INTERNATIONAL TELECOMMUNICATION UNION

ITU-T P.835

(11/2003)

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

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

Methods for objective and subjective assessment of quality

Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm

ITU-T Recommendation P.835

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

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

ITU-T Recommendation P.835

Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm

Summary

This Recommendation describes a methodology for evaluating the subjective quality of speech in noise and is particularly appropriate for the evaluation of noise suppression algorithms. The methodology uses separate rating scales to independently estimate the subjective quality of the Speech Signal alone, the Background Noise alone, and Overall Quality.

Source

ITU-T Recommendation P.835 was approved on 13 November 2003 by ITU-T Study Group 12 (2001-2004) under the ITU-T Recommendation A.8 procedure.

Keywords

Coded speech in background noise, noise preprocessor, noise suppression algorithm, speech quality evaluation, subjective testing.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications. 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 not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementors are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database.

CONTENTS

ITU-T Recommendation P.835

Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm

1 Scope

Typically, Noise Suppression Algorithms (NSA) operate on a noisy speech waveform and attempt to reduce the noise or background component without adversely affecting the speech or signal component of the waveform. This goal can often be realized for relatively low levels of noise suppression. For higher levels of noise suppression, however, NSAs often adversely affect the speech component as more noise is suppressed: there tends to be increasing degradation of the speech or signal component as more of the noise or background component is removed. In this situation, subjects can often become confused as to what they should be responding to in their ratings of the overall "quality" of the waveforms: while the background may have been improved because there is less noise present in the waveform, the speech signal may have been degraded in the process. In a single-scale rating method, the ACR, for example, each individual subject weights the signal and the background components in determining his ratings of overall speech quality. This weighting process introduces additional error variance in the subjects ratings of overall quality resulting in decreased reliability in those ratings. The methodology described in this Recommendation reduces the listener's uncertainty by requiring him to successively attend to and rate the waveform on: the *speech signal*, the *background noise*, and the *overall effect: speech + background*.

While this methodology has been shown to be reliable and valid for evaluating NSAs, it should not be restricted to testing NSA. The methodology can be used for the more general case of evaluating conditions of speech in background noise. It is particularly applicable in those cases where it is unknown whether a system includes a noise preprocessor.

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 Recommendation G.191 (2000), *Software tools for speech and audio coding standardization.*

– ITU-T Recommendation P.56 (1993), *Objective measurement of active speech level.*

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

– ITU-T Recommendation P.810 (1996), *Modulated noise reference unit (MNRU).*

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

3 Definitions

This Recommendation defines the following term:

3.1 dBov: dB relative to overload.

4 Abbreviations

This Recommendation uses the following abbreviations:

ACR	Absolute Category Rating
ANOVA	ANalysis Of VAriance
D/A	Digital-to-Analogue
MANOVA	Multiple ANalysis Of VAriance
MOS	Mean Opinion Score
NSA	Noise Suppression Algorithm
RMS	Root Mean Square
SNR	Signal-to-Noise Ratio
SPL	Sound Pressure Level

5 Experimental design

5.1 Speech material

5.1.1 Source speech material

The source speech material should be meaningful sentences representative of the language under test and including multiple speech samples for both male and female talkers.

5.1.2 Processing

Standard laboratory procedures shall be followed to ensure that the processed speech and noise type samples are mixed and filtered properly (see ITU-T Rec. G.191 (Software Tool library) and Appendix I).

5.1.3 Reference conditions

The reference conditions shall be selected to independently vary the signal and background ratings through their entire range of scale values. For example, speech in background noise should be varied along two dimensions, Speech-to-Noise Ratio (SNR) for varying the background ratings and MNRU for varying the signal ratings.

Figure 1 illustrates the relative independence of the signal score and the correlation of the overall score to the background score when MNRU is varied while keeping SNR constant.

Figure 2 illustrates the relative independence of the background score and the correlation of the overall score to the signal score when SNR is varied while keeping MNRU constant.

Figure 3 shows that the introduction of these combined reference conditions provide a full context within this two-dimensional perceptual space (signal/background).

