephone System

By C. E. SCHMIDT, L. R. RABINER, and D. A. BERKLEY

(Manuscript received March 20, 1978)

An increasing number of practical systems for speech communications have been proposed in the past few years. Such systems often must operate over both wideband channels and standard telephone connections. Thus, it is useful to be able to simulate the telephone channel as well as the other speech processing parts of the given system. This paper describes a digital network which provides a simple, controlled simulation of the properties of both the standard carbon microphone and the telephone transmission system. The simulation consists of a combination of nonlinear distortion, noise addition, and bandpass filtering. Both wideband and telephone signals were recorded simultaneously using a 2-channel A/D converter. The wideband signal was processed by the simulation system, whereas the telephone signal provided a reference signal for purposes of comparison. The parameters of the simulation were manually adjusted to provide optimum matches to several telephone links. The telephone simulation was then subjectively evaluated in two listening experiments. In both experiments, utterances from the simulations were paired with the corresponding telephone recording. In the first experiment, a group of listeners was asked to select the actual telephone recording from each pair of utterances. In the second experiment, a new group of listeners was asked to rank the similarity of the two utterances on a 1-to -10 scale. Results of the evaluations indicated that, for some sets of simulation parameters, the network provided a fairly good psychophysical simulation of a variety of telephone channels.

I. INTRODUCTION

In recent years, an increasing number of systems for speech communications have been proposed which must operate over both wideband channels and standard telephone connections.[1–5] Included among such systems are waveform coders, speech analysis-synthesis systems, and systems for man-machine communication by voice—e.g., speech recognizers or speaker verifiers. To investigate the capabilities of such systems in a controlled manner, it is useful to be able to simulate the effects of the telephone channel on the speech signal as well as the other signal processing parts of the system. The conventional approach to telephone simulations of speech processing systems is to repeatedly dial up a new line for each input utterance to obtain a reasonable distribution of lines. However, not only is this method clumsy, but it does not guarantee good statistical sampling of telephone lines. As an alternative, it would be desirable to substitute a controlled simulation of a telephone channel which attempted to model the system from the handset to the earphone—that is, the telephone carbon button, the telephone line, the switching, and the receiver.

Figure 1 is a set of block diagrams that illustrate the various types of simulations one can consider using. The simplest simulation, in Fig. 1a, is an "end-to-end" simulation in which the input signal (to either the telephone line or the black box simulation) is obtained directly from a talker, and the simulation output is sent directly to a listener. In such cases, the "black box" simulation need not in any way physically model the actual telephone system. It is sufficient for the black box to model only the psychophysically significant effects of the telephone system.

The type of simulation in Fig. 1b is a more demanding one in that the end result of the simulation is subjected to subsequent processing by a physical system prior to evaluation by either a listener or some form of measurement system (e.g., a computer). In order for the black box to be a good simulation of the telephone system, it must be a physical simulation of the relevant processing which actually occurs in the telephone system. For some, if not most cases, one would not expect an end-to-end simulation (such as that shown in Fig. 1a) to perform well in systems which require physical simulations.

The situations shown in Figs. 1c and 1d represent modified versions of the cases shown in Figs. 1a and 1b. For these cases, the input signal is preprocessed prior to the telephone line. *A priori*, one would expect that physical simulations of the telephone line would perform well in both situations (i.e., the systems of Figs. 1c and 1d). However, the end-to-end type of simulation would probably be most successful (because of the placement of nonlinear elements in the system) in the system of Fig. 1c (depending on the details of the preprocessing), whereas in the system of Fig. 1d it would often not be useful.

In this paper, we are primarily concerned with end-to-end simulations of the telephone system. We have chosen this alternative for two main reasons. The first, and perhaps most important, reason is the difficulty in obtaining a good physical characterization of the processing in the actual telephone system from the handset to the receiver. A previous attempt at a physical simulation of only the telephone trans-

