International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.501

Amendment 1

(06/2018)

SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Objective measuring apparatus

Test signals for use in telephonometry

Amendment 1

Recommendation ITU-T P.501 (2017) – Amendment 1

ITU International Telecommunication Union

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Voice terminal characteristics	Series	P.30 P.300
Reference systems	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 speech quality	Series	P.80
Methods for objective and subjective assessment of speech and video quality	Series	P.800
Audiovisual quality in multimedia services	Series	P.900
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100
Models and tools for quality assessment of streamed media	Series	P.1200
Telemeeting assessment	Series	P.1300
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400
Methods for objective and subjective assessment of quality of services other than speech and video	Series	P.1500

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

Recommendation ITU-T P.501

Test signals for use in telephonometry

Amendment 1

Summary

Recommendation ITU-T P.501 describes test signals that are applicable for several purposes in telephonometry. Recommendation ITU-T P.501 gives a wide variety of test signals starting with low complexity test signals up to test signals with a high degree of complexity incorporating many typical parameters of speech. Besides technical signals, such as sine waves or noise, more speech-like signals are described.

Recommendation ITU-T P.501 describes the principles of signal construction for each type of test signal. Characteristic properties, such as power density spectra, probability density functions or shaping filter responses, are shown.

Recommendation ITU-T P.501 gives an overview of 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 specific applications.

In order to avoid problems in creating the test signals described, all these test signals are freely available for download from the ITU-T test signals database.

Annex A proposes two test signals [a pseudo noise sequence (PN-sequence) with a low crest factor and a logarithmically distributed multi-sine wave] for the measurement of terminal coupling loss (TCL).

Annex B provides speech files and noise sequences to be used in combination with objective speech quality evaluation methods. This speech material does not replace the speech material found in Supplement 23 to the ITU-T P-series Recommendations.

Appendix I provides a description of the processing applied to the speech signals in clause 7.3.

This Recommendation includes an electronic attachment containing the set of freely available test signals described in the Recommendation.

Amendment 1 extends the applicability of the AM-FM test signal to super-wideband and fullband applications.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.501	1996-08-30	12	11.1002/1000/3635
2.0	ITU-T P.501	2000-05-18	12	11.1002/1000/5080
2.1	ITU-T P.501 (2000) Amd. 1	2004-05-14	12	11.1002/1000/7411
3.0	ITU-T P.501	2007-06-29	12	11.1002/1000/9065
4.0	ITU-T P.501	2009-12-14	12	11.1002/1000/10657
5.0	ITU-T P.501	2012-01-13	12	11.1002/1000/11459
5.1	ITU-T P.501 (2012) Amd. 1	2012-07-14	12	11.1002/1000/11686
5.2	ITU-T P.501 (2012) Amd. 2	2014-10-29	12	11.1002/1000/12330
5.3	ITU-T P.501 (2012) Amd. 3	2015-06-29	12	11.1002/1000/12515
6.0	ITU-T P.501	2017-03-01	12	11.1002/1000/13173
6.1	ITU-T P.501 (2017) Amd. 1	2018-06-13	12	11.1002/1000/13623

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. 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 Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

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.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T P.501

Test signals for use in telephonometry

Amendment 1

Editorial note: This is a complete-text publication. Modifications introduced by this amendment are shown in revision marks relative to Recommendation ITU-T P.501 (2017).

1 Scope

This Recommendation[1] describes test signals that 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 many typical parameters of speech. Besides technical signals, such as sine waves or noise, more speech-like signals are described.

The overview of 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 specific applications.

To avoid difficulties when creating the test signals described, all these signals are freely available for download from the ITU-T test signals database: http://www.itu.int/net/itu-t/sigdb/genaudio/Pseries.htm.

2 References

The following ITU-T 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; 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. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T G.122] Recommendation ITU-T G.122 (1993), *Influence of national systems on stability and talker echo in international connections*.

[ITU-T G.168] Recommendation ITU-T G.168 (2015), *Digital network echo cancellers*.

[ITU-T G.191] Recommendation ITU-T G.191 (2010), *Software tools for speech and audio coding standardization*.

[ITU-T P.50] Recommendation ITU-T P.50 (1999), *Artificial voices.*

[ITU-T P.56] Recommendation ITU-T P.56 (2011), *Objective measurement of active speech level*.

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

[ITU-T P.79] Recommendation ITU-T P.79 (2007), *Calculation of loudness ratings for telephone sets.*

---

[1] This Recommendation includes an electronic attachment containing the set of freely-available test signals described within the Recommendation. The electronic attachment is distributed with the base text, and not with Amendment 1.

[ITU-T P.340] Recommendation ITU-T P.340 (2000), *Transmission characteristics and speech quality parameters of hands-free terminals*.

[ITU-T P.800] Recommendation ITU-T P.800 (1996), *Methods for subjective determination of transmission quality*.

[ITU-T P.830] Recommendation ITU-T P.830 (1996), *Subjective performance assessment of telephone-band and wideband digital codecs.*

[ITU-T P.862] Recommendation ITU-T P.862 (2001), *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*.

[ITU-T P.862.3] Recommendation ITU-T P.862.3 (2007), *Application guide for objective quality measurement based on Recommendations ITU-T P.862, ITU-T P.862.1 and ITU-T P.862.2.*

[ITU-T P.863] Recommendation ITU-T P.863 (2014), *Perceptual objective listening quality assessment*.

[ITU-T P.863.1] Recommendation ITU-T P.863.1 (2014), *Application guide for Recommendation ITU-T P.863*.

3 Definitions

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 crest factor: Peak-to-RMS (root mean square) ratio of a signal.

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

3.2.3 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.

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

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

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

3.2.7 speech transmission index (STI): Index indicating the speech intelligibility, especially in reverberant conditions, derived from measuring the modulation transfer function.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

CSS Composite Source Signal

DC Direct Current

FFT Fast Fourier Transform

FIR Finite Impulse Response

IIR Infinite Impulse Response

LTI Linear Time-Invariant

HVAC Heating, Ventilation and Air Conditioning

mc Markov chain

MRP Mouth Reference Point

MSMP Markov Speech Model Process

MTF Modulation Transfer Function

PCM Pulse Code Modulation

PDF Probability Density Function

PN Pseudo Noise

RCV Receive

RMS Root Mean Square

SND Send

SSG Simulated Speech Generator

STI Speech Transmission Index

TCL Terminal Coupling Loss

5 Conventions

None.

6 Overview of test signals and typical applications

See Tables 6-1 to 6-5.

Table 6-1 – Linear, time-invariant systems (e.g., standard handset telephones without echo/noise cancelling)

	Sine wave (Note 1)	Noise (Note 2)	CSS (Note 3)	Probe tone	SSG	ITU-T P.50	ITU-T P.59	MSMP	Speech
Loudness ratings	✓	✓	✓	✓	✓	✓	✓	✓	✓
Frequency responses	✓	✓	✓	✓	✓	✓	✓	✓	✓
Listener sidetone/ talker sidetone	✓	✓	✓	✓	✓	✓	✓	✓	✓
Harmonic distortion	✓		✓						
Distortion		✓	✓						
Out-of-band signals	✓		✓						
Level measurements	✓	✓	✓	✓	✓	✓	✓	✓	✓
Delay measurements	✓	✓	✓		(✓)	✓		(✓)	✓

Table 6-1 – Linear, time-invariant systems (e.g., standard handset telephones without echo/noise cancelling)

	Sine wave (Note 1)	Noise (Note 2)	CSS (Note 3)	Probe tone	SSG	ITU-T P.50	ITU-T P.59	MSMP	Speech
Echo measurements		✓	✓	✓	(✓)	(✓)	(✓)	(✓)	(✓)

✓ Applicable

(✓) Applicable with some caution

NOTE 1 – Including modulated sine wave and Fourier spectra.

NOTE 2 – Including pink, white and switched noise.

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

Table 6-2 – Non-linear and/or time-variant systems (e.g., mobile phones)

	Sine wave (Note 1)	Noise (Note 2)	CSS (Note 3)	Probe tone	SSG	ITU-T P.50	ITU-T P.59	MSMP	Speech
Loudness ratings (long-term values)		(✓)	(✓)		(✓)	✓	(✓)	(✓)	✓
Loudness ratings (short-term values)		(✓)	✓						
Frequency responses (long-term values)		(✓)	(✓)		(✓)	✓	(✓)	(✓)	✓
Frequency responses (short-term values)		(✓)	✓						
Listener sidetone/ talker sidetone (long-term values)		(✓)	(✓)	(✓)	(✓)	✓	(✓)	(✓)	✓
Harmonic distortion			(✓)						
Distortion		(✓)	(✓)		(✓)		(✓)	(✓)	
Out-of-band signals			(✓)						
Level measurements		(✓)	(✓)		(✓)	✓	(✓)	(✓)	✓
Delay measurements		(✓)	(✓)			✓			✓
Echo measurements		(✓)	(✓)		(✓)	(✓)	(✓)	(✓)	(✓)

✓ Applicable

(✓) Applicable with some caution

NOTE 1 – Including modulated sine wave and Fourier spectra.

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

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

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

Table 6-3 – Hands-free telephones

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

✓ Applicable

(✓) Applicable with some caution

NOTE 1 – Including modulated sine wave and Fourier spectra.

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

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

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

Table 6-4 – Echo cancellers

	Sine wave (Note 1)	Noise (Note 2)	CSS (Note 3)	Probe tone	SSG	ITU-T P.50	ITU-T P.59	MSMP	Speech
Level measurements		(✓)	✓	(✓)	✓	✓	✓	✓	✓
Delay measurements		(✓)	✓						✓
Switching characteristics	(✓)	(✓)	✓		(✓)		(✓)	(✓)	✓
Background noise performance		(✓)	✓	(✓)	✓	✓	✓	✓	✓

Table 6-4 – Echo cancellers

	Sine wave (Note 1)	Noise (Note 2)	CSS (Note 3)	Probe tone	SSG	ITU-T P.50	ITU-T P.59	MSMP	Speech
Echo loss, terminal coupling loss		(✓)	(✓)	(✓)	✓	✓	✓	✓	(✓)
Double-talk performance		(✓)	✓		(✓)		(✓)	(✓)	✓
✓ Applicable (✓) Applicable with some caution NOTE 1 – Including modulated sine wave and Fourier spectra. NOTE 2 – Including pink, white and switched noise as well as modulated noise (MTF). NOTE 3 – Including various combinations of voiced sound and measurement signals (PN-sequence, sine, etc.).									

7 Types of test signals

Various test signals with varying levels of complexity are available and have been evaluated for different types of applications. According to the complexity level, different groups of signals can be identified and are listed in clauses 7.1 to 7.4.

7.1 Non-speech-like (fully artificial) signals

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

7.1.1 Deterministic signals

7.1.1.1 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 that might be modulated in amplitude or frequency is given in Equation 7-1:

$$s(t) = \left[A + \mu_{\text{am}} \cdot \cos\left(2\pi t \cdot f_{\text{am}}\right)\right] \cdot \cos\left[2\pi t \cdot f_0 + \mu_{\text{fm}} \cdot \sin\left(2\pi t \cdot f_{\text{fm}}\right)\right] \quad (7\text{-}1)$$

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 Equation 7-2:

$$s(t) = A\left(f_0 + st\right) \cdot \exp\left\{j\left[\pi t\left(f_0 + st\right) + \varphi_0\right]\right\} \quad (7\text{-}2)$$

s sweep rate

f_0 starting frequency

The logarithmic sweep is described by:

$$s(t) = A\left[f_0 \cdot 10^{(t/Td)}\right] \cdot \exp\left[j\left(\frac{2\pi f_0 Td}{\ln 10} \cdot 10^{(t/Td)} + \varphi_0 \right) \right] \quad (7\text{-}3)$$

Td time taken to sweep one octave

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

$$s(t) = \sum_n A_n \cdot \sin\left(2\pi t \cdot f_n + \varphi_n\right) \quad (7\text{-}4)$$

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 (PN) signal as described in clause 7.2.1.1 can be considered as a special form of an 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.

7.1.1.2 Application

Deterministic signals can always be used to determine the transfer characteristics of linear time-invariant (LTI) 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, for example, fast Fourier transform (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.

7.1.2 Random signals

7.1.2.1 Description

Random noise

Random noise can be determined by its statistical characteristics, the long-term power density spectrum, and one- and two-dimensional probability density function (PDF) or simply as a time history if 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 a 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.

7.1.2.2 Application

Random signals are typically used for LTI 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.

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

7.1.3 Combined random and deterministic signals

7.1.3.1 Description

By modulating random noise with a deterministic signal, one can obtain the modulation transfer function (MTF) of a system. If we take, for example, a noise signal with an average Intensity $\bar{I}$(t) and modulate it with a sinusoidal signal of the frequency *f* with a modulation index *m* = 1, we get a signal:

$$I(t) = \bar{I}(t) \cdot \left[1 + \cos(2\pi f t)\right] \tag{7-5}$$

7.1.3.2 Application

If this signal is applied to a system, the MTF [b-Steeneken, 1980] 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 clause 7.2.2).

7.2 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.

7.2.1 Composite source signals (composed signals in time)

7.2.1.1 Description

a) General considerations

When composing the composite source signal (CSS), the following three components were judged especially important [b-Gierlich, 1992]:

– 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 CSS consists of the following components:

1) Voiced signal produced from the "artificial voice" signal according to [ITU-T 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 [ITU-T P.340]. The duration of the signal amounts to 50 ms.

2) Pseudo noise signal:

The measurement signal is the 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 as important, but can also be determined.

The signal is produced as follows:

First, a complex spectrum is produced in the frequency domain according to Equation 7-6:

$$H(k) = W(k) \cdot \exp(\, j \cdot i_k \cdot \pi\,);\ k = -M/2, ..., M/2, \text{ without } 0;\ i_k \{+1, 0\},\ i_k = -i_{-k} \text{ random} \quad (7\text{-}6)$$

The index M is adjusted to the chosen FFT size (e.g., 2048 points). Equation 7-6 shows that the amount of the produced complex spectrum is constant for all frequencies if $W(k)$ is chosen to be 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 is transformed into the time domain by means of the inverse Fourier transform producing the following signal:

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

Thus, a signal is produced that is limited in time (corresponding to the chosen length of the Fourier transform) and that 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 value 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, other signals like *M*-sequences (maximum length sequences) or other sequences with perfect autocorrelation functions may be appropriate for special applications. For other applications, such as 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 ms to 150 ms.

NOTE 3 – The pause may be extended for measurements where 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°).

b) Calculation and analysis using a composite source signal

When using the CSS for measurements, the sequence of voiced sound, PN 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 root mean square (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 the sending direction and for the 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 the case of acoustical measurements, these input signals are measured at the mouth reference point (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 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 overlap has to be applied for the measurement signal at the input of the test object (MRP or digital interface) and measured signal at the output of the sending or receiving direction of the test object.
- Referring the Fourier transformed signal measured either at the output of the sending direction or the 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 CSSs, including the different types of measurement signal, are calculated as described above. Measurements are carried out using the measurement signal included in the CSS and using the calculations described in the relevant Recommendation.

7.2.1.2 Practical realization of a composite source signal for measurements up to 20 kHz (fullband)

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 16bit words with a sampling rate of 44.1 kHz.

Table 7-1 contains 134 ASCII values that have to be repeated 16 times to activate the system under test for a time of 48.62 ms. Read Table 7-1 in columns.

Table 7-1 – 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.

$$H(k) = \begin{cases} W(k) \cdot \exp\left(j \cdot i_k \cdot \pi\right);\ k = -928, ..., +928 \text{ without } 0, i_k \{+1, 0\}, \text{ random } i_k = i_{-k} \\ 0 \text{ else} \end{cases} \quad (7\text{-}8)$$

According to Equation 7-8, 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 kHz and 20 kHz. Each value $W(k)$ is 152 680. It is calculated such that levels within a bandwidth of 20 kHz are the same for the voiced signal and the PN-sequence.

NOTE 1 – As described for the narrow-band CSS, alternatively an 8192 point PN-sequence can be used.

c) Pause

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

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

See Figures 7-1 to 7-5.

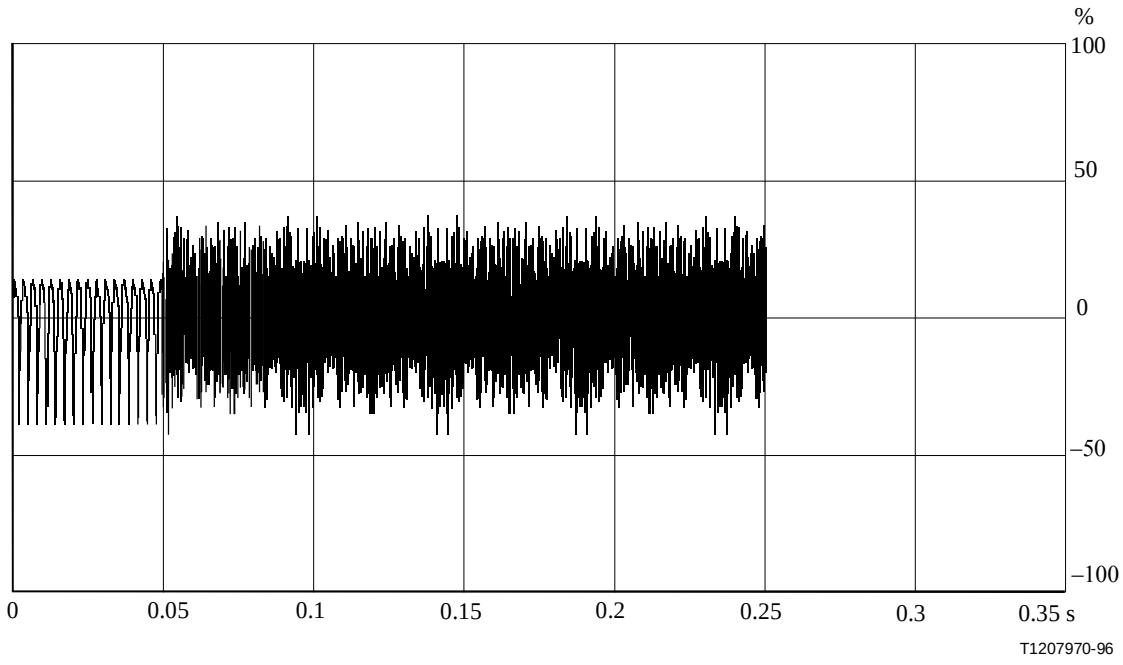


Figure 7-1 – Composite source signal, time signal

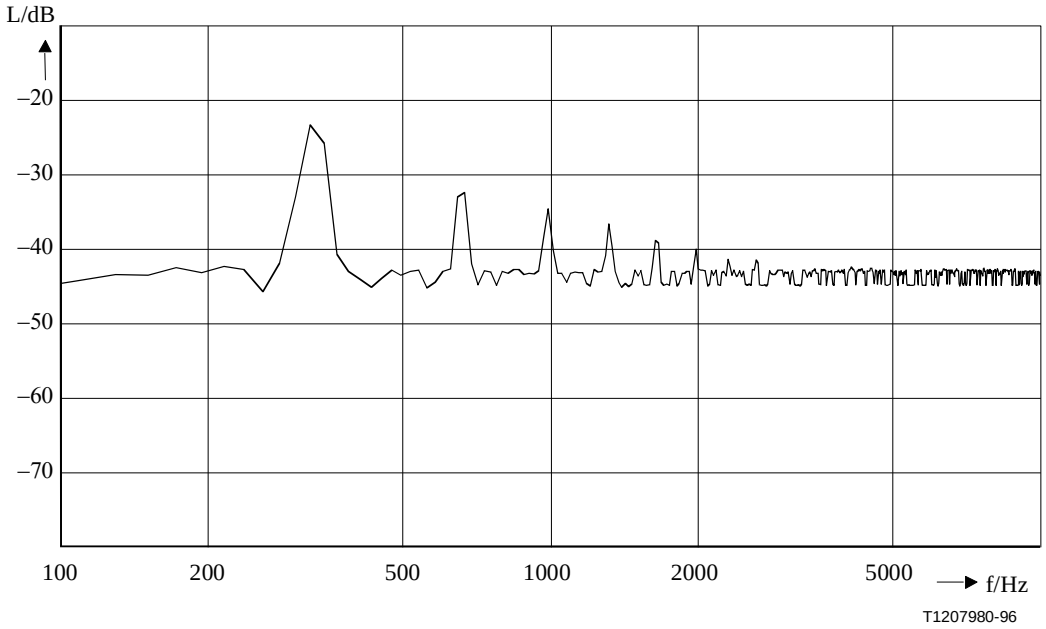


Figure 7-2 – Power density spectrum of the composite source signal (analysis window: Hanning)

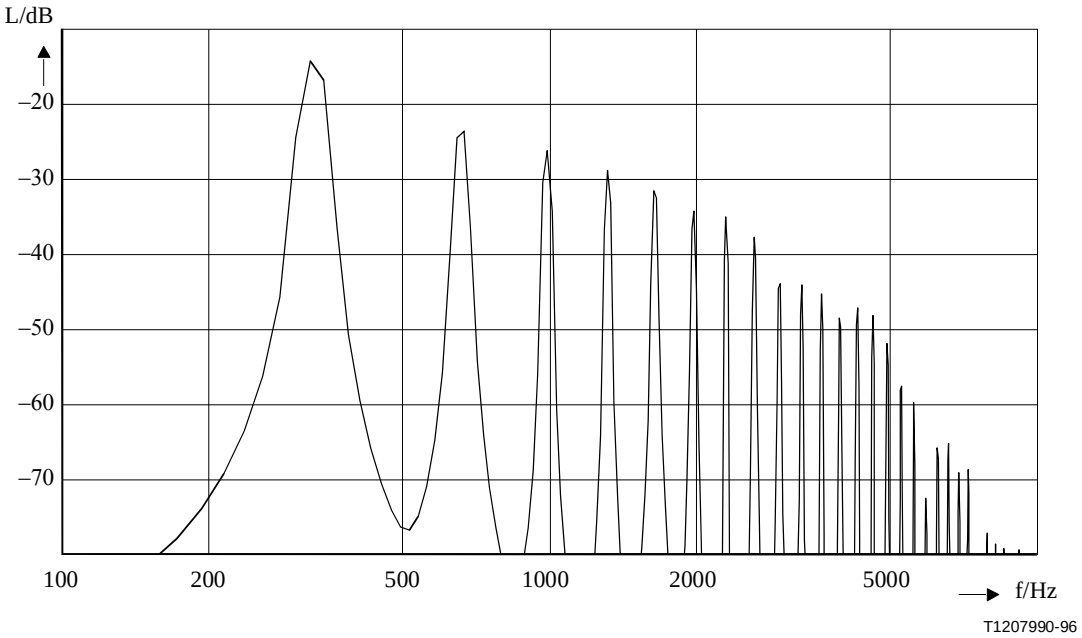


Figure 7-3 – Power density spectrum of the voiced signal (analysis window: Hanning)

