INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.501

(08/96)

SERIES P: TELEPHONE TRANSMISSION QUALITY

Objective measuring apparatus

Test signals for use in telephonometry

ITU-T Recommendation P.501

(Previously CCITT Recommendation)

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series P.10
Subscribers' lines and sets	Series P.30
	P.300
Transmission standards	Series P.40
Objective measuring apparatus	**Series P.50**
	P.500
Objective electro-acoustical measurements	Series P.60
Measurements related to speech loudness	Series P.70
Methods for objective and subjective assessment of quality	Series P.80
	P.800
Audiovisual quality in multimedia services	Series P.900

For further details, please refer to ITU-T List of Recommendations.

ITU-T RECOMMENDATION P.501

TEST SIGNALS FOR USE IN TELEPHONOMETRY

Summary

This Recommendation describes test signals which are applicable for several purposes in telephonometry. This Recommendation gives a wide variety of test signals starting with low complexity test signals up to test signals with a high degree of complexity incorporating a lot of typical parameters of speech. Besides technical signals such as sine waves or noise, more speech-like signals are described.

This Recommendation describes the principles of the signal construction for each type of test signal. Characteristic properties such as power density spectra, probability density functions or shaping filter responses are shown.

This Recommendation gives an overview about the typical application of the test signals described. This overview is a guideline giving general application rules. The detailed description of the application however should be found in the individual Recommendations describing the measurement procedures for a certain application.

In order to avoid problems in creating the test signals described, this Recommendation contains a CD-ROM where all test signals are stored.

Source

ITU-T Recommendation P.501 was prepared by ITU-T Study Group 12 (1993-1996) and was approved under the WTSC Resolution No. 1 procedure on the 30th of August 1996.

FOREWORD

ITU (International Telecommunication Union) is the United Nations Specialized Agency in the field of telecommunications. The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of the ITU. 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, establishes the topics for study by the ITU-T Study Groups which, in their turn, produce Recommendations on these topics.

The approval of Recommendations by the Members of the ITU-T is covered by the procedure laid down in WTSC Resolution No. 1 (Helsinki, March 1-12, 1993).

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

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

CONTENTS

Recommendation P.501

TEST SIGNALS FOR USE IN TELEPHONOMETRY

(Geneva, 1996)

1 Scope

This Recommendation describes test signals which are applicable for several purposes in telephonometry. A wide variety of test signals is given starting with low complexity test signals up to test signals with a high degree of complexity incorporating a lot of typical parameters of speech. Besides technical signals such as sine waves or noise, more speech-like signals are described.

The overview about the typical applications of the test signals described is a guideline giving general application rules. The detailed description of the application however should be found in individual Recommendations describing the measurement procedures for certain applications.

In order to avoid problems in creating the test signals described, this Recommendation contains a CD-ROM where all test signals are stored.

2 References

The following Recommendations, and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; all users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published.

– ITU-T Recommendation P.50 (1993), *Artificial voices*.

– ITU-T Recommendation P.59 (1993), *Artificial conversational speech*.

– ITU-T Recommendation P.340 (1996), *Transmission characteristics of hands-free telephones*.

3 Definitions and abbreviations

For the purposes of this Recommendation, the following definitions and abbreviations are used.

3.1 crest factor: Peak-to-RMS ratio of a signal.

3.2 composite source signal (CSS): Signal composed in time by various signal elements.

3.3 fast fourier transform (FFT)

3.4 modulation transfer function (MTF): Modulation signal, derived from the envelope of a test signal.

3.5 markov speech model process (MSMP): Speech simulating signal using trainable Markov-chains for the generation of a speech-like signal, taking into account the generation process of real speech.

3.6 pseudo noise sequence (PN-sequence): Pseudo-random noise with defined frequency-content, derived by inverse Fourier transformation of a predefined frequency spectrum.

3.7 simulated speech generator (SSG): Signal offering speech-like properties, constructed taking into account the generation process of real speech.

3.8 speech transmission index (STI): Index indicating the speech intelligibility especially in reverberant condition, derived from measuring the MTF.

4 Overview over test signals and typical applications

Linear, time invariant systems (e.g. handset telephones)									
	Sine wave (Note 1)	Noise (Note 2)	CSS (Note 3)	Probe tone	SSG	Rec. P.50	Rec. P.59	MSMP	Speech
Loudness ratings	x	x	x	x	x	x	x	x	x
Frequency responses	x	x	x	x	x	x	x	x	x
Listener sidetone/talker sidetone	x	x	x	x	x	x	x	x	x
Harmonic distortion	x		x						
Distortion		x	x						
Out-of-band signals	x		x						
Level measurements	x	x	x	x	x	x	x	x	x
Delay measurements	x	x	x		(x)	(x)		(x)	(x)
Echo measurements		x	x	x	x	x	x	x	x

x Applicable

(x) Applicable with some caution

NOTES

1 Including modulated sine wave and Fourier spectra.

2 Including pink, white and switched noise.

3 Including various combinations of voiced sound and measurement signals (PN-sequence, sine, etc.).

Non-linear and/or time variant systems									
	Sine wave (Note 1)	Noise (Note 2)	CSS (Note 3)	Probe tone	SSG	Rec. P.50	Rec. P.59	MSMP	Speech
Loudness ratings (long-term values)		(x)	(x)		(x)	(x)	(x)	(x)	(x)
Loudness ratings (short-term values)		(x)	x						
Frequency responses (long-term values)		(x)	(x)		(x)	(x)	(x)	(x)	(x)
Frequency responses (short-term values)		(x)	x						
Listener sidetone/talker sidetone (long-term values)		(x)	(x)	(x)	(x)	(x)	(x)	(x)	(x)
Harmonic distortion			(x)						
Distortion		(x)	(x)		(x)	(x)	(x)	(x)	(x)
Out-of-band signals			(x)						
Level measurements		(x)	(x)		(x)	(x)	(x)	(x)	(x)
Delay measurements		(x)	(x)						
Echo measurements		(x)	(x)		(x)	(x)	(x)	(x)	(x)

x Applicable

(x) Applicable with some caution

NOTES

1 Including modulated sine wave and Fourier spectra.

2 Including pink, white and switched noise as well as MTF.

3 Including various combinations of voiced sound and measurement signals (PN-sequence, sine, etc.).

4 Long-term values correspond to steady state behaviour of systems.
Short-term values correspond to dynamic behaviour of systems.

Hands-free telephones									
	Sine wave (Note 1)	Noise (Note 2)	CSS (Note 3)	Probe tone	SSG	Rec. P.50	Rec. P.59	MSMP	Speech
Loudness ratings (long-term values)		x	x	(x)	x	x	x	x	x
Loudness ratings (short-term values)		(x)	x						
Frequency responses (long-term values)		x	x	(x)	x	x	x	x	x
Frequency responses (short-term values)		(x)	x						
Harmonic distortion			x						
Distortion		(x)	x						
Out-of-band signals			x						
Level measurements		x	x	(x)	x	x	x	x	x
Delay measurements		x	x						
Switching characteristics	(x)	(x)	x	(x)	(x)	(x)	(x)	(x)	(x)
Reverberation measurements		x	x						
Echo path characteristics		(x)	(x)		(x)	(x)		(x)	(x)

x Applicable

(x) Applicable with some caution

NOTES

1 Including modulated sine wave and Fourier spectra.

2 Including pink, white and switched noise as well as modulated noise (MTF).

3 Including various combinations of voiced sound and measurement signals (PN-sequence, sine, etc.).

4 Long-term values correspond to steady state behaviour of systems.
Short-term values correspond to dynamic behaviour of systems.

Echo-cancellers									
	Sine wave (Note 1)	Noise (Note 2)	CSS (Note 3)	Probe tone	SSG	Rec. P.50	Rec. P.59	MSMP	Speech
Level measurements		(x)	x	(x)	x	x	x	x	x
Delay measurements		(x)	x						
Switching characteristics	(x)	(x)	x		(x)	(x)	(x)	(x)	(x)
Background noise performance		(x)	x	(x)	x	x	x	x	x
Echo loss, TCL		(x)	(x)	(x)	x	x	x	x	x
Double talk performance		(x)	x		(x)	(x)	(x)	(x)	(x)

x Applicable

(x) Applicable with some caution

NOTES

1 Including modulated sine wave and Fourier spectra.

2 Including pink, white and switched noise as well as modulated noise (MTF).

3 Including various combinations of voiced sound and measurement signals (PN-sequence, sine, etc.).

5 Types of test signals

Different test signals with different levels of complexity are available and have been evaluated for different types of applications. According to the complexity, different groups of signals can be identified and are listed as follows.

5.1 Non-speech-like (fully artificial) signals

The non-speech-like signals are the classical measurement signals. They can be divided into deterministic signals and random signals. The deterministic signals can be defined by a formula which fully describes the signal in time or frequency domain whereas the random signals can be described by the signal statistics and long-term spectrum.

5.1.1 Deterministic signals

5.1.1.1 Description

- sine wave;
- modulated sine wave;
- Fourier generated spectrum.

Deterministic signals for measurements in telephonometry are typically described in the frequency domain. The general description of a sinusoidal signal which might be modulated in amplitude or frequency is as follows:

$$s(t) = [A + \mu_{am} \cdot \cos(2\pi t \cdot f_{am})] \cdot \cos(2\pi t \cdot f_0 + \mu_{fm} \cdot \sin(2\pi t \cdot f_{fm})) \qquad [5\text{-}1]$$

A signal amplitude

μ_{am} modulation factor of amplitude modulation

f_{am} modulation frequency of amplitude modulation

f_0 carrier frequency

μ_{fm} modulation factor of frequency modulation

f_{fm} modulation frequency of frequency modulation

t time

A linear frequency sweep is described by:

$$s(t) = A(f_0 + st) \cdot e^{j(\pi t(f_0+st)+\varphi_0)} \qquad [5\text{-}2]$$

s sweep rate

f_0 starting frequency

The logarithmic sweep is described by:

$$s(t) = A(f_0 \cdot 10^{t/Td}) \cdot e^{j((2\pi f_0 Td/1n10)\cdot 10^{t/Td}+\varphi_0)} \qquad [5\text{-}3]$$

Td time taken to sweep one octave

A multiple sinusoidal spectrum can be described by its Fourier spectrum and can be denoted as follows:

$$s(t) = \sum_n A_n \cdot \sin(2\pi t \cdot f_n + \varphi_n) \qquad [5\text{-}4]$$

A_n amplitude of frequency component n

f_n frequency of frequency component n

φ_n phase of frequency component n

t time

A pseudo noise signal as described in 5.2.1.1 can be considered as a special form of a FFT-adapted multiple sinusoidal signal. Whereas discrete sine signals are normally applied at fixed levels independent of frequency, multiple sinusoidal and swept sine signals are often applied with a frequency weighting which matches the spectrum of speech more closely.

