SPEECH QUALITY ASSESSMENT OF COMPOUNDED DIGITAL TELECOMMUNICATION SYSTEMS; PERCEPTUAL DIMENSIONS.

Kim Tilgaard Petersen[†], *Steffen Duus Hansen*[‡], *John Aasted Sørensen*[‡]

[†]Tele Danmark A/S, Research and Development, DK-2630 Taastrup, Denmark
[‡]Department of Matematical Modelling
Technical University of Denmark, DK-2800 Lyngby, Denmark
Email ktp@tdk.dk

1. Abstract

Digital telecommunication networks may involve a multiple number of public switched telephone networks (PSTN), cellular and mobile systems and to some extent also satellite systems. Most of these networks contain non-linear speech coders and other speech algorithms which may degrade the overall end-to-end quality of speech. An important problem is how to assess the speech quality of such compounded systems.
The object of this paper is to describe the first stage of the construction of a proposed three-layer model for speech quality assessment. A subjective test of the speech quality of 16 different compounded transmission paths (mixtures of PCM, GSM full and half rate, DECT, CELP, LD CELP, FS10-16) is carried out by 40 subjects using 21 different rating scales. The main result of this paper is the test results which lead to the definition of four main perceptual dimensions to be used in the second layer of the proposed model[1].

2. Introduction

The performance of speech transmission in digital telecommunication networks could be investigated by extensive subjective tests. It is well-known that these tests are expensive and very time consuming and thus the need of a reliable objective speech quality measure is obvious.

Most of the approaches (e.g. [1]) suggested for objective speech quality assessment are based on an overall objective quality measure. This is calculated directly from the speech samples in order to classify the amount of degradation and distortion generated in a speech transmission path. This method may not be optimal for assessing speech quality in compounded telecommunication systems where perceptually different types of impairments occur. Alternatively, the overall objective speech quality measure may be calculated by weigthing of different perceptual dimensions in the human hearing. The basic model of such an approach is based on a proposed three-layer structure. In Figure 1 the overall model is shown.

The first layer describes the physical parameters, observable in the speech signal. This layer interfaces to the perceptual dimensions in the second layer. An overall objective measure is then obtained in the third layer by weighting the perceptual dimensions by a given function. By using such a model not only the total quality could be calculated but the reason for the degradation could also be investigated in the systems. This is important in relation to evaluation and planning of telecommunication networks.

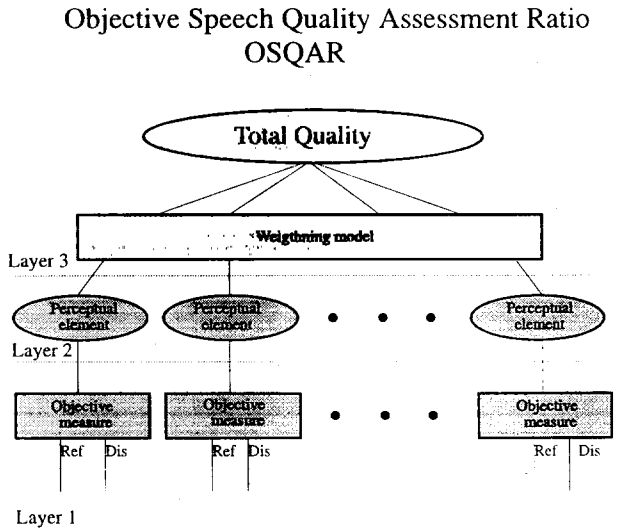

Figure 1: General model for predicting the speech quality in digital telecommunication systems.

The construction of an objective speech quality measure based on the above-mentioned principle is divided into two main phases. The purpose of the first phase is to determine the perceptual dimensions characterizing speech transmission quality in digital telecommunication systems and to obtain the relationship between these dimensions in forming the total quality measure. Then the goal of the second phase is to identify objectively the elements in the telecommunication systems, which influence speech quality, and to establish a relationship between these elements and the perceptual dimensions. The objective of this paper is to concentrate on the first phase.
The paper describes the result of a subjective test and factor analysis for identifying the main perceptual dimensions in a speech quality assessment system.

[1]The work is supported by Tele Danmark A/S and the Danish Academy of Technical Sciences.

