THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLIV JANUARY 1965 NUMBER 1

A Technique for Investigating On-Off Patterns of Speech

By PAUL T. BRADY

(Manuscript received July 6, 1964)

A study is made of certain properties of speech which are concerned with determining the presence of speech on a telephone circuit. A speech detector is constructed to yield an output of spurts and gaps, corresponding to the presence or absence of energy above a threshold. A computer program then attempts to correct this pattern for spurious noise operation and for gaps due to stop consonants, eventually yielding a pattern of talkspurts and pauses. Data reported here include the distributions of the spurts and gaps resulting from the detector as well as the distributions of talkspurts and pauses from the computer program. Studied here are the influence on these distributions of detector threshold variations as well as of parameter variations within the computer program. The gaps occurring within talkspurts retain their distribution over a range of thresholds, but the spurts do not. It appears that 200 msec forms a boundary between intersyllabic gaps and listener-detected pauses.

The detection technique developed here is considered to be an improvement over conventional methods, but still yields data whose significance is uncertain. It may be that a simple automatic speech detecting technique using fixed parameters is inadequate for some purposes.

I. INTRODUCTION

The object of this study is to investigate certain properties of speech pertinent to the problem of establishing the pattern of its presence and absence on a telephone circuit. Recent developments in telephony, such as the introduction of long circuits with appreciable delay and the increasing use of voice-operated devices, have prompted learning more about speech patterns, especially as they occur in conversation.[1,2,3]

Although the task of detecting the presence of speech may seem at first to be an almost trivial problem, it is in fact very difficult. Speech has a large dynamic range, and its level frequently falls into the noise, even during segments audible to a listener. In addition, momentary interruptions due in part to stop consonants (/p/, /t/, /k/, etc.) might cause a speech detector to indicate a silent interval whereas a listener would indicate a continuing flow of speech.

Most existing designs of speech detectors employ a slow release, or hangover, to bridge such gaps, but an error equal to the hangover time is made every time the person actually stops talking. The method of detection which was investigated in this study will hopefully avoid some of the pitfalls of conventional detectors.

II. THE DETECTION TECHNIQUE

The detection technique used here is a two-step process. Speech is first played through a speech detector, whose output is then processed by a computer program. These steps will be discussed separately.

2.1 *The Speech Detector — Spurts and Gaps*

A block diagram of the speech detector used in this study is shown in Fig. 1. The incoming signal is first amplified and then full-wave rectified. A threshold detector is set at this point to detect the presence of a voltage above some fixed value. The threshold detector triggers a flip-flop which is cleared 200 times per second by a clock. If the flip-flop is triggered in between clock pulses, a pulse will appear on the output when the flip-flop is cleared. That is, an output pulse indicates that at some time during the last 5 msec the speech energy crossed the threshold. A pulse is therefore an indication of an "on interval," and the absence of a pulse indicates an "off interval." The pulse train from the detector serves as the data for computer analysis.

The threshold width is the difference, in db, between the 1000-cps signal level just required to cause pulses to appear sporadically at the output, and the signal level required to maintain a constant train of pulses. It is about 1 db in this detector. The frequency response of the detector is flat over the voice range.

The detector is thus able to resolve speech into 5-msec segments, this

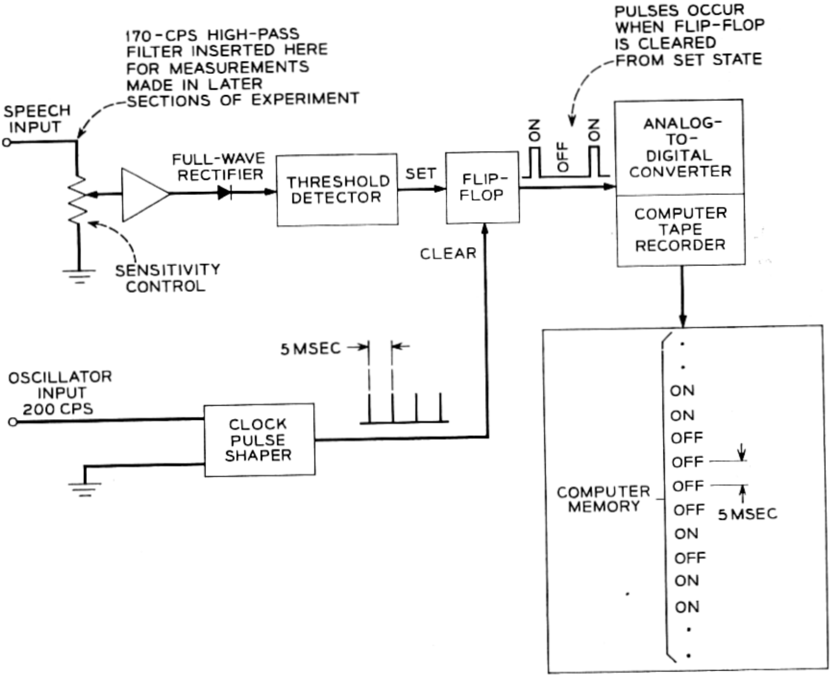

Fig. 1 — Apparatus to convert speech to on-off states.

interval being considered long enough to avoid audio-frequency fluctuations in the energy pattern. This is not exactly the same as a linear *RC* smoothing operation, but there is a rough correspondence. An *RC* network simply smoothes out fluctuations in the energy pattern, while our sampling circuit divides the energy pattern into 5-msec segments and for each segment produces a yes-no output, depending on the maximum amplitude in the segment. The *RC* circuit is much simpler to instrument, but the sampler is more suitable for subsequent computer analysis.

An *on interval* is defined as a 5-msec interval during which the speech energy exceeds the threshold at some time. In an *off interval*, the energy remains below the threshold during the entire interval.

A *spurt* is defined as an unbroken sequence of on intervals. A *gap* is an unbroken sequence of off intervals. By these definitions, therefore, speech can be transformed into a *spurt-gap pattern*. If a 1 represents an on interval and a 0 represents an off interval, the spurt-gap pattern could look like $\cdots$ 00111100011 $\cdots$.

2.2 *Talkspurts and Pauses*

We have already indicated that even while a person is talking his speech can still contain many gaps due to stop consonants and slight hesitations. To obtain a correspondence to the presence or absence of speech, we define a talkspurt and a pause.