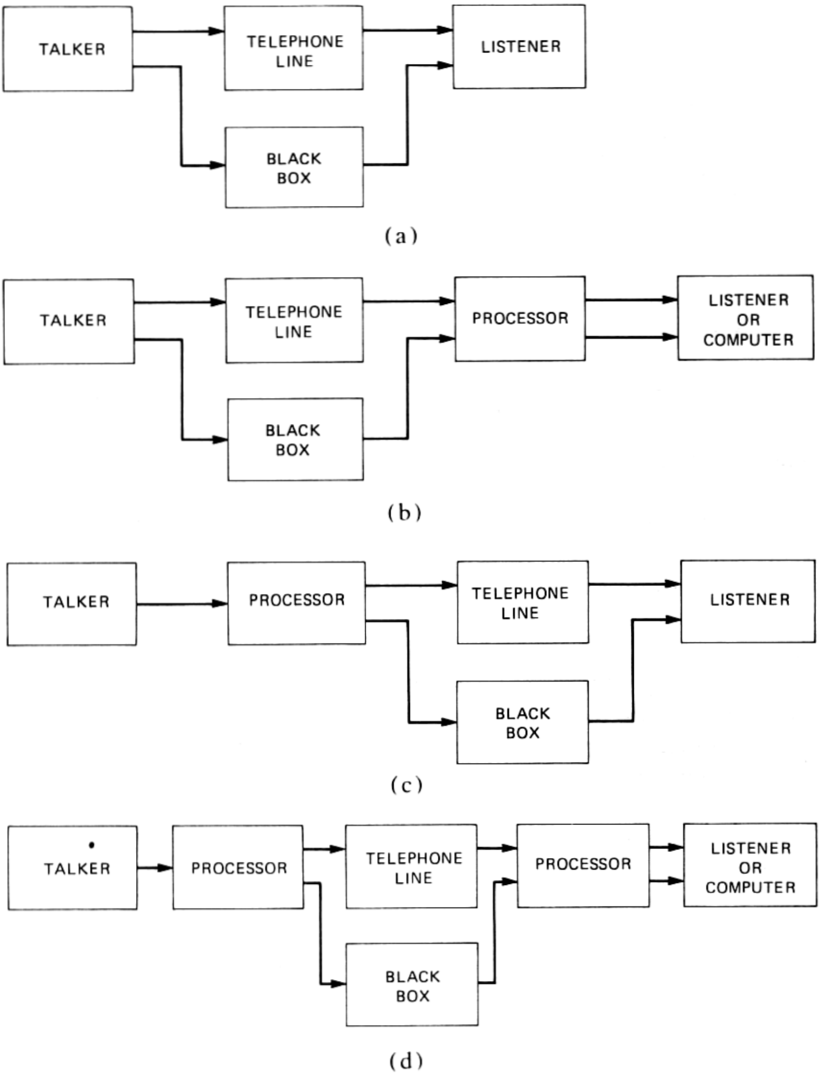

Fig. 1—Block diagrams of typical arrangements of telephone system simulations.

mission system was made at Lincoln Laboratories, based on a set of measurements of continental U.S. and European voice and data grade lines.[6,7] The simulation, run on the real-time signal processor at Lincoln Laboratories, modelled some transmission characteristics of the lines. The effects that were included in this simulation were linear filtering, quadrature distortion (phase jitter), carrier frequency offset, and various types of noise. Effects that occur in satellite transmissions, as well as echo suppression, cross talk, etc. were not included in the Lincoln simulation.[8] Although this simulation was quite sophisticated, the model is overly complex for many applications. For example, for local lines the amount of quadrature distortion and carrier frequency offset is negligible and hence need not be considered in the simulation. In addition, in the Lincoln simulation, no model was provided for the telephone handset itself—i.e., the carbon microphone. Although the characteristics of the carbon microphone are highly nonlinear[9] and not very well understood, a considerable measure of the "telephone" quality of speech is imparted by the carbon microphone. Thus, it is necessary for any end-to-end *or* physical simulation of the telephone system for use with speech input to model both the telephone carbon microphone and the transmission system.

A second reason for our interest in end-to-end simulations is that a number of simplifications can be made in the model, since we need only be concerned with aspects of the telephone system that are psychophysically significant. Hence, we can rely on both our knowledge of auditory and speech perception and past experience with systems that process telephone quality speech to aid in the selection of components of the model. In addition, we are free to investigate simpler models for an end-to-end simulation than could be justified for a physical simulation.

The purpose of this paper is to describe a simple digital network that provides an end-to-end simulation of the combined standard carbon microphone and the telephone transmission system. The network, implemented via digital simulation, has the flexibility of allowing the user to vary parameters of the model, thereby simulating a wide variety of telephone lines. By systematically varying these parameters, we have been able to match the characteristics of several different links and have obtained signals which perceptually have most of the "telephone quality" attributes.

The organization of this paper is as follows. In Section II, we describe the telephone model and the resulting simulation. In Section III, we describe a series of two experiments conducted to determine the effectiveness of the simulation in perceptually matching selected telephone links. In Section IV, we present the results of the experiments and discuss their significance.

II. SIMULATION OF THE TELEPHONE LINE

Figure 2 is a block diagram of how the telephone line simulation is organized. The overall line is modelled as a cascade of a model for the telephone handset (i.e., the carbon microphone) and a model for the telephone transmission channel. Ideally, the Lincoln Laboratories simulation would provide a sophisticated physical model for the telephone transmission. However, in line with our stated objectives, we chose to implement a considerably simpler model.

A more detailed block diagram of the telephone model and the system used to evaluate it is given in Fig. 3. The input to the model is assumed to be a speech signal, bandpass-filtered from 100 to 3000 Hz and sampled at a 10-kHz rate. The transmitter model consists of an interpolator, a nonlinearity (center clipper), and a decimator. The interpolator changes the sampling rate to 50 kHz, keeping the bandwidth the same as at the low rate through the use of a high-order, linear-phase, FIR digital filter.[10] The center clipper has a center clipping (CC) level which is a percentage of the peak signal level for each utterance. The decimator first filters the nonlinearly distorted signal to the original bandwidth and then reduces the sampling rate back to 10 kHz. The 5-to-1 change in sampling rate was sufficient to guarantee that the high-frequency, nonlinear distortion products of the center clipper would not affect the speech baseband.

If we call the input signal to the interpolator $x(n)$ and the output signal from the interpolator $y(n)$, then

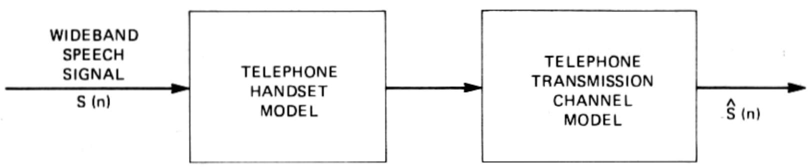

Fig. 2—Block diagram of telephone system model.

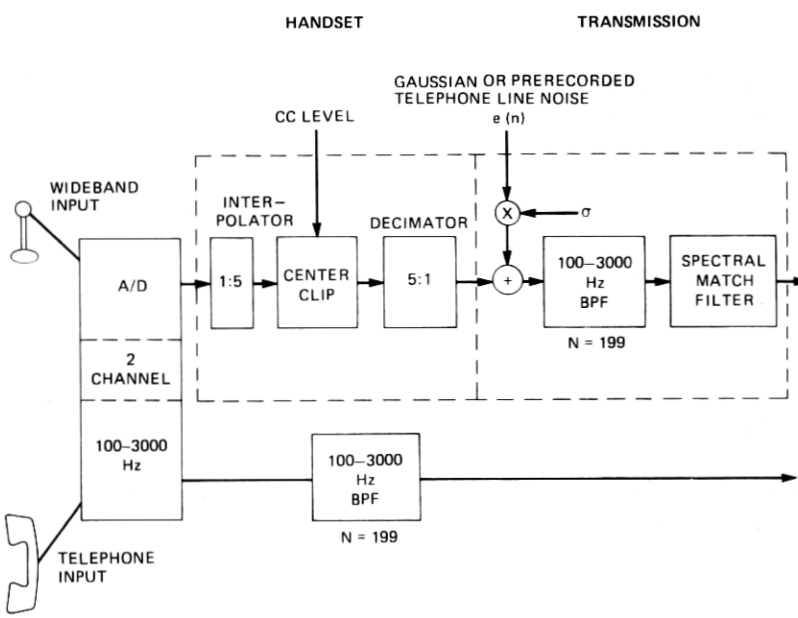

Fig. 3—Block diagram of experimental system used to test the simulation.