5.1.1.2 Application

Deterministic signals can always be used to determine the transfer characteristics of linear time invariant systems mainly in the frequency domain. Typically such signals are used to determine harmonic distortion and intermodulation distortion. The advantage of those signals is easy handling, determination of system parameters simply by level measurements.

The advantage of the logarithmic sweep is that the effective frequency resolution is more similar to the human ear frequency resolution.

The linear sweep allows the measurement result to be monitored and processed directly in the frequency domain as well as in the time domain using, e.g. FFT techniques. In addition, especially the linear sweep offers opportunities for suppression of unwanted noise and isolation of electrical or acoustical echoes.

Care needs to be taken for all discrete or multi-sinusoidal signals for use in measuring devices using adaptive techniques. The autocorrelation function of the measurement signal should not be periodically within the processing window of the device under test.

5.1.2 Random signals

5.1.2.1 Description

Random noise

Random noise can be determined by its statistical characteristics, the long-term power density spectrum, one dimensional and two dimensional probability density function or simply as a time history in case the random signal is sampled. The following signals are typically used in telephonometry:

white noise

– Frequency characteristics: lower and upper limit of the generated power density spectrum are defined by the application, typically the long-term power density spectrum is described.

– Probability density function: typically gaussian distribution with a crest factor of 11 dB ± 1 dB.

pink noise

– Frequency characteristics: the power density spectrum of the signal shows a decrease of 3 dB/octave, lower and upper limit of the generated power density spectrum are defined by the application, typically the long-term power density spectrum is described.

– Probability density function: typically gaussian distribution with a crest factor of 11 dB ± 1 dB.

For use in telephonometry, these signals are often modulated by rectangular modulating signal (ON/OFF-modulation). Typical modulating parameters are: 250 ms ON and 150 ms OFF. By this modulation the typical modulation of speech is simulated in a very simple way.

5.1.2.2 Application

Random signals are typically used for linear time invariant systems to determine broadband levels or levels in fractal octave bands. In addition such signals may be used to determine the transfer characteristics in the frequency domain such as frequency response or loudness ratings.

Using random signals for measurements long, averaging times (typically >10 s) are always required.

5.1.3 Combined random and deterministic signals

5.1.3.1 Description

By modulating random noise with a deterministic signal one can obtain the Modulation Transfer Function (MTF) of a system. If we take, e.g. a noise signal with an average Intensity $\overline{I}(\mathrm{t})$ and modulate it with a sinusoidal signal of the frequency f with a modulation index m=1, we get a signal:

$$I(\mathrm{t}) = \overline{I}(\mathrm{t}) \cdot (1 + \cos(2\pi f \mathrm{t})) \qquad [5\text{-}5]$$

5.1.3.2 Application

If this signal is applied to a system the Modulation Transfer Function (MTF) [2] can be measured by measuring the modulation index m(f) of the output of the system. By measuring and interpreting the MTF in the right way the speech intelligibility of a system can be predicted to a certain extent (see 5.2.2).

5.2 Speech-like signals

Speech-like signals exist up to various kinds of complexity. The degree of complexity is always related to the typical parameters of speech simulated by the speech-like signal. In general the classes of signals listed below can be found.

The building blocks for all speech-like signals are – besides other parameters – voiced and unvoiced sounds. In general it is required that a telephone or speech processing device using speech detectors is activated or should stay activated in the presence of these voiced or unvoiced sounds while other signals may prevent or cause an interruption of the transmission.

5.2.1 Composite Source Signals (Composed signals in time)

5.2.1.1 Description

a) *General considerations*

When composing the Composite Source Signal, the following three components were judged especially important:

– voiced signal to simulate voice properties;

– deterministic signal for measuring the transfer functions without statistical errors with constant power density spectrum of the excitation signal in the frequency domain to be measured;

– pause "signal" providing amplitude modulation.

The following features result:

i) short period of measurement;

ii) feeding in possibility of the test signal for the talking and listening direction at the same time (duplex operation).

The basic idea for using such a signal is to place the device under test in a well-defined, reproducible state for the period of measurement and to secure that the transfer functions of the device do not change appreciably during the actual measurement (quasi-stationarity). More precisely, the Composite Source Signal consists of the following components:

1) Voiced signal produced from the "artificial voice" signal according to Recommendation P.50

The voiced signal part of the CSS is the conditioning signal intended to activate possible speech detectors in voice controlled systems. The reason why the voiced signal has been chosen is that presumably all devices designed for speech transmission will quickly respond to a voiced sound. This signal is to activate the device under test for the direction of transmission to be measured. As the duration, beginning and end of the voiced signal are exactly known, this signal can also be used to measure the switching time for the direction of transmission under test. By means of the signal shape in the time domain the switching time and delay time of the entire system can be determined according to Recommendation P.340. The duration of the signal amounts to 50 ms.

2) Pseudo Noise Signal

The measurement signal is the Pseudo Noise (PN) signal presented after the voiced artificial speech sound. This signal has certain noise-like features. The magnitude of its Fourier transform is constant with frequency while the phase is changing. For measurements usually only the magnitude of the transfer function is of interest, the phase is not that important but can be determined as well.

The signal is produced as follows:

First a complex spectrum is produced in the frequency domain according to the following equation:

$$H(k) = W(k) \cdot e^{j \cdot i_k \cdot \pi} \text{ ; } k = -M/2, ..., M/2, \text{ without } 0;\ i_k\ \{+1, 0\},\ i_k = -i_{-k} \text{ random} \qquad [5\text{-}6]$$

The index M is adjusted to the chosen FFT size (e.g. 2048 points). The equation shows that the amount of the produced complex spectrum is constant for all frequencies if W(k) is chosen equal to 1 for all frequencies, whereas the phase may be π or 0 for each frequency, corresponding to a random sequence. However, to produce a different weighting in the frequency domain, W(k) can easily be adjusted in order to produce different spectra for the duration of the PN-sequence. Then, this spectrum will be transformed into the time domain by means of the inverse Fourier transform producing the following signal:

$$S(n) = \frac{1}{M} \sum_{k=-M/2, k\neq 0}^{M/2} H(k) \cdot e^{j2\pi \cdot n \cdot k/M} \text{ , } n = -M/2, ..., M/2 - 1; \qquad [5\text{-}7]$$

Thus, a signal is produced which is limited in time (corresponding to the chosen length of the Fourier transform) and which is adjusted to the chosen FFT size correctly. If a longer time sequence is wanted, the signal can be cycled. This method permits time sequences of any length.

The duration of this measurement signal amounts to about 200 ms by appropriate choice of M, the sampling rate and numbers of repetitions.

NOTE 1 – Typically the length of the FFT should be short for systems with high-time variant parameters such as companding techniques in order to get a good short-time estimation of the time variant transfer function. For systems incorporating adaptive techniques such as echo cancellers or noise cancellers, a higher number of M (close to 200 ms signal duration) may be appropriate in order to have the autocorrelation function of the measurement signal not periodically within the processing window of the device under test.

NOTE 2 – Instead of the pseudo-random noise signals also other signals like M-sequences (maximum length sequences) or other sequences with perfect autocorrelation functions may be appropriate for special applications. For other applications like distortion measurements, the PN-signal may be replaced by appropriate signals such as sine wave or narrow-band noise.

3) Pause

The pause has two purposes. An initial pause before applying any measurement signal is necessary to put systems with time variant transfer functions into a defined initial state. To this end, the pause should be as long as possible (> 1 s). If, however, the system is to be put into a constantly activated state (running speech-like), the intermediate pauses should be shorter (about 100 ms) to provide suitable amplitude modulation to the composite signal.

The pause of the CSS-sequence should be in the range of 100 - 150 ms.

NOTE 3 – The pause may be extended for measurements where a long-term behaviour after activation needs to be observed. In this case the pause may be prolonged up to several seconds depending on the measurement requirements.

In order to achieve a long-term offset free sequence, the repeated CSS-sequence should be inverted in amplitude (phase shift by 180 degrees).

b) *Calculation and analysis using a Composite Source Signal*

When using the CS-signal for measurements the sequence of voiced sound, pseudo noise signal and pause can be cycled. This means that after the pause the sequence starts again beginning with a voiced sound. Using this procedure sequences of any length may be produced.

1) Principle of acoustical and electrical calibration, test signal levels

Having created a sequence as described above this signal can be handled like a standard measurement signal, e.g. like the switched pink noise. The level calibration (acoustical and electrical) is done using the whole sequence including voiced sounds, PN-sequences and pauses. In principle, a standard RMS meter with a bandwidth of 20 kHz operating with "fast" averaging can be used. The preferred method however is to use FFT analysis for level calculations. The parameters for the FFT based calculation are:

– sampling rate according to the one chosen for signal generation (preferred 44.1 kHz or 48 kHz);
– FFT length according to the one chosen for signal generation;
– rectangular windowing;
– no overlap;
– averaging over the **whole (cycled) sequence**, including voiced sounds, PN-sequences, pauses;
– calculation of the level from the power density spectrum derived by the FFT calculation (integration of the levels over all frequency components).

2) Analysis parameters

For the measurement of transfer functions for sending direction and for receiving direction for loudness ratings etc., the sequence of voiced sound, PN-sequence and pause is cycled as well. The level of the complete sequence is adjusted in such a way that the overall level measured is according to the one specified as described above.

All measurements (analysis) are carried out **only during the PN-sequence**. For analysis of all transmission parameters in the frequency domain and loudness ratings, the measured and Fourier transformed signal has always to be referred to the Fourier transformed input signal using the same analysis parameters. In case of acoustical measurements, these input signals are measured at the MRP for measuring the sending characteristics. For the electrical measurements, the measured and Fourier transformed signal is referred to the input signal fed in either the digital or the analogue transmission path. This can be done using either a two-channel measurement or using techniques where the analysed input signal (measurement signal measured at the MRP or the digital interface) can be stored. For the analysis the following parameters are used:

– Sampling rate according to the one chosen for signal generation.
– FFT length according to the one used for signal generation, applied to the PN-sequence only.
– Rectangular windowing.
– Overlapping allowed between 0 and 99.9%. The same overlapping has to be applied for measurement signal at the input of the test object (MRP or digital interface) and measured signal at the output of sending or receiving direction of the test object.
– Referring the Fourier transformed signal measured either at the output of sending direction or receiving direction to the Fourier transformed signal at the corresponding excitation point (MRP, digital or analogue input).