A *talkspurt* is a time period which is judged by a listener to contain a sequence of speech sounds unbroken by a pause.

A *pause* is a time period which is judged by a listener to be a period of nontalking, other than one caused by a stop consonant, a slight hesitation, or a short breath.

2.3 *The Computer Program*

Spurts and gaps, as they come from the speech detector, are by definition physically measurable events, while talkspurts and pauses are determined subjectively. It is the function of the computer program to attempt to transform the spurt-gap pattern into a talkspurt-pause pattern. The program which was used actually performs two distinct functions:

(1) An attempt is made to obtain a talkspurt-pause pattern from the speech detector output. The manner in which this is done is described below. The original speech data are thus converted into "corrected data."

(2) The cumulative distribution functions of the durations of talkspurts and pauses are tabulated and plotted. Also computed are the per cent time speech is present and its converse, the per cent time speech is not present, as well as other data such as the mean and median talkspurt and pause lengths.

The procedure used by the computer program to obtain the corrected data is best described by an example. In Fig. 2, pattern (a) is a typical spurt-gap pattern produced by the speech detector. Each spurt and gap has, of course, a duration which is an integral multiple of 5 msec. The first step in data processing is to throw out all spurts which are less than or equal to a *throwaway time*. This is done because noise occasionally operates the speech detector for short periods, and the resulting spurts should be discarded. The throwaway operation produces pattern (b).

At this point, gaps less than or equal to a *fill-in-time* are filled in and considered as speech. This is an attempt to correct for the gaps due to stop consonants and other brief interruptions. It is hoped that pattern (c), obtained after fill-in, will correspond to talkspurts and pauses rather

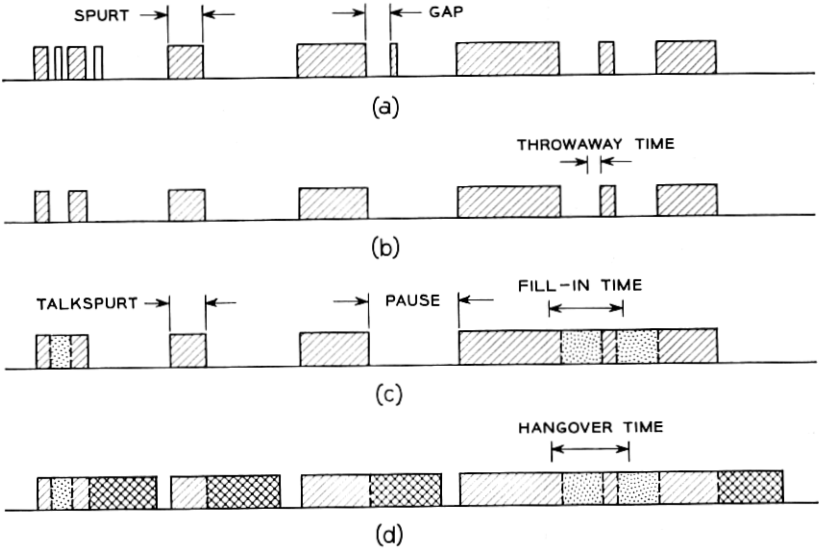

Fig. 2 — Process used by the computer to obtain talkspurts and pauses: (a) data from speech detector, (b) speech pattern after throwaway, (c) final pattern after fill-in, and (d) speech pattern if hangover were used instead of fill-in.

than to spurts and gaps. For convenience, this pattern will be labeled "talkspurts and pauses" even though such use of these terms is imprecise.

Notice that fill-in is not the same as hangover. With hangover, the beginning of *every* gap is filled in, regardless of the ultimate length of the gap. Pattern (d) would result if hangover were applied.

If the fill-in operation had preceded, rather than followed the throwaway operation, the resulting pattern would not have corresponded with pattern (c) of Fig. 2. The order of operation is, therefore, important. Now, the fill-in time, used to bridge stop consonants, should intuitively be much longer than the throwaway time, used to reduce noise effects. If the speech were filled in first, errors of fill-in time magnitude could occur as noise pulses were bridged to the adjacent speech. To avoid this problem, the throwaway operation is performed first.

2.4 *Discussion of Technique*

The process described above of using a speech detector with a fixed threshold in conjunction with a computer program with fixed throwaway and fill-in times was adopted here simply because it appeared to be a reasonable technique to use. No formal study was made in which several schemes were compared, as such an investigation did not seem warranted. The suitability of this technique for obtaining a talkspurt-pause pattern will be discussed later, after some of the resulting data have been examined.

III. PROCEDURE TO OBTAIN PARAMETER VALUES

3.1 *Outline of Method*

Values had to be determined for the speech detector threshold, throwaway time, and fill-in time. It was decided to make up a recording of "continuous speech," that is, speech which contained no noticeable pauses. This speech would be played many times into the detector, with the threshold set lower for each time. It was hoped that eventually the spurt and gap distributions would stabilize so that continued lowering of the threshold would contribute little to the detection of speech. This would establish a threshold.

To determine a throwaway time, the original (unedited) tapes would be examined for spurious noise. This noise would hopefully be of some maximum short duration and could be discarded with a throwaway time.

The fill-in time would finally be established by again processing the continuous speech, this time using the fixed threshold and throwaway time, and since the computer should ideally view the continuous speech as one long talkspurt, the fill-in time would be chosen equal to the longest observed gaps.

3.2 *Source of Speech — the Telephone Conversations*

Eight pairs of subjects were asked to hold telephone conversations over a special circuit. The circuit, illustrated in Fig. 3, was a four-wire circuit which had losses which simulated the effect of a long distance connection. Delay could be switched into one of the paths.*

The conversations were recorded on a two-channel tape recorder connected, as shown, to a level point representative of the zero TL point.† The two members of each pair of subjects were good friends

* The delay was included for use in a separate study. Some speech recorded on the delay circuit is analyzed here because, by doing so, twice as much continuous speech becomes available than if only the "standard" circuit were used. Also, note that although in this case a delay of 400 msec is inserted from B to A, the subjects cannot distinguish this condition from a 200-msec delay in each path, provided that they have no common time reference.