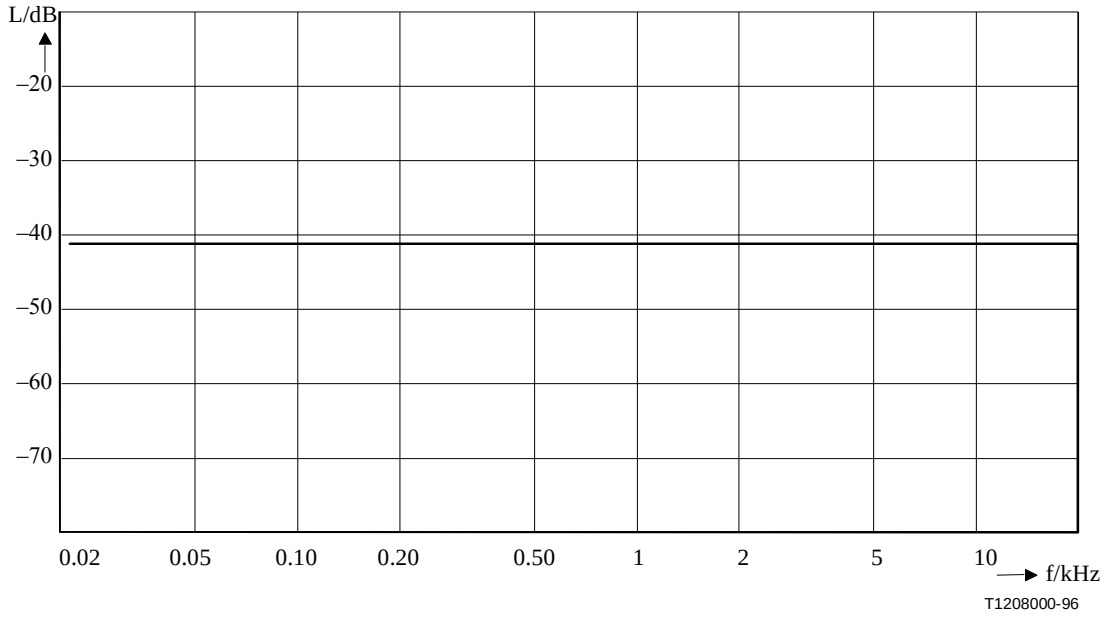


Figure 7-4 – Power density spectrum of the PN-sequence (analysis window: Rectangle)

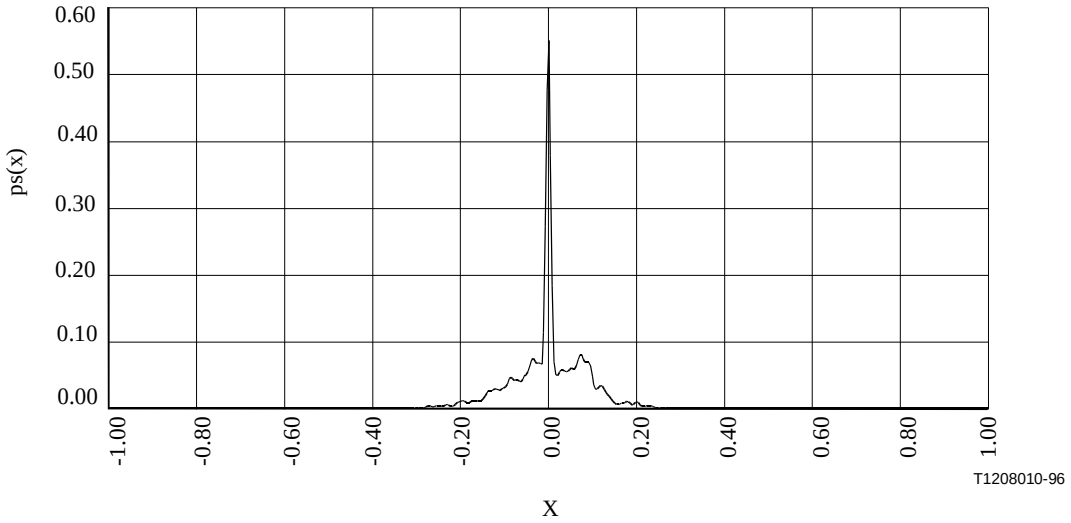


Figure 7-5 – Probability density function of the CSS according to clause 7.2.1.2

7.2.1.3 Application

The signal described in clause 7.2.1.2 may be applied to systems that behave non-linearly and that are time variant, but that can be considered for the short period of measurement to be in quasi-stationary conditions. Parameters in the frequency domain, such as frequency response and loudness ratings, 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 also be determined.

In general, the CSS represents a class of signal. 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 relevant application.

For adaptive systems that 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: $k = -3715, \ldots, 3715$ random without 0.

7.2.1.4 Fullband 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, a 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 a 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.

NOTE – For some applications, it may be desirable to generate a double-talk signal of length equal to that of the single-talk CSS. In order to achieve the same length of the double-talk CSS as the single-talk CSS, the artificial voice may be limited to 48.62 ms and the pause may be changed to 151.38 ms, while the pause for the single-talk CSS would also be 151.38 ms. Other combinations in time are possible.

a) Voiced signal

The voiced signal for double talk was chosen to have a different base frequency to that of the signal talk voiced signal. The values for the voiced signal for double talk can be found in Table 7-2. The level of this sound again is the same as that for single talk. Using a sampling rate of 44.1 kHz, 229 ASCII values represent 5.19 ms. Read Ttable 7-2 in columns.

Table 7-2 – ASCII values for the double-talk voiced signal (to be read in columns)

−64	2242	1614	166	391	926	727	406	561	2395	−4239	−4202
148	2310	1515	139	429	930	709	397	618	2299	−4797	−3601
341	2364	1416	121	462	935	690	388	686	2174	−5381	−3033
502	2413	1316	104	500	939	660	386	770	2033	−5974	−2506
637	2450	1212	96	534	935	645	379	856	1864	−6571	−2015
750	2474	1118	88	568	940	626	377	959	1663	−7170	−1553
856	2485	1023	88	606	935	604	377	1073	1439	−7779	−1120
973	2474	935	96	640	926	588	368	1190	1188	−8382	−726
1095	2446	852	97	676	925	571	367	1319	905	−8968	−364
1230	2408	762	109	706	911	555	413	1458	588	−9478	
1380	2363	686	124	735	901	538	422	1598	255	−9826	
1486	2322	615	139	762	887	513	413	1749	−73	−9925	
1591	2287	540	158	789	875	500	422	1894	−457	−9752	
1681	2243	477	178	814	857	483	426	2039	−852	−9318	
1763	2185	416	206	835	845	470	429	2179	−1263	−8667	
1844	2109	364	227	856	826	456	432	2310	−1693	−7903	
1927	2009	316	255	875	812	441	443	2413	−2153	−7087	
2014	1906	271	291	890	789	428	456	2477	−2643	−6287	
2097	1806	231	323	905	766	419	479	2488	−3152	−5541	
2170	1709	197	357	911	751	409	512	2461	−3687	−4849	

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

b) Random noise

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

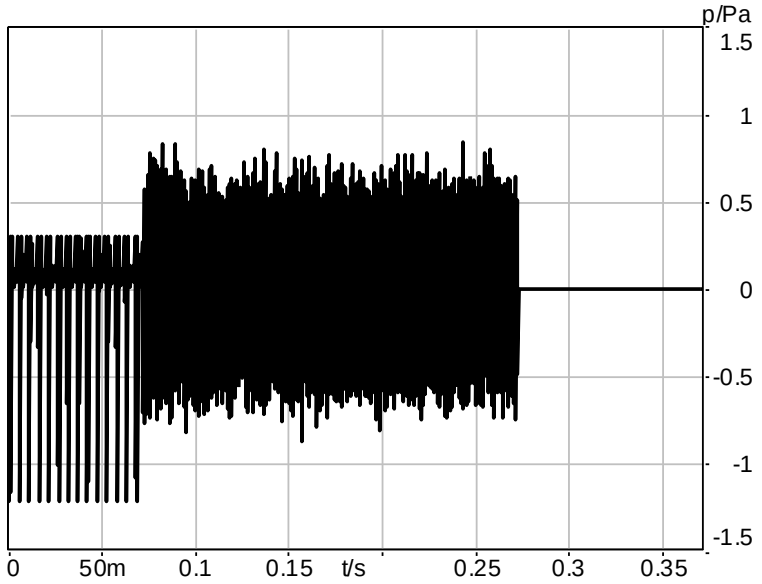


Figure 7-6 – Double-talk composite source signal, time signal

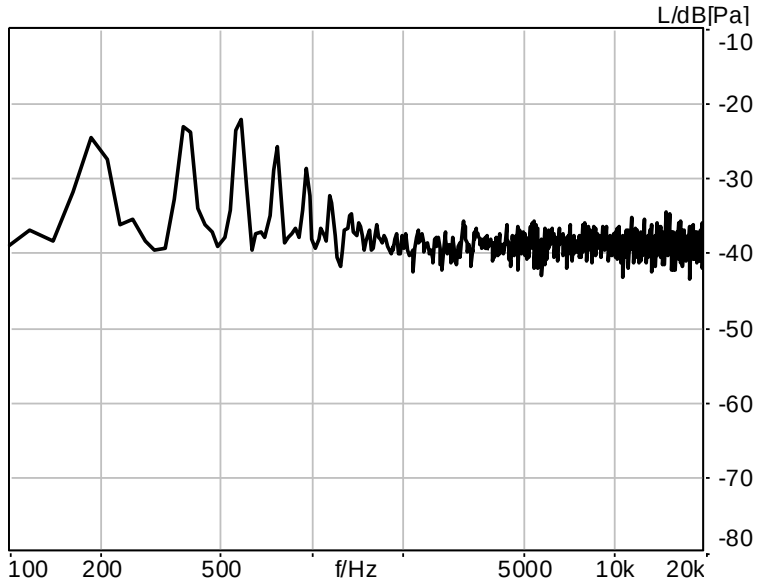


Figure 7-7 – Power density spectrum of the double-talk composite source signal (analysis window: Hanning)

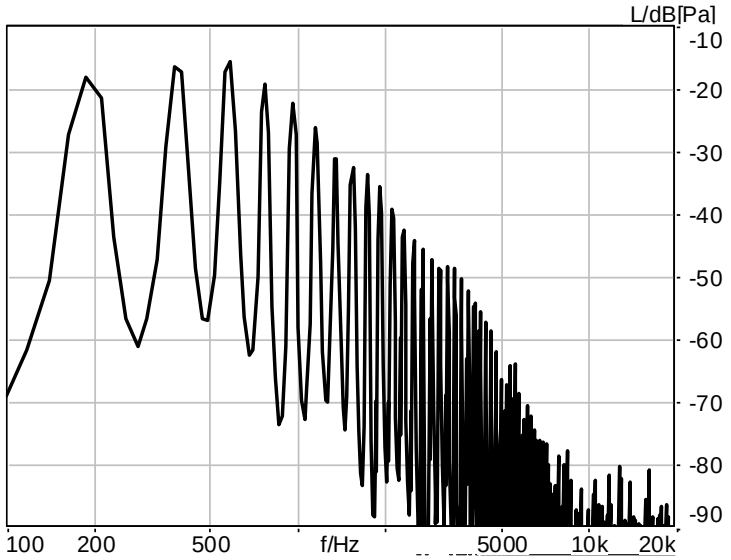


Figure 7-8 – Power density spectrum of the double-talk voiced signal (analysis window: Hanning)

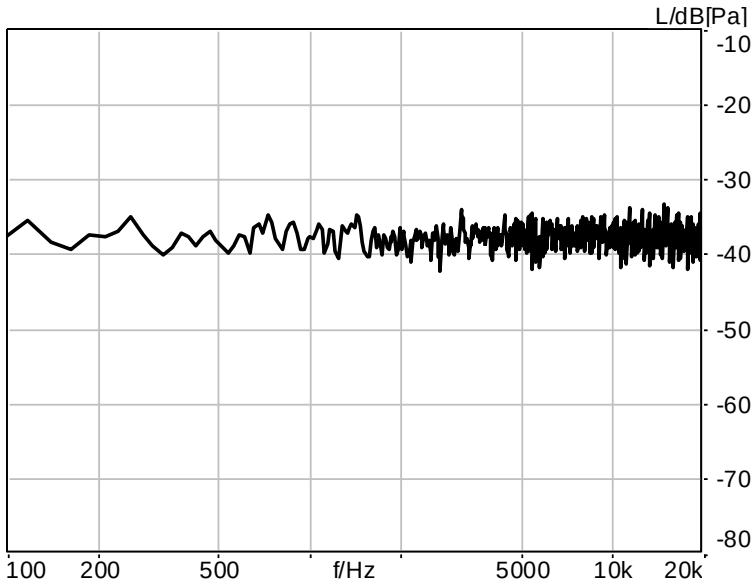


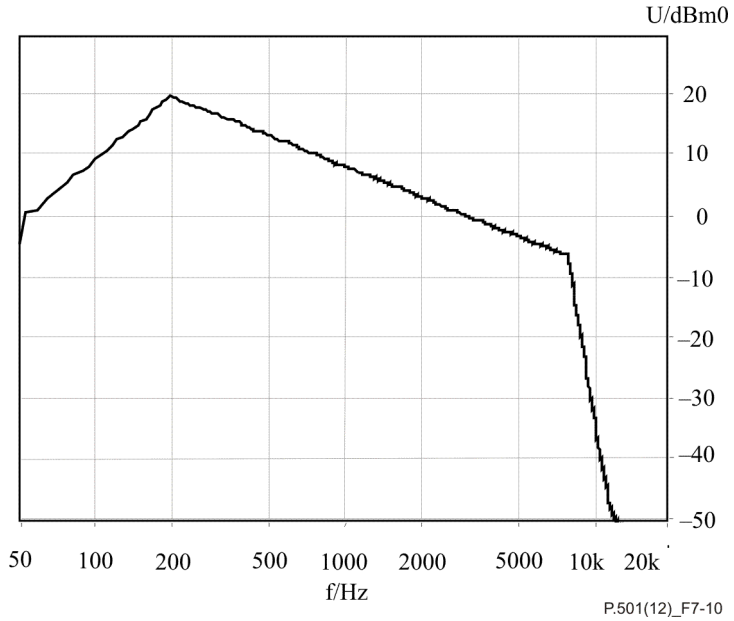
Figure 7-9 – Power density spectrum of the double-talk noise sequence (analysis window: Hanning)

7.2.1.5 Band limitation of fullband composite source signals and speech-like power density spectrum

Depending on the bandwidth of the transmission system, the CSSs should be band limited. In general, the band limitation is determined by the transmission system to which it is applied, e.g., 8 kHz for wideband transmission. The band limitation applied should be indicated.

In order to achieve a speech-like power density spectrum, a low pass filter with a slope of 5 dB/octave should be applied to the PN-sequence of the CSS for frequencies ≥200 Hz up to 10 kHz. For frequencies above 10 kHz, speech has no energy and, as a consequence, an adaptation of the power density spectrum in this frequency range is not required.

The filter that should be applied to the wideband PN-sequence of the CSS if a speech-like power density spectrum is desired for the test of wideband transmission systems is shown in Figure 7-10.



Filter corner frequencies

50 Hz	100 Hz	200 Hz	215 Hz	500 Hz	1 kHz	2.85 kHz	3.6 kHz	5 kHz	7.69 kHz	7.8 kHz	8.7 kHz
−.8 dB	9.2 dB	19.6 dB	18.8 dB	13.1 dB	8 dB	.4 dB	−1.3 dB	−3.5 dB	−6.6 dB	−7.8 dB	−20 dB

Figure 7-10 – Transfer function of the filter for band limiting the PN-sequence for wideband transmission

7.2.1.6 Narrow-band composite source signal with speech-like power density spectrum

7.2.1.6.1 Description

a) *Composite source signal for single talk*

1) Narrow-band voiced signal:

In Table 7-3, the ASCII values for the voiced signal described in clause 7.2.1.2, band-limited 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. Read Table 7-3 in columns.

Table 7-3 – ASCII values of the narrow-band 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 from 200 Hz to 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.

$$H(k) = \begin{cases} W(k) \cdot \exp(j \cdot i_k \cdot \pi); \; k = -928, \ldots, +928 \text{ except } 0, \; i_k \{+1, 0\}, \text{random } i_k = -i_{-k} \\ 0 \text{ else} \end{cases} \quad (7\text{-}9)$$

According to Equation 7-9, 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 21.5 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 152 680. 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:

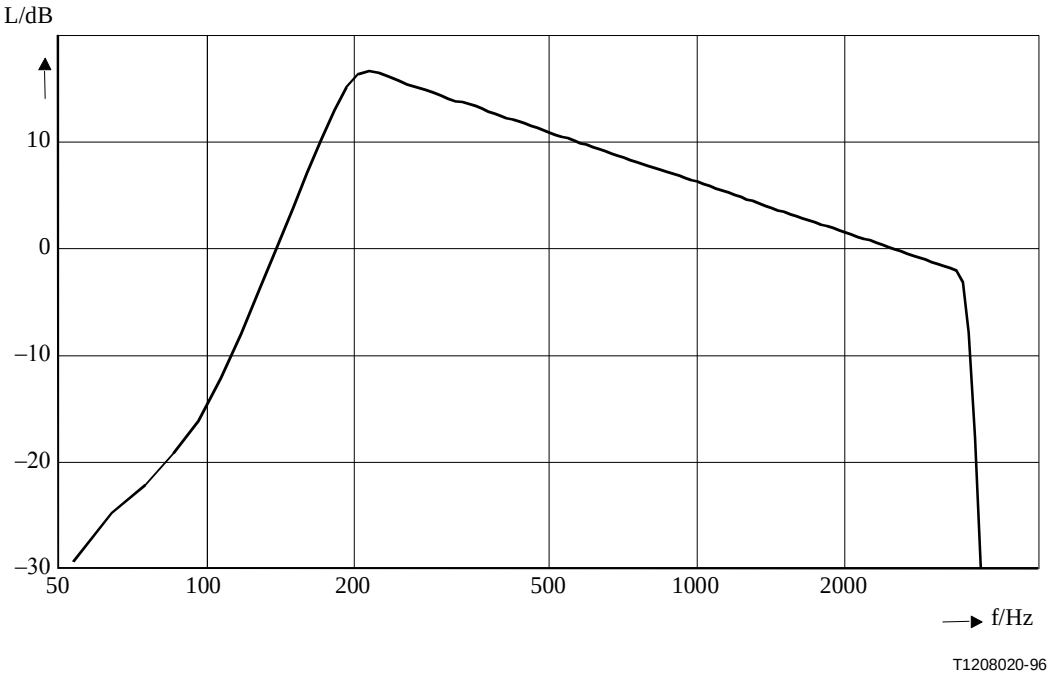
According to Equation 7-9, the time signal is calculated by inverse Fourier transformation. This sequence is repeated 1077 times to achieve a length of 200 ms for the PN-measurement sequence. The crest factor of the PN-sequence is 11 dB ± 1 dB.

According to the frequency resolution of 5.4 Hz (44.1 kHz/8192), there are 3715 FFT values in the frequency range between 0 kHz and 20 kHz. Each value $W(k)$ before filtering is 305 360. 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 narrow-band PN-sequence, the filter function shown in Figure 7-11 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 down-sampling, other sampling rates for the described sequence can be achieved. The interpolation filter used for up- and down-sampling should be close to an ideal rectangular filter. The stopband attenuation should be >60 dB, the passband ripple <±0.2 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 the 44.1 kHz sampling rate as described in Figure 7-11.



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

Figure 7-11 – Transfer function of the filter for band limiting the PN-sequence

b) Narrow-band 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, a 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 a 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 7-4. 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. Read Table 7-4 in columns.

Table 7-4 – ASCII-values for the narrow-band 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

Table 7-4 – ASCII-values for the narrow-band double-talk voiced signal

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, band limited at 20 kHz. The crest factor of the signal is 12 ± 1 dB. The RMS value of the band-limited random noise is chosen to be the same as the one for the voiced signal.

In order to band-limit the random noise between 200 Hz and 3.6 kHz, the filter function shown in Figure 7-11 is used. This ensures the same RMS value for the band-limited random noise.

See Figures 7-12 to 7-15.