Other measurements may require sinusoidal signals or different noise signals to measure different parameters, e.g. distortion. In this case the PN-sequence is replaced by the corresponding signal, e.g. sinusoidal frequency or narrow-band noise signal. The levels of the complete cycled Composite Source Signals including the different type of measurement signal are calculated as described above. Measurements are carried out using the measurement signal included in the CS-signal and using the calculations as described in the according standard.

5.2.1.2 Practical realisation of a Composite Source Signal for measurements up to 20 kHz

a) *Voiced signal to simulate voice properties*

The duration of the signal amounts to 48.62 ms. Within this period any speech detector should have recognized the voice and activated the system. The voiced signal can be described as a sequence of 16-bit words with a sampling rate of 44.1 kHz.

Table 1 contains 134 words which have to be repeated 16 times for activating the system under test for a time of 48.62 ms.

TABLE 1/P.501

Samples (ASCII values) of Part 1 of the CSS (to be read in columns)

–76	2098	3116	2930	2392	1824	1306	–3462	–7492	–2806	–626
112	2244	3158	2866	2410	1772	1170	–4024	–6414	–2844	–456
298	2360	3180	2808	2430	1742	968	–4590	–5334	–2888	–298
472	2456	3180	2764	2444	1750	702	–5154	–4428	–2898	–130
628	2538	3168	2728	2460	1760	394	–5716	–3772	–2846	
776	2626	3146	2686	2472	1762	76	–6298	–3360	–2698	
916	2730	3132	2632	2452	1736	–244	–6912	–3128	–2460	
1068	2824	3122	2572	2398	1684	–594	–7556	–3002	–2166	
1234	2904	3108	2496	2300	1624	–968	–8194	–2924	–1846	
1398	2964	3096	2432	2178	1572	–1384	–8719	–2870	–1544	
1572	2996	3076	2382	2068	1516	–1846	–8998	–2830	–1274	
1752	3032	3038	2362	1976	1460	–2356	–8898	–2800	–1032	
1932	3072	2992	2368	1892	1390	–2898	–8378	–2792	–818	

b) *Pseudo noise signal*

The parameters for the PN-sequence are:

Sampling rate 44.1 kHz, 16-bit word length, length of Fourier transform 2048 points.

$$\mathrm{H(k)} = \begin{cases} W(k) \cdot e^{j \cdot i_k \cdot \pi}; & \mathrm{k} = -928, \ldots, +928 \text{ without } 0,\ \mathrm{i_k}\,\{+1,\ 0\},\ \text{random } \mathrm{i_k} = \mathrm{i_{-k}} \\ 0 \text{ else} \end{cases} \qquad [5\text{-}8]$$

According to the above described Formula [5-8] the time signal is calculated by inverse Fourier transformation. This sequence is repeated 4307 times to achieve a length of 200 ms for the PN-measurement sequence. The crest factor of the PN-sequence is 11 dB ± 1 dB.

According to the frequency resolution of 21.533 Hz (44.1 kHz/2048) there are 928 FFT values in the frequency range between 0 and 20 kHz. Each value W(k) is 152680. It is calculated such that levels within a bandwidth of 20 kHz are the same for the voiced signal and the PN-sequence.

c) *Pause*

The pause is used as described in the general description of the CS-signal. The length of the pause amounts to 101.38 ms in order to achieve a signal duration of exactly 350 ms.

NOTE – By appropriate up or downsampling other sampling rates for the described sequence can be achieved. The interpolation filter used for up and downsampling should be close to an ideal rectangular filter. The stopband attenuation should be > 60 dB, the passband ripple < ±0.2 dB.

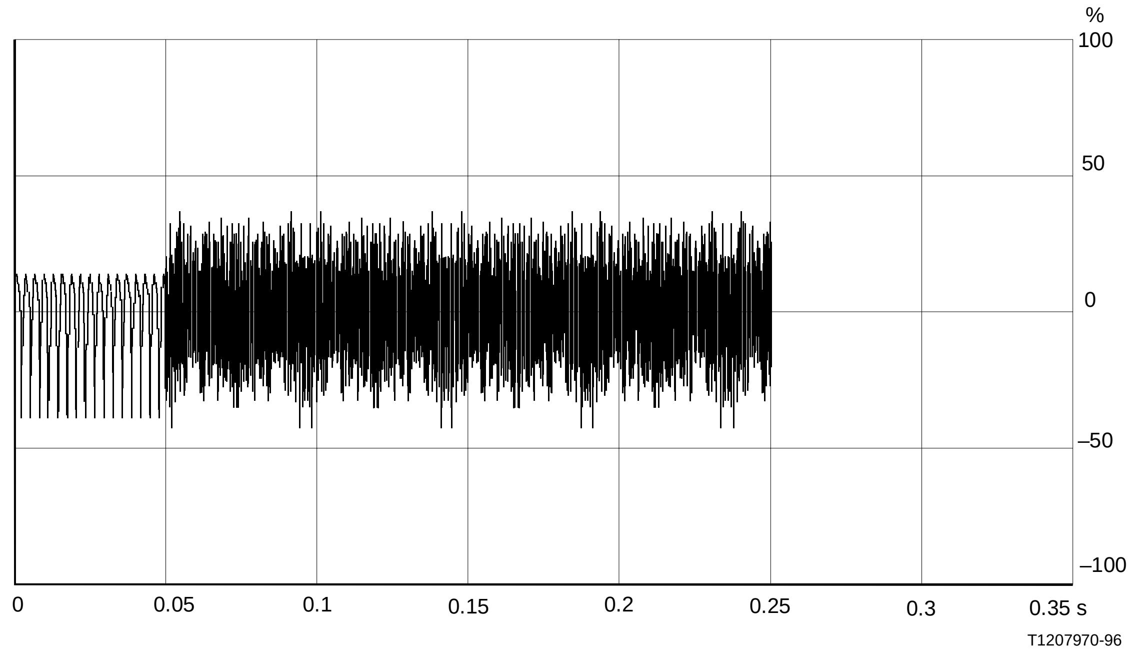

FIGURE 1/P.501

Composite Source Signal, time signal

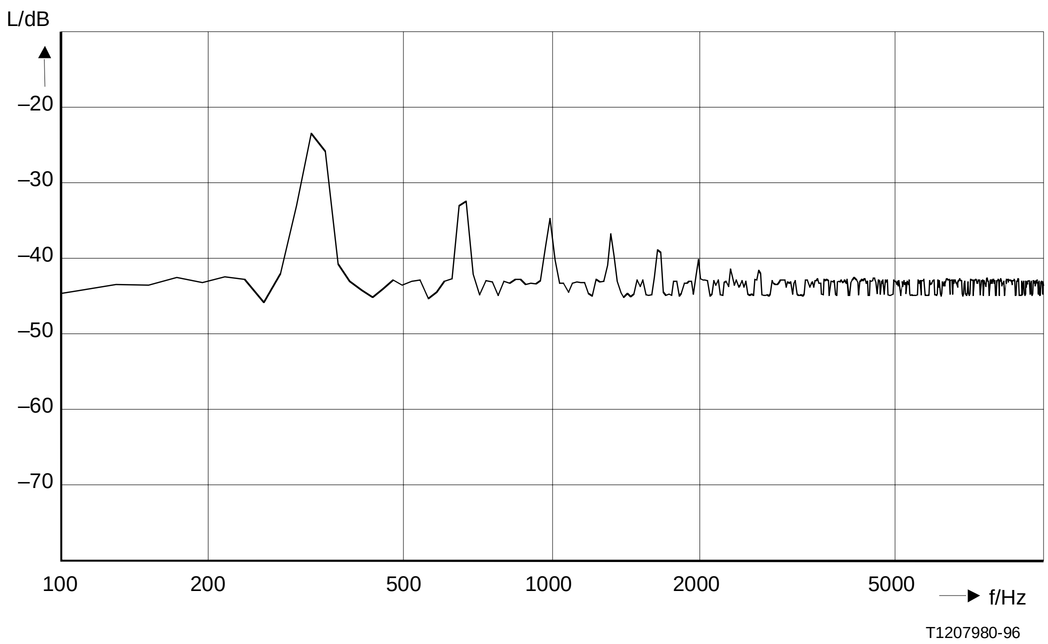

FIGURE 2/P.501

Power density spectrum of the Composite Source Signal (analysis window: Hanning)

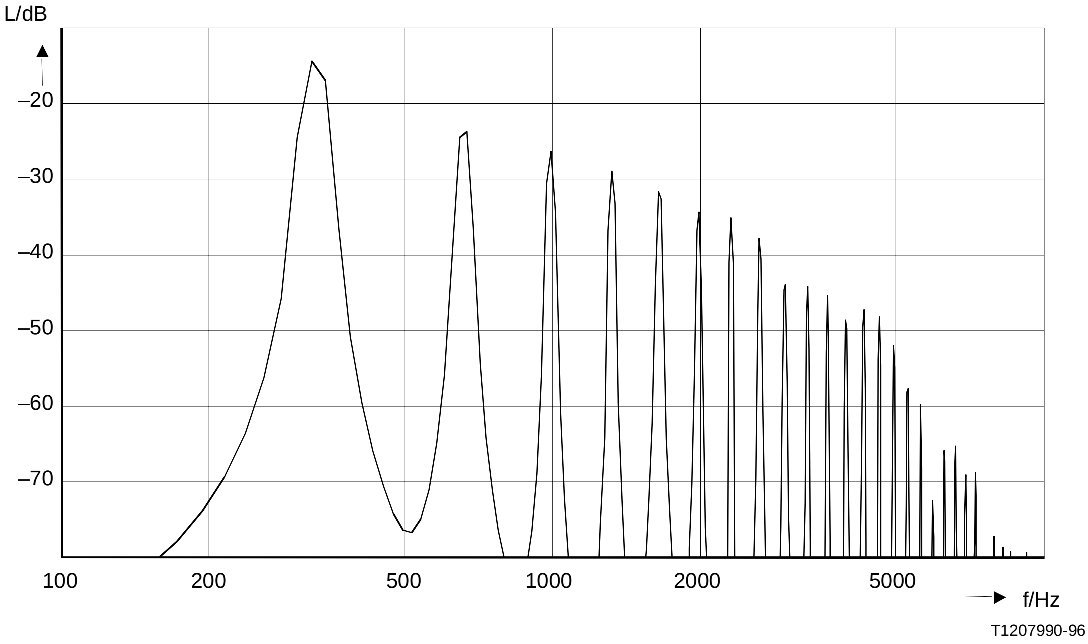

FIGURE 3/P.501

Power density spectrum of the voiced signal (analysis window: Hanning)