† The zero transmission level point is a point to which all level points in a toll system can be referred. It is analogous to citing altitude by referring to height above sea level.

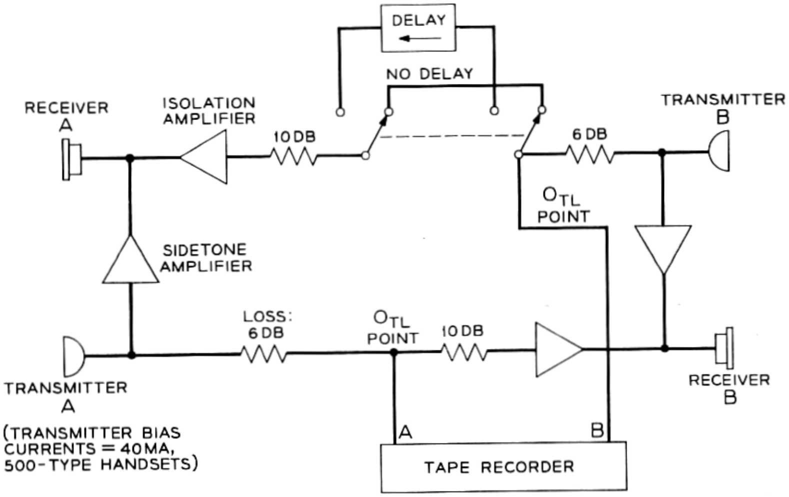

Fig. 3 — Circuit over which subjects talked.

who had many mutual interests. They received instructions to engage in active conversation about any topic they desired. They were told that their conversations would be recorded for use in "speech analysis work," but no other details were furnished. In the opinion of the experimenter, the subjects conversed readily and easily. There is no apparent evidence that their knowledge they were being recorded had a constraining effect on their conversations, but the reader should be aware that the experimental conditions differed in certain respects from those encountered in real life. The speech of the subjects was, however, certainly of a conversational nature and did not result from any formal preparation, as would occur with readings from printed matter.

The eight pairs of subjects consisted of four pairs of women and four of men. Each pair was allowed about a minute of warm-up time before the recorder was started. Then they talked for seven minutes over a standard circuit, followed by seven minutes over the 400-msec delay circuit. (One pair of girls spoke for only 3.3 minutes on the standard circuit and did not talk on the delay circuit.)

3.3 *Preparing Continuous Speech Tapes*

Each recording of each person's speech was edited by the author so that only segments of continuous speech, containing no pauses, were allowed to remain. This procedure was, of course, arbitrary and the edited tape could not possibly be exactly duplicated, even by the experimenter. There were, however, certain rules which the experimenter tried to follow:

(1) In the edited tape there were no obvious long conversational pauses.

(2) Pauses for breathing were excluded, unless the breaths were very short and embedded in a train of speech. This event occurred only a few times in all the samples.

(3) Brief comments ("of course," "really?") and other short sounds ("uh-huh," "um") were excluded, as those were too difficult to handle in the splicing operation.

(4) Similarly, stuttering and other meaningless speech was excluded.

(5) Very low-level speech, which was difficult to hear, was excluded. This was an infrequent occurrence in speech which was part of an attempt to converse. Most very low-level speech consisted of sighs and other such remarks and could be ruled out by (3) and (4), above.

On the average, seven minutes of conversation of one person was reduced to about 55 seconds of continuous speech. The average length of a speech segment uninterrupted by tape splices was 2.13 seconds. In all, there were 27.4 minutes of continuous speech, made up of 773 tape segments.

To show that the edited tape did indeed consist only of continuous speech, a panel of six people was invited to listen to all of the samples. The listeners were provided with tally sheets, and were requested to make a checkmark whenever they felt that the speaker paused, or that there was any break in his conversation. The panel members were asked to be severe — if they detected any hesitancy in the speaker's voice, or any gap which they felt could be used as an opportunity to interrupt, they were told to count this as a pause.

On the average each listener indicated 14.7 pauses in the 27.4 minutes of recorded speech. Computer analysis of this speech (discussed later) showed that thousands of gaps did occur in the continuous speech. Thus, a negligible number of the gaps in the edited tape were judged by the listeners to be conversational pauses.

Although the continuous speech tapes contain virtually no pauses, the tapes do not by any means contain *all* of the continuous speech which occurred in the conversation. The data of this study are therefore not representative of *all* the spurts and gaps that were present in the original speech, but rather of a large number of them. Gaps which occurred in the neighborhood of pauses were usually excluded, since the editing process tended to select speech away from the beginnings and endings of talkspurts.

IV. RESULTS USED TO SET PARAMETERS

4.1 *Threshold*

The speech detector threshold should ideally be chosen low enough to pick up almost all of the speech signal and high enough to avoid noise operation. To see how low a threshold was required for the first criterion, the effect of threshold variation on the continuous speech gap and spurt distributions was studied. If a point were reached where the data remain substantially unchanged as the threshold is lowered, then that threshold would be considered sufficiently low to cause operation on most of the speech.

Volume measurements were made of the 16 samples of continuous speech taken from the conversations made on the no delay circuit. A Daven volume level indicator, Model 1866, was used to obtain VU readings, a commonly accepted indication of speech volume.[4,5] The readings for the samples ranged from −29.9 VU (weakest talker) to −16.5 VU (loudest talker). The loudest, softest, and median speakers were selected from both the male and female talkers, thus providing six speech samples for analysis. Each of these samples was played through the speech detector four times, with the threshold set at −44, −40, −36, and −32 dbm,* for each time respectively. The lowest (most sensitive) threshold was chosen as −44 dbm, since greater sensitivity would have aggravated noise problems resulting from low-level tape hiss.

The four samples from each speaker were analyzed to see the effect of threshold variation on gap and spurt distributions. A fill-in time of 10 msec was provided to eliminate the introduction of gaps due to tape splices.† The throwaway time was zero.

The set of gap distributions of subject AD, volume = −24.1 VU, is typical of the subjects, and is shown in Fig. 4. The abscissa is the length of the gap and the ordinate is the per cent of gaps which are less than the abscissa. For example, when the threshold was set at −44 dbm, 50 per cent of the gaps were less than 28 msec long. There are no gaps equal to or less than 10 msec, because those that existed were filled in.