$$y(n) = v(n) * h(n) = \sum_{m=0}^{L-1} h(m) v(n-m),$$

where

$$\begin{aligned} v(n) &= x(n/5) \qquad n = 0, \pm 5, \pm 10, \cdots \\ &= 0 \qquad \text{otherwise,} \end{aligned}$$

and $h(n)$ is an L-point FIR linear phase lowpass filter with bandwidth 3 kHz. Thus $y(n)$ and $x(n)$ have the same frequency spectrum in the band $0 \leq f \leq 3$ kHz.

The center clipper has an input-output characteristic of the form

$$\begin{aligned} \hat{y}(n) &= y(n) \quad \text{if} \quad |y(n)| > C_n \\ &= 0 \qquad \text{otherwise,} \end{aligned}$$

where

$$C_n = k \cdot \max_n [|x(n)|]$$

and n ranged over all samples in the utterance. (In the experiment, k was a variable that was investigated.)

To eliminate the high-frequency distortion products in $\hat{y}(n)$, due to the nonlinearity (i.e., the center clipper), the signal $\hat{y}(n)$ was again filtered by the lowpass filter $h(n)$ to give the output $z(n)$, computed as

$$z(n) = \sum_{m=0}^{L-1} h(m) \hat{y}(n-m).$$

The decimated signal, $\hat{x}(n)$, was obtained by retaining every fifth sample of $z(n)$, i.e.,

$$\hat{x}(n) = z(5n).$$

Efficient signal processing techniques were used to implement both the interpolator and the decimator.[10]

The telephone transmission model consisted simply of three components: an additive wideband noise whose amplitude was variable; a fixed-FIR, linear-phase, bandpass filter; and a variable digital filter which provided a spectral shaping that could match any desired shape. The noise used in the simulation was one of two types—either a wideband Gaussian noise or a prerecorded telephone line noise. The spectral matching filter was implemented as a cascade of two digital filters. The first filter was a 25-point, FIR, linear-phase filter which provided a gross spectral match to the signal spectrum of an average telephone line—i.e., it provided an approximation to the general non-flat frequency weighting in the system. The second filter was a 255-point, FIR, minimum-phase filter which provided a detailed match of the simulation spectrum to the spectrum of a specified telephone line.

The lower path of the system shown in Fig. 3 represents a direct recording of a telephone signal at the receiver of the called party—i.e., after the speech signal has been transduced by the carbon microphone of the transmitter, sent over a link to a central office, and returned to a different telephone at the same location where the call was initiated. A 2-channel, analog-to-digital converter was used to simultaneously digitize the original speech signal (from a high-quality microphone) and the resulting telephone speech signal. Both signals were bandpass-filtered from 100 to 3000 Hz and digitized at 10-kHz rates.

Using the model of Fig. 3, a series of investigations were made to see to what extent the signal at the output of the simulation could perceptually match the telephone signal recorded simultaneously with the original wideband signal. In the next section, we describe these investigations.

III. EXPERIMENTAL INVESTIGATIONS

To evaluate how well the system shown in Fig. 3 could model the perceptual characteristics of various telephone lines, a series of recordings was made under the following conditions:

(*i*) Two speakers—one male, one female.

(*ii*) Two sentences.

(*iii*) Three telephone links.

- (*a*) Single PBX loop within Bell Laboratories (Murray Hill to Murray Hill). This is a standard Centrex line.
- (*b*) Double PBX loop (Murray Hill to Whippany to Murray Hill).
- (*c*) Double PBX loop (Murray Hill to Holmdel to Murray Hill).

The double PBX should be typical of local exchange carrier transmission—i.e., the transmission path typically contains channel bank filters (analog or digital) and some form of companding.[8]

To illustrate a typical recording, Fig. 4 is a plot of the long-term average spectrum for one of the wideband sentences. Included in this plot are both the long-time average spectrum (computed using a 1024-point FFT analysis), and a cepstrally smoothed representation of the long-time average spectrum.[11] The speech spectrum is seen to fall by about 40 dB at 5 kHz.

Similarly, Fig. 5 is a plot of the long-time average spectrum (and its cepstrally smoothed representation) for the telephone recording which corresponds to the wideband recording of Fig. 4. The most significant difference in the spectra of the respective signals is the reduced bandwidth of the telephone recording. However, careful comparisons between spectra shows significant differences in the region from 100 to 3000 Hz. The spectral matching filter was intended to model these differences.

Preliminary informal experimentation with the model of Fig. 3 for the above set of recording, indicated that clipping levels below 1

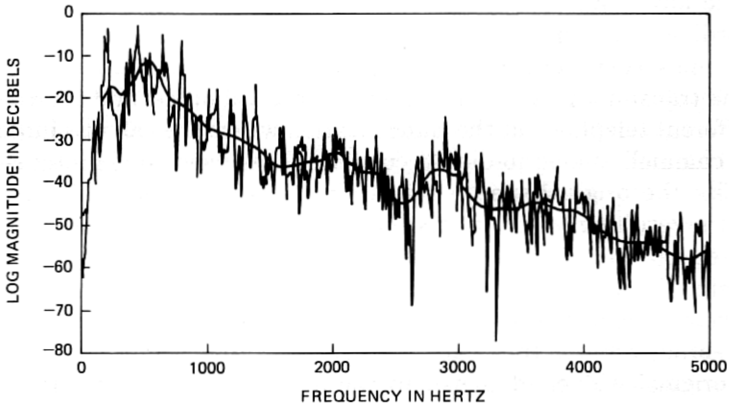

Fig. 4—Long-time average spectrum (irregular curve) and cepstrally smoothed version (smooth curve) for a typical wideband speech signal.