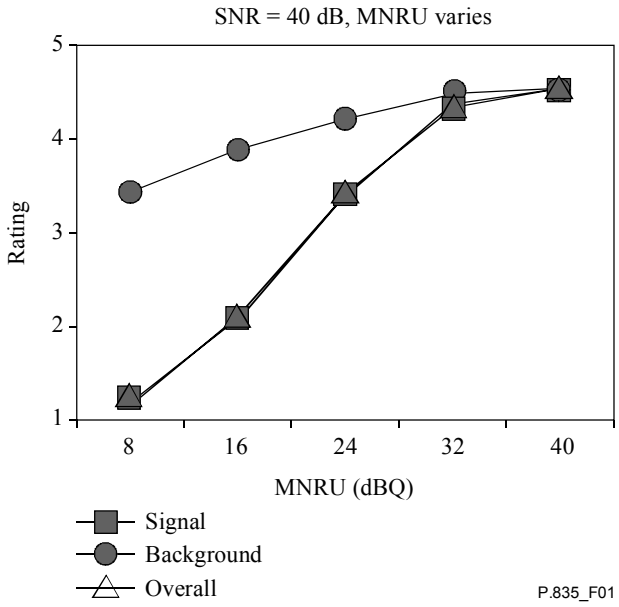


Figure 1/P.835 – Reference condition: SNR constant, MNRU varies

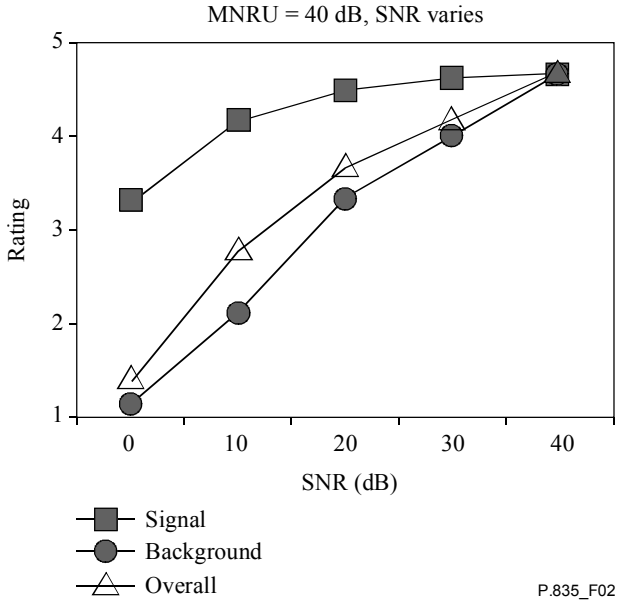


Figure 2/P.835 – Reference condition: MNRU constant, SNR varies

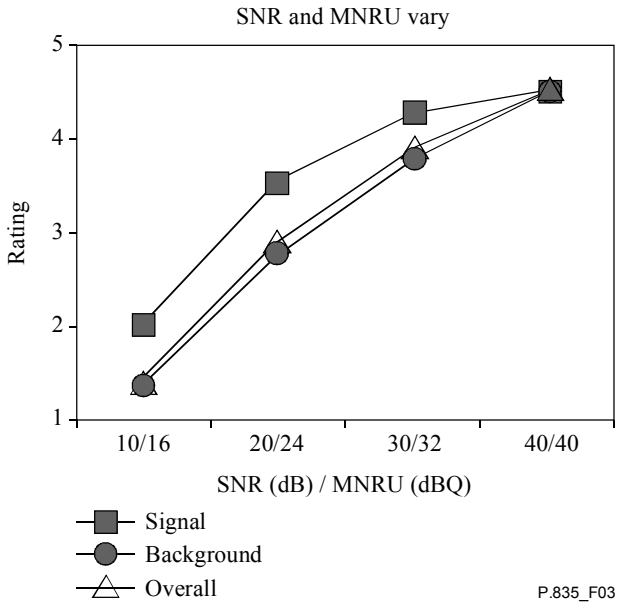


Figure 3/P.835 – Reference condition: SNR and MNRU vary

5.1.4 Speech sample presentation

Each trial contains a three-sentence sample of speech laid out in the general format illustrated in the example given in Figure 4. Each sample is comprised of three sub-samples, where each sub-sample is followed by a silent voting period. In the example shown in the figure, each sub-sample is approximately 4 s in duration including 1 s of background noise alone, 2 s of speech + noise, 1 s of background, and an appropriate silent voting interval. In practice, the sub-samples should be as long as necessary for the subjects to make reliable ratings. For the first two sub-samples, listeners rate either the signal **or** the background depending on the rating scale order specified for that trial. For the signal, subjects are instructed to attend ***only*** to the ***speech signal*** and rate the speech on the five-category distortion scale shown in Figure 5. For the background, subjects are instructed to attend ***only*** to the ***background*** and rate the background on the five-category intrusiveness scale shown in Figure 6. For the third sub-sample in each trial, subjects are instructed to listen to the speech + background and rate it on the five-category overall quality scale shown in Figure 7, the Mean Opinion Score (MOS) used with the ACR.

To control for the effects of rating scale order, the order of the rating scales shall be balanced across the experiment, i.e., scale order should be "Signal, Background, Overall Effect" for half of the trials, and "Background, Signal, Overall Effect" for the other half. Furthermore, rating scale order should be counter-balanced across listening panels.