The curves are very much alike, except possibly the −32-dbm curve. At first glance, it appears that for reasonably low thresholds, the threshold value is not critical for measuring speech. A fill-in time of about 130 msec would bridge all the gaps for any of the chosen threshold values.

* That is, db re 1 milliwatt into 600 ohms, so that zero dbm equals 0.775 volts rms.

† A typical splice in tape traveling at 15 ips causes the level of a tone to drop about 6 db for about 8 msec. A fill-in time of 10 msec bridges any gap caused by this momentary level drop.

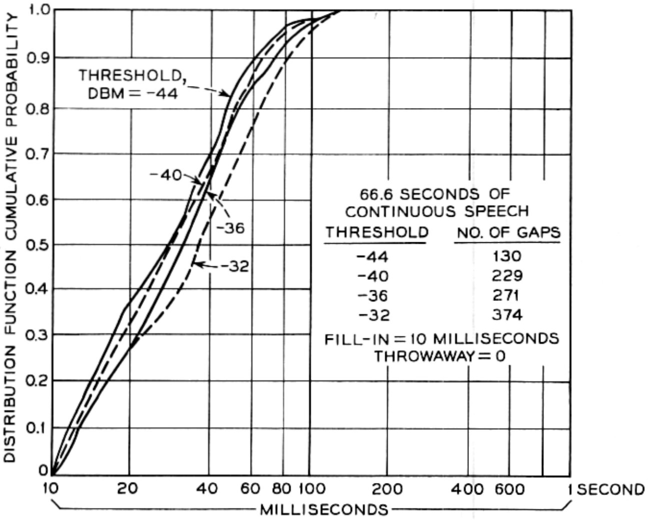

Fig. 4 — Gap distribution for subject AD.

But the gap distribution tells only part of the story. The total number of gaps almost triples as the threshold is raised from −44 to −32 dbm, going from 130 to 374 gaps in a continuous speech sample 66.6 seconds long.

The spurt distributions are plotted in Fig. 5. These change appreciably as the threshold is varied, showing shorter spurts with the higher thresholds. Also, whereas the low threshold indicates energy present 94 per cent of the time, the higher threshold yields only 75 per cent, as found from other results of the computer analysis.

Examination of the gap distributions of the six talkers whose speech was analyzed at four different thresholds indicates that gap distributions remain fairly stable for any threshold which is at least 12 db below the VU level.* As the threshold increases above this level, the gaps generally become noticeably longer. The 12-db value is only an estimate, which was arrived at by visual inspection of the data. It is a conservative estimate; a somewhat higher threshold would probably suffice for most of the speakers. However, until more data can be obtained for better analysis, 12 db will be used as a rule of thumb.

Although the gap distribution may stabilize as the threshold is lowered, the spurt distribution does not, and neither does the per cent time

* A similar analysis of the speech samples of all speakers also shows very little effect on gap distributions as the threshold is varied, as long as the threshold remains fairly low.

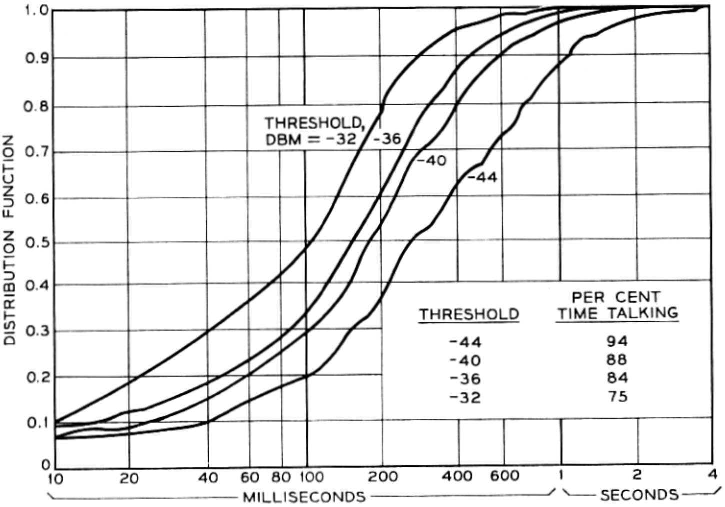

Fig. 5 — Spurt distribution for subject AD.

talking. It appears, then, that the threshold should be set as low as possible in order to pick up most of the speech.

The threshold cannot be set so low as to cause excessive noise operation. To estimate a lower bound, the original unedited tapes were played and the experimenter compared the detector output with his impressions of the sound on the tape. For thresholds much below −40 dbm there was noticeable detector triggering on breathing, moving the telephone handset, and other spurious noises. It was found that passing the speech through a high-pass filter helped considerably in reducing such noise operation. An SKL filter was used with a cutoff frequency of 170 cps. The filter has unity gain at 1000 cps.

Even with the filter in place, however, setting the threshold much below −40 db still seemed to result in some unwanted noise operation, and −40 dbm became the arbitrary choice for the fixed threshold. This is still a fairly sensitive value; echo suppressors, for example, commonly operate at −32-dbm sensitivity and adequately detect speech.* (Of course, the suppressors are equipped with hangover, which helps bridge subthreshold gaps. The comparison is still within reason, since fill-in will later be applied to our present data.)

The threshold is thus an unavoidable compromise between too much noise and too much lost speech. The value was chosen arbitrarily be-

* Echo suppressors are voice-operated devices which insert losses into transmission paths to alleviate the irritating effects of echoes.[2]

cause the data did not indicate an optimum setting, contrary to our initial expectations.

4.2 *Throwaway Time*

The throwaway operation is intended to remove spurts generated by short bursts of noise. The throwaway time should be long enough to eliminate most of the spurious noise operation, but short enough to avoid discarding much speech. Electrical disturbances from various sources, such as the tape recorder, constitute a type of spurious noise. From listening to the tapes, it appears that this noise consists of an occasional impulse of fairly low level. This would normally register as one pulse on the speech detector and show up as a 5-msec spurt. The impulse of noise is, of course, smeared out by the tape recording process, and if the impulse occurs near the end of a sampling interval, it may be wide enough to bridge two intervals, causing a 10-msec spurt to register. Visual observation of the detector output shows that every so often, one or two speech detector pulses appear when the talker appears to be silent. As a first guess, 10 msec seems reasonable as a throwaway time.