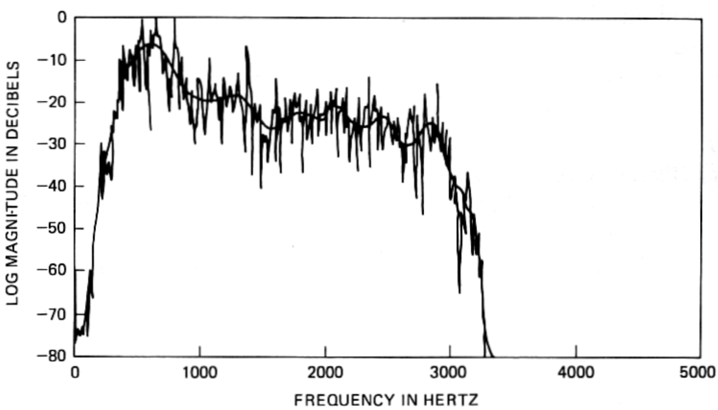

Fig. 5—Long-time average spectrum (irregular curve) and cepstrally smoothed version (smooth curve) for a typical telephone speech signal.

percent had no significant effect on the speech, whereas those above 2 percent produced excessive distortion in the speech. Thus, a choice of 3 clipping levels was made, namely 1 percent, 1½, and 2 percent of the peak signal level of the signal throughout the utterance.

The range of noise gains (σ in Fig. 3) was also determined by informal listening. For the Gaussian case, it was found that noise levels corresponding to signal-to-noise ratios in the range from 30 to 40 dB were optimal. As such, two noise gains were chosen corresponding to approximately 30- and 40-dB signal-to-noise ratios. For the prerecorded telephone line noise, the original signal was obtained at the level of a normal single PBX loop by recording directly off the telephone line with no speech present. Two gain levels of $\sigma = 1$ and $\sigma = 2$ were used in the experiment.

Once the center clipping and noise level ranges were determined, the remaining factor in the simulation was the design of the spectral match filter. To formulate an idea of the requirements on the spectral match filter, plots of the smoothed, long-time spectral differences between the digitally filtered telephone signal and the signal at the output of the bandpass filter in the model were obtained. Figure 6 shows one of these plots. In the region from 100 to 3000 Hz, the real telephone signal spectrum was from 0 to 15 dB above the simulation spectrum. The variation in this spectral difference among speakers, sentences, and telephone links was not very large. Thus, a rough approximation to this spectral difference was used to design a 25-point, FIR, linear-phase digital filter which provided a gross spectral match (to within ±5 dB) of the simulation spectrum to the telephone spectrum in the range from 100 to 3000 Hz. Figure 7 shows the frequency response of a typical gross spectral match filter. The filter provides a fairly good match near the peaks of the difference curve but is significantly worse near the valleys.

Although informal listening comparisons between the simulation output (after the gross spectral match filter) and the telephone input

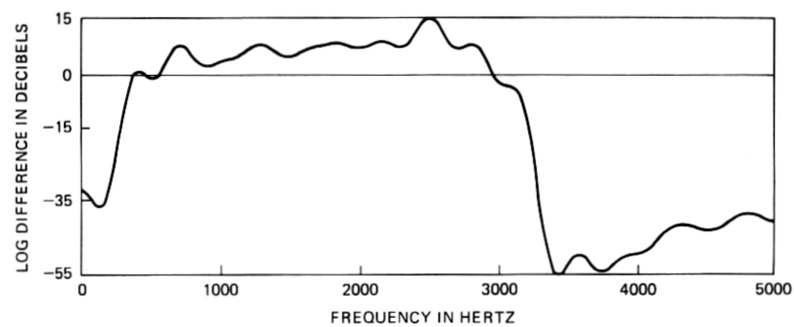

Fig. 6—Long-time spectral difference between a telephone signal and the simulation output after simple bandpass filtering.

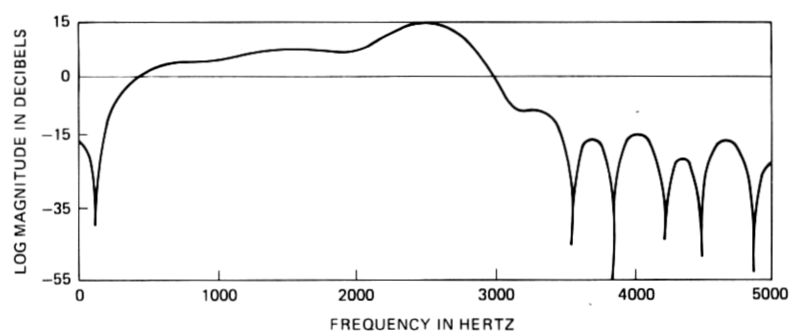

Fig. 7—Frequency response of a typical gross spectral match filter.

indicated a strong similarity between the simulations and the actual telephone speech, there were equally significant differences in both temporal and spectral detail. To assess to what degree the spectral differences were significant, a second spectral match filter was implemented using a high-order (255-point, FIR minimum phase filter) match to the difference spectrum. Figure 8 shows the frequency response to a typical fine spectral match filter, and Fig. 9 shows the resulting spectral difference. As seen in this figure, spectral deviations of about ¼ dB are obtained in the range from 100 to 3000 Hz.