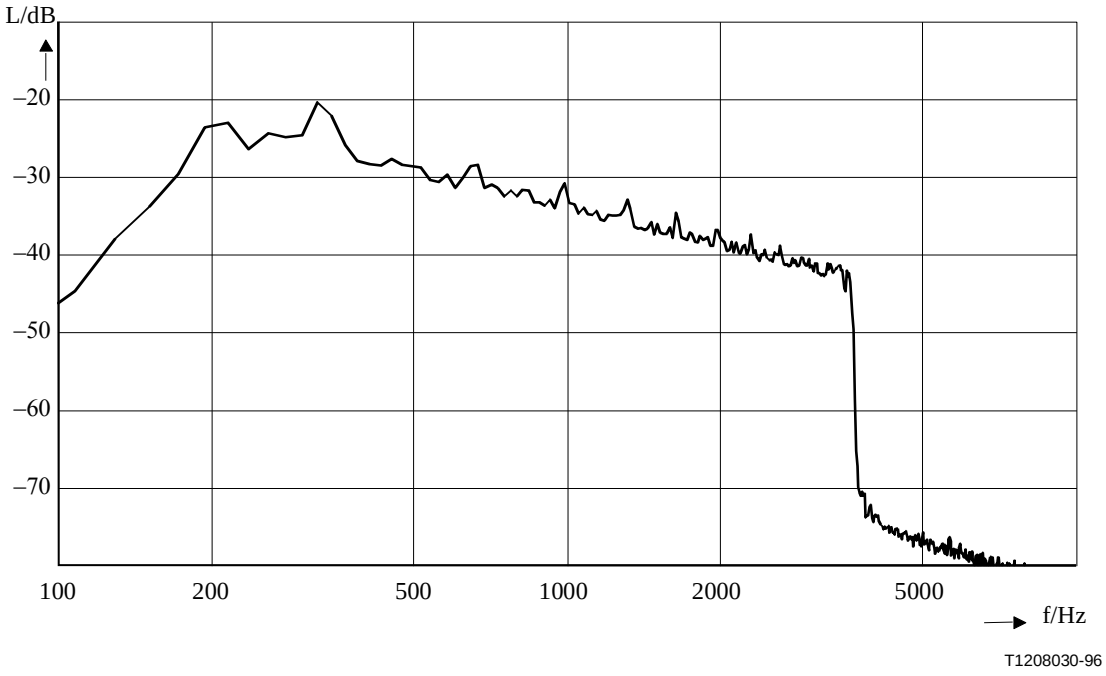


Figure 7-12 – Power density spectrum of the narrow-band CS signal (single-talk signal, analysis window: Hanning)

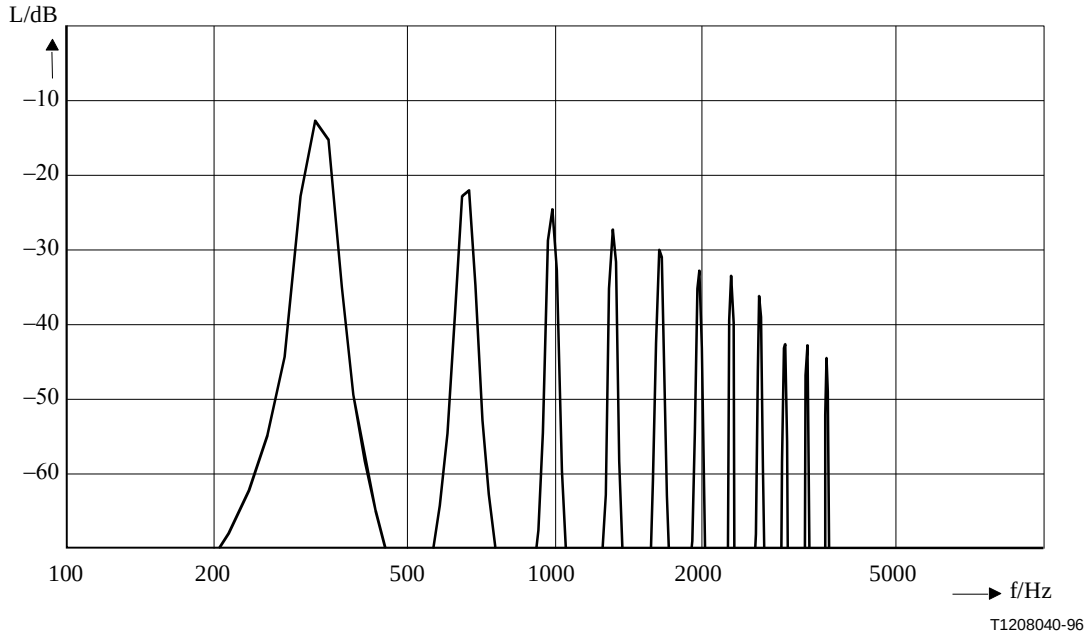


Figure 7-13 – Power density spectrum of the narrow-band voiced signal (single-talk signal, analysis window: Hanning)

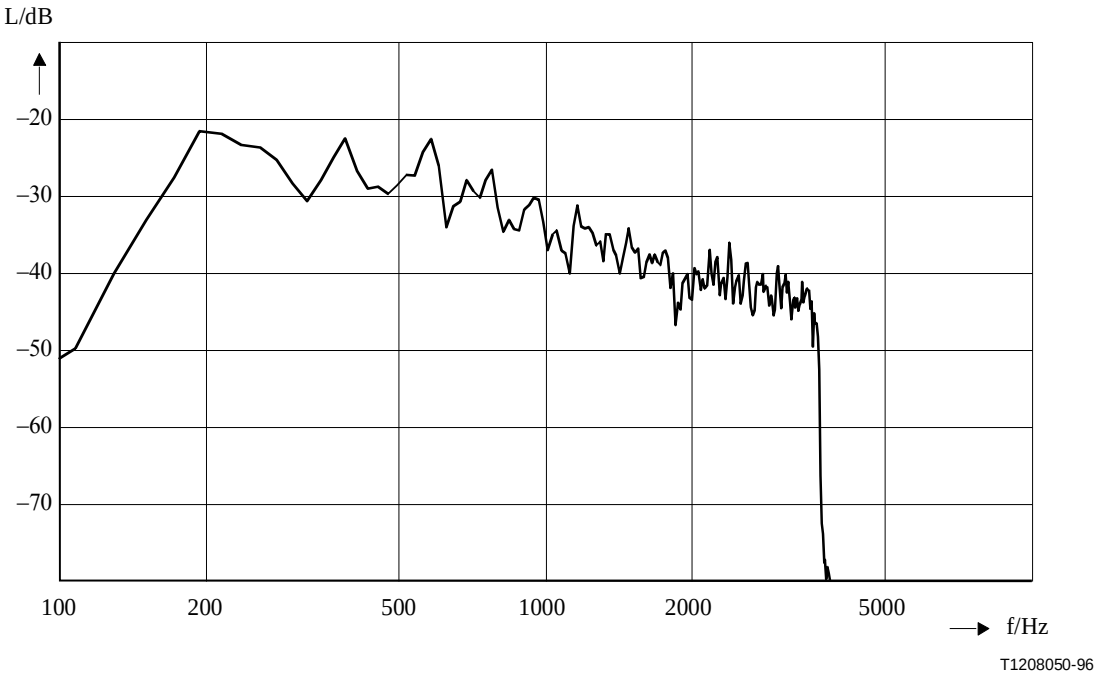


Figure 7-14 – Power density spectrum of the narrow-band double-talk CSS (analysis window: Hanning)

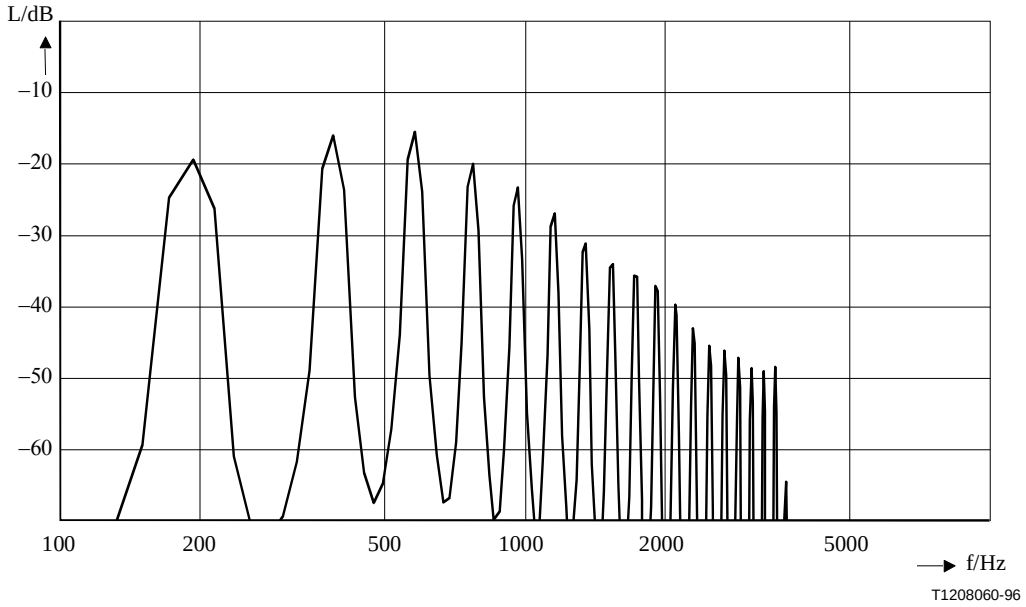


Figure 7-15 – Power density spectrum of the narrow-band double-talk voiced signal (analysis window: Hanning)

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

7.2.1.6.2 Application

The application of the narrow-band CSSs for single talk as well as for double talk is for all measurements where narrow-band systems need to be measured in non-linear and time variant operation 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 narrow-band CSS for single-talk measurements shall be used. In the case of measurements in double-talk conditions, the double-talk signal shall be used in the double-talk direction, whereas the single-talk signal is fed in the far-end direction.

7.2.2 Speech-like modulated noise

7.2.2.1 Description

As pointed out in clause 7.1.3, the MTF can be used to measure the speech intelligibility of a system. By modulating octave band filtered noise, the MTF can be obtained in different octave bands. With correct weighting of the modulation indices in each octave band and over different modulation frequencies, a speech transmission index ($0 \leq STI \leq 1$) that has a 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 to be 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.

7.2.2.2 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 generally, a broad class of non-linear distortions;
– automatic gain control;

– reverberation.

7.2.3 Composed signals in frequency (probe tone technology)

7.2.3.1 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.

7.2.3.2 Application

Typically, this method is used to determine average (long-term) characteristics of a system. If, for example, the average frequency response under realistic operating conditions (presence of reverberation and noise) is to be measured, a series of linear sinusoidal sweeps may be used. Often, the analytical signal is a single tone that 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.

7.2.4 Voice-like composed signals in frequency

7.2.4.1 Description

The evaluation of systems during double talk requires the separation of the double-talk sequences after passing the system under test. The signal described below emulates orthogonal voiced sounds of speech taking into account the following speech properties:

– excitation signal provides typical speech frequency components;

– the voiced sound modulation is voice-like, in frequency and amplitude;

– the signals are orthogonal.

The signal generation block is shown in Figure 7-16.

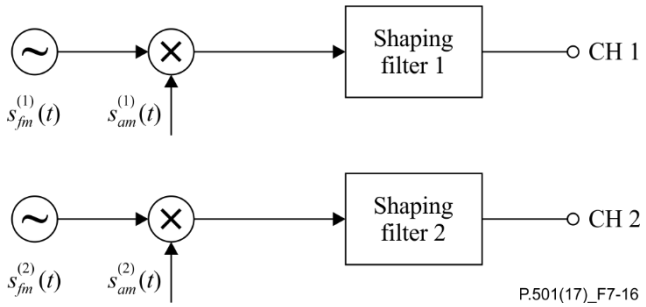


Figure 7-16 – Two channel test signal generation for double-talk evaluations based on AM-FM signals

$$s_{\text{fm}}^{(1,2)}(t) = \sum_n A^{(1,2)}(n) \cdot \cos[2\pi f_0^{(1,2)}(n) \cdot t + \mu_{\text{fm}}^{(1,2)}(n) \cdot \sin\left(2\pi f_{\text{fm}}^{(1,2)} \cdot t\right)] \qquad \text{n = 1,2,...}$$

$$\text{where } \mu_{\text{fm}}^{(1,2)}(n) = \frac{\Delta f^{(1,2)}(n)}{f_{\text{fm}}^{(1,2)}}$$

$$s_{\text{am}}^{(1,2)}(t) = [1 + \mu_{\text{am}}^{(1,2)} \cdot \cos\left(2\pi f_{\text{am}}^{(1,2)} \cdot t\right)]$$

Three parameters are chosen in a frequency-independent manner: $f_{\text{fm}}^{(1,2)} = 1$ Hz, $\mu_{\text{am}}^{(1,2)} = \frac{2}{3}$ and $f_{\text{am}}^{(1,2)} = 3$ Hz. The remaining two parameters are given in Table 7-5. Both shaping filters are identical: a low pass with a cut-off frequency of 250 Hz and a slope of 5dB/octave.

The amplitudes $A^{(1,2)}(n)$ that determine the signal levels may be chosen according to the application. Average measurement levels may be chosen, i.e., to −4.7 dB_{Pa} [send (SND)] and −20 dB_V [receive (RCV)] for terminal testing. The test signal may be embedded in speech or speech-like sequences.

Table 7-6 lists the corresponding properties for wideband applications.

This signal is used for super-wideband and fullband applications as well.

NOTE – In human speech, energy above 7 kHz is present mostly during non-voiced parts and to a lesser extent during voiced parts. The AM-FM signal simulates the harmonic content of voiced parts of human speech and its 7 kHz frequency range is considered sufficient also in super-wideband and fullband, for the specific purpose of separating simultaneous talkers, e.g., during double-talk evaluation.

Table 7-5 – Properties of a specific double-talk sequence, modulated in amplitude and frequency for narrow-band applications

Sending direction		**Receiving direction**	
$f_0^{(1)}$ **(Hz)**	$\pm\Delta f^{(1)}$ **(Hz)**	$f_0^{(2)}$**(Hz)**	$\pm\Delta f^{(2)}$ **(Hz)**
250	±5	270	±5
500	±10	540	±10
750	±15	810	±15
1000	±20	1080	±20
1250	±25	1350	±25
1500	±30	1620	±30
1750	±35	1890	±35
2000	±40	2160	±35
2250	±40	2400	±35
2500	±40	2650	±35
2750	±40	2900	±35
3000	±40	3150	±35
3250	±40	3400	±35
3500	±40	3650	±35
3750	±40	3900	±35
(4000	±40)	(4150	±35)

Figures 7-17 and 7-18 show time signal and frequency distributions when applying such a test signal.

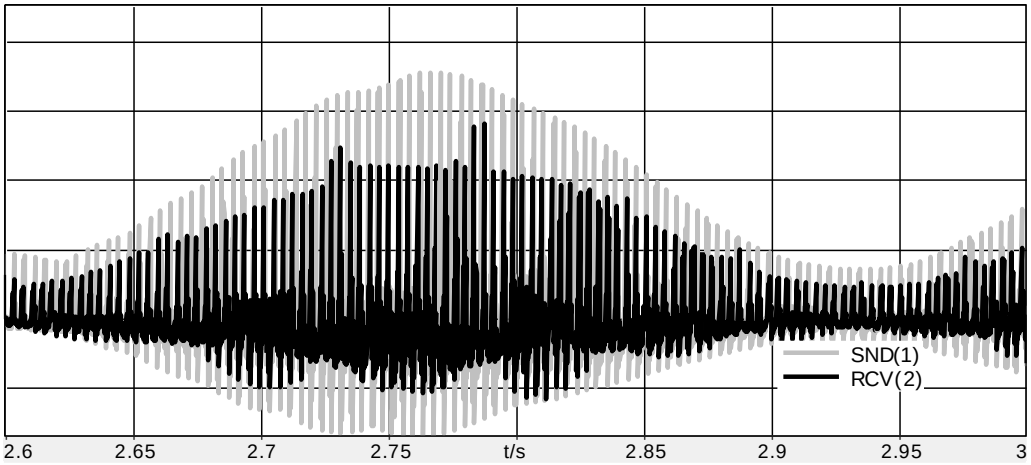


Figure 7-17 – Specific double-talk signal, time domain; Light grey: Signal in sending direction; Black: Signal in receiving direction

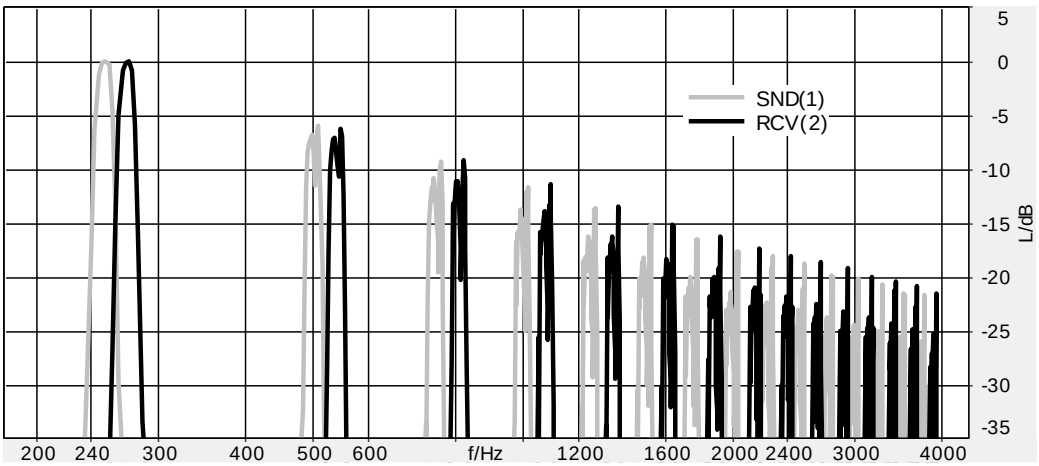


Figure 7-18 – Spectrum of the specific double-talk signal; Light grey: Signal in sending direction; Black: Signal in receiving direction

Table 7-6 – Properties of a specific double-talk sequence, modulated in amplitude and frequency for wideband applications

Sending direction		**Receiving direction**	
$f_0^{(1)}$ **(Hz)**	$\pm\Delta f^{(1)}$ **(Hz)**	$f_0^{(2)}$**(Hz)**	$\pm\Delta f^{(2)}$ **(Hz)**
125	±2.5	180	±2.5
250	±5	270	±5
500	±10	540	±10
750	±15	810	±15
1 000	±20	1 080	±20
1 250	±25	1 350	±25
1 500	±30	1 620	±30
1 750	±35	1 890	±35

Table 7-6 – Properties of a specific double-talk sequence, modulated in amplitude and frequency for wideband applications

Sending direction		Receiving direction	
$f_0^{(1)}$ (Hz)	$\pm\Delta f^{(1)}$ (Hz)	$f_0^{(2)}$(Hz)	$\pm\Delta f^{(2)}$ (Hz)
2 000	±40	2 160	±35
2 250	±40	2 400	±35
2 500	±40	2 650	±35
2 750	±40	2 900	±35
3 000	±40	3 150	±35
3 250	±40	3 400	±35
3 500	±40	3 650	±35
3 750	±40	3 900	±35
4 000	±40	4 150	±35
4 250	±40	4 400	±35
4 500	±40	4 650	±35
4 750	±40	4 900	±35
5 000	±40	5 150	±35
5 250	±40	5 400	±35
5 500	±40	5 650	±35
5 750	±40	5 900	±35
6 000	±40	6 150	±35
6 250	±40	6 400	±35
6 500	±40	6 650	±35
6 750	±40	6 900	±35
7 000	±40		

In order to extract the signal needed for evaluation from the double-talk signal measured (e.g., to extract an echo signal component), either a specific filter setting or a specific post-processing of the FFT analysis is required, since the spectrum of the signal, as well as of the double-talk signal, is a kind of comb filter spectrum where a specific modulation is applied. The mid-frequency, $f_0(n)$, of any frequency component, the corresponding frequency modulation, $\Delta f(n)$, as well as the filter shapes or the windowing function of the Fourier transformation need to be taken into account. If the filter approach is used, the bandwidth of each filter should be constructed in such a way that:

$$f_{\text{lower}}(n) = f_0(n) - \Delta f(n)$$

$$f_{\text{upper}}(n) = f_0(n) + \Delta f(n)$$

The same applies to analyses derived from Fourier transformations of the measured echo signal. Here, the frequency "smearing" effect of the windowing function needs to be taken into account. In order to have a sufficient separation between the echo signal and the double-talk signal in the low frequency domain, a minimum FFT length of 8 k (sampling rate 48 kHz) should be chosen.

The principle of the analysis is shown in Figure 7-19 for the extraction of an echo signal. The signal extracted may also be the double-talk signal in order to evaluate signal level variations during double talk.

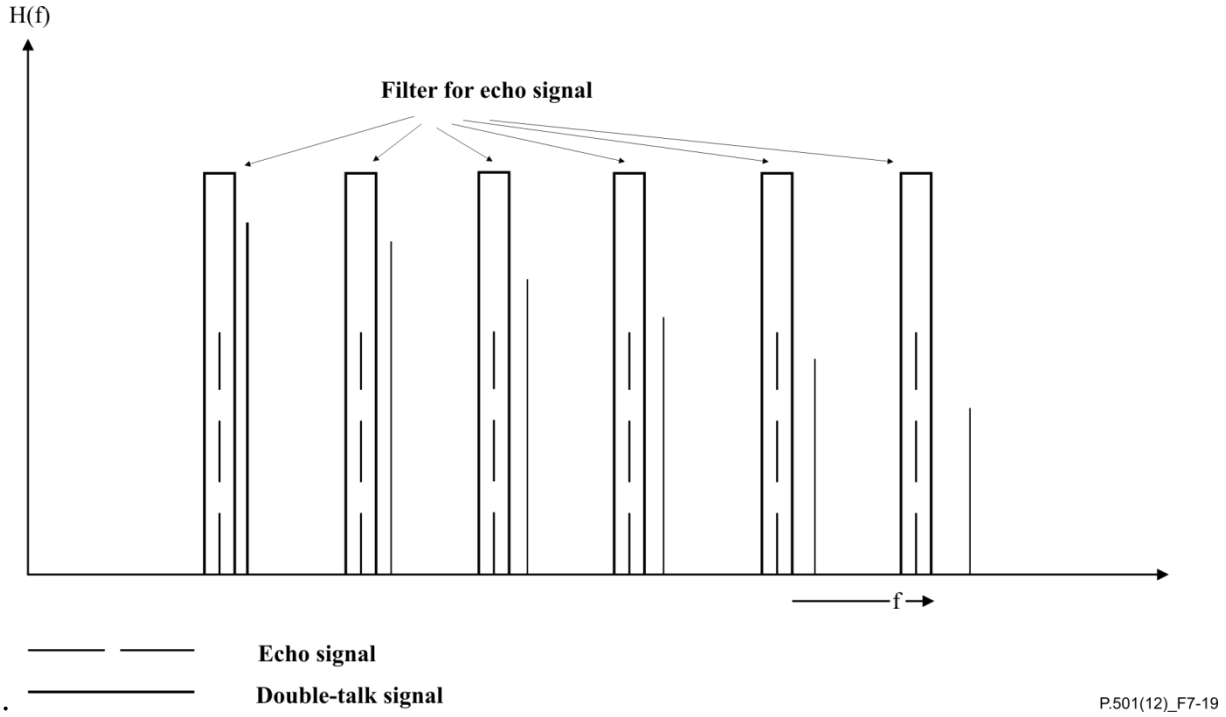


Figure 7-19 – Extraction of the echo components of the double-talk signal (schematic)

After applying the appropriate filter or FFT analysis, the post-processing according to the task to be performed is possible. For example, in order to determine the echo loss using this methodology, the measured signal is referred to the excitation signal. Weighted TCL, TCL_w, may be calculated according to [ITU-T G.122].

NOTE – The resolution in the low frequency domain requires either a filter with a long impulse response or a FFT analysis with high resolution. Hence, the time variant echo loss in the low frequency domain can be determined only with a certain amount of accuracy. A good separation between echo signal and double-talk signal is possible only up to a certain extent.

7.2.4.2 Application

Typically, this method is used for evaluations during double talk where evaluation is required under real double-talk conditions while the double-talk signal is present during the analysis. The typical applications are the evaluations of echo loss, echo loss variation or level variation of the double-talk signal under double-talk conditions. If a time variant echo loss or level variation during double talk is to be evaluated, the echo level (double-talk signal level) calculated from the components as described above or the echo spectrum (double-talk signal spectrum), displayed as a spectrogram, can be used.

In any case, the signals are fed simultaneously into the far-end as well as into the near-end direction. Any delay in the system under test should be taken into account for the signal insertion, as well as for the signal analysis. If the measured signal is referred to the input signal, the input signal should be time aligned by taking into account the actual delay between the measured and input signals.

7.2.5 Complex composed signals

7.2.5.1 Simulated speech generator

7.2.5.1.1 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 7-20. The resultant signal is shaped to approximate the long-term frequency spectrum of speech, as shown in Figure 7-21. Figure 7-22 shows the amplitude distribution of the simulated speech generator.