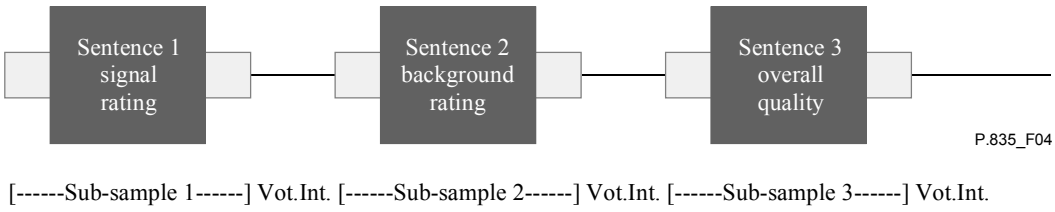


Figure 4/P.835 – Example of the timing of the speech materials in a P.835 trial

NOTE 1 – Experiments have shown that the sequence duration may be 4 s or 8 s, without influencing the results. The use of the shorter duration reduces the overall test duration.

NOTE 2 – Experiments have shown that sentences 1, 2 and 3 in Figure 4 may be the same in a complete sequence or may be different. This factor does not influence the results.

5.2 Listening session

5.2.1 Listeners

At least 32 naïve listeners shall participate in the tests.

All the listeners shall be native speakers of the language used for the test and no listener shall have participated in a subjective experiment in the previous three months.

5.2.2 Audio presentation

Audio presentation shall comply with the guidelines given in ITU-T Rec. P.800. These guidelines include the listening system, listening levels, test duration and listening environment.

5.2.3 Instructions and rating scales

Listeners shall receive written instruction in the rating tasks to be performed in the methodology. The instructions are provided in text form to avoid ambiguity and differences across experiments and across listening panels within an experiment. The instructions should show examples of the three rating scales involved in the methodology. Examples of the three rating scales in English are shown in Figure 5 for the Speech Signal rating, Figure 6 for the Background Noise rating, and Figure 7 for the Overall Quality rating. The rating scales and category descriptors in languages other than English should provide a close translation of those shown in the example figures.