To see if a throwaway time could be determined from the data, the original unedited speech tapes were processed with a detector sensitivity of −40 dbm, using the high-pass filter. An arbitrary fill-in time of 150 msec was chosen to bridge over most of the short gaps occurring within talkspurts, leaving long pauses suitable for searching for spurious noises. Thus, if a 5-msec spurt is observed it is a relatively isolated event, for the fill-in operation has presumably bridged over such short spurts occurring within talkspurts. The eight conversations analyzed lasted a total of 51.56 minutes, representing 103.12 minutes of speech by the 16 subjects.

The results showed that of a total of 3375 spurts observed, 953, or 27 per cent, were 5 or 10 msec long. A 10-msec throwaway time would eliminate this excess of very short spurts. Should the throwaway time be greater? It turned out that an excess of spurts did not occur in other short spurt regions, such as 15 to 20 msec (61 spurts) 25 to 30 msec (45 spurts), or 35 to 40 msec (41 spurts). Since there is no reason to believe that spurts longer than 10 msec are due to circuit noise rather than the conversants, a 10-msec throwaway time was chosen.

Now, there are several types of noise which a 10-msec fill-in time could never eliminate. A great deal of the noise on the tapes was caused by the conversants; breathing, coughing, etc., generally operated the speech detector, sometimes for long durations (at least 100 msec). There seems to be little possibility of distinguishing talker-generated noises from speech solely on the basis of their on-off patterns. However, it is reasonable to include such sounds in the speech analysis because:

(1) These noises, although they carry little speech information, are generally audible to the listener and are a legitimate part of the conversational exchange.

(2) If there are any voice-operated devices on a communications circuit, these will be influenced just as much by coughing, etc., as by speech. If the results of this study are ever to be used to predict the behavior of these devices, then the data must include all sounds, speech or otherwise.

4.3 *Fill-In Time*

The fill-in time should be chosen just long enough to bridge the longest gaps in the continuous speech, having applied the throwaway operation. The distribution of the long gaps in the continuous speech gap distribution is shown in Table I.

From the data of Table I, there appears to be no obvious setting for the fill-in time which should be used to bridge the gaps. A fill-in time in excess of 250 msec would be required to bridge *all* the gaps, but it seems unreasonable to pick a value based on the one or two longest gaps. Because the distribution trails off smoothly without an obvious breakpoint, we arbitrarily select 200 msec as a fill-in time. There are two justifications for this choice:

(1) 200 msec is an easily remembered number. By its very nature, it appears to be rounded off, and therefore an approximation, which of course it is.

(2) When the panel of listeners monitored the continuous speech tapes, each member detected, on the average, 15 pauses in the speech.

TABLE I — DISTRIBUTION OF GAPS GREATER THAN 150 MSEC FOR ALL TALKERS*

Length, msec	Number	Per Cent
155,160	20	0.44
165,170	12	0.26
175,180	8	0.18
185,190	4	0.09
195,200	8	0.18
205–225	5	0.11
230–250	3	0.07
>250	2	0.04

Total number of gaps = 4537

* Continuous speech, 16 talkers, both circuit conditions. Throwaway = 10 msec, fill-in = 10 msec, threshold = −40 dbm; 170-cps high-pass filter.

If the 15 longest gaps are thrown out of the Table I distribution, the longest remaining gap is, by happy coincidence, 200 msec long. (This does not imply that the 15 longest gaps are those particular ones which the subjects called pauses.)

4.4 *Additional Gap and Spurt Data*

Although the continuous speech data already reported were sufficient for purposes of setting threshold and fill-in, additional data were obtained which may be of interest to some researchers. Fig. 6 is a plot of the gap distribution for the 27.4 minutes of continuous speech for all 16 talkers over both circuit conditions. The speech detector threshold was set at −43 dbm. No filter was used prior to the detector.

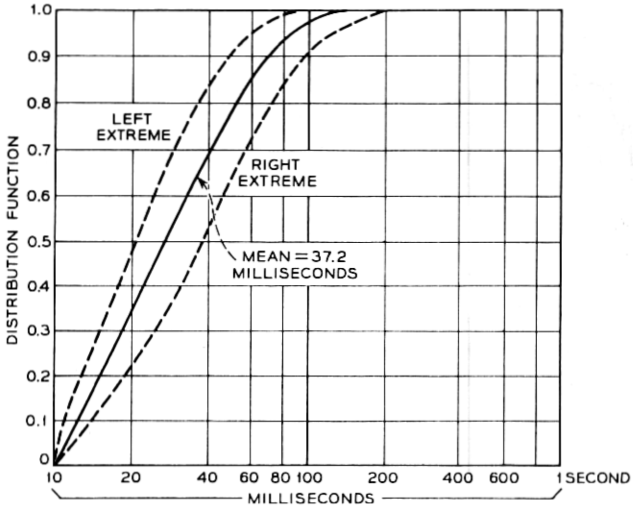

Fig. 6 — Continuous speech gap distribution: solid curve is the distribution for the entire speech sample for all talkers; dashed curves are composites of the boundaries of the curves for individual speech samples. Conditions: 16 speakers; 27.4 minutes of continuous speech; both delay and no-delay circuit conditions; speech detector threshold −43 dbm with no filter; throwaway time 0; fill-in 10 msec; 3754 gaps.

To determine the variations in the gap distribution among the different talkers, the individual gap distributions for all 30 samples* were plotted on one graph. The left boundary of this graph is shown as the "left extreme" curve of Fig. 6. This curve is not necessarily the curve for a single sample but is a composite of all curves which happen to fall on the boundary. The same applies to the "right extreme" curve.

* Sixteen talkers on the standard circuit, fourteen on 400-msec delay.

Visual inspection of the curves for individual talkers shows no apparent over-all differences between male and female talkers.

Fig. 7 shows the continuous speech spurt distribution. The spurt distribution might be considered representative of syllabic bursts, but care must be taken in drawing conclusions from these data since the graph is strongly influenced by the threshold setting (as shown in Fig. 5) and also because of a possible tape splice effect, described as follows.