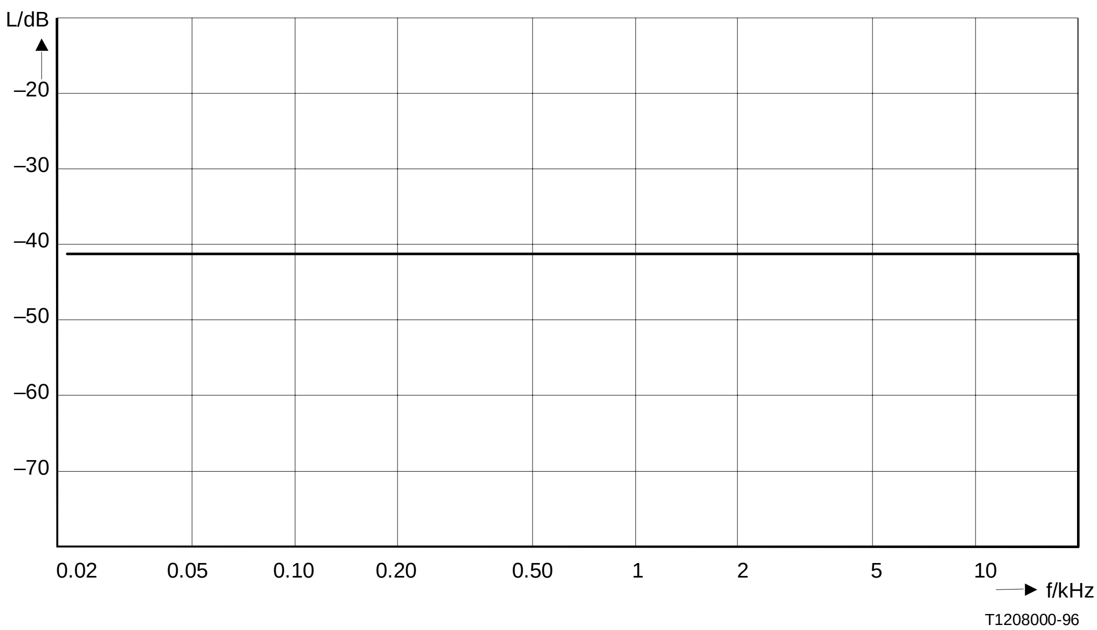

FIGURE 4/P.501

Power density spectrum of the PN-sequence (analysis window: Rectangle)

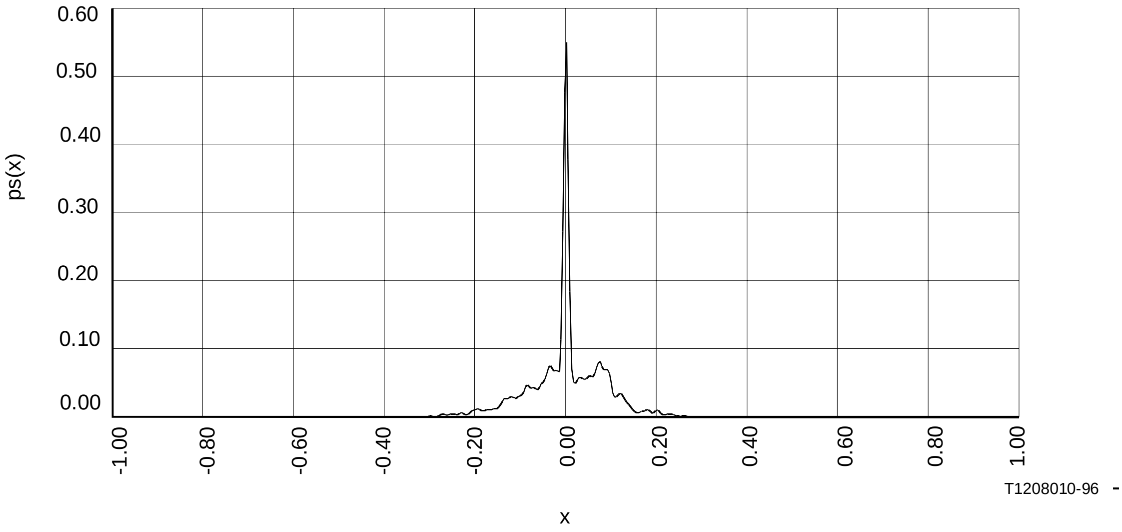

FIGURE 5/P.501

Probability density function of the CS-signal according to 5.2.1.2

5.2.1.3 Application

The above described signal may be applied to systems which behave non-linearly and are time variant but which can be considered for the short period of measurement to be in quasi stationary conditions. Parameters in the frequency domain such as frequency response, loudness ratings etc., as well as parameters in the time domain such as switch-on times, can be determined. If a signal for distortion measurement is inserted instead of the PN-sequence (sinusoidal signals or narrow-band noise) those parameters can be determined as well.

In general the CS-signal represents a class of signals. If for a special application longer parts of the voiced sound are required, the sequence of the voiced sound may be repeated until the required signal length is achieved. The same procedure can be applied to the PN-sequence and the pause. If such special applications are desired, the procedure of signal composition should be described in the according application.

In case of adaptive systems which change their transmission properties depending on the signal characteristics a low correlated signal is of advantage. For such systems the Fourier transformation length should be extended to approximately 200 ms (e.g. 8192 point FFT instead of 2048 point FFT). The signal analysis and generation parameters have to be adjusted accordingly: k = -3715, …, 3715 random without 0.

5.2.1.4 Bandlimited Composite Source Signal with speech-like power density spectrum

5.2.1.4.1 Description

a) *Composite Source Signal for single talk*

1) Bandlimited voiced signal

In Table 1, the ASCII values for the voiced signal described in 5.2.1.2, bandlimited between 200 Hz and 3.6 kHz can be found. According to a sampling rate of 44.1 kHz, the 134 ASCII values amount to 3.04 ms. The values are to be read in columns; see Table 2:

TABLE 2/P.501

ASCII values of the bandlimited voiced signal

–155	948	3224	4000	3129	1440	241	–888	–1853	–6137	–3474
276	1362	3370	4043	3043	1310	190	–957	–2121	–6560	–2508
517	1741	3500	4034	2914	1146	103	–1034	–2414	–6948	–1595
578	2043	3569	3974	2750	965	–9	–1103	–2707	–7301	–802
491	2276	3603	3862	2560	776	–138	–1146	–3017	–7568	
302	2422	3603	3724	2353	603	–267	–1181	–3319	–7732	
86	2500	3595	3577	2155	448	–388	–1190	–3612	–7758	
–103	2552	3586	3439	1991	345	–491	–1198	–3913	–7620	
–207	2595	3595	3336	1853	276	–569	–1215	–4224	–7310	
–198	2655	3638	3267	1750	250	–638	–1259	–4560	–6810	
–60	2758	3724	3224	1672	250	–698	–1327	–4922	–6155	
190	2896	3819	3198	1603	267	–759	–1457	–5301	–5344	
543	3060	3922	3172	1534	267	–813	–1629	–5715	–4439	

The values of the voiced signal in the frequency range 200 Hz - 3.6 kHz are again calculated such that the RMS value of the voiced signal and the PN-sequence are equal. The sequence is repeated 16 times to achieve a length of 48.62 ms.

2) Pseudo noise signal generated using 2048 point FFT

The parameters for the PN-sequence are:

Sampling rate 44.1 kHz, 16-bit word length, length of Fourier transform 2048 points.

$$\mathrm{H(k)} = \begin{cases} W(k) \cdot e^{j \cdot i_k \cdot \pi}; \quad \mathrm{k} = -928, \ldots, +928 \text{ except } 0,\ \mathrm{i_k}\ \{+1,\ 0\},\ \text{random } \mathrm{i_k} = -i_{-k} \\ 0 \text{ else} \end{cases} \qquad [5\text{-}9]$$

According to the above described Formula [5-7] the time signal is calculated by inverse Fourier transformation. This sequence is repeated 4307 times to achieve a length of 200 ms for the PN-measurement sequence. The crest factor of the PN-sequence is 11 dB $\pm$ 1 dB.

According to the frequency resolution of 21.5 Hz (44.1 kHz/2048) there are 928 FFT values in the frequency range between 0 and 20 kHz. Each value W(k) (before filtering) is 152680. It is calculated such that levels within a bandwidth of 20 kHz are the same for the voiced signal and the PN-sequence.

3) Pseudo noise signal generated using 8192 point FFT

According to the above described Formula [5-7] the time signal is calculated by inverse Fourier transformation. This sequence is repeated 1077 times to achieve a length of 200 ms for the PN-measurement sequence. The crest factor of the PN-sequence is 11 dB $\pm$ 1 dB.

According to the frequency resolution of 5.4 Hz (44.1 kHz/8192) there are 3715 FFT values in the frequency range between 0 and 20 kHz. Each value W(k) before filtering is 305360. It is calculated such that levels within a bandwidth of 20 kHz are the same for the voiced signal and the PN-sequence.

In order to achieve the same RMS value for the bandlimited PN-sequence the filter function shown in Figure 6 must be applied. The filter is chosen such, that the levels of the filtered and the unfiltered PN-sequence are equal.

NOTE 1 – By appropriate up- or downsampling other sampling rates for the described sequence can be achieved. The interpolation filter used for up- and downsampling should be close to an ideal rectangular filter. The stopband attenuation should be > 60 dB, the passband ripple < ± 0.2 dB.