An alternative measure of the spectral similarity between two utterances is the LPC distance or the log likelihood ratio as proposed by Itakura.[12] This measure shows, on a *frame-by-frame* basis, the log spectral difference between two utterances. Figure 10 shows a pair of plots of the LPC distance between the original wideband recording and the original telephone input. The LPC distance is not exactly symmetrical. Thus, the plots show both distances associated with the pair of

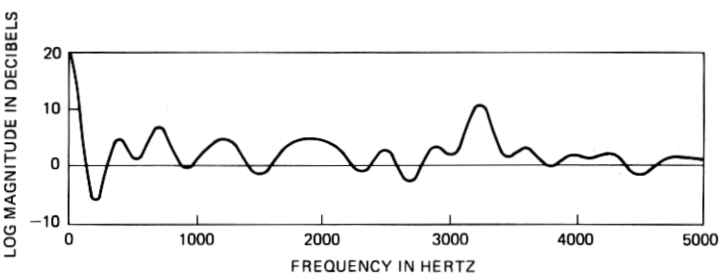

Fig. 8—Frequency response of a typical fine spectral match filter.

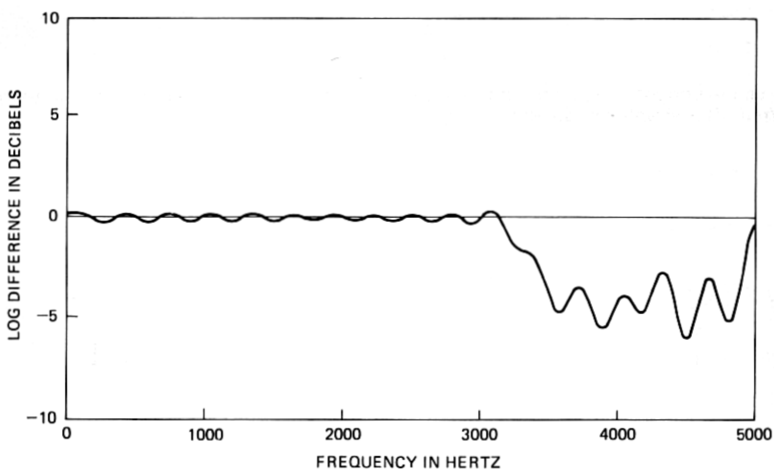

Fig. 9—Long-time spectral difference between a telephone signal and the final simulation output.

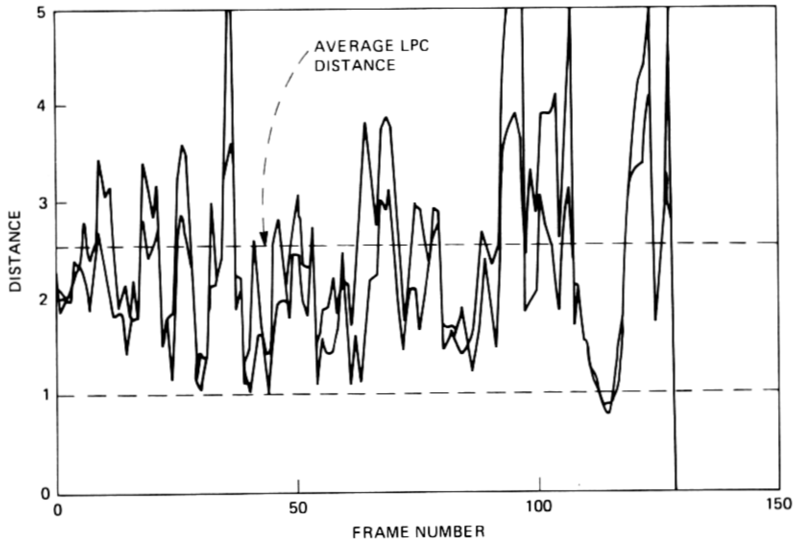

Fig. 10—Plots of frame-by-frame LPC distances between an original wideband signal and the corresponding telephone signal.

utterances. The dotted line at the bottom (at a distance of about 1) is a threshold for which the spectral differences between frames is perceptually significant.[13] From Fig. 10, it is seen that, for almost every frame in the utterance, the distance was significant. The average LPC distance for this example was about 2.6.

In contrast, Fig. 11 shows a pair of plots of the LPC distance between the original telephone input and the output after the gross spectral match filter (Fig. 11a) and after the fine spectral match filter (Fig. 11b). The average LPC distance is about 0.66 for the plots of Fig. 11a and 0.57 for the plots of Fig. 11b. Both these distances are well below the perceptually significant threshold. However, it is still seen in both Figs. 11a and 11b that for a number of frames, the LPC distance exceeds the perceptual threshold. These differences are due to both temporal and short-time spectral differences between the utterances which the simulation is incapable of handling.

Based on the informal observations and the objective measurements described above, further evaluation of the simulation was achieved through two subjective experiments. The first was an A-B comparison between the output of the model and the actual telephone signal to see how well listeners could identify the actual telephone signal. The second experiment was a similarity ranking test between the simulation output and the actual telephone signal.

Table I provides a summary of the experimental factors included in the tests. We have already discussed the first six factors in the test; namely, speakers, sentences, transmission links, center clipping levels,

