INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.59

(03/93)

TELEPHONE TRANSMISSION QUALITY
OBJECTIVE MEASURING APPARATUS

ARTIFICIAL CONVERSATIONAL SPEECH

ITU-T Recommendation P.59

(Previously "CCITT Recommendation")

FOREWORD

The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of the International Telecommunication Union. The ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Conference (WTSC), which meets every four years, established the topics for study by the ITU-T Study Groups which, in their turn, produce Recommendations on these topics.

ITU-T Recommendation P.59 was prepared by the ITU-T Study Group XII (1988-1993) and was approved by the WTSC (Helsinki, March 1-12, 1993).

____

NOTES

1 As a consequence of a reform process within the International Telecommunication Union (ITU), the CCITT ceased to exist as of 28 February 1993. In its place, the ITU Telecommunication Standardization Sector (ITU-T) was created as of 1 March 1993. Similarly, in this reform process, the CCIR and the IFRB have been replaced by the Radiocommunication Sector.

In order not to delay publication of this Recommendation, no change has been made in the text to references containing the acronyms “CCITT, CCIR or IFRB” or their associated entities such as Plenary Assembly, Secretariat, etc. Future editions of this Recommendation will contain the proper terminology related to the new ITU structure.

2 In this Recommendation, the expression “Administration” is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

CONTENTS

Recommendation P.59

ARTIFICIAL CONVERSATIONAL SPEECH

(Helsinki, 1993)

1 Introduction

The signal described here reproduces the on-off temporal characteristics of human conversational speech for characterizing speech processing systems which have speech detectors, such as loudspeaker telephones, echo control devices, digital circuit multiplication equipment (DCME), packet systems, and asynchronous transfer mode (ATM) systems. This signal reflects parameters of human conversation such as the length of the talk-spurt, pause, double talk, and mutual silence. The following chapters describe these characteristics and a method of generating artificial conversational speech.

NOTES

1 The artificial voice described in Recommendation P.50 is a single-channel signal without pauses and is used for objective measurements of speech processing systems and devices in which the conversational environment is not relevant, such as speech codecs.

2 The artificial conversational speech described in this Recommendation generates the artificial voice described in Recommendation P.50 during talk-spurts.

2 Characteristics of human conversational speech

The durations and rates of talk-spurt and pause vary according to the measurement conditions. The following specifies two values for each parameter in conversational speech. One is based on measurement of speech without hangover time, while the other is from that with hangover time.

2.1 Characteristics measured without hangover time

The characteristics described below were derived from Reference [1].

1) *Talk-spurt characteristics*

The probability density function (pdf) of talk-spurt duration is modelled by two weighted geometric pdfs:

$$f_1(k) = C_1(1 - U_1)\ U_1^{k-1} + C_2(\ 1-U_2\ )\ U_2^{k-1}, k = 1, 2, 3, \ldots$$

where

$$C_1 = 0.60278 \quad U_1 = 0.92446$$
$$C_2 = 0.39817 \quad U_2 = 0.98916$$

Every increment of the variable k is equal to 5 ms. The cumulative distribution function of talk-spurt durations is shown in diagram a) of Figure 1. The average talk-spurt duration is 227 ms.

2) *Pause characteristics*

The pdf of pause duration is also modelled by two weighted geometric pdfs:

$$f_p(k) = D_1(1 - W_1)\ W_1^{k-1} + D_2(1 - W_2)\ W_2^{k-1}, k = 1, 2, 3, \ldots$$

where

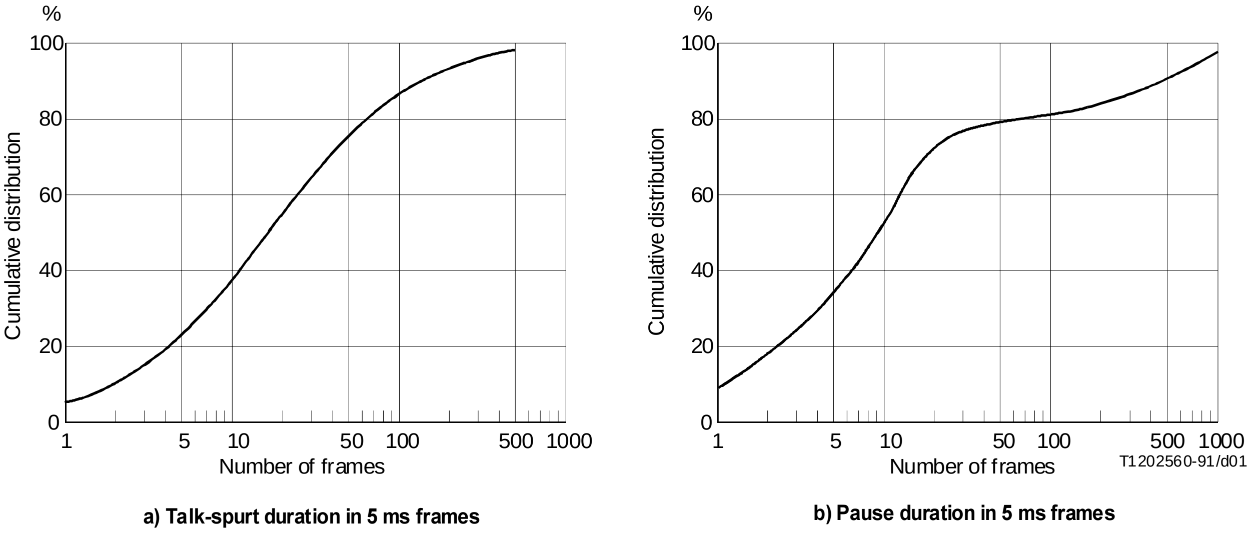
$$D_1 = 0.76693 \quad W_1 = 0.89700$$
$$D_2 = 0.23307 \quad W_2 = 0.99791$$