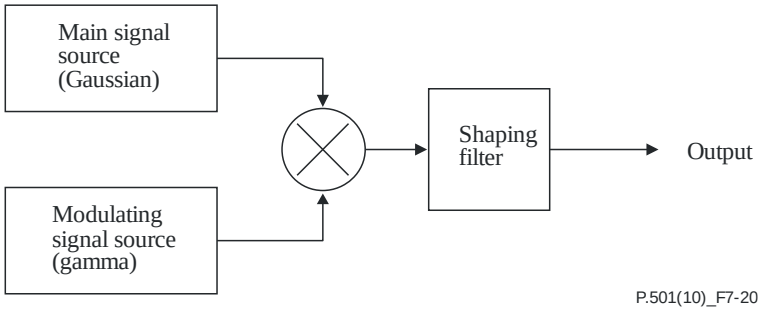


Figure 7-20 – Block diagram of simulated speech generator

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 randomized within and between the segments uniformly from 0° to 360°, in order to randomize 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 eliminate any transients that 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 that 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 waveform 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.

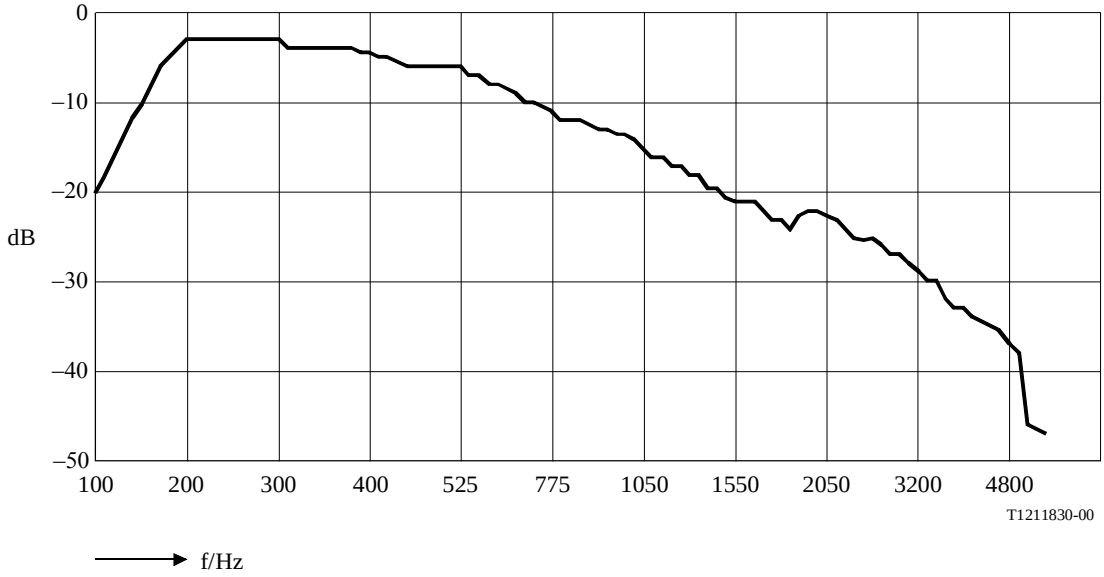


Figure 7-21 – Spectrum generated by the simulated speech generator

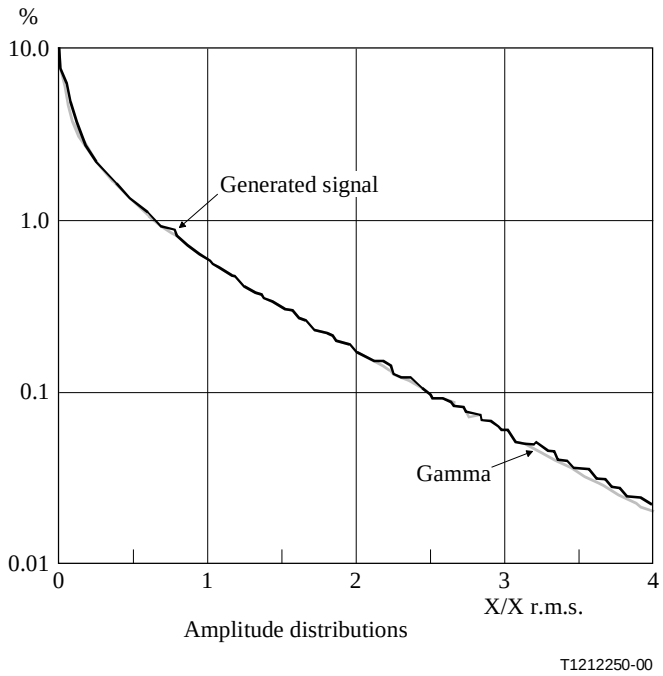


Figure 7-22 – Amplitude distribution of the simulated speech generator

7.2.5.1.2 Application

The SSG signals may be used if a "typical" speech sequence is required for the measurement. Similar to the ITU-T 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, such as frequency responses or loudness ratings, need to be calculated from the measurements.

7.2.5.2 Artificial voice [ITU-T P.50]

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

7.2.5.3 Artificial conversational speech [ITU-T P.59]

[ITU-T P.59] describes an artificial conversational speech signal, of complex composition, offering talk spurts, etc., in addition to a double-talk sequence.

7.2.5.4 Speech-model process controlled by discrete Markov chains

7.2.5.4.1 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) [b-Serafat, 1996] (where SIRP stands for spherically invariant random process). As mentioned in [b-Serafat, 1996] and [b-Halka, 1993], there were some deviations between the long-term spectrum and the spectrum of the signal envelope of MSIRP and those of natural telephone speech (300–3400 Hz). Due to improvements in the model, the spectra of the signal envelopes 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)

Figure 7-23 illustrates the generation procedure of MSMP as a block diagram. As Figure 7-23 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 PDF and the autocorrelation of the resulting process separately.

The time variant properties of the resulting process are controlled by trained Markov chains (mcs) 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 MC that is responsible for the pitch value in this frame (mc-pitch). This trained MC 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 generalized MC is controlled that produces a natural index sequence for choosing one of 50 formant filters for this frame (mc-formant). This generalized MC (mc-formant) works as the hidden part of a hidden Markov model, 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 low-pass 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 point, we have to identify a special property of product processes: the amplitude PDF of the product processes is generally symmetric. However, 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 $n(t)$. A random process with the desired amplitude PDF is thus obtained. 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:

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

A Gaussian process is thus obtained that 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 that 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 the 16th order. The coefficients are obtained from a codebook, which is optimized with the generalized Lloyd algorithm [b-Linde, 1980] 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 7-24 shows a comparison of spectra of speech and MSMP while Figure 7-25 illustrates a comparison of the PDFs of speech and MSMP. Figure 7-26 is a comparison of the spectra of signal envelopes of speech and MSMP.

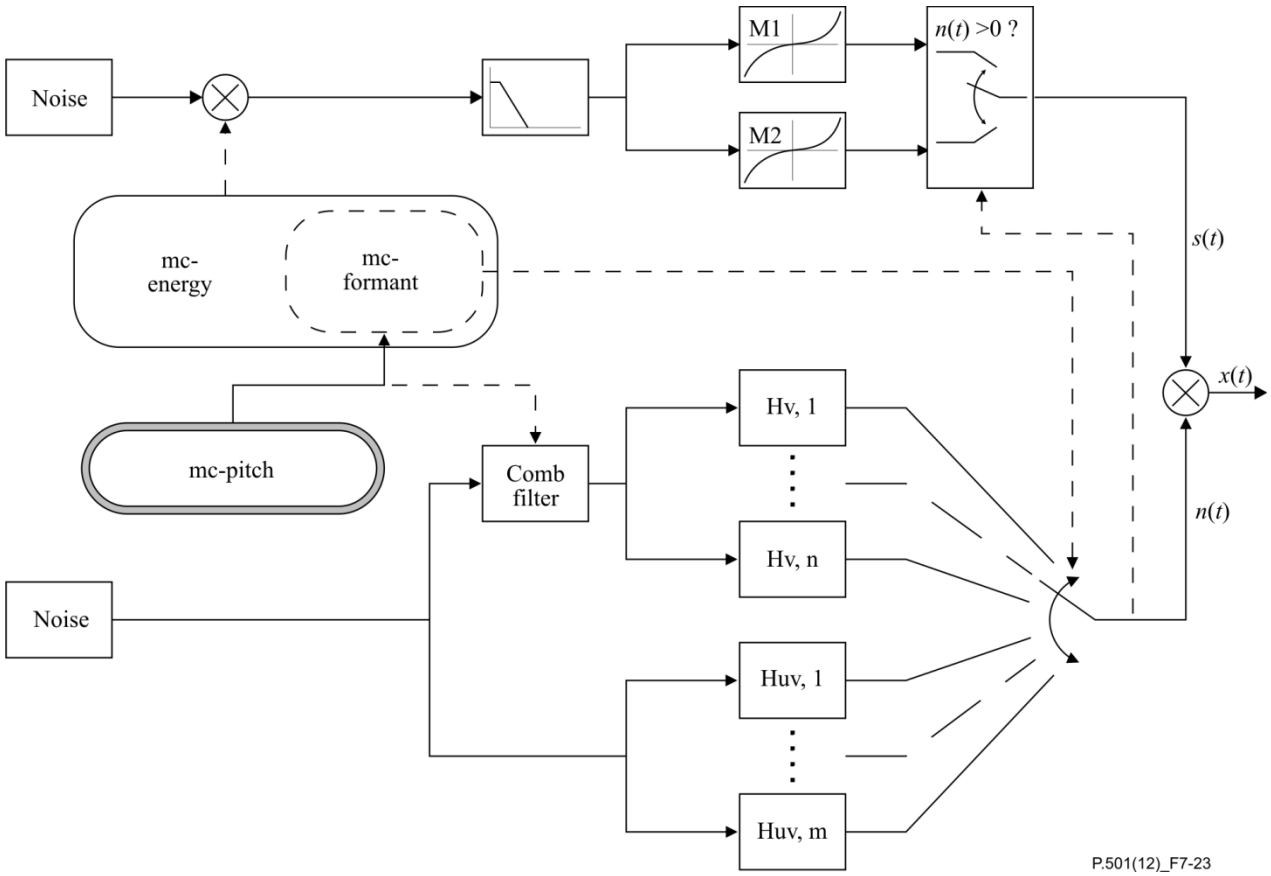


Figure 7-23 – Block diagram of the MSMP generator

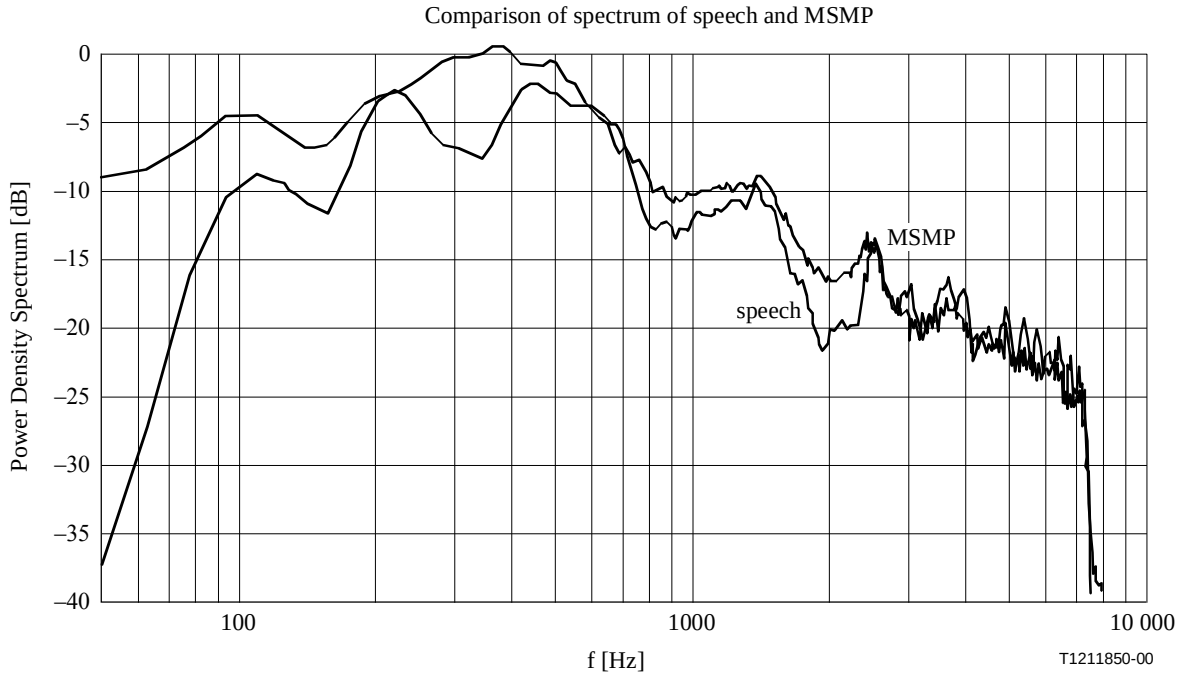


Figure 7-24 – Comparison of the spectra of speech and MSMP

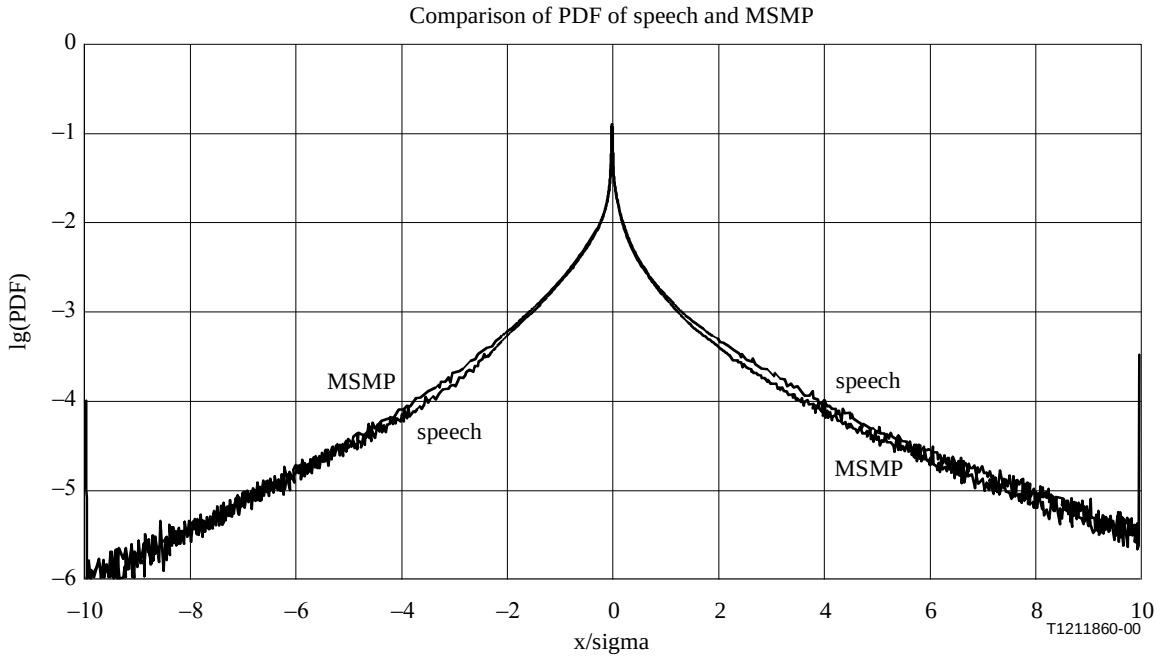


Figure 7-25 – Comparison of the PDFs of speech and MSMP

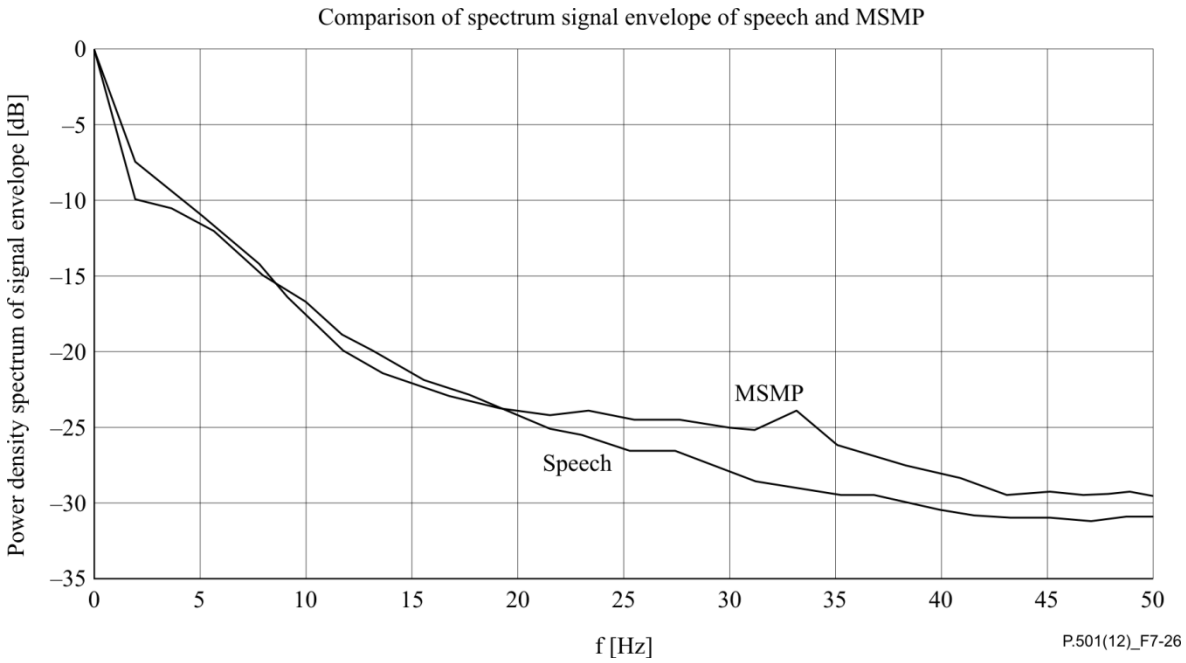


Figure 7-26 – Comparison of the spectra of signal envelopes of speech and MSMP

7.2.5.4.2 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 MCs. The MSMP allows prescription of not only the long-term spectrum and PDF, but also inclusion of 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 (such as male/female and weak/strong variation of characteristics).

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

7.3 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 traditional concepts of frequency response, distortion and signal-to-noise ratio are no longer directly applicable or even relevant. However, behaviour of the device under test can be investigated using real speech and applying known analysis techniques that also could be used in conjunction with artificial or speech-like signals. Another option is to judge the overall quality by a model of human auditory perception. With such a model, a "perceptual frequency response function" or special measures, e.g., based on psychoacoustic parameters (perceptual distortion measures), can be defined. However, such methods are not part of this Recommendation.

In this clause, a variety of speech sequences for different types of application are given. Since these sequences take into account different conversational situations, it could be desirable to concatenate single instances of all test sequences starting with conditioning sequences and to do post-analysis of the pre-recorded material. Such a procedure does not only ensure a close to real conversation test, it also in many cases will lead to good repeatability and to significantly reduced testing time. However, it should be noted that there are exceptions to this procedure, such as testing of the adaptation behaviour of echo cancellers, noise cancellers and the investigation of initial switching.

7.3.1 Reference speech samples

The reference sample set includes six female and six male native British English-speaking adults, speaking from the Harvard list of phonetically balanced sentences [b-IEEE No.297].

NOTE 1 – If there are differences in test results received when using different languages, the test result received with British English shall be the normative one.

The sample set of the first contributor include pulse code modulation (PCM), 16 bit, 48 kHz monaural formatted recordings of individual sentences from four male and four female speakers recorded in anechoic conditions with a free-field microphone. The samples provided by the second contributor, including two male and two female speakers with the same format, have similar auditory characteristics to the ones from the first contributor. These samples were provided as sample-pairs with digital zeros between, from which individual sentences or words were extracted.

The range of spectra for sentence samples chosen for the single-talk sequence are shown for each female speaker in Figure 7-27 and each male speaker in Figure 7-28. The between-sample average crest factor is 17.42 dB (min.; 14.20 dB, max.; 21.79 dB) with an average activity factor of 95.77% (min.; 93.72 %, max.; 98.45%) measured with the default settings of the sv56demo application provided with [ITU-T G.191].

NOTE 2 – It should be noted that, as seen in Figures 7-27 and 7-28, the samples of the first contributor exhibit an increase in noise around ~21 kHz. It can be speculated that this is noise coming from the computer monitor used to present the sentences to speakers. In practice, this is not seen to be an issue, as it is mostly masked during speech activity.

The spectra of the concatenated samples of the same female and male samples are shown in Figure 7-29, along with the spectra of the 16 kHz-sampled male and female artificial speech samples provided with [ITU-T P.50].

Finally, the spectrum of concatenation of all the new samples is shown in Figure 7-30, again with the 16 kHz-sampled male and female samples provided with [ITU-T P.50].

The original samples had shown some low frequency energy below the fundamentals of the speech. This was an inherent part of the samples from the recording environment, e.g., heating, ventilation and air conditioning (HVAC) or external vehicular noise. For some applications, this may be problematic, e.g., automated level normalization may be biased by low frequencies or direct current (DC) components. For this reason, a high-pass filter removing frequency components close to 0 Hz was applied (see clause I.1).

These samples, as well as other sentences and individual words from the same speakers, have been edited into sample sequences defined in clauses 7.3.2 to 7.3.8.