FIGURE 6/P.501

Transfer function of the filter for bandlimiting the PN-sequence

TABLE 3/P.501

Table of filter corner frequencies

50 Hz	100 Hz	200 Hz	215 Hz	500 Hz	1 kHz	2.85 kHz	3.6 kHz	3.66 kHz	3.68 kHz
–25.8 dB	–12.8 dB	17.4 dB	17.8 dB	12.2 dB	7.2 dB	0 dB	–2 dB	–20 dB	–30 dB

For adaptive systems such as echo cancellers a longer PN-sequence may be preferable in order not to have correlated measurement signals within the adaptation window. For those systems the FFT length should be extended to 8192 points when using 44.1 kHz sampling rate as described in Table 3.

b) *Bandlimited Composite Source Signal for double talk*

The double talk sequence is generated in the same way as the single talk signal. However, the times of the voiced signal and the pause are slightly different in order to achieve a typical double talk condition with two signals applied at the same time, signal present only in one channel, voiced signals present on both sides as well as voiced signals and unvoiced signals present at the same time in the different channels. The correlation between single talk signal and double talk signal is low. This is achieved by choosing a different voiced signal with a different pitch frequency and a random noise signal instead of the PN-sequence. The duration of the voiced signal is 72.69 ms, the duration of the random noise signal is 200 ms and the duration of the pause amounts to 127.31 ms. This results in a total length of 400 ms.

1) *Voiced signal*

The voiced signal for double talk was chosen to have a different base frequency than the signal talk voiced signal. The values for the voiced signal for double talk can be found in Table 4. The level of this sound again is the same as the one for single talk. Using a sampling rate of 44.1 kHz 229 ASCII values represent 5.19 ms. The table is to be read in columns:

TABLE 4/P.501

ASCII-values for the bandlimited double talk voiced signal

–198	1146	–8292	4827	5853	1422	–1293	–810	–690	–1052	–621
–112	871	–8715	5094	5715	1224	–1302	–793	–724	–1043	–560
–9	560	–9077	5344	5560	1026	–1293	–767	–767	–1043	–509
103	233	–9370	5594	5387	819	–1267	–741	–793	–1052	–457
233	–121	–9542	5827	5215	603	–1250	–698	–819	–1060	–397
388	–491	–9542	6043	5043	388	–1233	–672	–845	–1060	–345
543	–871	–9361	6215	4879	181	–1224	–638	–853	–1060	–276
724	–1250	–8956	6344	4732	9	–1224	–603	–871	–1052	–207
896	–1638	–8327	6413	4586	–181	–1224	–595	–879	–1034	–112
1060	–2043	–7465	6422	4439	–328	–1224	–586	–888	–1017	
1233	–2465	–6396	6379	4276	–448	–1215	–595	–896	–991	
1388	–2896	–5163	6310	4086	–543	–1198	–603	–922	–957	
1517	–3345	–3827	6215	3870	–629	–1172	–621	–948	–931	
1638	–3819	–2448	6120	3629	–707	–1129	–629	–974	–905	
1747	–4310	–1103	6051	3370	–784	–1077	–938	–1009	–888	
1810	–4810	155	6000	3086	–871	–1026	–638	–1026	–862	
1845	–5319	1293	5991	2801	–948	–974	–638	–1052	–845	
1845	–5836	2241	5991	2534	–1026	–922	–638	–1069	–819	
1802	–6353	3034	6000	2267	–1112	–888	–638	–1077	–793	
1707	–6853	3655	6008	2034	–1181	–871	–638	–1069	–767	
1569	–7353	4138	5991	1819	–1241	–845	–647	–1060	–724	
1379	–7836	4517	5939	1612	–1276	–828	–664	–1060	–672	

In order to achieve the required length of 72.69 ms the values are to be repeated 14 times.

2) *Random noise*

The random noise is chosen as a white gaussian noise bandlimited at 20 kHz. The crest factor of the signal is 12 ± 1 dB. The RMS value of the bandlimited random noise is chosen to be the same as the one for the voiced signal.

In order to bandlimit the random noise between 200 Hz and 3.6 kHz the filter function shown in Figure 6 is used. This ensures the same RMS value for the bandlimited random noise.

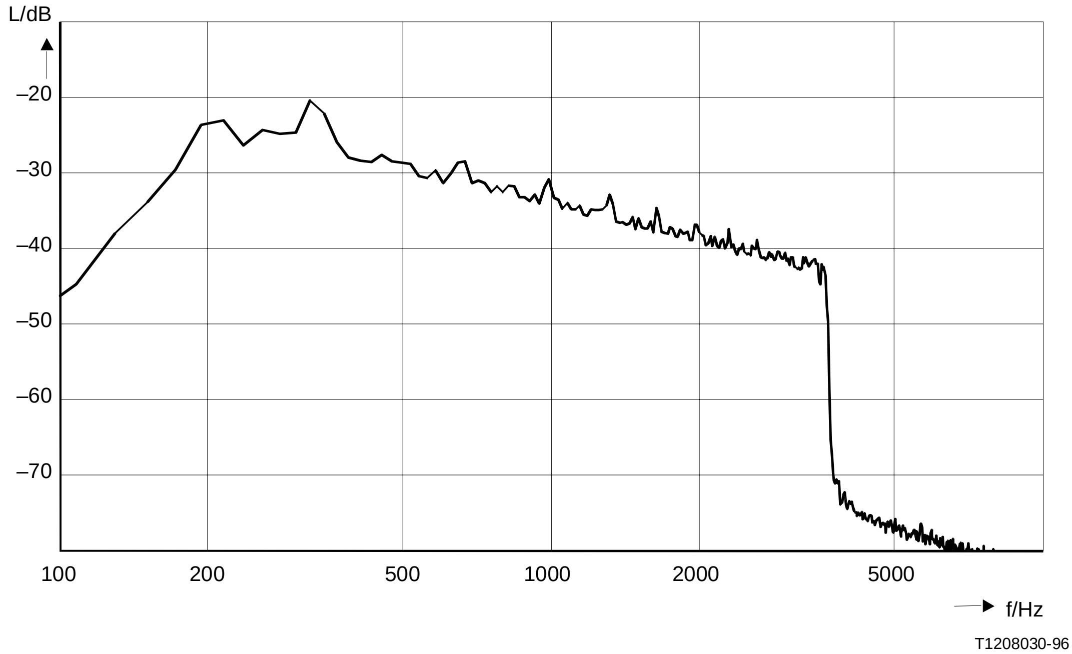

FIGURE 7/P.501

Power density spectrum of the bandlimited CS-signal (single talk signal, analysis window: Hanning)

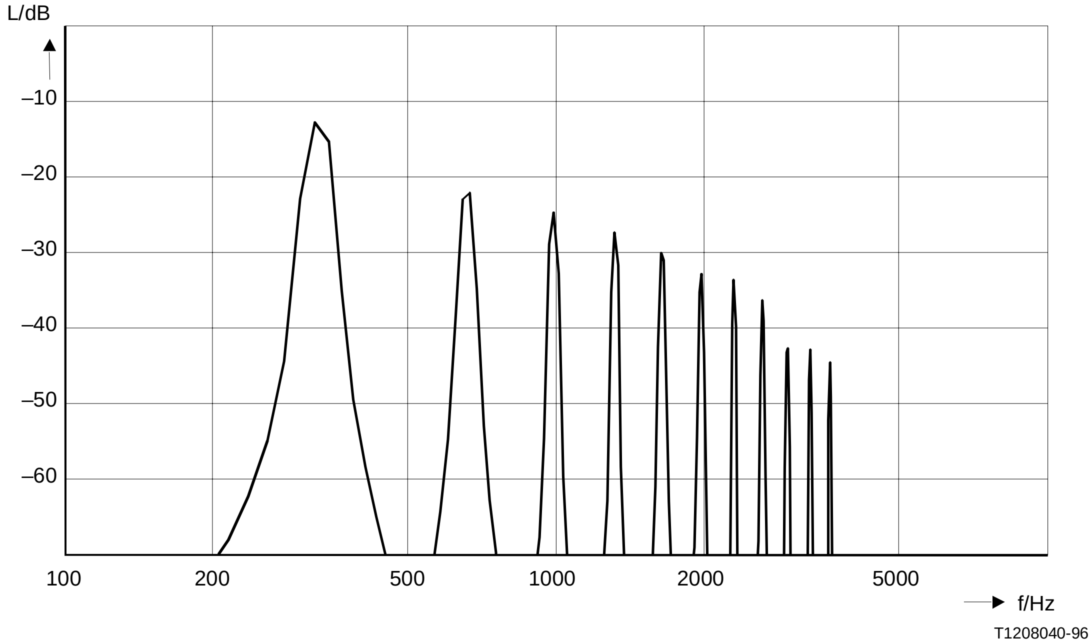

FIGURE 8/P.501

Power density spectrum of the bandlimited voiced signal (single talk signal, analysis window: Hanning)

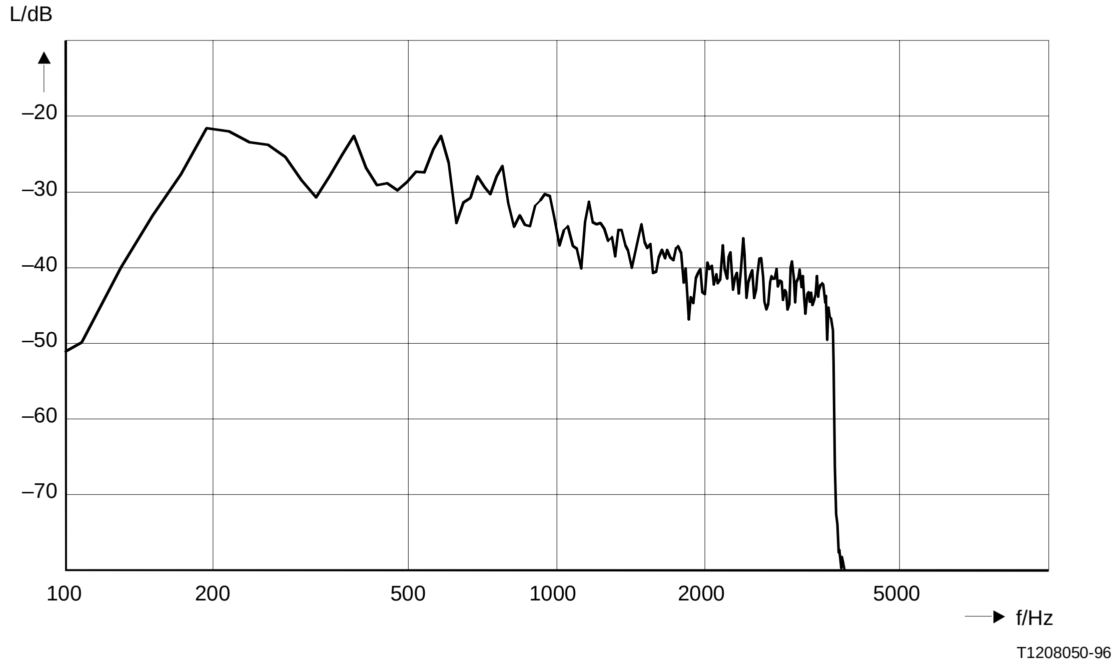

FIGURE 9/P.501

Power density spectrum of the bandlimited double talk CS-signal (analysis window: Hanning)