A probabilistic analysis of the influence of tape splices on the speech reveals that of the more than 770 splices which were present, about 130 of them involved a gap. Thus, 130 of the 3754 gaps (or 3.5 per cent) were affected by splices, probably only slightly influencing the gap distribution of Fig. 6.

It turns out, however, that almost all of the splices had a spurt on one side or both (virtually all of the 130 splices involving a gap had a spurt on the other side). This means that over 20 per cent of the total number of spurts were affected in some way by tape splices.

Because of the influence of artifacts on the spurts, further analysis of the spurt distribution (such as obtaining the range among subjects) was not carried out.

V. TALKSPURT AND PAUSE DISTRIBUTIONS FOR CONVERSATIONAL SPEECH

5.1 *Data from Conversations*

This section is an illustration of the results obtained from analysis of conversational speech. Fig. 8 is a plot of the distribution of talkspurts and pauses for all 16 speakers engaging in eight conversations over the standard circuit. The conversations lasted almost 52 minutes, yielding about 103 minutes of conversation data for all subjects. Some of the more interesting statistical measures of these data are shown in Table II.

The results shown here are included only to illustrate the speech measuring technique developed in this study. The conversations were artificially induced in a test-room atmosphere in which the subjects knew they were being recorded. There is no assurance that the statistical measures shown in Table II and in Fig. 8 are representative of those which would be obtained on real telephone calls.

5.2 *A Comparison with Other Studies*

Many other studies have been concerned with measuring some of the statistical properties of speech patterns. Three of these studies are selected for comparison with our results.

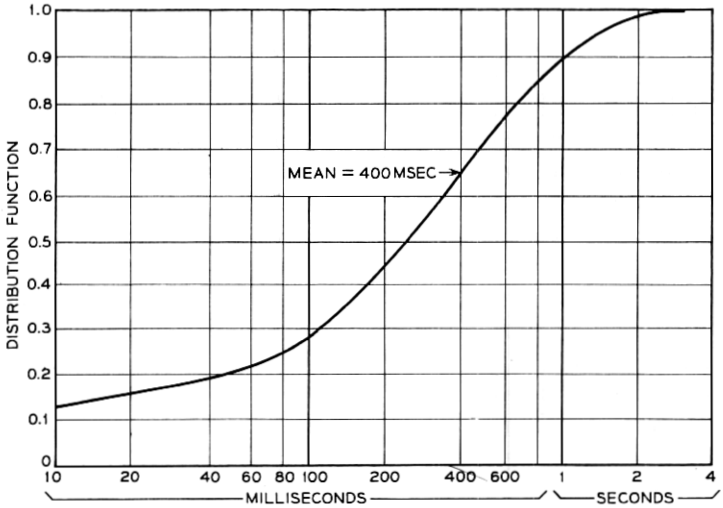

Fig. 7 — Continuous speech spurt distribution. Conditions: 16 speakers; 27.4 minutes of continuous speech; both delay and no-delay circuit conditions; threshold −43 dbm; throwaway time 0; fill-in 10 msec; 3754 spurts.

Norwine and Murphy[6] observed "talkspurts" and "pauses" in oscillograph recordings of telephone conversations made on a special circuit between New York and Chicago. The circuit had a round-trip delay of almost 600 msec and was equipped with echo suppressors. Detailed observations were made from about 4400 feet of graph paper, representing 51 calls with a total duration exceeding 13,000 seconds. The Norwine and Murphy definition of talkspurt is significantly different from ours and is included here.

"A talkspurt is speech by one party, *including his pauses*, which is preceded and followed . . . by speech from the other party perceptible to the one producing the talkspurt." (Italics mine.)

Since Norwine and Murphy include pauses in their talkspurts, their talkspurts are much longer than ours, and a direct comparison of their distributions with ours is inappropriate. It is possible, however, to apply a correction on one or two statistics and make a fair comparison. Considering the mean talkspurt length, Norwine and Murphy indicate a value of 4.3 seconds for 2845 talkspurts. These talkspurts include, however, 2811 pauses with a mean of 0.73 seconds. One may then calculate that there were about 10,200 seconds of speech composed of 2845 + 2811 = 5656 "shortened talkspurts." (For each pause inserted, a new talkspurt is created.) The new average talkspurt length is 1.8 seconds, which compares with our average of 1.34 seconds.

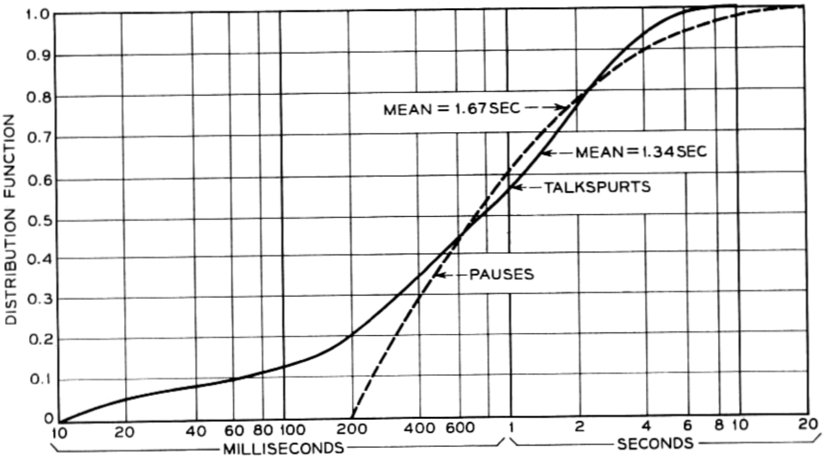

Fig. 8 — Talkspurt and pause distributions for all speakers combined; there were 51.56 minutes of conversational speech: therefore 16 speakers talked for 103.12 minutes. Conditions: no delay; threshold −40 dbm; throwaway 10 msec; fill-in 200 msec; 170-cps high-pass filter in detector circuit; 2042 talkspurts, 2054 pauses.