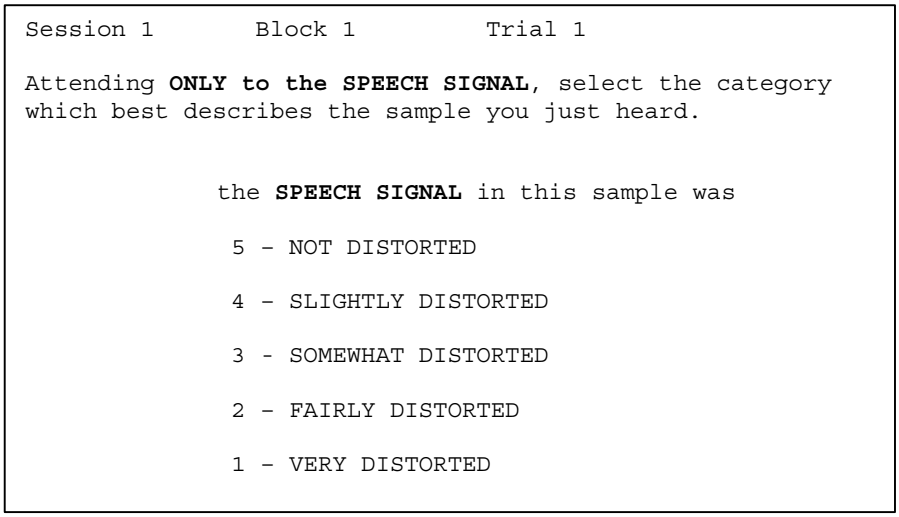


Figure 5/P.835 – Speech signal rating scale

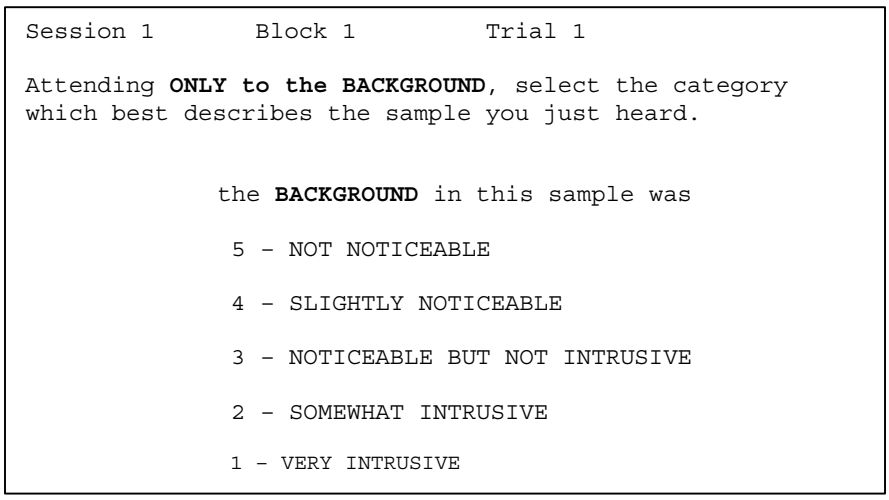


Figure 6/P.835 – Background noise rating scale

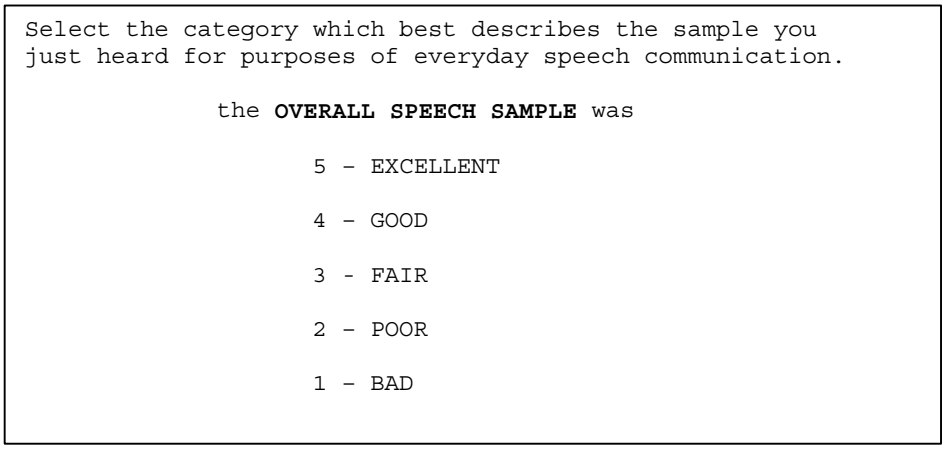


Figure 7/P.835 – Overall quality rating scale (same as the MOS rating scale) used in the ACR procedure (see ITU-T Rec. P.800)

An example of an Instructions sheet is given in Appendix II, in the case of the order "Signal, Background noise, Overall Quality". It shall be adapted in the case of the order "Background noise, Signal, Overall Quality".

5.2.4 Voting process and data collection

Push-button score boxes or other suitable means shall be used to collect votes from the subjects. Voting is only permitted following the completed presentation of each voting stimulus. Listeners are required to register responses prior to the subsequent presentation of a new stimulus. The scale to be used by subjects ("Speech signal distortion" or "Background noise intrusiveness" or "Overall quality") should be made apparent for each sub-sample presentation.

5.3 Data analysis

5.3.1 Analysis methods

Depending on the experimental design, t-tests, Tukey's test, ANOVA, or MANOVA shall be conducted, as appropriate.