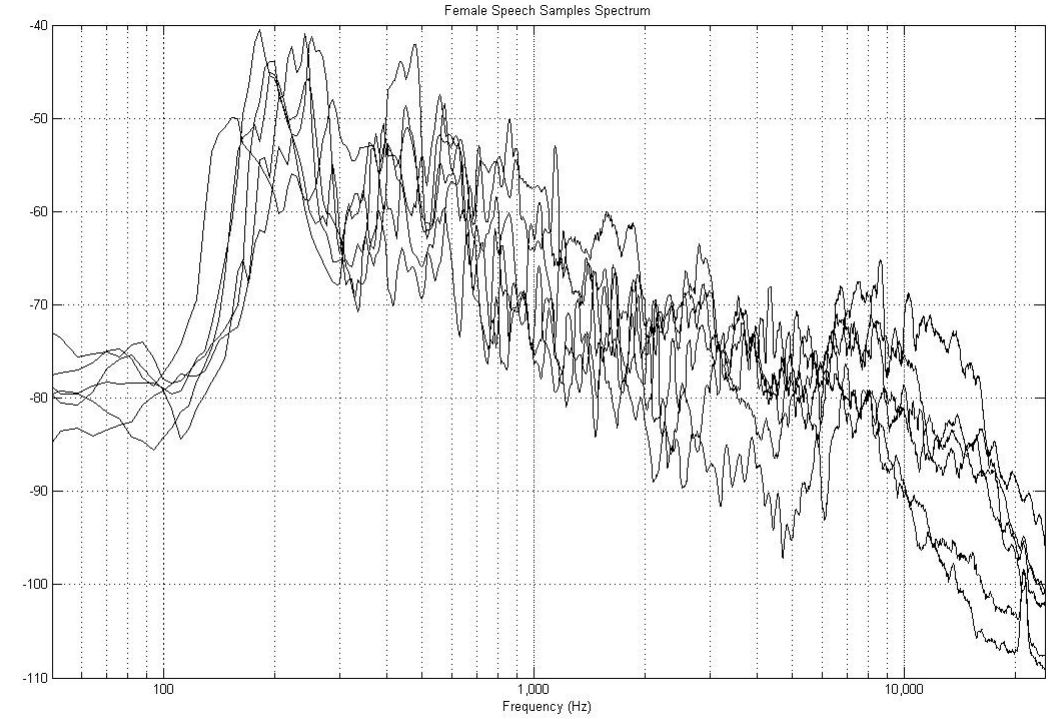


Figure 7-27 – Spectra (sampled at 1/24th octave) of individual female speech samples

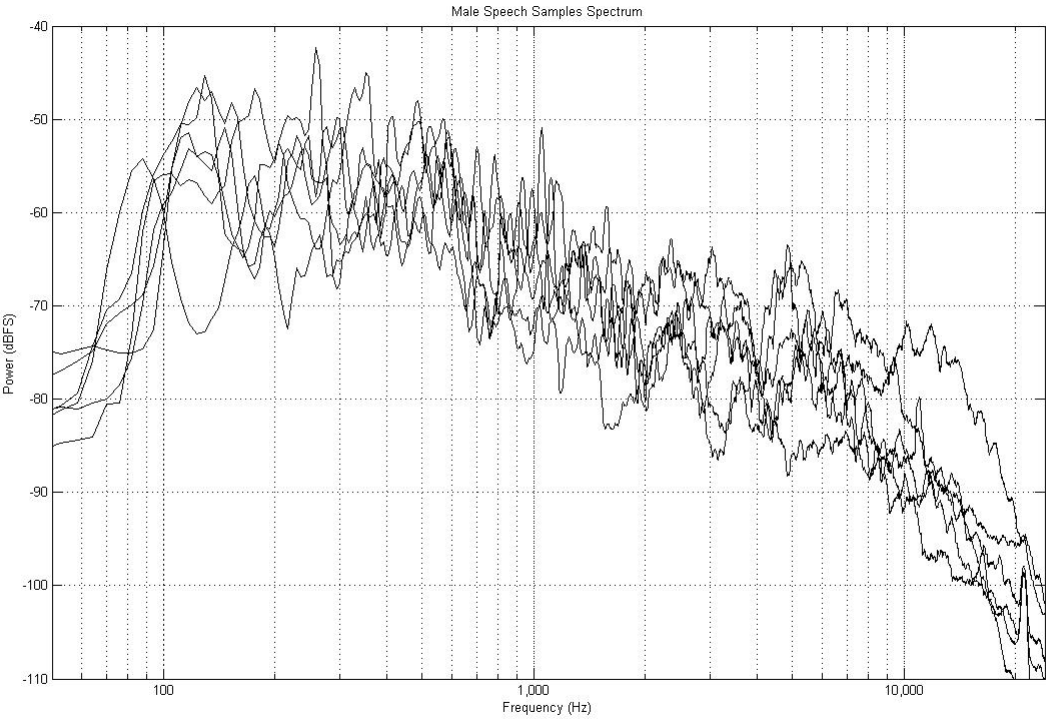


Figure 7-28 – Spectra (sampled at 1/24th octave) of individual male speech samples

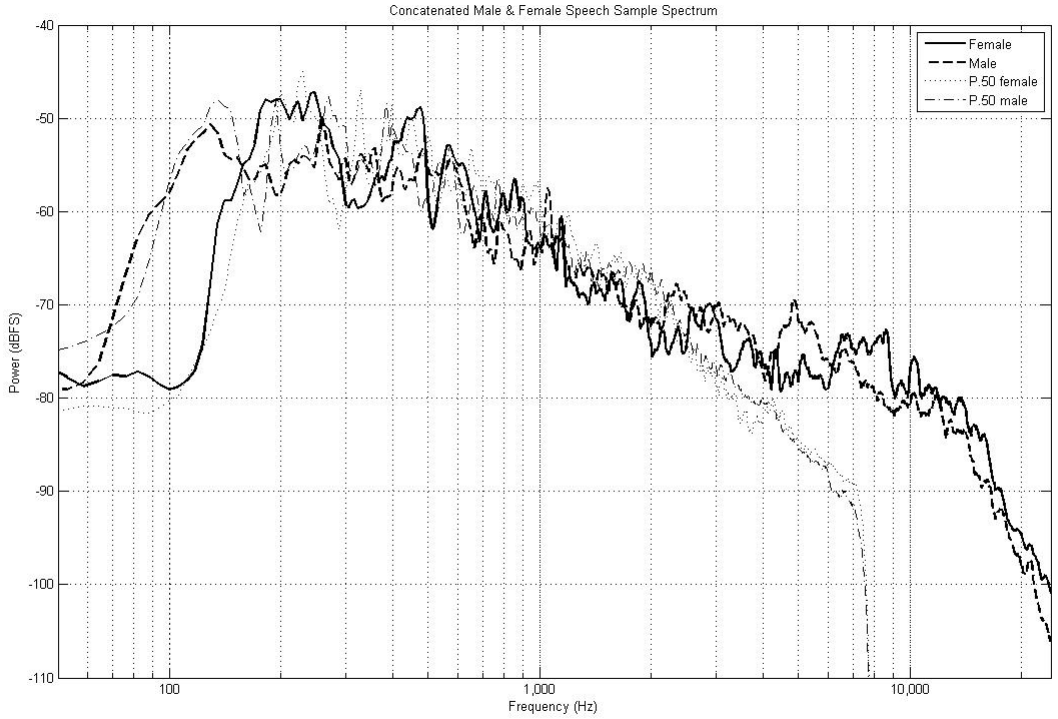


Figure 7-29 – Spectra (sampled at 1/24th octave) of all female and all male speech samples along with separate male and female ITU-T P.50 artificial voice

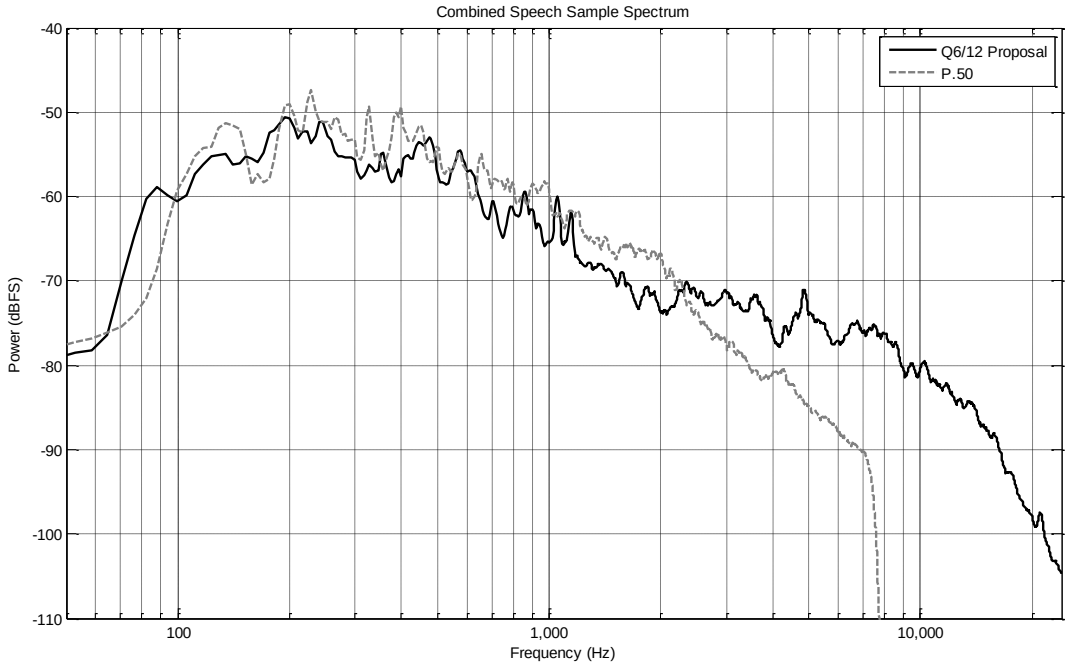


Figure 7-30 – Spectra (sampled at 1/24th octave) of all speech samples along with ITU-T P.50 artificial speech

7.3.2 Speech signals for single-talk testing

7.3.2.1 Description

A typical "single-talk" sequence of sentences spoken by all 12 speakers from the reference speech samples is shown in Figure 7-31. The sequence, lasting ~35.4 s, is created using three males (M1–M3), three females (F1–F3), the remaining three males (M4–M6) and the remaining three females (F4–F6), with each speaker speaking a unique sentence. A silence period of 0.5 s is inserted between each sentence as well as at the beginning and end of the sequence.

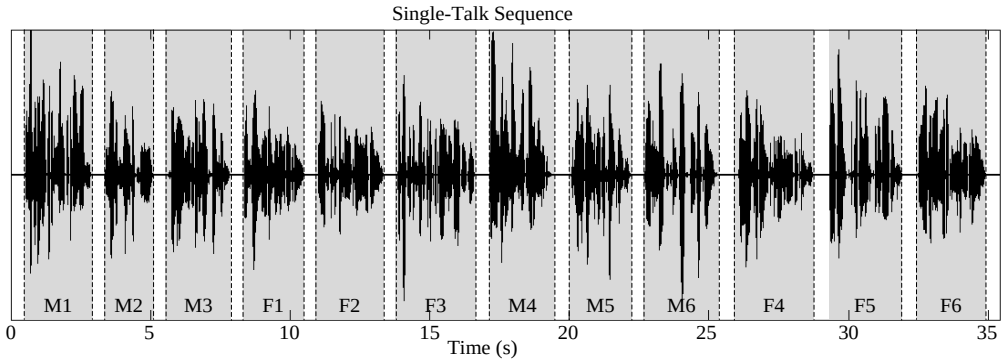


Figure 7-31 – Single-talk test sequence using six male (M) and six female (F) speakers with a pause of 0.5 s at the beginning, end and between individual samples (total duration of ~35.4 s)

NOTE – If a reduced measurement time is desired, the test sentences might be concatenated with no pause between them.

7.3.2.2 Application

This sequence is intended for use in the determination of long-term transmission characteristics under single-talk conditions, such as frequency response or loudness rating. However, the sequence may also be applied, for example, to the investigation of automatic gain control behaviour or compression.

The sequence may be used if a typical speech sequence is required for the measurement. This is typically the case for systems that behave non-linearly, are time variant and may react to artificial signals differently from speech.

The test sequence may be shortened if a shorter measurement duration is required and it is confirmed that the system reaction is either not different to the system reaction when applying the complete test sequence or the overall properties given by the test sequence are of minor importance to the measurement. Under such conditions, the first six sentences of the sequence may be sufficient.

If the system reaction to different speakers is of specific interest, the analysis could be made on a per-sentence basis. Such application may better show the time variant system reaction. However, due to the insufficient spectral and temporal representation of the language properties when using just one sentence for analysis, differences compared with the average behaviour must be expected, even for LTI systems under test.

If the measured signal is referenced against the input signal, the input signal should be time aligned by taking into account the actual delay between the measured and input signals.

7.3.3 Compressed speech signals for testing

7.3.3.1 Description

A dynamically compressed version of the real speech sequence for single-talk measurements of clause 7.3.2 is also provided. This version applies pre-equalization and hard limiting of the signal to achieve highly uniform energy across the audio frequency band. See Figure 7-32.

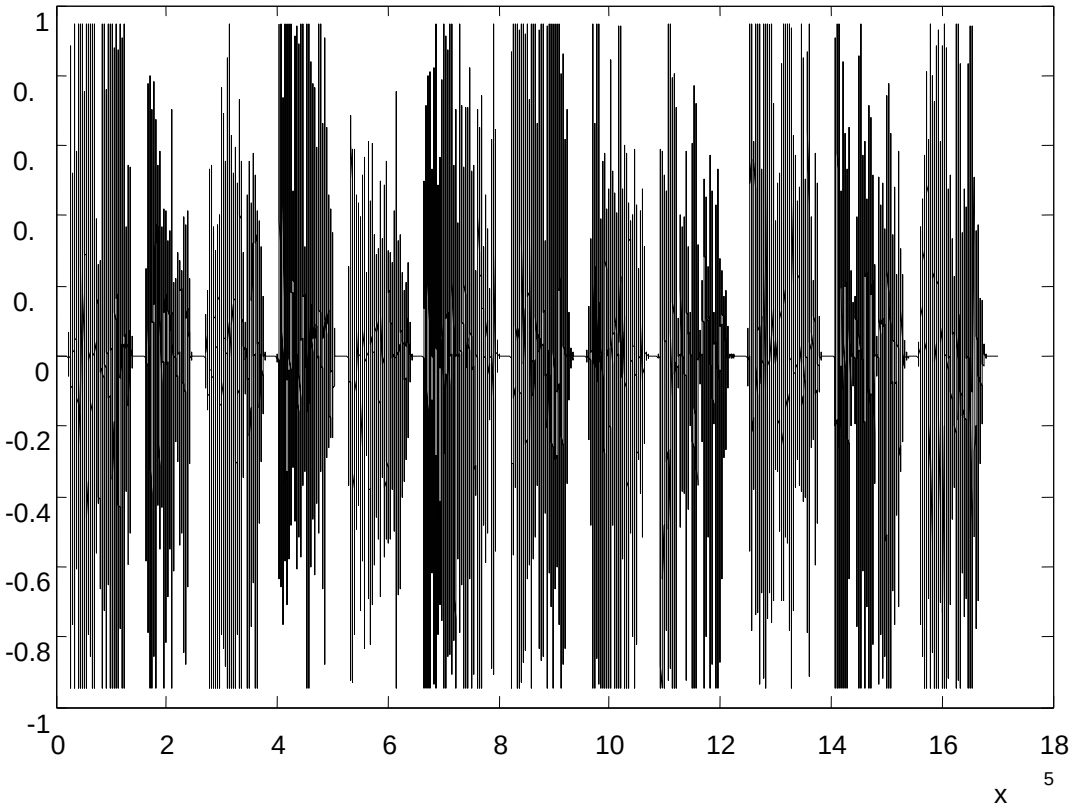


Figure 7-32 – Time history of the dynamically compressed, spectrally enhanced real speech signal

7.3.3.2 Application

The signal can be applied for one-way, echo cancellation tests where high signal-energy is necessary for adequate echo loss measurements, and artificial test signals fail to work properly with the speech coding or speech enhancement under test. Appropriate filtering should be used for narrowband, wideband and super wideband applications.

Although this signal is based on real-speech, and suitable for most applications, the artificially generated spectral distribution is not representative of a long-term real-speech spectral distribution, and it may represent unexpected behaviour for certain classes of speech codecs and speech enhancement algorithms.

7.3.4 Short words for activation (temporal) tests

7.3.4.1 Description

A short word for activation, representing a typical, short utterance of people in real conversations is shown in Figure 7-33. The sequence consists of the word "five" spoken by M1.

Figure 7-34 shows an activation sequence. It consists of the concatenation of the sequence shown in Figure 7-33, adding a pause of about 0.4 s and increasing the amplification of each following word by 1 dB. In the beginning and at the end a pause of 0.5 s added. By this process, an activation sequence is created covering 21 dB of dynamic range.

NOTE – The signal level is determined for each utterance individually using, for example, the default settings of the sv56demo application provided with [ITU-T G.191].

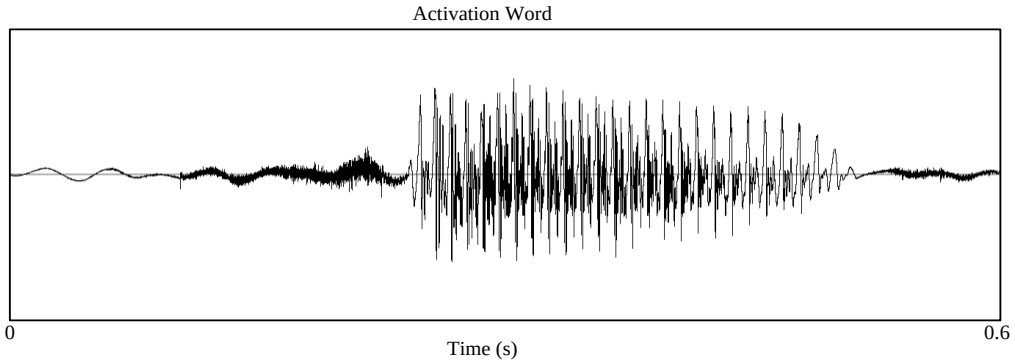


Figure 7-33 – Isolated word for activation (total duration of ~0.6 s)

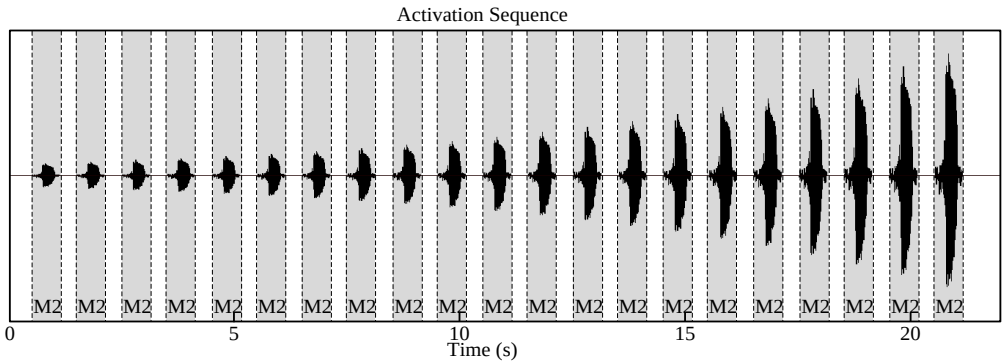


Figure 7-34 – Activation test sequence constructed using the isolated word and repeating it, applying 1 dB additional amplification for each word (total duration of ~22 s)

7.3.4.2 Application

Typically, short words for activation are used for two purposes. Either they are used for conditioning a system under test in advance to the tests or they are used to investigate the system behaviour with regard to activation performance, switching characteristics or minimum activation level. The isolated word for activation can be used not only as is, but also for concatenation with other sequences. Alternatively, it can be repeated, e.g., with different levels as shown for the activation sequence given in Figure 7-34. This activation sequence is specifically chosen to investigate the minimum activation level of a system under test.

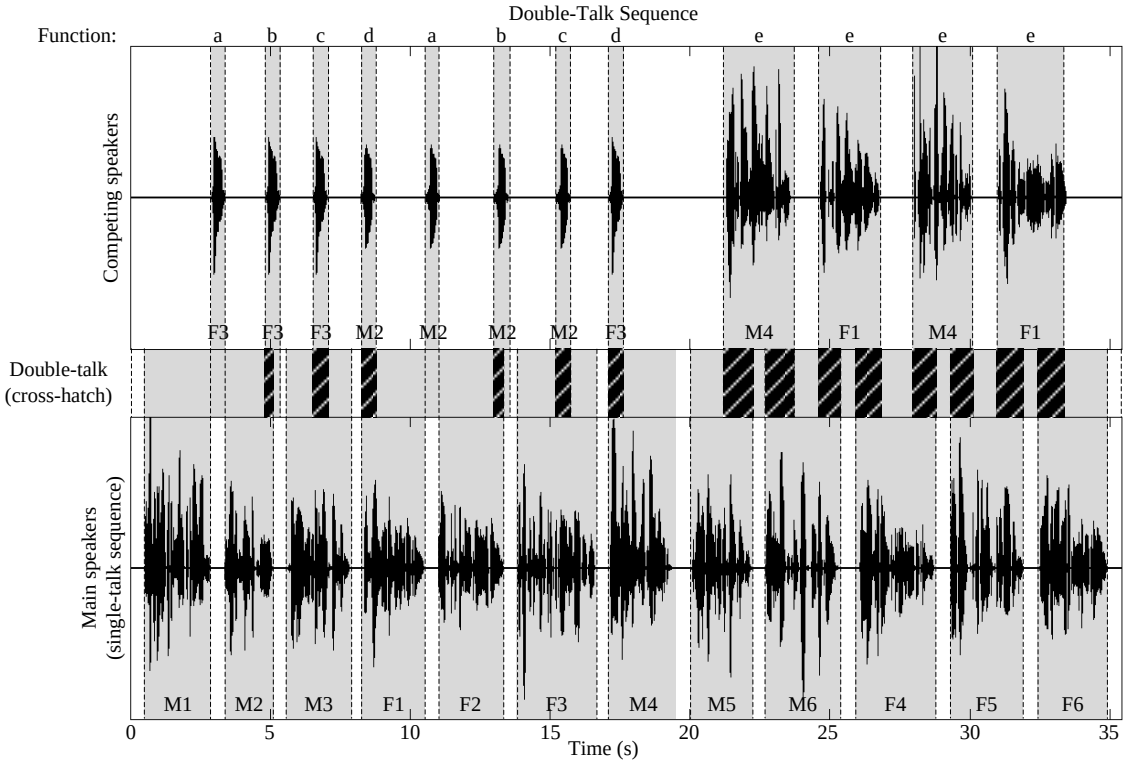
However, the use of these sequences is not limited to this application. Another typical application of these sequences is the evaluation of the system behaviour (e.g., switching characteristics, echo performance) in special conversational situations.

Any delay in the system under test should be taken into account for the signal insertion, as well as for the signal analysis. If the measured signal is referenced against the input signal, the input signal should be time aligned by taking into account the actual delay between the measured and input signals.

7.3.5 Speech signals for double-talk testing

7.3.5.1 Description

A "double-talk" sequence representing typical double-talk scenarios in real conversations is shown in Figure 7-35. This uses the single-talk sequence described in clause 7.3.1, shown in the lower pane, as the main speech and an additional competing speaker sequence, shown in the upper pane.



NOTE – Cross-hatched areas between the upper and lower panes show periods of double talk.

Figure 7-35 – Double-talk test sequence using the single-talk sequence and competing speech serving different functions (a to e)

The competing-speaker sequence includes single words (the word "five") spoken by speakers F3 and M2 during the first half of the sequence, followed by full sentences by speakers F1 and M4 during the second half of the sequence. No speaker is competing with himself or herself during the sequence.