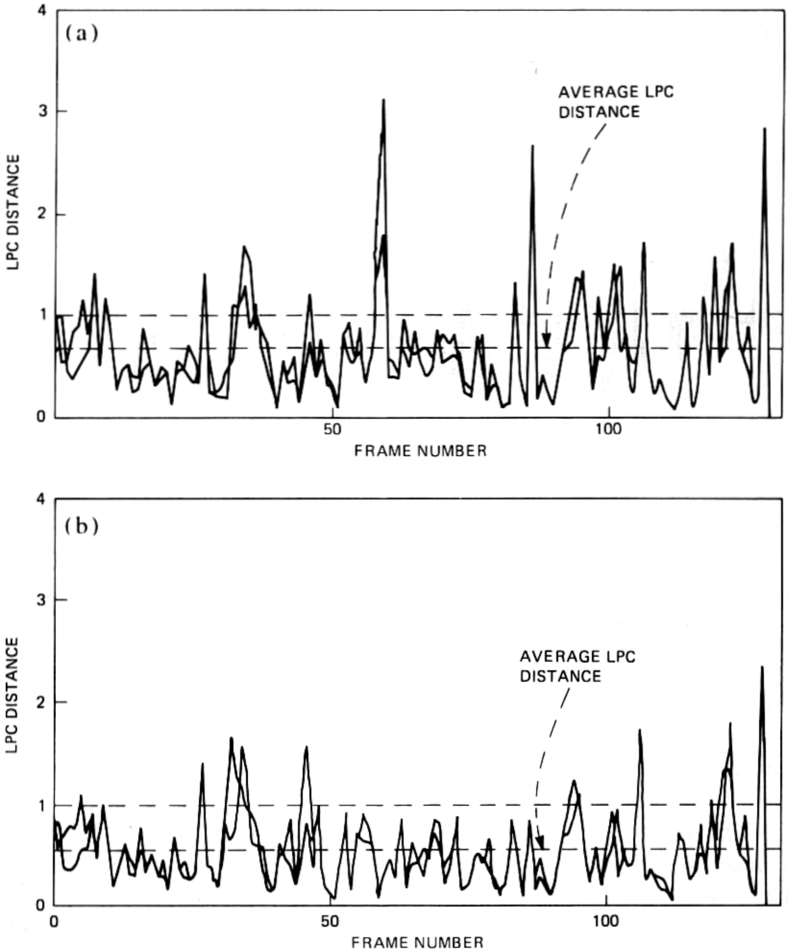

Fig. 11—Plots of frame-by-frame LPC distances between the simulation output and the corresponding telephone signal; (a) after the gross spectral match filter; (b) after the fine spectral match filter.

Table I—Summary of experimental factors in the listening experiment

Experimental Factors
1. Speakers (2-1 male, 1 female).
2. Sentences (2).
3. Transmission links (MH-MH, WH-MH, HO-MH).
4. Center clipping levels (3).
5. Types of noise (2).
6. Noise levels (2).
7. Levels of filtering (3).
8. Listeners (8).
Experiments
1. A-B comparison for identifying actual telephone line.
2. Similarity ranking between simulation and actual telephone line (scale 1–10).

types of noise, and noise levels. The seventh factor was the three levels of filtering; namely, a flat bandpass filter alone, a combination of a flat bandpass filter and the gross spectral match filter, and the triple combination of the flat bandpass filter, the gross spectral match filter, and the fine spectral match filter. The last factor was listeners (eight were used). In both experiments, the simulation utterances were paired with the corresponding telephone recordings in a random ordering. For the A-B test, a simple choice was required of the listeners as to which of the pair was the actual telephone signal. In the ranking tests, the listeners were asked to rank the similarity of the pair on a scale of 1 to 10 where 1 was most similar and 10 was not at all similar. Results of these experiments are given in the next section.

IV. RESULTS

For the telephone identification experiment (the A-B comparison test), a score of 1 was given if the simulation was identified as the telephone, and a score of 0 was given if the listener correctly chose the actual telephone recording. Data were collected over a 1-week period. An analysis of variance of the results indicated that the only significant factors in the experiment were listeners and types of filtering. The fact that the center clipping level was not a significant experimental factor was not surprising since the *range* of this parameter had been carefully chosen to be reasonable for typical telephone lines. Although the noise level factor was not significant, the type of noise was treated as an independent factor since the data for each type of noise were obtained in separate runs, and this factor is an interesting one from the simulation point of view. We shall see that there were surprisingly small differences in this factor.

It was not surprising that the factors of speakers and sentences were not significant; however, it was somewhat unexpected that the transmission link factor was not significant. This result implies that, to a first order, a linear system provides a good (or a uniformly bad, which seems very unlikely) approximation to differences in transmission.

The average identification scores, as a function of level of filtering, were computed from the raw data and are shown in Fig. 12a. Figure 12b shows a more detailed set of results in which the listener scores are individually plotted for the Gaussian noise case. For these plots, a mean score of 0.5 indicates that the listener could not identify the actual telephone recording, and scores close to 0 indicate the listener could always identify the actual telephone line. The brackets in Fig. 12a indicate the range for one standard deviation. The extreme variability in mean scores across listeners can readily be seen in Fig. 12b. However, it can be seen that, for all but two of the listeners, for both noise types, the mean identification scores got larger (headed towards chance identification) as the complexity of filtering increased. Fur-