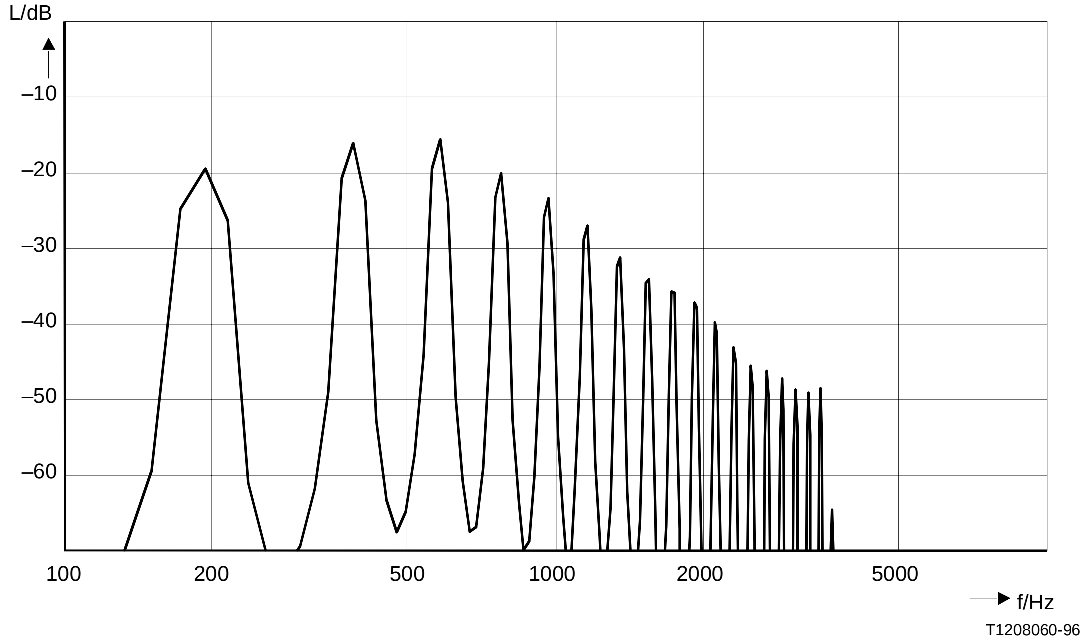

FIGURE 10/P.501

Power density spectrum of the bandlimited double talk voiced signal (analysis window: Hanning)

NOTE 2 – By appropriate up- or downsampling other sampling rates for the described sequence can be achieved. The interpolation filter used for up- and downsampling should be close to an ideal rectangular filter. The stopband attenuation should be > 60 dB, the passband ripple < ± 0.2 dB.

5.2.1.4.2 Application

The application of the bandlimited Composite Source Signals for single talk as well as for double talk is for all measurements where bandlimited systems need to be measured working non-linear and time variant and requiring the typical long-term power density spectrum of speech. The typical application is the measurement of speech echo cancellers in the network. For all one-directional measurements the bandlimited CS-signal for single talk measurements shall be used. In case of measurements in double talk conditions the double talk signal shall be used in double talk direction, whereas the single talk signal is fed in the far-end direction.

5.2.2 Speech-like modulated noise

5.2.2.1 Description

As pointed out in 5.1.3 the MTF can be used to measure the speech intelligibility of a system. By modulating octave band filtered noise one can obtain the MTF in different octave bands. With correct weighting of the modulation indices in each octave band and over different modulation frequencies, a speech transmission index ($0 \le STI \le 1$) that has high correlation to the speech intelligibility of a system can be obtained. The STI can be measured using a signal composed of a number of simultaneously modulated noise bands. The long-term power density spectrum is chosen equal to the power density spectrum of speech. Using the right modulation a signal is created that reflects the temporal characteristics of running speech. The STI has proven to be a good predictor of speech intelligibility for a wide range of distortions.

5.2.2.2 Application

The STI can be used to measure intelligibility of speech that is corrupted by the following distortions:

- noise;
- bandpass filtering;
- peak clipping and more in general a broad class of non-linear distortions;
- automatic gain control;
- reverberation.

5.2.3 Composed signals in frequency (probe tone technology)

5.2.3.1 Description

In order to determine the transmission characteristics of dynamically varying telephone systems, it may be necessary to apply a proper (speech-like) conditioning signal simultaneously with a suitable analytical test signal. Therefore, it is essential that:

– The analytical signal is applied at a level at which its influence on the dynamic behaviour of the telephone under test is insignificant. This requirement may imply a need for lengthy averaging in order to obtain sufficient measurement accuracy. For dynamically changing systems this leads to a kind of "average" transfer characteristic which does not take into account short-time effects.

– The correlation between the conditioning signal and the analytical test signal is kept to a minimum. Often this may be accomplished by a simple spectral separation of the two signals.

In general the relationship between the actual condition of the system under test and the measurement signal is not clear for the reason that the measurement signal is uncorrelated to the activation signal.

5.2.3.2 Application

Typically this method is used for determining average (long-term) characteristics of a system. If for example the average frequency response under realistic operating conditions (presence of reverberance and noise) is to be measured, a series of linear sinusoidal sweeps may be used. Often the analytical signal will be a single tone which may also be used, for example for the measurement of temporal gain variation caused by the conditioning signal and measured at the frequency of the single tone.

5.2.4 Complex composed signals

5.2.4.1 Simulated Speech Generator (SSG)

5.2.4.1.1 Description

1) *General description*

To generate a signal approximating the amplitude distribution of speech, a main signal having a gaussian distribution is modulated by a specially-tailored modulating signal, as shown in Figure 11. The resultant signal is shaped to approximate the long-term frequency spectrum of speech, as shown in Figure 12.

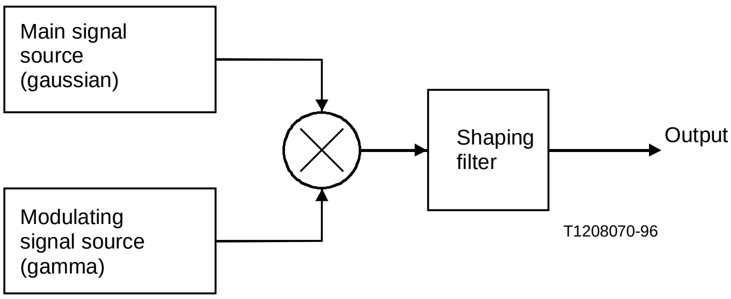

FIGURE 11/P.501

Block diagram of simulated speech generator

2) *Main signal*

The main signal consists of eight 1024 point pseudo-random noise segments. Each segment has the same magnitude spectrum but a different phase spectrum with the phase randomised within and between the segments uniformly from 0 to 360 degrees, in order to randomise the interaction between the intermodulation products of the harmonically related spectral components. The duration of each segment is 80 ms and they are merged with each other through a raised cosine window with an additional 80 ms merging segment between them. The simultaneous fade-out of the previous segment and the fade-in of the following segment eliminates the transients which would occur at the segment boundaries. The complete main signal thus consists of eight pseudo-random segments interleaved with eight merging segments, each of 80 ms duration having a total length of 1.28 s. The desired frequency shaping to approximate an average speech spectrum is provided by a simple filter at the output.

3) *Modulating signal*

Measurements show that a gamma distribution with parameter m = 0.545 provides a good approximation to the instantaneous amplitude distribution of continuous speech. The syllabic characteristics can be represented by a low-pass response which is practically flat up to about 4 Hz (which may be considered the -3 dB point) followed by -6 dB per octave roll-off.

The final waveshape of the modulating signal was derived empirically from the gamma distribution. Varying the period of this pulse in a pseudo-random manner and adjusting its rise and fall time ratio results in a satisfactory approximation to the spectrum of the modulation envelope of real speech.

4) *Combined signal*

In order to extend the repetition time of the final signal and to spread more evenly the maxima of the modulating signal over the repeated sequence of the gaussian signal, the ratio between the sampling clock frequencies of both signals was chosen to be 4/255. Thus the clocking frequency of the main signal is 12 800 Hz, and the clock frequency for the modulating signal is about 200.8 Hz. The repetition times are: 1.28 s for the gaussian signal, 10.2 s for the modulating signal and 326.4 s for the final modulated signal.