The competing samples serve different double-talk functions, defined as functions "a" to "e" above the upper pane of Figure 7-35. The functions are:

a) competing word within a speech pause;

b) competing word partially masked;

c) competing word fully masked within a sentence;

d) competing word fully masked coincident with the start of a sentence;

e) sentence masking another sentence.

These are meant to represent possible double-talk situations in normal conversation. The area between the upper and lower pane of Figure 7-35 shows the periods during which double-talk happens as cross-hatched patches. The competing sequence can be used either as a send signal or a receive signal in testing.

7.3.5.2 Application

Typically, these sequences are used for evaluations during double talk where evaluation is required under real double-talk conditions while the double-talk signal is present during the analysis. The typical applications are the evaluation of switching characteristics during double talk, echo loss, spectral echo loss, echo loss variation or level variation of the double-talk signal under double-talk conditions.

In any case, the signals are fed simultaneously into the far-end as well as into the near-end direction. Any delay in the system under test should be taken into account for the signal insertion, as well as for the signal analysis. If the measured signal is referenced against the input signal, the input signal should be time aligned by taking into account the actual delay between the measured and input signals.

The channels can be swapped if the double-talk signal should be applied to the opposite channel.

7.3.6 Speech sequences for echo performance testing

7.3.6.1 Description

In general, high frequency echo components are more annoying than lower frequency echo components. This is especially important for wideband and fullband echo testing. The major impairment typically occurs if such high frequency echo components reach the user's ear unmasked. To take this effect into account, the two speech samples shown in Figures 7-36 and 7-37 can be used for diagnostic purposes. Especially for wideband, super-wideband and fullband connections, it is important that a test signal provide excitation energy in the high frequency range above 3.5 kHz. The two samples have been selected due to the fricative sounds during and especially at the end of the utterances as indicated by the yellow circle in Figures 7-36 and 7-37. Additionally the male speaker also provides a considerable amount of energy in the low frequency range below 150 Hz (see red oval).

Sentence 1, male speaker:
"The birch canoe slid on the smooth planks"

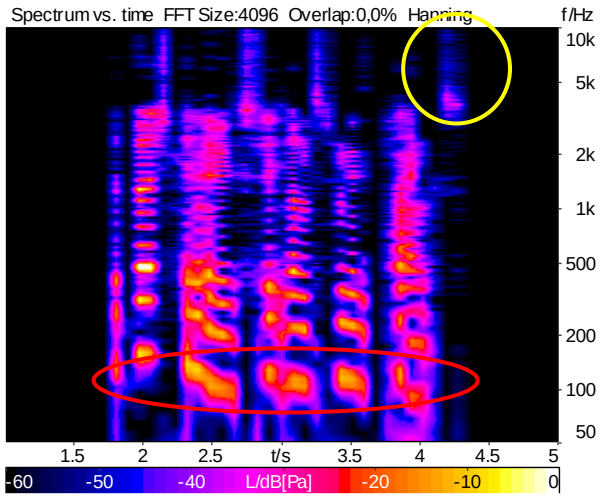


Figure 7-36 – Spectrogram of echo test sentence 1, male speaker (male 2, clause B.3.1)

Sentence 2, female speaker:
"The hogs were fed chopped corn and garbage"

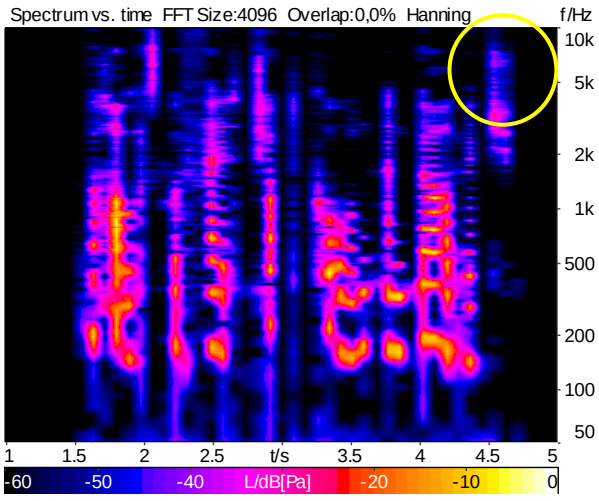


Figure 7-37 – Spectrogram of echo test sentence 2, female speaker (female 1, clause B.3.1)

7.3.6.2 Application

The sentences may be used if a speech sequence is required specifically focusing on echo impairments. This is typically the case for systems with adaptive signal processing, including echo

cancellation, which behave non-linearly and are time variant and which may react to artificial signals differently from speech.

This sequence is intended for use in the determination of echo impairments, especially high frequency echo components. Typical measurements could be spectral echo loss, temporal echo loss or perceptually based procedures for determining echo impairments.

If the measured signal is referenced against the input signal, the input signal should be time aligned by taking into account the actual delay between the measured and input signals.

7.3.7 Conditioning speech sequences

7.3.7.1 Description

Speech sequences typically used at the beginning of a communication that can be used to place devices into a typical or steady condition are shown in Figures 7-38 and 7-39. These include a single male speaker in one channel and a single female speaker in the other, one of which can be assigned as a send signal with the other being the receive signal.

Figure 7-38 is a long sequence (duration of 23.5 s) and Figure 7-39 is a shorter version (duration of 10 s) using the same speakers (M4 and F1 from the single-talk sequence) but different sentences. Sentences II and III from the long sequence are the same as sentences I and II for the short sequence. The male sentence I in the long sequence is actually a single word ("Grace") taken from a larger sentence.

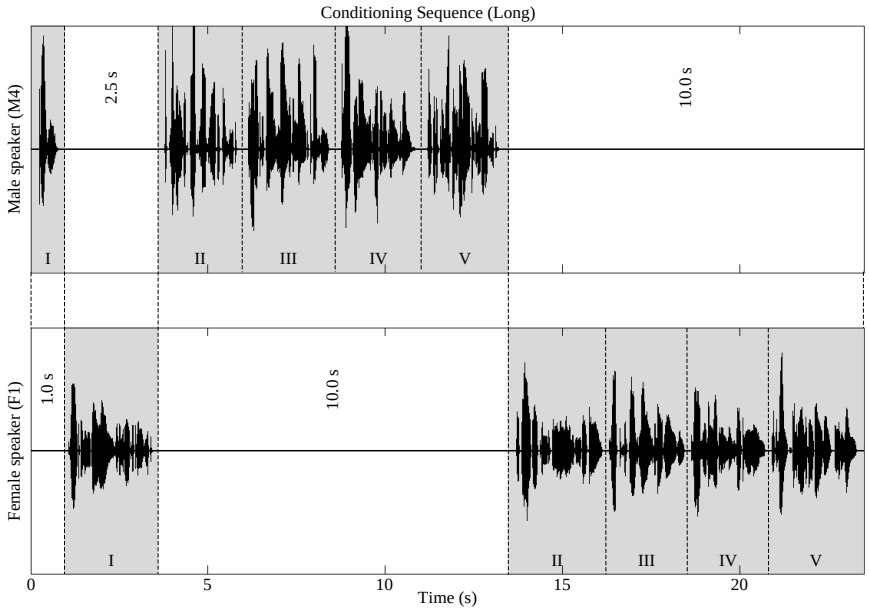


Figure 7-38 – Long conditioning sequence using sentences (I – V) from a single male and female speaker (total duration 23.5 s)

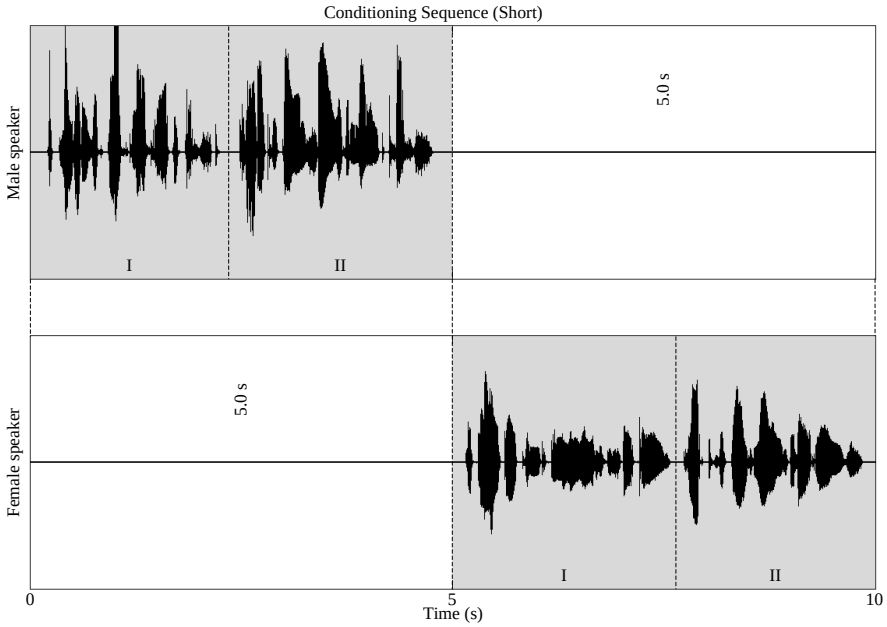


Figure 7-39 – Short conditioning sequence using sentences (I and II) from a single male and female speaker (total duration 10 s)

7.3.7.2 Application

Typically, these sequences are used to condition a system before testing. However, the use of these sequences is not purely limited to this application. Another typical application of these sequences is the evaluation of the system behaviour (e.g., switching characteristics, echo performance) during the initial phase of a call.

In any case, the signals are fed simultaneously into the far-end as well as into the near-end direction. Any delay in the system under test should be taken into account for the signal insertion, as well as for the signal analysis. If the measured signal is referenced against the input signal, the input signal should be time aligned by taking into account the actual delay between the measured and input signals.

The channels can be swapped, depending on the desired way of activating the send and receive directions.

7.3.8 Filters for limiting the speech test signal bandwidth

The speech signals provided are fullband signals. If these signals have to be applied to systems requiring a low-pass-filtered signal, the fullband signal has to be band limited accordingly. This is required, for example, for analogue or digital access points providing no or insufficient low-pass filtering. The filter amplification is 0 dB. The required filter characteristics for the band limitation of the different transmission bandwidths as defined in ITU-T are given in Table 7-7.

Table 7-7 – Filter characteristics for band limiting the fullband speech signals

	f_{-3dB}		f_{cutoff}	
Narrowband (NB)	3600 Hz	> −3 dB	4000 Hz	< −80 dB
Wideband (WB)	7200 Hz	> −3 dB	8000 Hz	< −80 dB
Super-wideband (SWB)	14400 Hz	> −3 dB	16000 Hz	< −80 dB

Any digital filter fulfilling these requirements can be used. In [ITU-T G.191], filters can be found fulfilling these requirements if appropriate up-/down-sampling of the original signal is applied.

NOTE – The signal level for the different transmission bandwidths is always determined from the band-limited signal.

7.4 Additional languages

Besides the speech signals described in clause 7.3, other speech sequences from different languages may be used, as described in this clause. The speech samples described here follow the same construction principle as those described in clause 7.3. However, it is known that different signals may have different impacts on signal processing in modern non-linear and time variant signal processing and may lead to different measurement results. Therefore, if there are differences in test results received when using different languages, the test result received with British English shall be normative.

7.4.1 Chinese speech samples

7.4.1.1 Chinese reference speech samples

The format of recordings is PCM, 16 bit, 48kHz sampled and monaural.

The mean active speech level is adjusted to approximately −26 dBov.

Phonetic balance

Mandarin Chinese characters can be analysed phonetically and represented by 22 consonants, 36 vowels and 4 tones. Efforts have been made to ensure that the frequency of occurrence of those components is similar to daily oral communication.

Word familiarity

The words chosen in the speech material are easy to use and not hard to understand. The selected speech material includes common words in daily use.

Activity factor and crest factor for single-talk sentences

See Table 7-8.

Table 7-8 – Activity factor and crest factor

Speaker	Activity factor (%)	Average (%)	Crest factor (dB)	Average (dB)
F1	98.682	97.38	19.243	16.69
F2	98.059		14.955	
F3	96.776		14.560	
F4	96.248		15.577	
F5	98.179		15.571	
F6	97.759		15.473	
M1	97.944		15.857	
M2	94.704		16.500	
M3	97.777		19.836	
M4	97.739		14.146	
M5	98.446		19.489	
M6	96.247		19.016	

Speech spectra

See Figures 7-40 to 7-43.

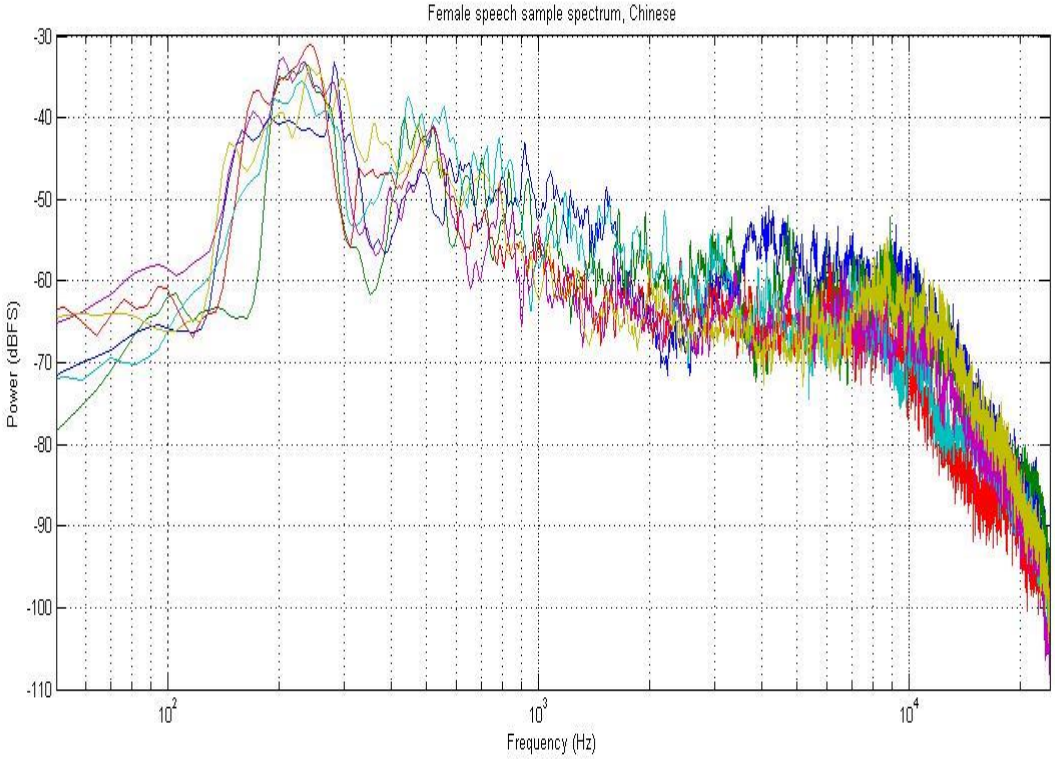


Figure 7-40 – Female individual spectra

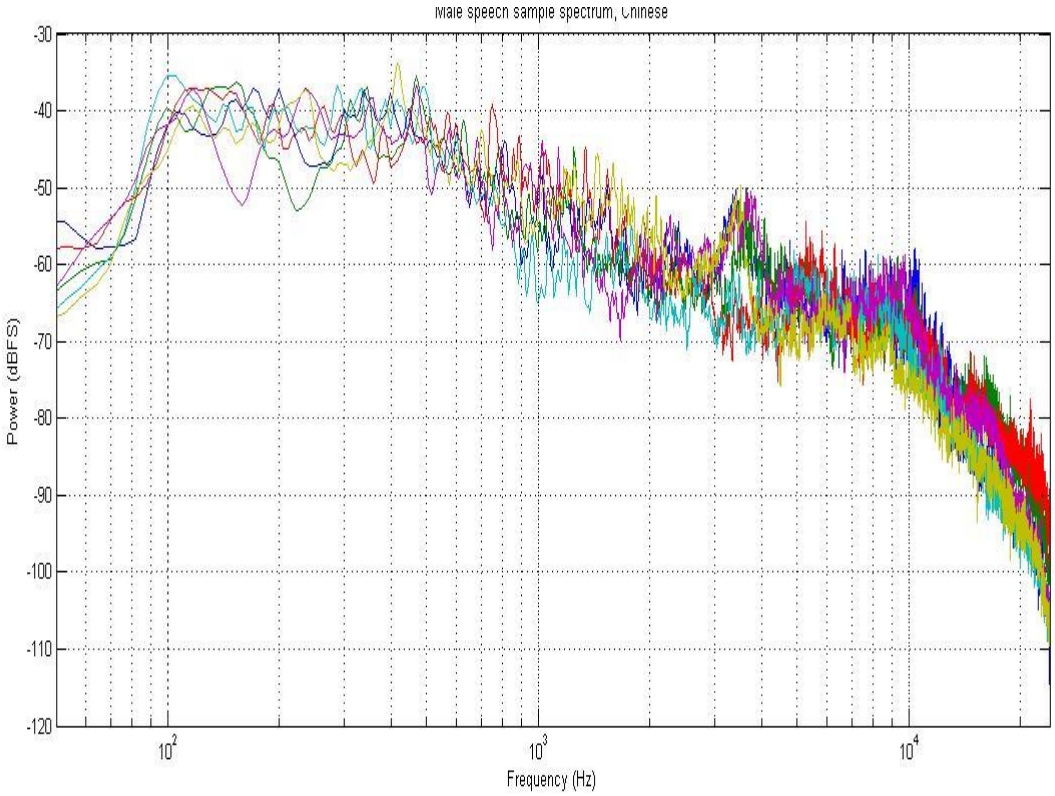


Figure 7-41 – Male individual spectra

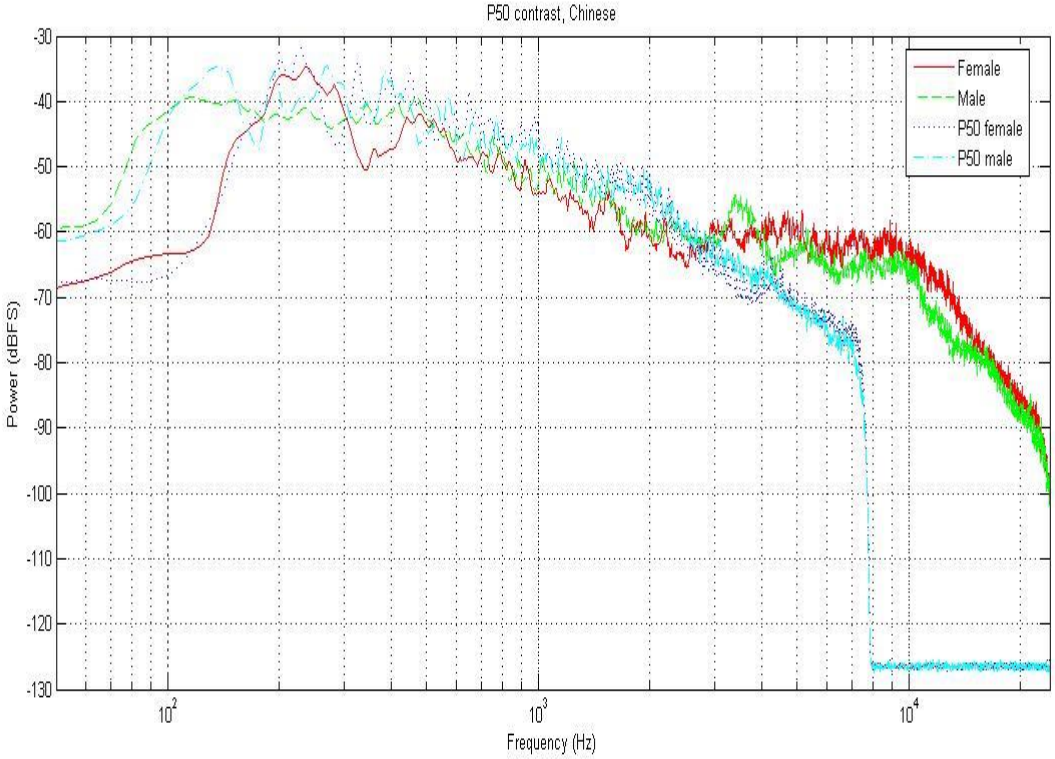


Figure 7-42 – Spectra of all female and all male speech samples along with separate male and female ITU-T P.50 artificial voices

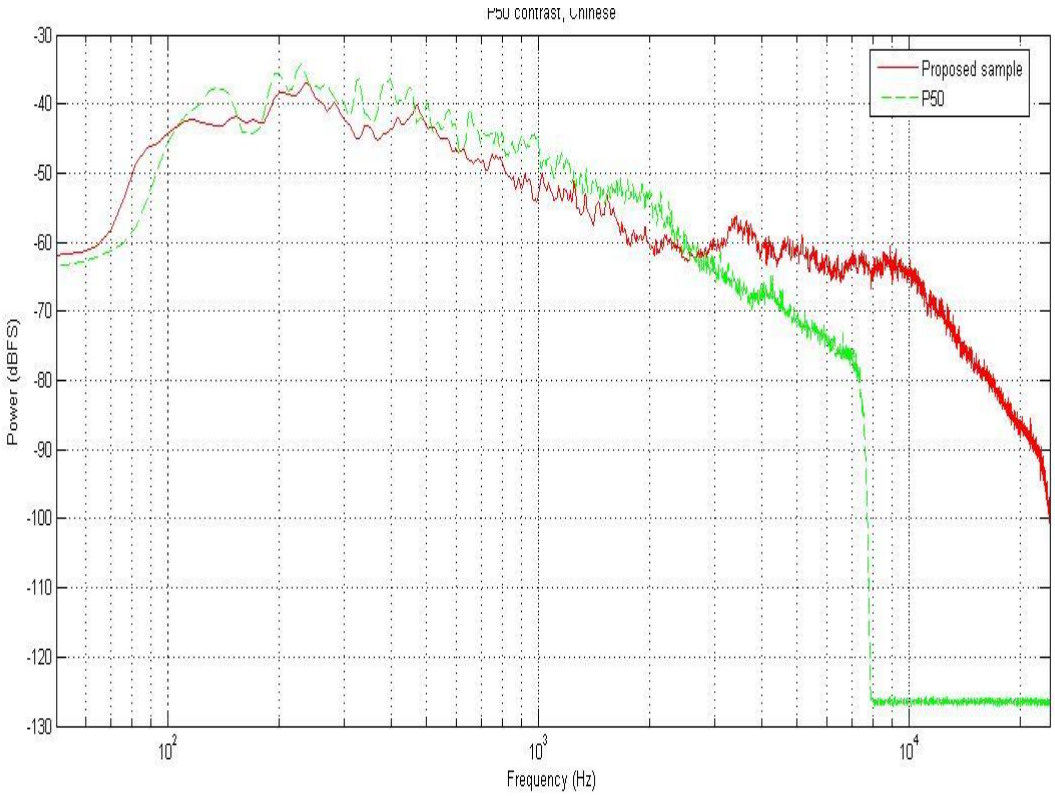


Figure 7-43 – Spectra of all speech samples along with ITU-T P.50 artificial speech

7.4.1.2 Chinese single-talk speech sequence

The single-talk speech contains 12 sentences spoken by six male and six female native Chinese speakers. A typical "single-talk" sequence of sentences spoken by all 12 speakers from the reference speech samples is shown in Figure 7-43. The sequence, lasting ~35.7 s, is created using three males (M1–M3), three females (F1–F3), the remaining three males (M4–M6) and the remaining three females (F4–F6), with each speaker speaking a unique sentence. A silence period of 0.5 s is inserted between each sentence, as well as at the beginning and end of the sequence.