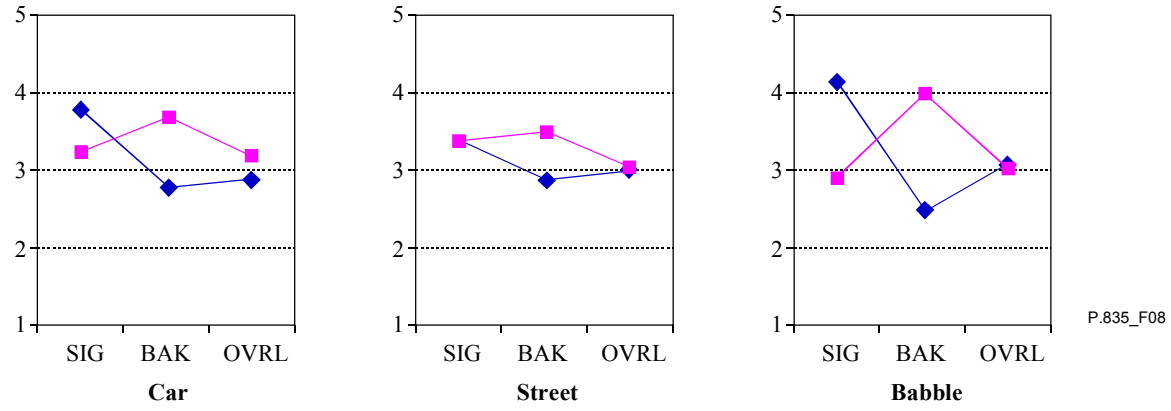
5.4 Presentation and interpretation of results

5.4.1 Summary results

Summary results should include, at a minimum mean ratings and standard deviations for all talkers and for male and female talkers. Other summary statistics, e.g., confidence intervals, should be included as appropriate for the experiment.

5.4.2 Score profiles (Signal, Background, Overall)

While the primary result for this methodology is the Overall Quality score, the score profiles, i.e., the combination of Signal, Background, and Overall scores, provide important information for the subjective quality of a specific system or condition.



SIG Signal
BAK Background
OVRK Overall

Figure 8/P.835 – Score profiles

Appendix I

Procedure for proper mixing of speech and noise samples

I.1 General

The procedure for mixing of speech and noise samples is shown in Figure I.1. The various components of the procedure are described in the following subclauses.

I.2 Parameters

In addition to the choice of source material, mixing conditions are defined in terms of three parameters:

- Speech level. This parameter is expressed in dBov and is the level of the filtered and level-normalized speech measured using the P.56 algorithm.
- Background noise level. This parameter is the RMS level of the filtered background noise.
- SNR. This is signal-to-noise ratio expressed in dB, defined as the ratio of the P.56 speech level to the RMS level of the filtered and level-normalized background noise.

I.3 Speech and background noise files

The speech and background noise input files should be recorded using a flat frequency response.

I.4 Speech and noise input filters

The two input filters simulate the response of a handset to speech and noise respectively. The choice of handset response may depend on the application of interest, for example, the typical response of a mobile handset will be different to that of a fixed-line handset.

In simple simulations, the speech and noise filters may have the same response, for example, the modified IRS specification in ITU-T Rec. P.830. In more sophisticated simulations, the two filters may be different, recognizing the fact that handsets may have a different response to near-field speech and a diffuse noise field.

A set of filter implementations is provided in the ITU-T Software Tool library (ITU-T Rec. G.191).

I.5 P.56 speech level adjustment

The level of the filtered speech file should be adjusted such that its level measured using the method described in ITU-T Rec. P.56 equals the target value, for example –26 dBov. The P.56 speech level measurement excludes periods of silence from the level calculation. A software implementation of process is provided in the ITU-T Software Tool library (ITU-T Rec. G.191).

I.6 Basic noise level adjustment

The level of the filtered noise file should be adjusted such that its RMS level provides the desired SNR when combined with the speech level. Care should be taken that the filtering process does not produce unexpected results with signals that contain a large low-frequency component, such as vehicle noise.