The cumulative distribution function of pause duration is shown in diagram b) of Figure 1.

3) *Activity factor*

The average pause duration of 596 ms, combined with the 227 ms average talk-spurt duration, yields a long-term speech activity factor of 27.6 per cent.

NOTE – This value is measured by a meter without hangover. However, if a meter conforming to Recommendation P.56 is used, a higher activity factor is to be expected (see Table 1).



a) Talk-spurt duration in 5 ms frames

b) Pause duration in 5 ms frames

FIGURE 1/P.59

Cumulative distribution of talk-spurt and pause durations (without hangover time)

2.2 Characteristics measured with hangover time

Table 1 lists the values of feature parameters in human conversational speech. These values were obtained by averaging the values reported in References [1]-[4].

TABLE 1/P.59

Temporal parameters in conversational speech (average for English, Italian, and Japanese)

Parameter	Duration (s)	Rate (%)
Talk-spurt	1.004	38.53
Pause	1.587	61.47
Double talk	0.228	6.59
Mutual silence	0.508	22.48

The cumulative distribution function of talk-spurt duration is approximated by an exponential function and that of pause durations is approximated by a constant-plus-exponential. That is, for talk-spurt:

$$\text{Pts}(t) = 1 - \exp(-\text{Ats} \cdot t)$$

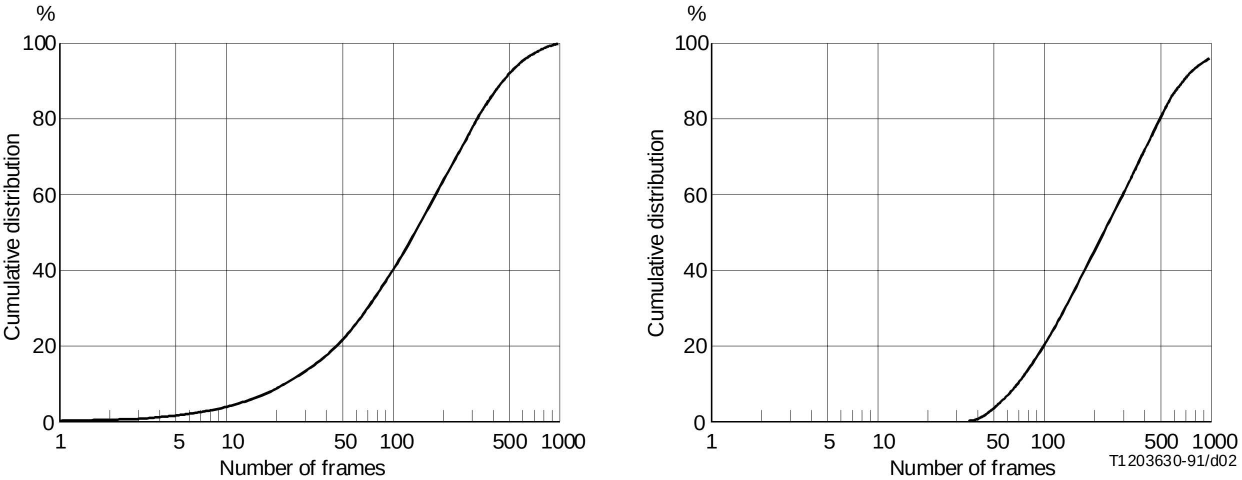
$\text{Ats} = 1/\bar{\ },\text{Tts},$ $\bar{\ },\text{Tts}$: average talk-spurt duration,

and for pause,

$$\text{Pps}(t) = \begin{cases} 0 & \text{for } 0 \le t \le 0.2 \\ 1 - \exp[-\text{Aps}(t - 0.2)] & \text{for } t > 0.2 \end{cases}$$

$\text{Aps} = 1/(\bar{\text{T}}\text{ps} - 0.2)$ $\bar{\text{T}}\text{ps}$: average pause duration.