If it can be confirmed that the system under test reacts in the same way when the sequence contains fewer sentences or measurement duration is of more importance, a shortened sequence is also feasible.

See Figure 7-44.

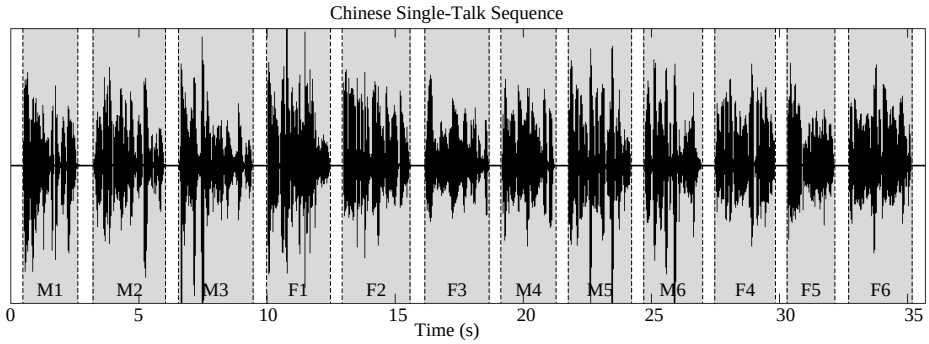
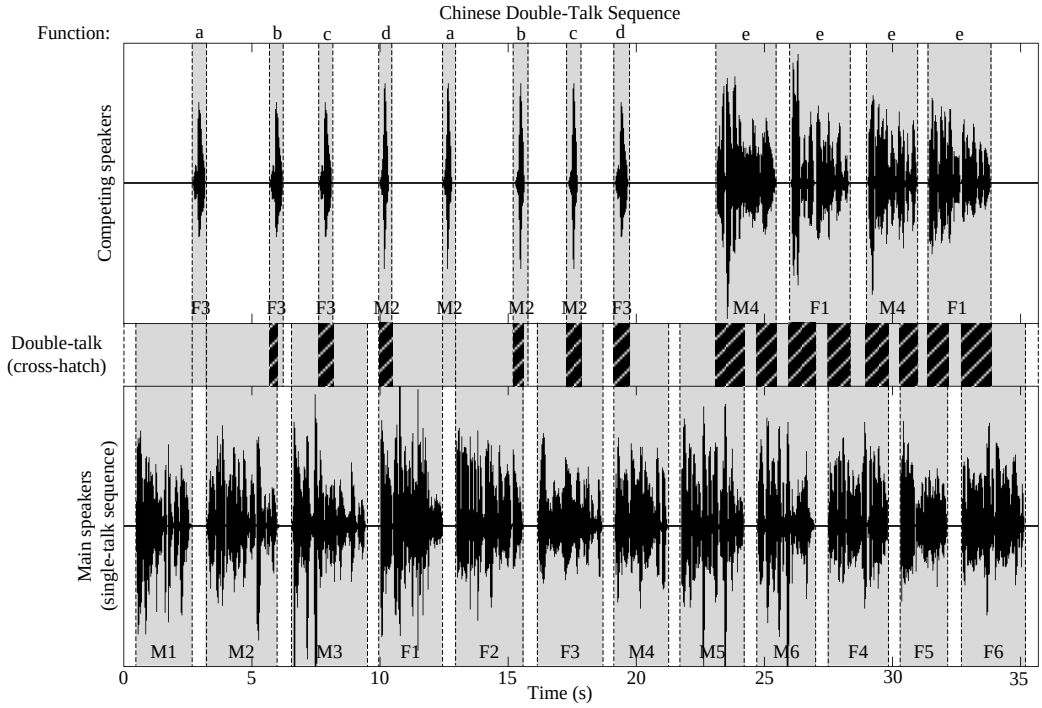


Figure 7-44 – Single-talk test sequence using six male and six female Chinese speakers with a pause of 0.5 s at the beginning, end and between individual samples (total duration of ~35.7 s)

7.4.1.3 Chinese double-talk speech sequence

See Figure 7-45.



NOTE – Cross-hatched areas between the upper and lower panes show periods of double talk.

Figure 7-45 – Double-talk test sequence using the single-talk sequence and competing speech serving different functions (a–e)

7.4.1.4 Chinese conditioning speech sequence

See Figures 7-46 and 7-47.

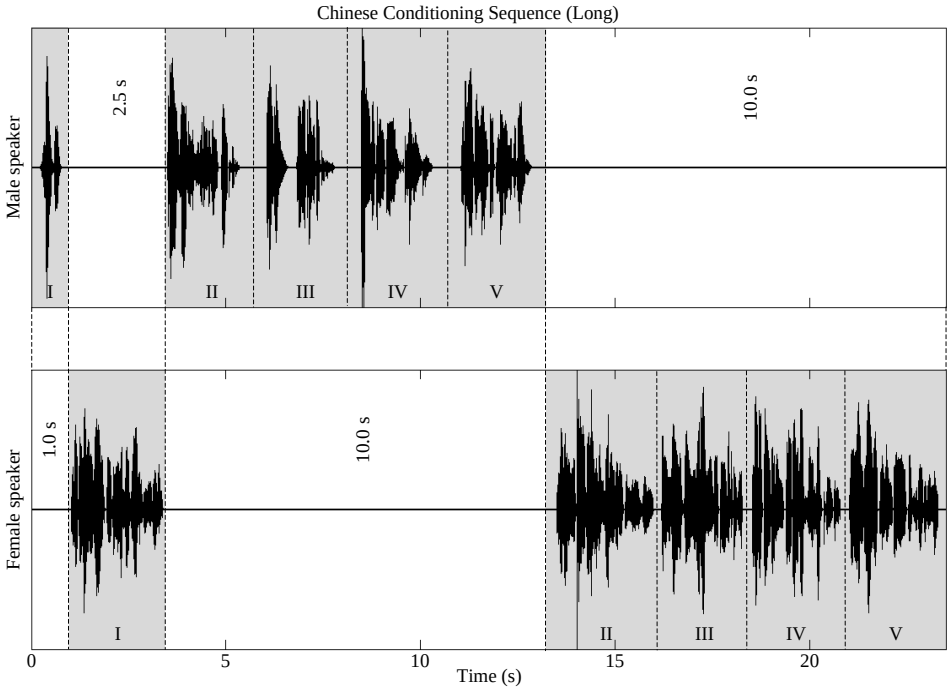


Figure 7-46 – Long conditioning sequence using sentences (I-V) from a single male and a single female speaker (total duration 23.5 s)

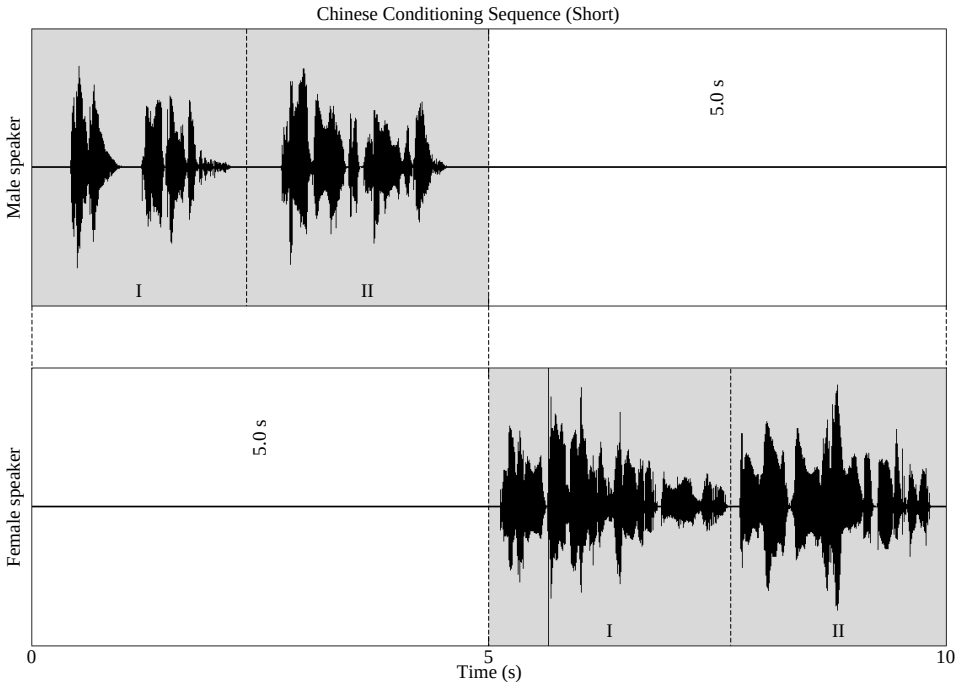


Figure 7-47 – Short conditioning sequence using sentences (I and II) from a single male and a single female speaker (total duration 10 s)

Annex A

Test signals for terminal coupling loss tests

(This annex forms an integral part of this Recommendation.)

For the measurement of terminal coupling loss (TCL), a PN-sequence with a low crest factor and a logarithmically distributed multi-sine wave are equally well applicable. Both provide TCL measurements with high dynamic range, typically >58 dB.

For non-linear and/or time variant systems, it must be ensured that the equipment under test is under "steady state conditions". Depending on the task, for example, echo cancellers should be fully converged. This can be achieved by using training sequences, for example, using artificial voice (as described in [ITU-T P.50]), CSSs (as described in clause 7.2.1 and in [ITU-T G.168], or other speech-like test signal before inserting the actual test signal.

PN-based test signal

The actual test signal is a PN-sequence according to this Recommendation with a length of 4096 points (for the 48 kHz sampling rate) and a crest factor of 6 dB. The duration of the test signal is 1 s. The test signal level is −3 dBm0.

Sinusoidal-based test signal

When using a logarithmically spaced multi-sine test signal, it is defined as:

$$s(t) = \sum_i \{[A + \mu_{\text{am}} \cos(2\pi t \times f_{\text{am}})] \times \cos(2\pi t \times f_{0i})\}$$

with:

A = 0.5

f_{am} = 4 Hz

μ_{am} = 0.5

f_{0i} = 250 Hz × $2^{(i/3)}$

i = 1.12

The test signal level is adjusted to −3 dBm0.

Annex B

Speech files and noise sequences

(This annex forms an integral part of this Recommendation.)

B.1 General

The signals provided on the ITU-T test signals database that are associated with this annex are recorded at various locations by different parties who kindly provided the sequences. All sequences are stored as *.wav files, no calibration for the individual signals is provided. For signals where the original signal level is known, this is indicated in the signal description. In general, users of the test signals have to find a suitable digital amplification in order to achieve the required signal level for their application – for the test sentences as well as for the noise sequences. General guidance on speech signal levels can be found in [ITU-T P.800] and [ITU-T P.79], further guidance and tools for speech processing can be found in [ITU-T G.191].

B.2 Description of the recording procedure used for speech signals

The following general guideline was given for the recording of the tests sentences:

"ITU-T SG12 wishes to extend ITU-T P.501 and include speech files for various languages to be used in combination with objective speech quality evaluation methods. It is the purpose to have a comprehensive set of speech sentences that can be used worldwide and that help to give comparable results when used in conjunction with ITU-T recommended objective speech quality evaluation procedures. It is not the aim to replace the speech material found in [b-ITU-T P-Sup.23], which is only available under special non-disclosure agreement conditions due to commercial issues and cannot be used for the purposes described in this Recommendation.

The speech material for ITU-T P.501 should be limited to:

– Four sentence pairs spoken by four different speakers (two male and two female).
– The test sentences should be phonetically balanced.
– The duration should be about 8 s for each sentence.
– The recordings should be made in quiet and mostly non-reverberant conditions. Studio conditions would be ideal.
– The recordings should be sufficiently undistorted and noise free. Care should be taken neither to overload the recording device nor to come close to the noise floor of the recording device. A signal-to-noise ratio of more than 50 dB is desirable.
– The microphone should be positioned at a distance of about 30-50 cm from the mouth in order to avoid any distortions from proximity effects."

B.3 Test sentences

All speech samples are processed such that the levels measured using a speech level voltmeter according to [ITU-T P.56] are equal.

B.3.1 Chinese (fullband)

Female 1:

仓库的后面是一间小屋。太阳从东方升起来。

Cang ku de hou mian shi yi jian xiao wu. Tai yang cong dong fang sheng qi lai.

Female 2:

哈尔滨在中国的最北面。厨房的桌子上摆好了早餐。

Ha'e'bin zai zhong guo de zui bei mian. Chu fang de zhuo zi shang bai hao lie zao can.

Male 1:

药店一直都关着门。妈妈在另外一个房间里休息。

Yao dian yi zhi dou guan zhe men. MaMa zai ling wai yi ge fang jian li xiu xi.

Male 2:

他每次来都背着沉重的包。他看起来不过十八九岁。

Ta mei ci lai dou bei zhe chen zhong de bao. Ta kan qi lai bu guo shi ba jiu sui.

B.3.2 Dutch (fullband)

Female 1:

Dit produkt kent nauwelijks concurrentie.

Hij kende zijn grens niet.

Female 2:

Ik zal iets van mijn carriere vertellen.

Zijn auto was alweer kapot.

Male 1:

Zij kunnen de besluiten nehmen.

De meeste mensen hadden het wel door.

Male 2:

Ik zou liever gaan lopen.

Willem gaat telkens naar buiten.

B.3.3 English (fullband)

Female 1:

These days a chicken leg is a rare dish.

The hogs were fed with chopped corn and garbage.

Female 2:

Rice is often served in round bowls.

A large size in stockings is hard to sell.

Male 1:

The juice of lemons makes fine punch.

Four hours of steady work faced us.

Male 2:

The birch canoe slid on smooth planks.

Glue the sheet to the dark blue background.

B.3.4 English (American)

Female 1:

We need grey to keep our mood healthy.

Pack the records in a neat thin case.

Female 2:

The stems of the tall glasses cracked and broke.

The wall phone rang loud and often.

Male 1:

The shelves were bare of both jam or crackers.

A joy to every child is the swan boat.

Male 2:

Both brothers were the same size.

In some form or other we need fun.

B.3.5 Finnish (fullband)

Female 1:

Ole ääneti tai sano sellaista, joka on parempaa kuin vaikeneminen.

Suuret sydämet ovat kuin valtameret, ne eivät koskaan jäädy.

Female 2:

Jos olet vasara, lyö kovaa. Jos olet naula pidä pääsi pystyssä.

Onni tulee eläen, ei ostaen.

Male 1:

Rakkaus ei omista mitään, eikä kukaan voi sitä omistaa.

Naisen mieli on puhtaampi, hän vaihtaa sitä useammin.

Male 2:

Sydämellä on syynsä, joita järki ei tunne.

On opittava kärsimään voidakseen elää.

B.3.6 French (fullband)

Female 1:

On entend les gazouillis d'un oiseau dans le jardin.

La barque du pêcheur a été emportée par une tempête.

Female 2:

Le client s'attend à ce que vous fassiez une réduction.

Chaque fois que je me lève ma plaie me tire.

Male 1:

Vous avez du plaisir à jouer avec ceux qui ont un bon caractère.

Le chevrier a corné pour rassembler ses moutons.

Male 2:

Ma mère et moi faisons de courtes promenades.

La poupée fait la joie de cette très jeune fille.

B.3.7 German

Female 1:

Zarter Blumenduft erfüllt den Saal.

Wisch den Tisch doch später ab.

Female 2:

Sekunden entscheiden über Leben.

Flieder lockt nicht nur die Bienen.

Male 1:

Gegen Dummheit ist kein Kraut gewachsen.

Alles wurde wieder abgesagt.

Male 2:

Überquere die Strasse vorsichtig.

Die drei Männer sind begeistert.

B.3.8 German (fullband)

Female 1:

Im Fernsehen wurde alles gezeigt,

Alle haben nur einen Wunsch.

Female 2:

Kinder naschen Süßigkeiten.

Der Boden ist viel zu trocken.

Male 1:

Mit einem Male kam die Sonne durch.

Das Telefon klingelt wieder.

Male 2:

Sekunden entscheiden über Leben.

Flieder lockt nicht nur die Bienen.

B.3.9 Italian (fullband)

Female 1:

Non bisogna credere che sia vero tutto quello che dice la gente. Tu non conosci ancora gli uomini, non conosci il mondo.

Dopo tanto tempo non ricordo più dove ho messo quella bella foto, ma se aspetti un po' la cerco e te la prendo.

Female 2:

Questo tormento durerà ancora qualche ora. Forse un giorno poi tutto finirà e tu potrai tornare a casa nella tua terra.

Lucio era certo che sarebbe diventato una persona importante, un uomo politico o magari un ministro. Aveva a cuore il bene della società.

Male 1:

Non bisogna credere che sia vero tutto quello che dice la gente tu non conosci ancora gli uomini, non conosci il mondo.

Dopo tanto tempo non ricordo più dove ho messo quella bella foto ma se aspetti un po' la cerco e te la prendo.

Male 2:

Questo tormento durerà ancora qualche ora. Forse un giorno poi tutto finirà e tu potrai tornare a casa nella tua terra.

Lucio era certo che sarebbe diventato una persona importante, un uomo politico o magari un ministro, aveva a cuore il bene della società.

B.3.10 Japanese (fullband)

Female 1:

彼は鮎を釣る名人です。

Kare wa ayu wo tsuru meijin desu.

古代エジプトで十進法の原理が作られました。

Kodai ejipto de jusshinhō no genri ga tsukuraremashita.

Female 2:

読書の楽しさを知ってください。

Dokusho no tanoshisa wo shitte kudasai.

人間の価値は知識をどう活用するかで決まります。

Ningen no kachi wa chishiki wo dō katsuyō suruka de kimarimasu.

Male 1:

彼女を説得しようとしても無駄です。

Kanojo wo settoku shiyōtoshitemo mudadesu.

その昔ガラスは大変めずらしいものでした。

Sono mukasi garasu wa taihen mezurashii monodeshita.

Male 2:

近頃の子供たちはひ弱です。

Chikagoro no kodomo tachi wa hiyowa desu.

イギリス人は雨の中を平気で濡れて歩きます。

Igirisujin wa ameno nakawo heikide nurete arukimasu.

B.3.11 Polish

Female 1:

Pielęgniarki były cierpliwe.

Przebiegał szybko przez ulicę.

Female 2:

Ona była jego sekretarką od lat.

Dzieci często płaczą kiedy są głodne.

Male 1:

On był czarującą osobą.

Lato wreszcie nadeszło.

Male 2:

Większość dróg było niezmiernie zatłoczonych.

Mamy bardzo entuzjastyczny zespół.

B.3.12 Spanish (American)

Female 1:

No arroje basura a la calle.

Ellos quieren dos manzanas rojas.

Female 2:

No cocinaban tan bien.

Mi afeitadora afeita al ras.

Male 1:

Vé y siéntate en la cama.

El libro trata sobre trampas.

Male 2:

El trapeador se puso amarillo.

El fuego consumió el papel.

B.4 Noise sequences

Two types of noise sequences are provided on the CD-ROM:

– Noise sequences recorded binaurally using a freefield equalized artificial head according to [b-ITU-T P.58].

– Noise sequences recorded monaurally with a single microphone.

B.4.1 Binaural noise recordings

Train

Noise in a railway station while a train is entering the station.

Average level (whole signal): 70 dB_{SPL}(A).

Traffic

Traffic noise recorded at a crossing.

Average level (whole signal): 70 dB_{SPL}(A).

Bus

Noise recorded in a bus while driving.

Average level (whole signal): 66 dB_{SPL}(A).

Kids

Children recorded while playing in a room.

Average level (whole signal): 78 dB_{SPL}(A).

Medium size car

Noise at constant driving conditions (100 km/h) in a medium size car.

Average level (whole signal): 67 dB_{SPL}(A)

Car_bin1_FFeq

Car interior noise, car driving, radio on (speech programme).

Con_bin1_FFeq

Construction noise, impulse type noise (hammering), sawing noise.

Met_bin1_FFeq

Metro train arriving to the station.

Off_bin1_FFeq

Office noise, fans, typing, phone ringing, noise from chair.

Rai_bin1_FFeq

Railway station, echoing surroundings, speech, shoes clacking.

Res_bin1_FFeq

Restaurant, babble, water, dishes.

B.4.2 Monaural noise recordings

Cafeteria

Typical cafeteria noise.

In car

Noise inside a typical medium size car.

Street

Typical street noise.

Car_mono1_30s

Car interior noise, car driving, radio on (speech programme).

Con_mono1_30s

Construction noise, impulse type noise (hammering), sawing noise.

Met_mono1_30s

Metro train arriving to the station.

Off_mono1_30s

Office noise, fans, typing, phone ringing, noise from chair.

Rai_mono1_30s

Railway station, echoing surroundings, speech, shoes clacking.

Res_mono1_30s

Restaurant, babble, water, dishes.

Annex C

Speech files prepared for use with ITU-T P.800 conformant applications and perceptual-based objective speech quality prediction

(This annex forms an integral part of this Recommendation.)

C.1 General

The signals provided in Annex C are based on the full-band speech samples in Annex B. The speech samples have been modified/cleaned up using the following procedure:

– aligning the initial pause prior to the first speech activity to about 500 ms

– aligning the pause in between the two sentences to about 1000 ms

– aligning all speech samples to a file length of 8 s.

– a version filtered with an SWB bandpass as in [ITU-T G.191] (14 kHz bandpass) is included

– a flat narrowband version derived by downsampling using the HiQ lowpass as defined in [ITU-T P.863.1] is included

– a narrowband version filtered with an IRS send mod filter as in [ITU-T P.830] is included.

The samples in Italian and Finnish consist of more than two sentences. For the samples in these two languages, two sentences have been chosen and combined into a typical sentence pair.

The aligned sample French Male 1 has a shorter pause than 1000 ms in between the sentences because of the long spoken sentences.

In all speech samples, the amount of active speech is >3.2 s, as recommended in [ITU-T P.863.1].