More recently, measurements were made on calls placed on the Atlantic cable via TASI.* "Spurts" and "gaps" were determined by the TASI speech detector operation, and results were tabulated for about 600 commercial calls. The TASI speech detector has an operate (pickup) time of 5 msec and a deferred hangover of 240 msec. A deferred hangover means that the full hangover value is applied only for longer spurts (60 msec or greater) but for very short spurts the hangover is shorter (minimum of 25 msec). The sensitivity is −40 dbm re 0TL.

Although the TASI detector uses hangover rather than fill-in and a delayed operate time rather than throwaway, the values used for the measuring parameters are very similar to ours, especially since the sensitivity is the same. The TASI data indicate an activity period (fraction of time the TASI detector is operated) of 48 per cent averaged over all calls. This compares with 44.3 per cent obtained in our study, as shown in Table II. In addition, the mean TASI "spurt" length is 1.3 seconds, compared with our value of 1.34 seconds.

A distribution of TASI talkspurts is shown in Fig. 9 and was taken from an unpublished report by H. Miedema,[8] printed here with the author's permission. Fig. 9(a) shows both the talkspurt distribution

* TASI is essentially a bank of voice-operated switches which connects a subscriber to a channel only when he is actually talking. Thus 72 people may talk over 36 channels.[7]

Table II — Some Statistical Measures of Eight Conversations

Total conversation time	51.56 min
Per cent time both speakers are silent	18.2%
Per cent time circuit in use by either or both speakers	81.8%
Per cent time double talking	7.2%
Per cent time the average conversant talks (obtained by dividing the total time all speakers A talk plus the total time of all B talk by 103.12 minutes)	44.3%
Number of talkspurts (all 16 speakers)	2042
Median talkspurt length	0.77 sec
Mean talkspurt length	1.34 sec
Number of pauses	2054*
Median pause	0.72 sec
Mean pause	1.67 sec
Circuit conditions: 500-type sets, four-wire connections, 16-db loss between speakers, speech recorded at 0TL point, 6 db from each transmitter.	
Measurement parameters: speech detector threshold set at −40 dbm re 0TL, throwaway time = 10 msec, fill-in time = 200 msec, 170-cps high-pass filter.	

* There are more pauses than talkspurts because several of the eight conversations began and ended with a pause for each of the speakers.

obtained in our study and the distribution of TASI speech detector spurts for U. S. talkers. Also included is a "corrected" curve, in which the 10-msec pickup time is added to each talkspurt and the hangover time is subtracted. Since the hangover time is variable, the correction depends on the talkspurt length. The net effect is to shift the curve left 15 msec for short spurts (25-msec hangover minus 10-msec pickup) and 230 msec for long spurts. This widens the discrepancy between our results and the TASI data, making the TASI talkspurts appear shorter. Some of the short TASI spurts, however, may have been due to line noise on the trunk. The magnitude of this effect cannot be determined, since there are at present no available data on TASI noise operation.

Fig. 9(b) shows the pause distributions. Again, the original pause curve must be corrected for hangover. We will assume that a hangover of 240 msec preceded each gap (and it did for all gaps following the 90 per cent of the talkspurts exceeding 60 msec). The 10-msec pickup time will reduce the correction to 230 msec. The resulting curve is in this case very close to ours.

Finally, J. F. Agnello of Ohio State University made an analysis of the gaps which occur in spoken text, and he studied the effect on the distributions of varied text (prose, poetry, single sentences).[9] He differentiated between *intraphase pauses* (gaps) and *interphase pauses* (pauses) and had the speakers listen, as a group, to their own speech, indicating on the printed text whenever they detected a pause. A "pause timer"

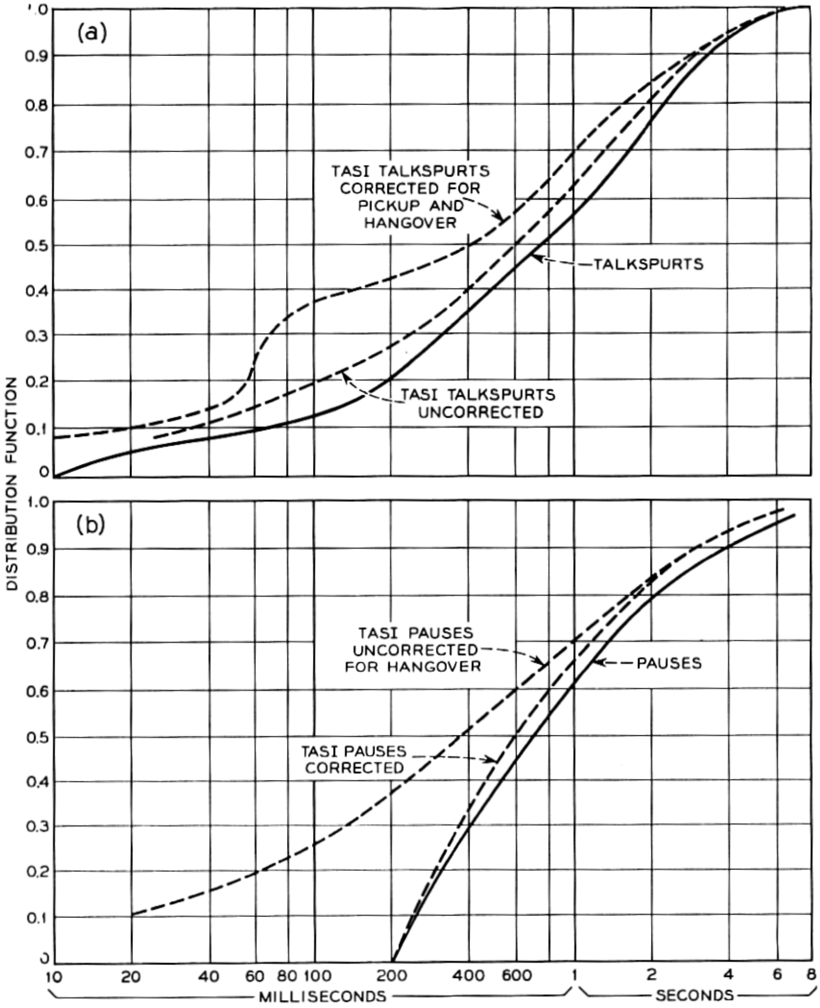

Fig. 9 — A comparison of talkspurt and pause data with TASI data: (a) talkspurts, (b) pauses.

detected and classified by duration 887 pauses exceeding 150 msec in about 40 minutes of speech, and the listeners indicated 710 pauses. Agnello concludes that "the minimal detected pause was estimated . . . to be 190 msec." He also concludes on other grounds that the longest *intraphase pause* was 200 msec long. These results are in excellent agreement with our choice of 200 msec as a fill-in time.