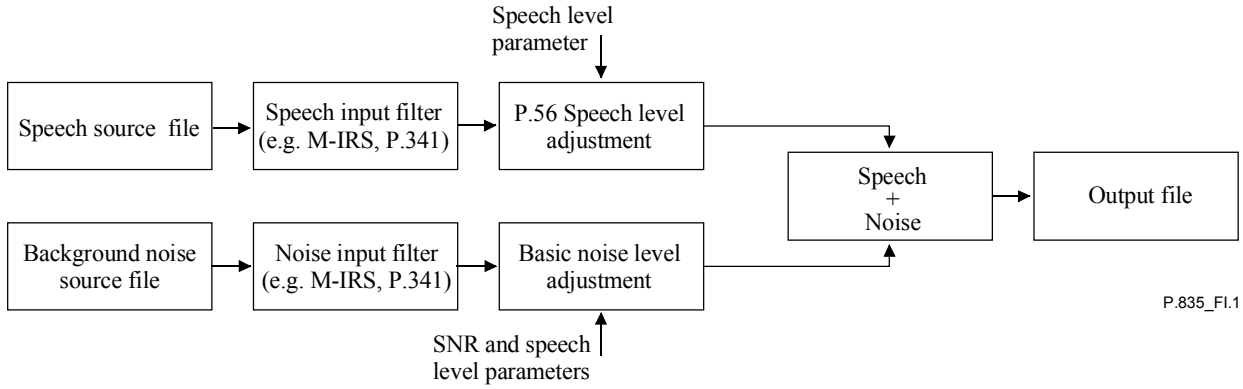


Figure I.1/P.835 – Procedure for mixing speech and background noise files

Appendix II

Example of Instructions to subjects

In this experiment you will be rating the quality of sound samples involving speech in background noise. Each trial will include three 4-second sub-samples where each sub-sample is a sentence in a noisy background. Within each trial you will give three ratings, one for **each** sentence or sub-sample.

For one sentence in each trial you will be instructed to attend **only to the speech signal** and rate how distorted the **speech signal** sounds to you. You will use the rating scale shown in the figure below to register your ratings of the speech signal. Your task will be to choose the numbered phrase from the list that best describes your opinion of the **SPEECH SIGNAL ALONE** and then enter the corresponding number on your keyboard, followed by the <Enter> key.

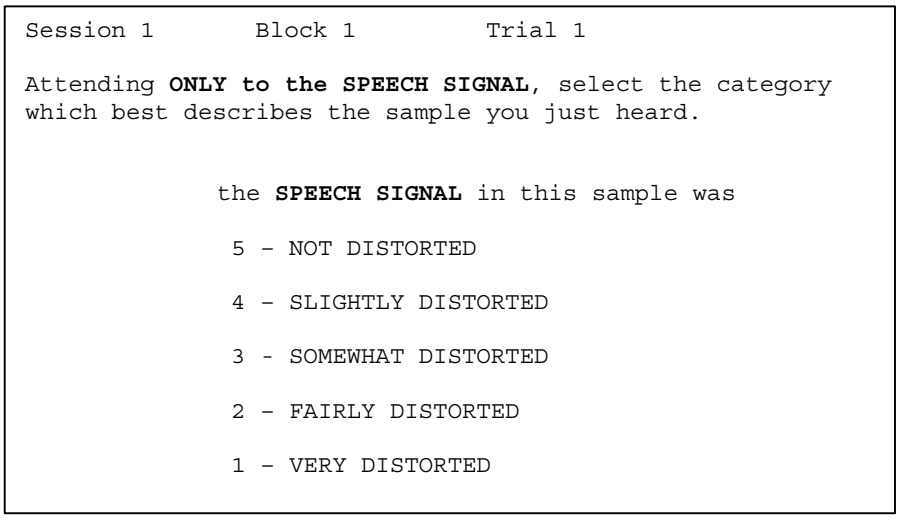


Figure II.1/P.835 – Signal rating scale

For another sentence in each trial you will be instructed to attend **only to the background** and rate how noticeable or intrusive the **background** sounds to you. You will use the rating scale shown in the figure below to register your ratings of the background. Your task will be to choose the numbered phrase from the list that best describes your opinion of the **BACKGROUND ALONE** and then enter the corresponding number on your keyboard, followed by the <Enter> key.