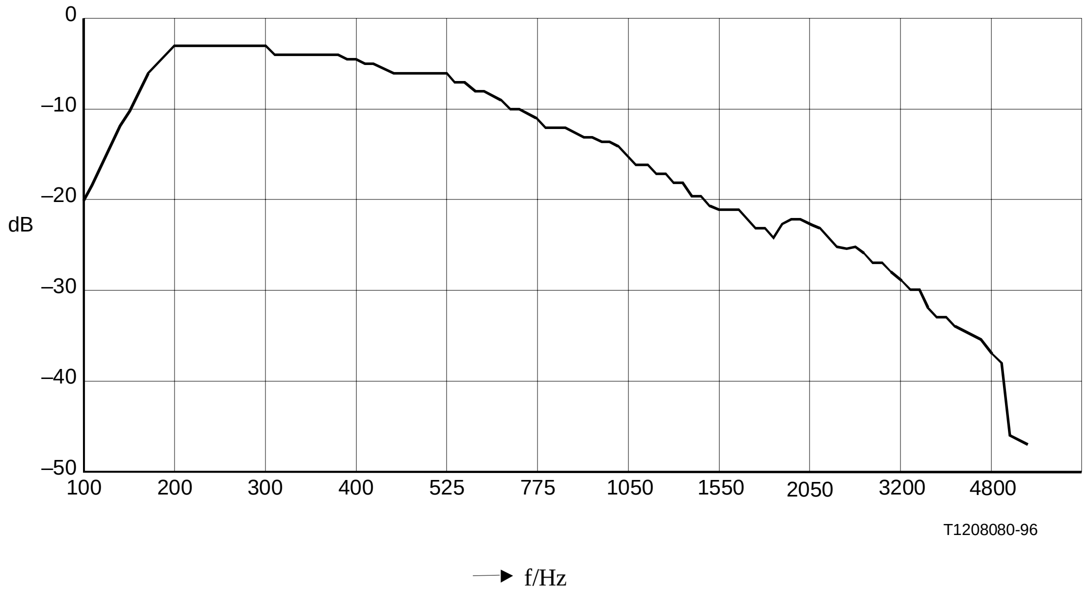

FIGURE 12/P.501

Spectrum generated by the simulated speech generator

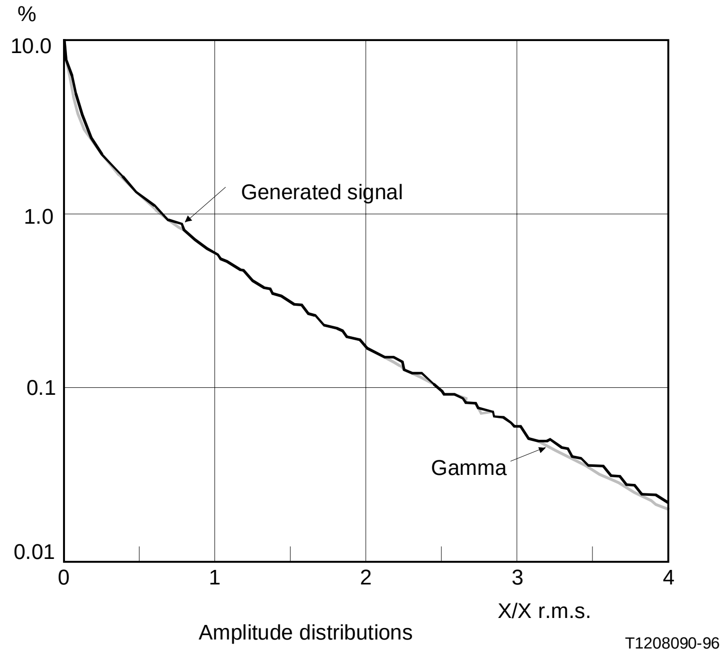

FIGURE 13/P.501

Amplitude distribution of the simulated speech generator

5.2.4.1.2 Application

The SSG signals may be used if a "typical" speech sequence is required for the measurement. Similar to the P.50 signal the SSG signal represents speech by generating typical parameters of natural speech by a defined process. Compared to real speech no specific language or specific voice is simulated.

In general, long averaging times (>10 s) are required if parameters like frequency responses or loudness ratings need to be calculated from the measurements.

5.2.4.2 Artificial voice ITU-T P.50

The most used complex composed speech-like signal in telephonometry is the artificial voice as described in Recommendation P.50.

5.2.4.3 Artificial conversational speech ITU-T P.59

Recommendation P.59 describes an artificial conversational speech signal, complex composed, offering talk spurts etc., and a double talk sequence as well.

5.2.4.4 Speech-model process controlled by discrete Markov Chains (MSMP)

5.2.4.4.1 Description

1) *General considerations*

In the following there is a brief description of the present version of a speech-model process MSMP (Markov Speech Model Process) proposed as a test signal for wideband speech processing applications. This test signal is an extension of the known model for the narrow-band case (MSIRP) [3]. As mentioned in earlier publications [3] and [4], there were some deviations between the long-term spectrum and the spectrum of the signal envelope of MSIRP and those of natural telephone speech (300 Hz - 3400 Hz). Due to the improvements of the model, the spectrum of the signal envelope of MSIRP and MSMP are adapted more closely to that of narrow-band or wideband speech.

2) *Generation procedure of Markov Speech Model process (MSMP)*

The generation procedure of MSMP is given in the block diagram of Figure 14. As this block diagram shows, the MSMP is constructed as the product of a gaussian process ***n(t)*** and of a process ***s(t)***. This concept renders it possible to adjust the amplitude probability density function (PDF) and the autocorrelation of the resulting process separately.

The time variant properties of the resulting process are controlled by trained Markov chains to achieve natural formant and pitch structures. The decision, whether a frame of 20 ms duration is voiced (v) or unvoiced (uv), is carried out by the Markov chain (mc) which is responsible for the pitch value in this frame (mc-pitch). This trained Markov chain produces a natural sequence of 33 different pitch values. One of these values is 0 Hz indicating that the current frame is an unvoiced one. Depending on this decision, a generalised Markov chain is controlled which produces a natural index sequence for choosing one of 50 formant filters for this frame (mc-formant). This generalised Markov chain (mc-formant) works as the hidden part of a HMM, which produces a natural sequence of gain terms and specifies the short-time energy of the product process (mc-energy).

The upper branch controls the PDF of the resulting process. A lowpass filter produces a slowly varying random process. This filter is excited by a weighted gaussian white random process. The weighting factors are obtained from mc-energy and are constant during each frame. At this place we have to point to a special property of product processes: The amplitude PDF of the product processes are generally symmetric. But natural speech has an asymmetric amplitude PDF. Therefore we use two different non-linear mappings to achieve the desired PDF of the process ***s(t)*** and switch between the outputs of these non-linear mappings depending on the sign of the process ***n(t)***. So we obtain a random process with the desired amplitude PDF. This PDF is formed such that the multiplication of the process ***s(t)*** with a gaussian process yields the desired PDF of natural speech.

The lower branch controls the autocorrelation of the process ***x(t)***. In the case of an unvoiced frame the formant filter is excited by an uncorrelated gaussian process with zero mean and unit variance. In voiced frames the excitation could be modelled by periodically spaced impulses. It has to be taken into consideration that the introduction of a pitch frequency causes a quasi-periodical structure in the resulting process. This is a contradiction *a priori* because the resulting process has to be a random process. A compromise can be found by using a comb filter with the transfer function

$$H(z) = \alpha \Big/ \left(1 - a_k z^{-k_0}\right).$$

So, we obtain a gaussian process which has a quasi-line structure in its spectrum as the excitation of the formant filter during the voiced frames. Note that natural speech also has just an approximate spectral line structure, which corresponds well to the spectral structure of the output of the comb filter. The distances of these spectral lines represent the pitch frequency, which is determined by k_0. The value of k_0 is changed by mc-pitch, but it is constant during each voiced frame. The switching between voiced or unvoiced regions is carried out in two steps, to achieve a smoothed transition. Thus the sharpness factor a_k of the used comb filter is equal to 0.6 for the first and the last frame of each region and 0.95 for all other cases. Finally, the formant filters are specified by a set of 50 lattice filters of sixteenth order. The coefficients are obtained from a codebook, which is optimised with the generalised Lloyd algorithm [5] for codebook design. For a smooth switching between the formant filters the actual filtering coefficients are updated every 2 ms by linearly interpolating between the two sets of filter coefficients of neighbouring frames.