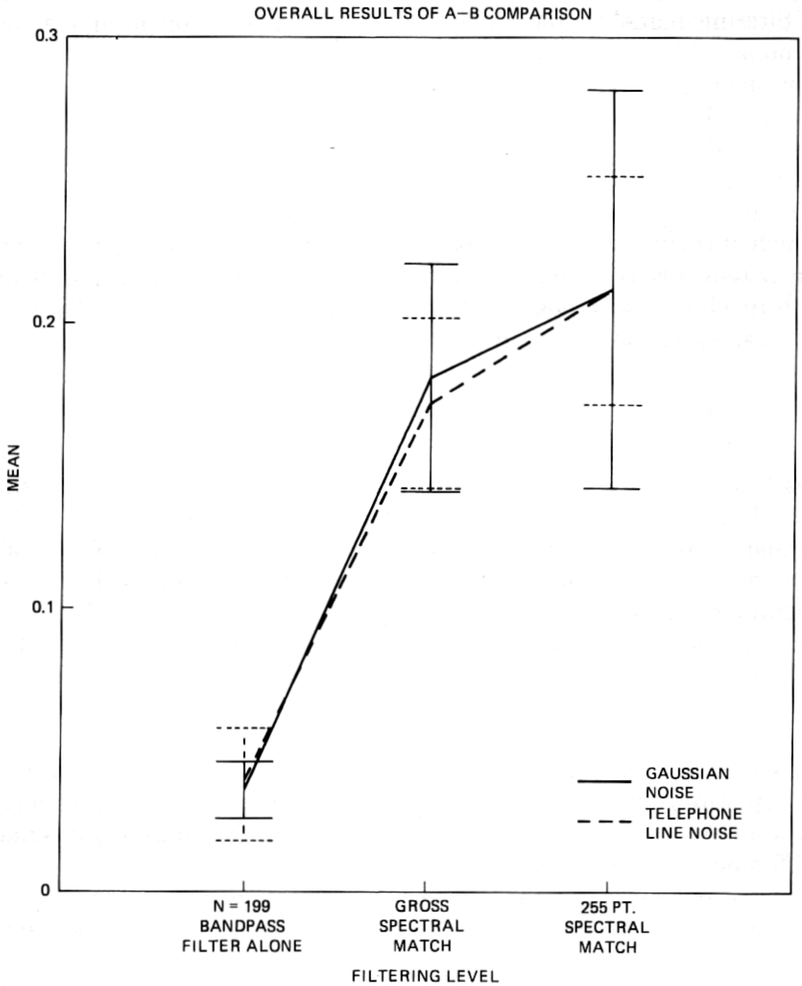

Fig. 12a—Mean identification scores for experiment 1 as a function of filtering level averaged across all conditions.

thermore, for the fine spectral match, the mean identification score was quite close to the score for the gross spectral match filter, thereby indicating the small improvement obtained with the final filter. The standard deviation of the results using the telephone noise was somewhat smaller than for the comparable Gaussian noise. The mean identification score of 0.22 indicates that, even for the best cases, listeners could distinguish between an actual telephone recording and the simulation output approximately 50 percent of the time in an A/B test. However, such a test is a very severe one as any flaw in the simulation will immediately cue the listener as to which of the pair of

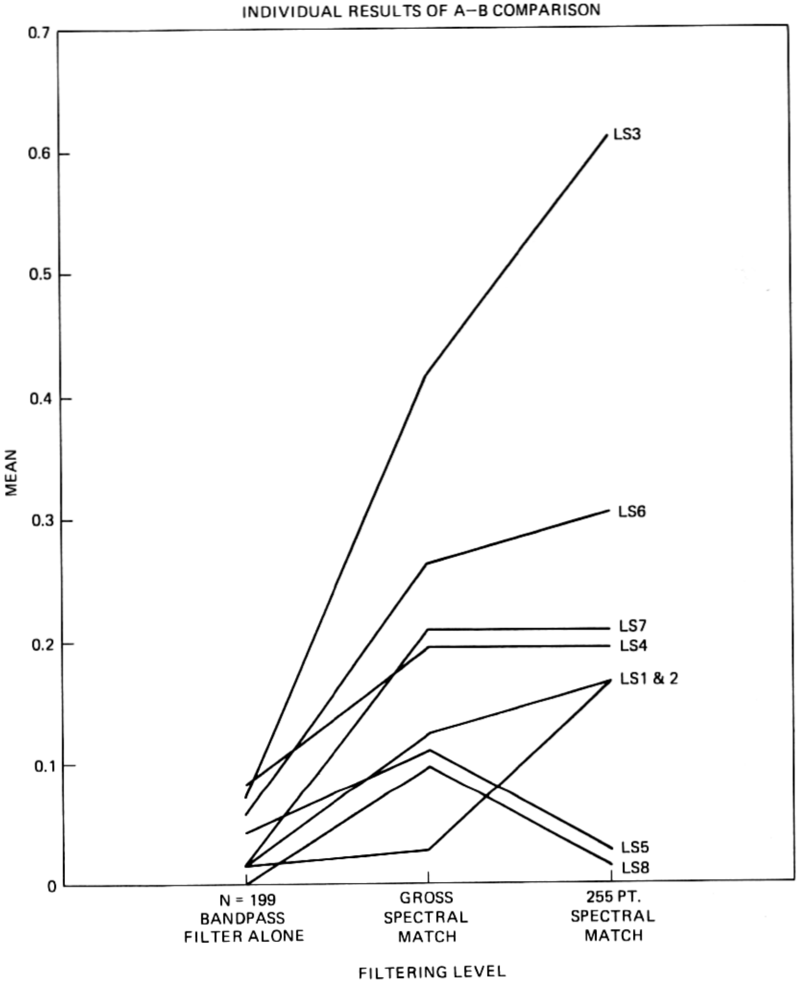

Fig. 12b—Mean identification scores for experiment 1 as a function of listener using Gaussian noise.

utterances is the simulation. Thus, the average score of 0.22 is actually an encouraging one in such a test.

For the similarity ranking experiment, a score of 1 indicated high similarity between sentences, whereas a score of 10 indicated large differences between sentences. For this experiment, a new group of 8 listeners was used and test data were again recorded over a 1-week period. An analysis of variance of the results again showed that the only significant experimental factors were listeners and filtering complexity. Again, however, results are shown for both types of noise. Figure 13 shows a plot of the overall mean ranking score as a function