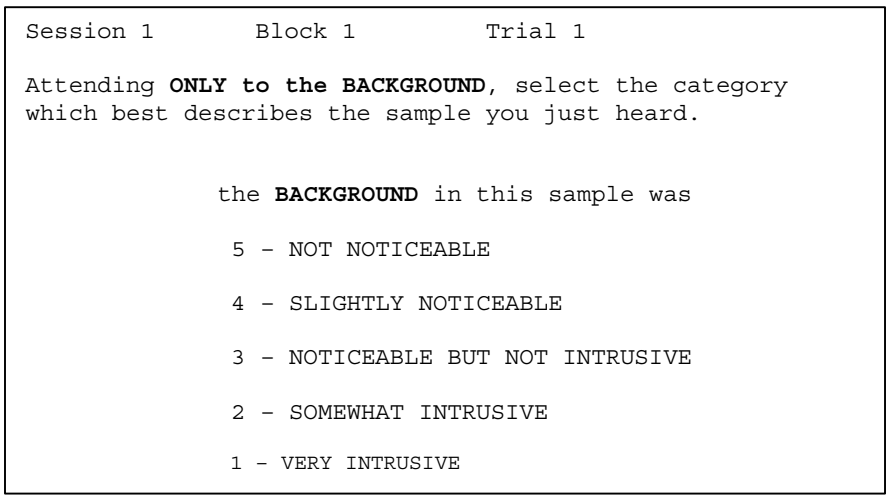


Figure II.2/P.835 – Background rating scale

For the third sentence in each trial you will be instructed to attend to the entire sample (both the speech signal and the background) and rate your opinion of the **OVERALL QUALITY** of the sample for purposes of everyday speech communication.

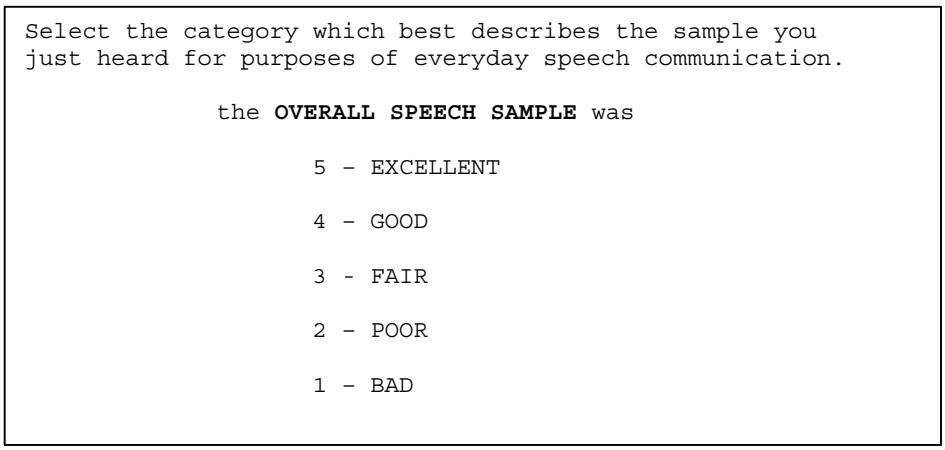


Figure II.3/P.835 – Overall quality rating scale

The experiment will involve two test sessions separated by a short rest period. In one test session you will rate the **signal** for the first sentence, the **background** for the second sentence, and the **overall effect** for the third sentence. In the other session, the order of the ratings will be **background**, then **signal**, then **overall effect**.

Before the first test session you will have a practice block of 8 trials to familiarize you with the rating tasks. The practice block will be followed by 4 test blocks of 18 trials each (approximately 22 minutes). After a short rest period you will have the second test session which will also take approximately 22 minutes (4 blocks of 18 trials each). Each test block begins with a short tone. The test sessions will be intense and will require your complete attention throughout the session in order to keep up with the speech samples and the rating tasks required of you.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series B	Means of expression: definitions, symbols, classification
Series C	General telecommunication statistics
Series D	General tariff principles
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	Construction, installation and protection of cables and other elements of outside plant
Series M	TMN and network maintenance: international transmission systems, telephone circuits, telegraphy, facsimile and leased circuits
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling
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 and open system communications
Series Y	Global information infrastructure, Internet protocol aspects and Next Generation Networks
Series Z	Languages and general software aspects for telecommunication systems

Geneva, 2004