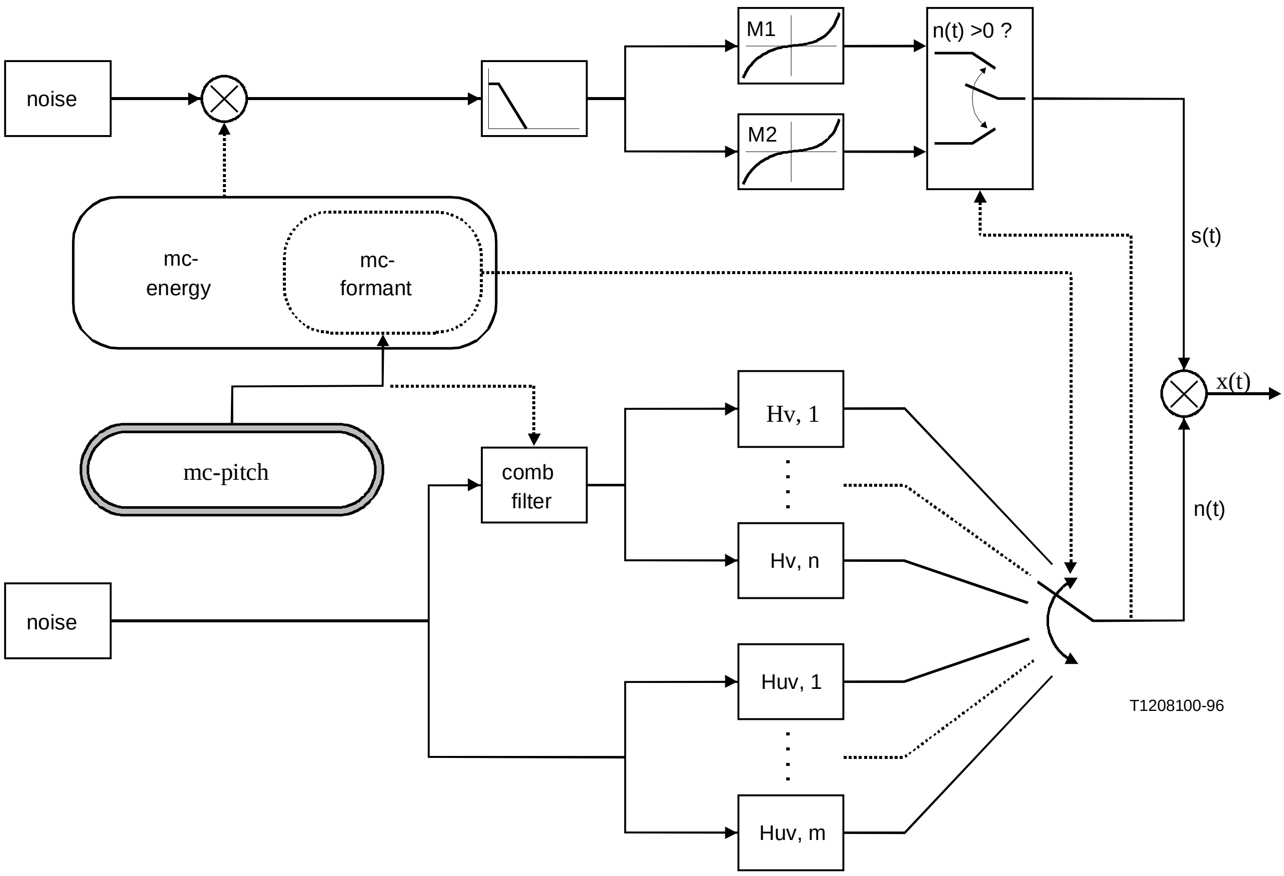

FIGURE 14/P.501

Block diagram of the MSMP generator

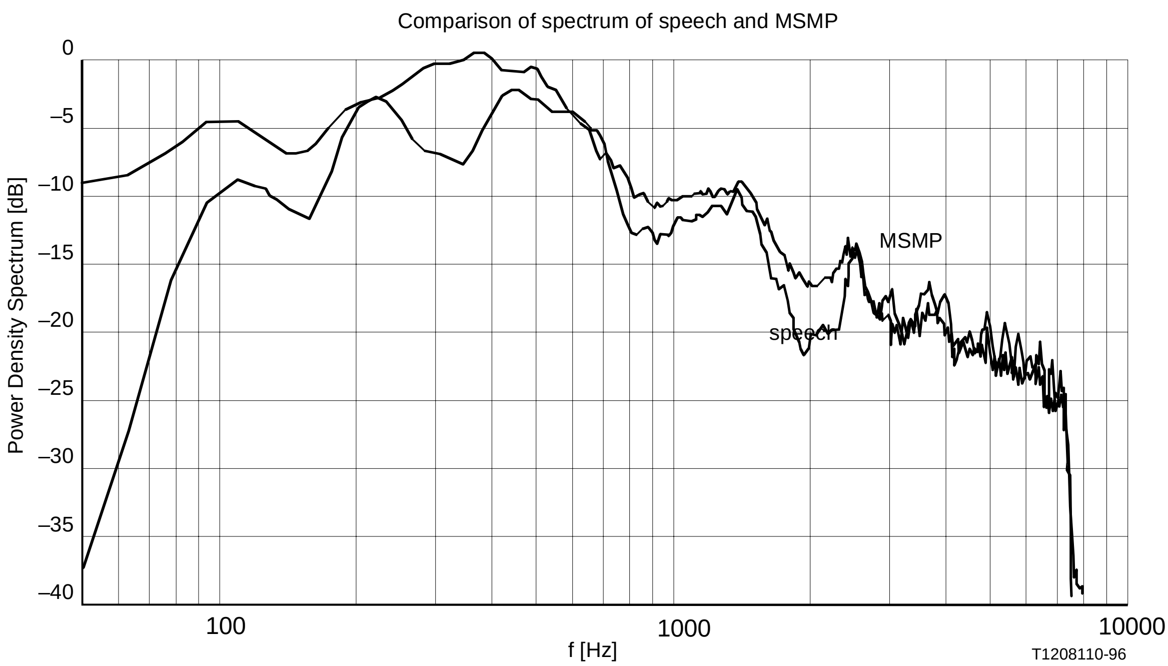

FIGURE 15/P.501

Comparison of spectrum of speech and MSMP

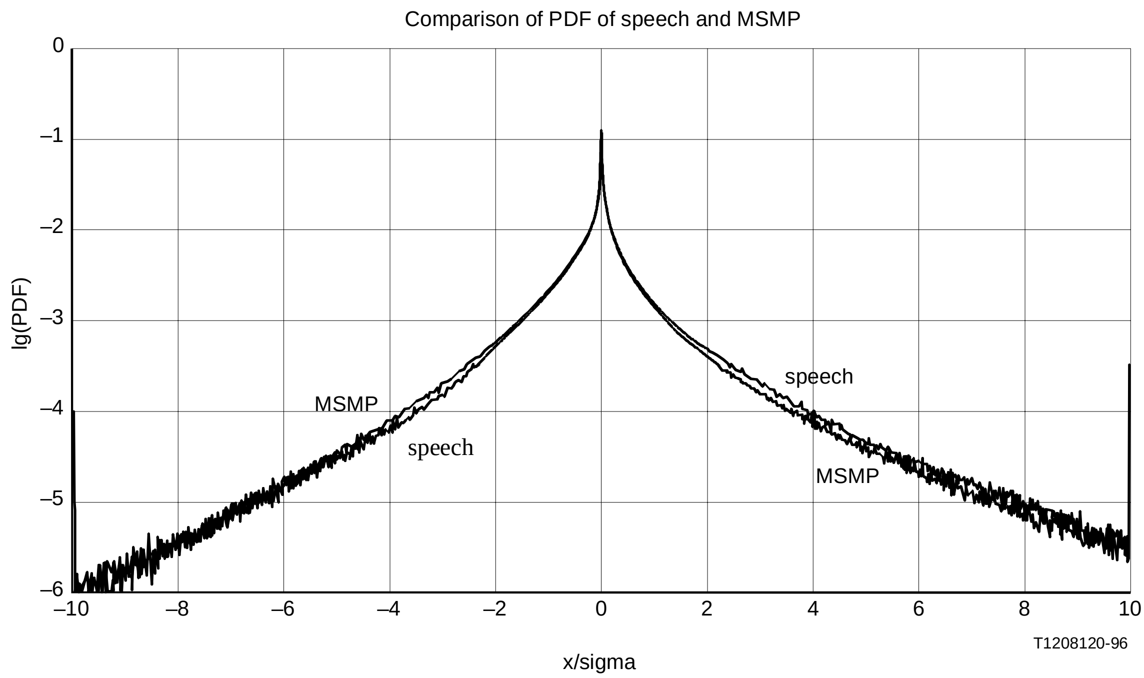

FIGURE 16/P.501

Comparison of PDF of speech and MSMP

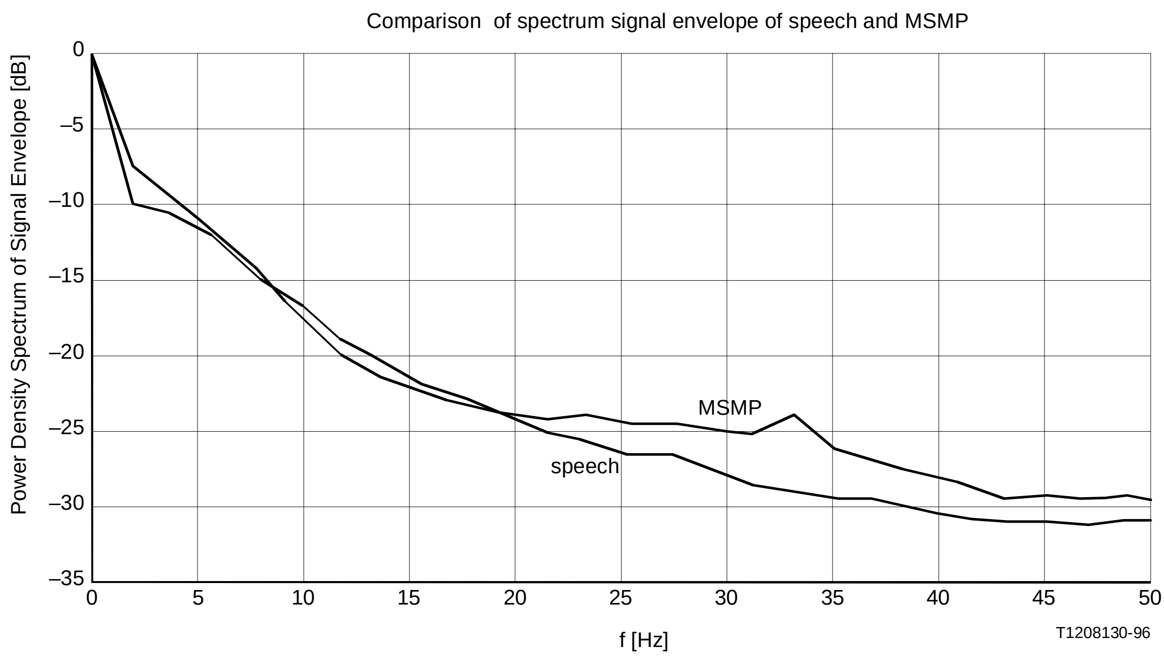

FIGURE 17/P.501

Comparison of spectrum of signal envelope of speech and MSMP

5.2.4.4.2 Application

MSMP signals may be used if a "typical" speech sequence is required for the measurement. The above concept is applicable to produce a test signal, which is adapted either to one special language (e.g. English or German) or to a mixture of several languages. In this order we only have to use another speech data set for the training of the Markov chains. The MSMP allows to prescribe not only the long-term spectrum and PDF, but also to include the important short-time characteristics of "natural" formant and pitch changes. This model eliminates, on the one hand, the dependencies on the speech material in its applications, but, on the other hand, it renders it possible to introduce required special features of certain speech (like male/female, weak/strong variation of characteristics etc.).

In general long averaging times (>10 s) are required if parameters like frequency responses or loudness ratings need to be calculated from the measurements.

5.3 Speech signals

For some applications real speech signals (one way or conversational) can be used. If the device under test shows a strong non-linear and/or time varying behaviour the concepts of frequency response, distortion and signal-to-noise ratio are no longer directly applicable or even relevant. One can however investigate the behaviour of the device under test using real speech fragments and judge the overall quality by a model of the human auditory perception. With such a model a "perceptual frequency response function" or special measurers, e.g. based on psychoacoustic parameters (perceptual distortion measures) can be defined.

Bibliography

[1] GIERLICH (H.W.): A measurement technique to determine the transfer characteristics of hands-free telephones, *Signal Processing*, Vol. 27, No. 3, pp. 281-300, 1992.

[2] STEENEKEN (H.J.M.), HOUTGAST (T.): A physical method for measuring speech transmission quality, *J. Acoust. Soc. Am.*, Vol. 67, No. 1, pp. 318-326, January 1980.

[3] ITU-T COM 12-14, *Speech and Speech-Model Signals: A Comparison*, ITU-T Study Group 12, Geneva, May 1993.

[4] HALKA (U.), HEUTE (U.): Speech-Model Processes Controlled by Discrete Markov-Chains, *Proc. Asilomar Conf. Sig. Syst. Comp.*, pp. 1196-1200, Pacific Grove, United States, 1993.

[5] LINDE (Y.), BUZO (A.), GREY (R.M.): An Algorithm for Vector Quantizer Design, *IEEE Transactions on Communications*, Vol. COM-28, No. 1, January 1980.

ITU-T RECOMMENDATIONS SERIES

Series A	Organization of the work of the ITU-T
Series B	Means of expression
Series C	General telecommunication statistics
Series D	General tariff principles
Series E	Telephone network and ISDN
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media
Series H	Transmission of non-telephone signals
Series I	Integrated services digital network
Series J	Transmission of sound-programme and television signals
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	Maintenance: international transmission systems, telephone circuits, telegraphy, facsimile and leased circuits
Series N	Maintenance: international sound-programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality**
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminal equipment and protocols for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks and open system communication
Series Z	Programming languages