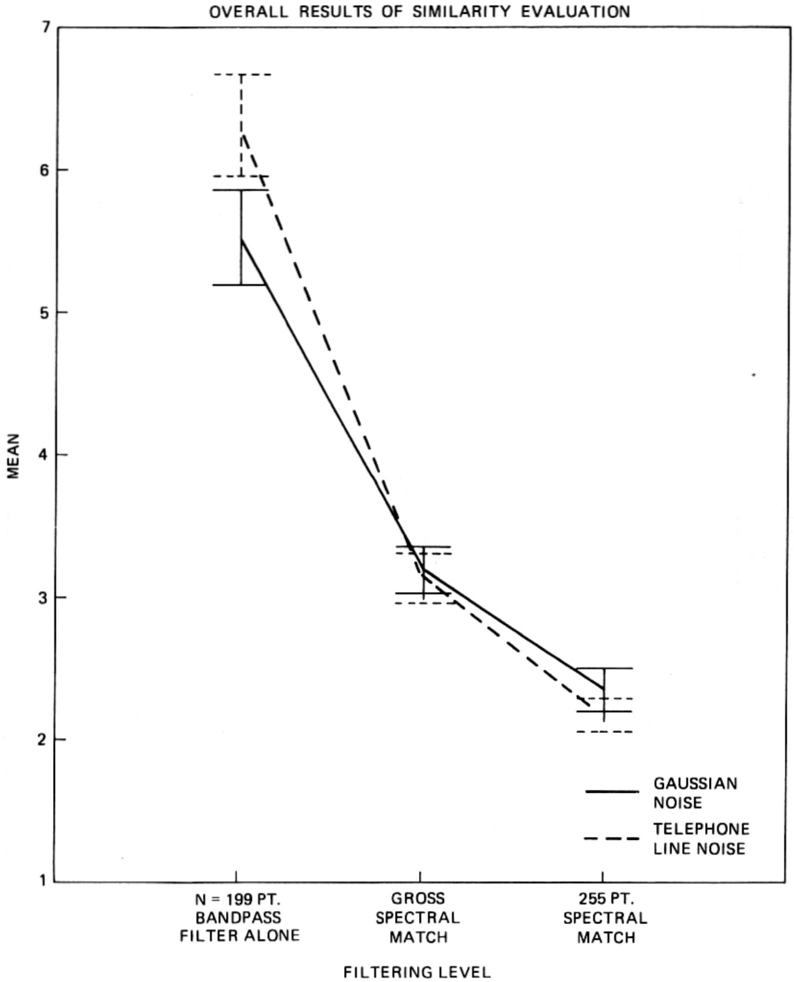

Fig. 13—Mean identification scores for experiment 2 as a function of filtering level averaged across all conditions.

of the level of filtering. It can be seen that the average ranking score was about 3.2 for the gross spectral match filter, whereas it was 2.3 for the fine spectral match filter, thereby indicating a very strong similarity between the simulation output and a real telephone recording. It is also seen that the ranking scores are very similar for both the Gaussian and prerecorded telephone noises, again showing this is not a critical feature of the simulation. The results presented in this figure confirm that the simulation produced an output which listeners considered highly similar to actual telephone recordings.

V. SUMMARY

The results of the experimental evaluations indicated that a fairly simple simulation of a telephone channel can provide a good approximation (in a perceptual sense) to a variety of actual telephone lines. Informal listening indicated that, to give speech the "telephone-like" quality (as opposed to matching an individual telephone line), a general spectral matching filter along with the center clipping nonlinearity was adequate. Although the best results were obtained with a combination of a gross spectral match filter and a fine spectral match filter, the results obtained with the low-order gross spectral match filter alone were quite good and undoubtedly would be adequate for a variety of applications.

The intended purpose of this simulation was to provide a digital network which would be controlled in such a way as to make it interchangeable with a real telephone line for perceptual testing and evaluating speech communications systems. The next step in evaluating the applicability of this model to other situations is to compare objective performance on a speech communication system over real telephone lines with results using the simulation. If these comparisons are favorable, then models such as the one proposed here should be useful for testing and evaluating speech processing systems without the need for extensive evaluation over actual telephone lines.

In summary, we have proposed a simple digital network which provides an end-to-end simulation model of a telephone line from the handset to the receiver. Subjective evaluations indicate that a good match to a variety of telephone lines can be obtained.

REFERENCES

1. J. L. Flanagan, "Computers that Talk and Listen: Man-Machine Communication by Voice," Proc. IEEE, *64,* No. 4 (April 1976), pp. 405–415.
2. A. E. Rosenberg, "Automatic Speaker Verification: A Review," Proc. IEEE, *64,* No. 4 (April 1976), pp. 475–487.
3. L. R. Rabiner, "On Creating Reference Templates for Speaker Independent Recognition of Isolated Words," IEEE Trans. on Acoustics, Speech, and Signal Proc., *ASSP-26,* No. 1 (February 1978).
4. A. E. Rosenberg, "Evaluation of an Automatic Speaker-Verification System Over Telephone Lines," B.S.T.J., *55,* No. 6 (July–August 1976), pp. 723–744.
5. L. R. Rabiner, C. E. Schmidt, and B. S. Atal, "Evaluation of a Statistical Approach to Voiced-Unvoiced-Silence Analysis for Telephone-Quality Speech," B.S.T.J., *56,* No. 3 (March 1977), pp. 455–482.
6. S. Seneff, "A Real-Time Digital Telephone Simulation on the Lincoln Digital Voice Terminal," Technical Note 1975–65, M.I.T. Lincoln Laboratory, 30 December 1975.
7. U. S. Government Study of Continental U.S. Voice Grade and Data Lines, unpublished work.
8. Bell Telephone Laboratories, *Transmission Systems for Communications,* 1970.
9. R. W. Kett, "Carbon Microphones for Communication," Proceedings I.R.E.E. Australia (April 1964), pp. 250–256.
10. R. W. Schafer and L. R. Rabiner, "A Digital Signal Processing Approach to Interpolation," Proc. IEEE, *61* (June 1973), pp. 692–702.
11. A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, "Nonlinear Filtering of Multiplied and Convolved Signals," Proc. IEEE, *56* (August 1968), pp. 1264–1291.
12. F. Itakura, "Minimum Prediction Residual Applied to Speech Recognition," IEEE Trans. on Acoustics, Speech, and Signal Proc., *ASSP-23,* No. 1 (February 1975), pp. 67–72.
13. M. R. Sambur and N. S. Jayant, "LPC Analysis/Synthesis From Speech Inputs Containing Quantizing Noise or Additive White Noise," IEEE Trans. on Acoustics, Speech, and Signal Proc., *ASSP-24,* No. 6 (December 1976), pp. 488–494.