Both characteristics are shown in Figure 2.



a) Talk-spurt duration in 5 ms frames

b) Pause duration in 5 ms frames

FIGURE 2/P.59

Cumulative distribution of talk-spurt and pause durations
(with hangover time)

3 Method of generating artificial conversational speech

Talk-spurts and pauses are generated according to the state transition model shown in Figure 3, in which P_1, P_2, and P_3 denote transition probabilities expressed in per cent. The artificial voice described in Recommendation P.50 is generated during a talk-spurt.

Tst (single talk duration), Tdt (double talk duration), and Tms (mutual silence duration) vary according to the following equations. The times in these equations are expressed in seconds.

$$\text{Tst} = -0.854 \ln(1 - x_1)$$

$$\text{Tdt} = -0.226 \ln(1 - x_2)$$

$$\text{Tms} = -0.456 \ln(1 - x_3)$$

$0 < x_1, x_2, x_3 < 1$: Random variables with uniform distribution.

If the pause duration is less than 200 ms, the model chooses either the single talk or mutual silence state with probabilities of 50% until the pause duration exceeds 200 ms. The values of P_1, P_2, and P_3 are 40, 50, and 50, respectively. The total duration of artificial conversational speech must be at least 10 minutes to comply with the characteristics specified in 2.2.

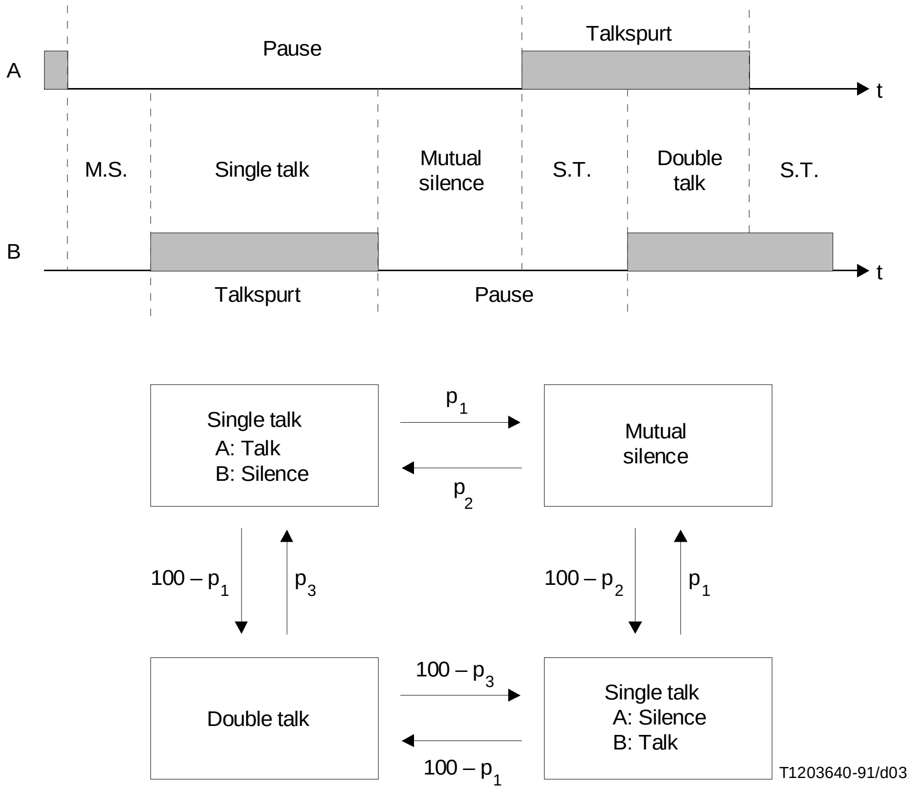


FIGURE 3/P.59

State transition model for conversation

References

[1] LEE (H.H.), UN (C.K.): A study of on-off characteristics of conversational speech, *IEEE Trans. on Comm.*, Volume COM-34, No. 6, pp. 630-637, June 1986.

[2] BRADY (P.T.): A statistical analysis of on-off patterns in 16 conversations, *BSTJ*, pp. 73-91, January 1968.

[3] CCITT Contribution COM XII-20, *On-off characteristics of conversational speech* (CSELT), Study Period 1989-1992.

[4] CCITT Contribution Delayed D.42 (WP XII/1), *Collecting procedure for on-off characteristics of conversational speech in telecommunication* (NTT), Study Period 1989-1992.

[5] CCITT Contribution Delayed COM-64 (WP XII/1), *Generation of artificial voice with pauses* (NTT), Study Period 1989-1992.