The original speech samples in this Recommendation show different noise floors; individual files have muted pauses, others have an artificially added/inserted low noise floor or the original recording noise.

Therefore, as a second step, the leading and trailing pauses, as well as the pauses between the sentences have been muted manually without hurting active speech areas. After muting the pauses, the whole file was mixed with a low-levelled white Gaussian noise of −85 dB (OVL).

NOTE – The collection of French samples show an original noise floor of >−75 dB (OVL). This noise floor is perceptible during active speech and may have an influence on subjective and objective scoring methods.

The collection of Finnish samples shows a high noise floor too [>−70 dB (OVL) partially] and Male 2 has a perceptible tonal component that is also present during active speech and may also have an influence on subjective and objective scoring.

Note that the speech sample in Dutch is only available up to SWB 14 kHz, there is no full-band version provided.

All sequences are stored as *.wav files; no calibration for the individual signals is provided. All signals are calibrated to the same level. In general, users of the test signals have to find a suitable digital amplification in order to achieve the required signal level for their application – for the test sentences as well as for the noise sequences. General guidance on speech signal levels can be found in [ITU-T P.800] and [ITU-T P.79]; further guidance and tools for speech processing can be found in [ITU-T G.191].

C.2 Test sentences

All speech samples are processed so that the level measured using a speech level voltmeter according to [ITU-T P.56] are equal. The signals are available with 8 kHz and 48 kHz sampling rates.

Naming convention:

P501_C_xyz_flat_08k.wav

P501_C_xyz_IRS_08k.wav

P501_C_xyz_SWB_48k.wav

P501_C_xyz_FB_48k.wav

C.2.1 Dutch (fullband)

Female 1:

Dit produkt kent nauwelijks concurrentie.

Hij kende zijn grens niet.

Female 2:

Ik zal iets van mijn carriere vertellen.

Zijn auto was alweer kapot.

Male 1:

Zij kunnen de besluiten nehmen.

De meeste mensen hadden het wel door.

Male 2:

Ik zou liever gaan lopen.

Willem gaat telkens naar buiten.

C.2.2 Chinese (fullband)

Female 1:

仓库的后面是一间小屋。太阳从东方升起来。

Cang ku de hou mian shi yi jian xiao wu. Tai yang cong dong fang sheng qi lai.

Female 2:

哈尔滨在中国的最北面。厨房的桌子上摆好了早餐。

Ha'e'bin zai zhong guo de zui bei mian. Chu fang de zhuo zi shang bai hao lie zao can.

Male 1:

药店一直都关着门。妈妈在另外一个房间里休息。

Yao dian yi zhi dou guan zhe men. MaMa zai ling wai yi ge fang jian li xiu xi.

Male 2:

他每次来都背着沉重的包。他看起来不过十八九岁。

Ta mei ci lai dou bei zhe chen zhong de bao. Ta kan qi lai bu guo shi ba jiu sui.

C.2.3 British English

Female 1:

These days a chicken leg is a rare dish.

The hogs were fed with chopped corn and garbage.

Female 2:

Rice is often served in round bowls.

A large size in stockings is hard to sell.

Male 1:

The juice of lemons makes fine punch.

Four hours of steady work faced us.

Male 2:

The birch canoe slid on smooth planks.

Glue the sheet to the dark blue background.

C.2.4 Finnish

Female 1:

Ole ääneti tai sano sellaista, joka on parempaa kuin vaikeneminen.

Suuret sydämet ovat kuin valtameret, ne eivät koskaan jäädy.

Female 2:

Jos olet vasara, lyö kovaa. Jos olet naula pidä pääsi pystyssä.

Onni tulee eläen, ei ostaen.

Male 1:

Rakkaus ei omista mitään, eikä kukaan voi sitä omistaa.

Naisen mieli on puhtaampi, hän vaihtaa sitä useammin.

Male 2:

Sydämellä on syynsä, joita järki ei tunne.

On opittava kärsimään voidakseen elää.

C.2.5 French

Female 1:

On entend les gazouillis d'un oiseau dans le jardin.

La barque du pêcheur a été emportée par une tempête.

Female 2:

Le client s'attend à ce que vous fassiez une réduction.

Chaque fois que je me lève ma plaie me tire.

Male 1:

Vous avez du plaisir à jouer avec ceux qui ont un bon caractère.

Le chevrier a corné pour rassembler ses moutons.

Male 2:

Ma mère et moi faisons de courtes promenades.

La poupée fait la joie de cette très jeune fille.

C.2.6 German

Female 1:

Im Fernsehen wurde alles gezeigt.

Alle haben nur einen Wunsch.

Female 2:

Kinder naschen Süßigkeiten.

Der Boden ist viel zu trocken.

Male 1:

Mit einem Male kam die Sonne durch.

Das Telefon klingelt wieder.

Male 2:

Sekunden entscheiden über Leben.

Flieder lockt nicht nur die Bienen.

C.2.7 Italian

Female 1:

Non ricordo più dove ho messo quella bella foto,

ma se aspetti un po'

Female 2:

Questo tormento durerà ancora qualche ora.

Aveva a cuore il bene della società.

Male 1:

Tu non conosci ancora gli uomini,

ma, se aspetti un po', la cerco e te la prendo.

Male 2:

Questo tormento durerà ancora qualche ora.

Aveva a cuore il bene della società.

C.2.8 Japanese

Female 1:

彼は鮎を釣る名人です。

Kare wa ayu wo tsuru meijin desu.

古代エジプトで十進法の原理が作られました。

Kodai ejipto de jusshinhō no genri ga tsukuraremashita.

Female 2:

読書の楽しさを知ってください。

Dokusho no tanoshisa wo shitte kudasai.

人間の価値は知識をどう活用するかで決まります。

Ningen no kachi wa chishiki wo dō katsuyō suruka de kimarimasu.

Male 1:

彼女を説得しようとしても無駄です。

Kanojo wo settoku shiyōtoshitemo mudadesu.

その昔ガラスは大変めずらしいものでした。

Sono mukasi garasu wa taihen mezurashii monodeshita.

Male 2:

近頃の子供たちはひ弱です。

Chikagoro no kodomo tachi wa hiyowa desu.

イギリス人は雨の中を平気で濡れて歩きます。
Igirisujin wa ameno nakawo heikide nurete arukimasu.

Annex D

Speech files composed of a pair of sentences spoken by a male and a female speaker

(This annex forms an integral part of this Recommendation.)

D.1 General

This nanex provides speech samples that follow the temporal structure and technical requirements given in [ITU-T P.863.1] and [ITU-T P.862.3]. In contrast to the speech samples defined in Annexes B and C, each sample (sentence pair) is composed of one sentence spoken by a male and one by a female speaker. The samples given in this annex are especially prepared for use in automated speech quality measurement systems in combination with perceptual based objective quality measurement prediction, as in [ITU-T P.862] and [ITU-T P.863].

Subjective and objective scores obtained for a given scenario depend also on the speech sample, and more on the speaker and gender. This leads to systematic differences in quality scoring depending on the speech sample used. The applied composition of a male and a female speaker minimizes the gender dependency of measurement results, as usually observed in auditory tests and objective quality prediction.

In particular, for mobile field testing, sentence pairs consisting of one male and one female sentence are commonly used in practice. This annex provides a set of composed sentence pairs in different languages spoken by different speakers. Some of the speech samples are based on the fullband speech samples in Annex C.

The speech samples in this annex are targeted at applications where many scores have to be obtained in a minimum time in order to track fast changes or instabilities, e.g., in mobile networks. To enable a high frequency of measurements, these samples are only 6 s in length. For other applications, longer sequences of sentences as found in Annex C, and described in [ITU-T P.863.1] and [ITU-T P.862.3], should be used.

A composed speech sample is provided for the following languages:

- Dutch*
- British English
- German*
- Finnish*
- French*
- Italian*
- Chinese (Mandarin)
- American English

NOTE – * indicates samples composed of material given in Annex C of this Recommendation

Note that the speech samples in Dutch and English are only available up to SWB 14 kHz, there is no full-band version provided.

Each of these male/female composed samples balances the systematic bias between male and female voices as known for [ITU-T P.862] and [ITU-T P.863]. Additionally, the sentences and speakers have been selected to match mean opinion score predictions for typical codec conditions that can be observed as averages over larger sets of speech samples. The processing procedure and presentation scheme follow [ITU-T P.863.1] exactly.

Each sample is 6 s in length, it has a leading and a trailing pause as well as a pause in between the two sentences that meet the requirements of [ITU-T P.863.1] and [ITU-T P.862.3]. The noise floor in the speech pauses is <−85 dBOv (A) r.m.s., but not digital silence.

All sequences are stored as *.wav files, no calibration for the individual signals is provided. All signals are calibrated to the same level. In general, users of the test signals have to find a suitable digital amplification in order to achieve the required signal level for their applications – for the test sentences, as well as for the noise sequences. General guidance on speech signal levels can be found in [ITU-T P.800] and [ITU-T P.79]; further guidance and tools for speech processing can be found in [ITU-T G.191].

D.2 Test sentences

All speech samples are processed so that the levels measured using a speech level voltmeter according to [ITU-T P.56] are equal. The signals are available with 8 kHz and 48 kHz sampling rates.

The basis for all versions of the test samples is the fullband version of the samples. The super-wideband version (SWB) is processed by applying the 14 kHz lowpass as in [ITU-T G.191], 14 kHz bandpass], the narrowband version (flat 08k) is derived by downsampling using the HiQ lowpass as defined in [ITU-T P.863.1], at the narrowband IRS version (IRS 08k) and IRS send mod filter acc. to [ITU-T P.830] and realized in [ITU-T G.191] was applied.

Naming convention:

P501_D_xyz_fm_flat_08k.wav

P501_D_xyz_fm_IRS_08k.wav

P501_D_xyz_fm_SWB_48k.wav

P501_D_xyz_fm_FB_48k.wav

D.2.1 Dutch

Zijn auto was alweer kapot. (*Female 2*)

Zij kunnen de besluiten nemen. (*Male 1*)

D.2.2 British English

The glow deepened in the eyes of the sweet girl. (*Female*)

The lamp shone with a steady green flame. (*Male*)

D.2.3 Finnish

Ne eivät koskaan jäädy. (*Female 1*)

On opittava kärsimään voidakseen elää. (*Male 2*)

D.2.4 French

On entend les gazouillis d'un oiseau dans le jardin. (*Female 1*)

Ma mère et moi faisons de courtes promenades. (*Male 2*)

D.2.5 German

Im Fernsehen wurde alles gezeigt. (*Female 1*)

Sekunden entscheiden über Leben. (*Male 2*)

D.2.6 Italian

Non ricordo più dove ho messo quella bella foto. (*Female 1*)

Tu non conosci ancora gli uomini. (*Male 1*)

D.2.7 Chinese (Mandarin)

我愿意送她回去(*Female*)

北京近来很寒冷 (*Male*)

Wǒ yuànyì sòng tā huíqù (Female)

Běijīng jìnlái hěn hánlěng (Male)

D.2.8 American English

The frosty air passed through the coat. (*Female*)

The hogs were fed chopped corn and garbage. (*Male*)

Appendix I

Description of the processing applied to the speech signals in clause 7.3

(This appendix does not form an integral part of this Recommendation.)

I.1 Filter for DC removal

High-pass filtering to significantly reduce the frequencies close to 0 Hz was applied (see Figure I.1). For this purpose, an infinite impulse response (IIR) second-order Butterworth high-pass filter with a −3 dB point at 10 Hz was used to ensure that the phase response and group delay are constant in the speech frequencies.

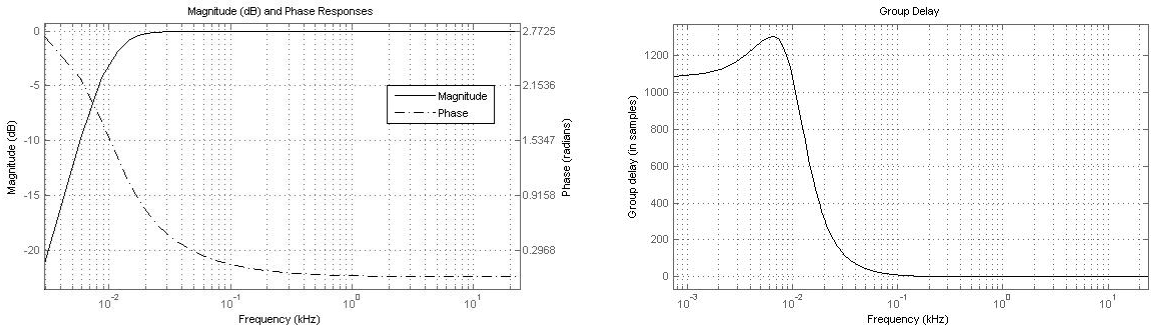


Figure I.1 – Magnitude, phase response and group delay for an IIR second-order high-pass filter (–3 dB at 10 Hz)

The coefficient files and batch scripts to apply them are included in the package reference in Note 1 described in [b-McGill, 2010].

I.2 Creation of the single-talk speech sequence

The single-talk sample was created from the original reference recordings using applications from the set of tools in [ITU-T G.191] and [b-McGill, 2010]. The chain of processing for this includes:

a) (Optional high-pass filtering of reference samples using the FiltAudio application – see clause I.3)

b) Extraction of single sentences from sentence-pair samples where necessary and conversion from *.wav to *.raw format using the CopyAudio application.

c) Amplitude fading at the beginning and end of each sample over 240 samples using the astrip application's "smooth" option.

d) Active speech level normalization of each sample to −26 dBov using the sv56demo application.

e) Generation of a 0.5 s dither-like silence signal from Gaussian noise with a standard deviation of 2^{16} using the GenNoise application.

f) Concatenation of all samples in sequence and conversion to *.wav format using the CopyAudio application.

I.3 Example high-pass filter designs

Recordings of real speech, even when made in reference environments, such as isolation booths or anechoic rooms, can have large amounts of low-frequency energy not sourced from the speaker. This is typically from (or conducted through) HVAC systems required for the speaker to spend a significant amount of time in the environment or mechanical coupling to the outside environment.

Digital filtering is typically applied, when found necessary, to get rid of this. During the drafting process of the sequences described in this appendix, the author created two different filters to deal with specific problems, although these should be viewed as optional. All filters were designed using the fdtool function of Mathworks' Matlab software. The filter coefficients were exported to text files and applied using the *AFsp* FiltAudio tool.

For the samples, there was a degree of what seemed to be ventilation noise below approximately 70 Hz. To deal with this, a finite impulse response (FIR) high-pass shelving-type filter having an attenuation of 33 dB between 60 Hz and 30 Hz was designed. The performance of this filter is shown in Figure I.2.

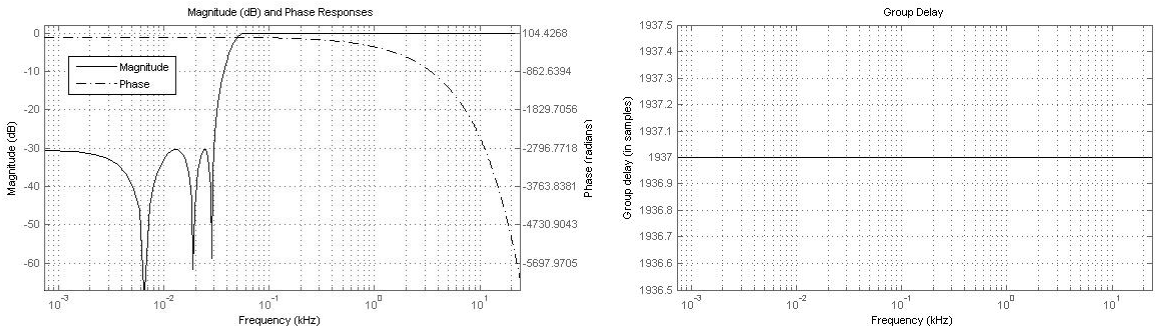


Figure I.2 – Magnitude, phase response and group delay for an FIR high-pass shelving-type filter having an attenuation of 33 dB between 60 Hz and 30 Hz

There is a further high-pass filtering option in order to significantly reduce the frequencies close to 0 Hz. For this purpose, an IIR second order Butterworth high-pass filter with a −3 dB point at 10 Hz is used to ensure that the phase response and group delay are constant in the speech frequencies. See Figure I.3.

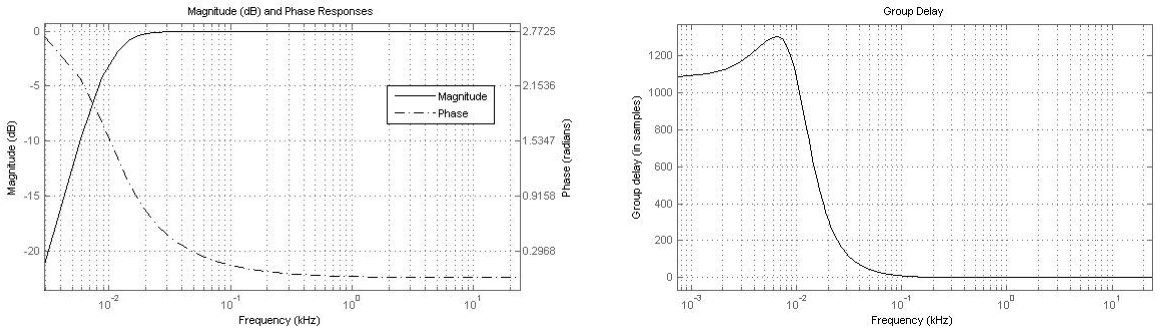


Figure I.3 – Magnitude, phase response and group delay for an IIR second order high-pass filter (−3 dB at 10 Hz)

Bibliography

[b-ITU-T P.58] Recommendation ITU-T P.58 (2013), *Head and torso simulator for telephonometry*.

[b-ITU-T P-Sup.23] ITU-T P-series Recommendations – Supplement 23 (1998), *ITU-T coded-speech database*.

[b-Gierlich, 1992] Gierlich, H.W.(1992). A measurement technique to determine the transfer characteristics of hands-free telephones. *Signal Processing* **27**(3), pp. 281–300.

[b-Halka, 1993] Halka, U., Heute, U. (1993). Speech-model processes controlled by discrete Markov-chains. In: *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, 1993, pp. 1196–1200, Pacific Grove, United States, (1993).

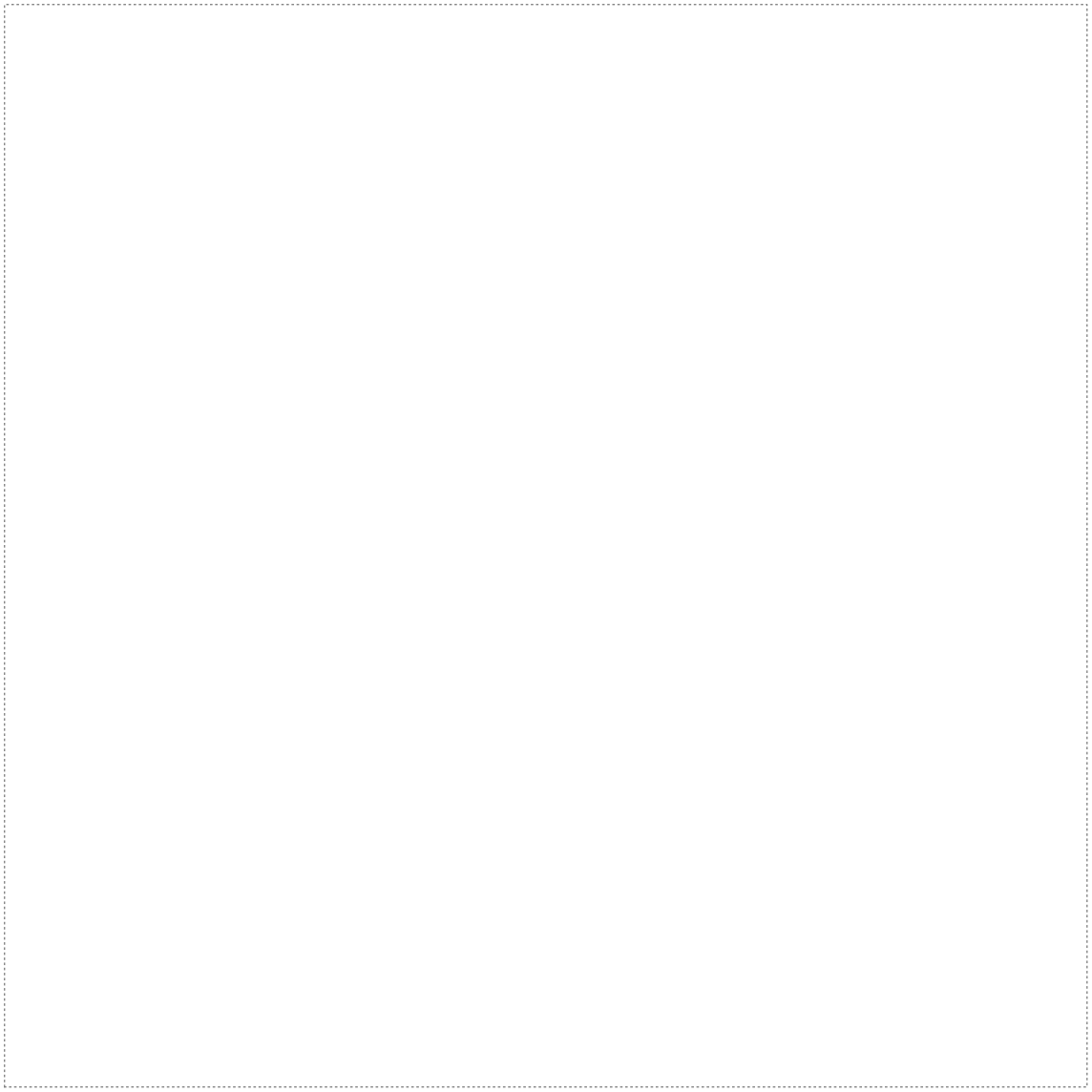
[b-IEEE No.297] IEEE No. 297 (1969), *Recommended practice for speech quality measurements* .

[b-Linde, 1980] Linde, Y., Buzo, A., Grey, R.M. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communications*, **COM-28**(1), 84–95.

[b-McGill, 2010] *AFsp Audio File Programs and Routines*, Software Package 9.0 (2010). McGill University Telecommunications and Signal Processing Laboratory.

[b-Serafat, 1996] Serafat, M.R., Heute, U. (1996). A wide-band speech-model process as a test signal for objective quality assessment. In: *IEEE International Symposium on Circuits and Systems*, 1996. ISCAS '96, Connecting the World ,Vol. 2, pp. 61-64.

[b-Steeneken, 1980] Steeneken, H.J., Houtgast, T. (1980) A physical method for measuring speech-transmission quality. *J. Acoust. Soc. Am.* **67**, 318–326.



SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2018