VI. DISCUSSION

6.1 *Suitability of Technique*

The question which now presents itself is whether or not the method developed here for speech detecting is a good technique. In answering this question, we are immediately faced with another one, that of defining what is meant by a "good technique." The answer to the latter question is, of course, a function of the intended purpose of the detector.

It was originally hoped that the computer would yield speech patterns such that segments of talkspurts would be bridged together and spurious activity would be ignored. One way of evaluating this criterion is to look at the curves of Fig. 8 and see if these results are consistent with our goals.

On one hand, the detection technique seems to be a meaningful indication of perceived speech because of the many points of correspondence between the Fig. 8 data and that of other studies. For example, although our data do not exactly duplicate the TASI detector operation, there are many similarities, and people are able to converse fluently over TASI.

On the other hand, the talkspurt data of Fig. 8 seem unreasonable because of the presence of numerous short spurts. Note that 10 per cent are less than 60 msec long. It is unlikely that an utterance of such short duration would be considered by a listener to be a "talkspurt." What, then, are these sounds?

On listening to the tapes, one observes that there *are* occasional short utterances, such as those produced when a person begins to talk but suddenly realizes that he has been interrupted. These are sometimes as short as a few pitch periods. Short sounds are often produced by parting the lips quickly, or by clicking the tongue against the roof of the mouth. Another source of short sounds is the low-level speech which triggers the speech detector only intermittently. An informal experiment was carried out in which a speech detector was hooked up to cut out any speech exceeding a threshold, leaving only the subthreshold speech audible. (The detector consisted of a relay with 5-msec pickup and hangover times.) The three people who listened to the speech noticed that occasionally whole phrases would get through. This was an infrequent event, however, for sensitivities as low as −40 dbm.

The author is convinced that these short sounds are indeed produced by the talkers and are not extraneous noises. They may be fragmentary parts of talkspurts, or clicks, or whatever you will, but their sources are indistinguishable to the computer. Within the framework of the detection technique adopted here there are two things which can be done with them:

(1) Throw them out with some arbitrary throwaway time. This introduces the possibility of discarding real speech, as well as the difficulty of establishing a suitable throwaway time.

(2) Keep them as is, realizing that there is some doubt that they represent meaningful speech, and hope that they will not hamper further work with speech analysis. The author has chosen to adopt this second approach.

6.2 *Relationship to the Acoustics of Speech Production*

The curves presented in Fig. 8 are obtained from gross statistics of over 100 minutes of speech. No effort has been made to refine and classify the data into subheadings such as "plosives" or "glottal consonants," etc. Accordingly, it is not possible, given only the curves shown here, to explain the shape of the curves in terms of the acoustics of speech production. This could easily be a study unto itself, and, indeed, studies of this nature presently constitute an important phase of speech research by acousticians and phoneticians who are expert in this field.

If anyone wishes to examine our results in the light of data in the speech literature, he must note that our speakers were not trained and made no effort to talk clearly and precisely. In addition, they spoke over telephones, not high-quality microphones. This procedure is quite different from that used in most speech studies, in which the speech is carefully manicured to approach an "ideal" sound.

VII. CONCLUSION

This study has shown that even small changes in the measuring technique can produce noticeable effects on the results. Data on speech dynamics must therefore include a detailed description of the technique used to measure the speech. Without this, the results must be regarded as unrepeatable.

Many unforeseen difficulties arose with the proposed technique of using a speech detector with a fixed threshold followed by a computer program with simple throwaway and fill-in operations. The setting of the detector threshold is a compromise between excessive noise and too little speech operation, and errors of both kinds must be expected. Although a fill-in time of 200 msec seems fairly well established, the throwaway time of 10 msec is not adequate to remove many short spurts, some of which may represent legitimate talkspurts, while others do not. Indeed, it may turn out that a simple automatic speech detecting technique involving fixed parameters is inadequate for some purposes, and a considerably more sophisticated method must be employed.

What therefore originally began as a seemingly straightforward attempt to build a speech detector has instead exposed a problem of far more difficulty than was first imagined. The data obtained here have hopefully shed light on some aspects of the problem, but further study is required before a completely satisfactory solution is obtained.

VIII. ACKNOWLEDGMENTS

I wish to thank Mrs. Nancy Shrimpton for writing the computer program used to process the speech and, in addition, for assisting in the data analysis. I would also like to thank Miss Kathryn McAdoo for making the VU measurements on the continuous speech samples.

REFERENCES

1. Emling, J. W., and Mitchell, D., The Effects of Time Delay and Echoes on Telephone Conversations, B.S.T.J., **42,** Nov., 1963, p. 2869.
2. Brady, P. T., and Helder, G. K., Echo Suppressor Design in Telephone Communications, B.S.T.J., **42,** Nov., 1963, p. 2893.
3. Riesz, R. R., and Klemmer, E. T., Subjective Evaluation of Delay and Echo Suppressors in Telephone Communications, B.S.T.J., **43,** Nov., 1963, p. 2919.
4. McAdoo, K. L., Speech Volumes on Bell System Message Circuits — 1960 Survey, B.S.T.J., **42,** Sept., 1963, p. 1999.
5. Beranek, L. L., *Acoustic Measurements*, Wiley, New York, 1949, p. 504.
6. Norwine, A. C., and Murphy, O. J., Characteristic Time Intervals in Telephone Conversation, B.S.T.J., **17,** April, 1938, p. 281.
7. Miedema, H., and Schachtman, M. G., TASI Quality — Effect of Speech Detectors and Interpolation, B.S.T.J., **41,** July, 1962, p. 1455.
8. Miedema, H., unpublished report presenting TASI speech detector data, 1960.
9. Agnello, J. G., A Study of Intra- and Inter-Phrasal Pauses and Their Relationship to the Rate of Speech, Ohio State University Ph.D. Thesis, 1963.