1	GSM Full Rate - CELP 8 kBit/s - GSM Full Rate
2	Bit quantisation
3	DECT - CELP 8 kBit/s - DECT
4	DECT - CELP 8 kBit/s - GSM Full Rate
5	Amplitude modulation - 60 %
6	PCM
7	GSM Full Rate - PCM - GSM Full Rate
8	FS10-16
9	LD CELP - LD CELP - DECT
10	CELP 8 kBit/s - CELP 8 kBit/s - DECT
11	PCM - CELP 8 kBit/s
12	LPC modification
13	PCM - GSM Half Rate
14	DECT - DECT - DECT - DECT
15	PCM - GSM Full Rate
16	Bandpass filtering (400 Hz-3000 Hz)

Table 1: Transmission path used for subjective test.

3. Transmission Systems - Test Stimuli

Initially the 16 kHz sampled speech signal was downsampled to an 8 kHz sampling rate and filtered by using a modified IRS (Intermediate Reference System) transmitting frequency characteristic[8]. Before being presented to the listener, the speech signal was likewise filtered using a modified IRS receive filter characteristic [8].

The test stimuli were produced as speech signals transmitted through fixed point simulated versions of different standardized speech coding algorithms in cascade. Also included are four artificially generated distortions, each designed for accessing a specific perceptual dimension. These are bit quantization noise, 60 % amplitude envelope modulation, LPC (Linear Prediction Coding) parameter modification within 10 % and bandpass filtering (400 Hz - 3000 Hz). The seven main speech coding algorithms used in the subjective test are:

- 64 kbit/s PCM (Pulse Coded Modulation) [2]
- GSM full rate coder - 13 kbit/s RPE-LTP (Residual Pulse Exited - Long Term Predictor) [6]
- GSM half rate coder - 5.8 kbit/s CELP (Code Excited Linear Prediction) [5]
- 32 kBit/s ADPCM (Adaptive Differential Pulse Coded Modulation) algorithm [3]
- 4.8 kbit/s FS10-16 (Federal Standard) speech coder [9]
- 8 kbit/s CELP coder [7]
- 13 kbit/s Low Delay CELP coder [4]

The speech coding algorithms are cascaded in order to form 12 different transmission paths which together with the four artificially generated distortions form 16 different transmission paths. The stimuli are listed in Table 1 in the order of presentation to the listeners.

4. Test Procedure

The subjects are selected from Tele Danmark A/S and cover the ages from 22 to 51 years. A total number of 40 subjects (29 male and 11 female) were participating in the test.

The quality of 16 different transmission paths was rated on 21 different scales, derived from the DAM (Diagnostic Acceptability Measure) test, [10]. The DAM test originally applied ten separate rating scales for the speech signal, seven rating scales for the background noise and three rating scales for the total quality - a total of 20 scales. As also the echo effect is an important aspect of the transmission quality, this scale was added resulting in 11 rating scales concerned with the speech signal. All rating scales are presented in Table 2.

1	Irregular speech signal (Ir)
2	Bass speech signal (Bass)
3	Thin speech signal (Thin)
4	Rasping speech signal (Rasp)
5	Treble speech signal (Treb)
6	Synthetic speech signal (Synt)
7	Babbling speech signal (Babb)
8	Intermittent speech signal (Int)
9	Nasal speech signal (Nas)
10	Interrupted speech signal (Itr)
11	Echo speech signal (Echo)
12	Hissing background (Hiss)
13	Chirping background (Chirp)
14	Roaring background (Roar)
15	Crackling background (Crac)
16	Humming background (Hum)
17	Rumbling background (Rum)
18	Bubbling background (Bub)
19	Intelligibility (I)
20	Pleasantness (P)
21	Total Quality (TQ)

Table 2: Rating scales used for subjective test in order to define perceptual dimensions.

The listening test was performed binaurally using headphones. As listening method a relative dual-stimulus procedure was selected as: Reference signal - test stimuli. As reference signal a 64 kbit/s PCM coding was selected because the transmission quality of such a connection is often considered as standard PSTN quality. The subject is asked to rate the degree of degradation in the test stimuli relative to the reference.

Furthermore, the 16 transmission paths are augmented with four training sequences of perceptually different transmission paths and initially presented to the listener. This leads to a total of 20 test stimuli. One male and one female speaker (Danish) are selected for the test, resulting in a total of 40 different test stimuli to be rated on each scale.

The rating is performed on each rating scale presenting all 40 test stimuli before the next scale is selected. Ten subjects are rating all 40 test stimuli on a given rating scale. As it may not be pratically possible to allow a given subject to rate 40 test stimuli on 21 different rating scales, the subjects were divided into four groups. Each group rates the test stimuli on 4 - 5 rating scales.

The rating form consists of a continuous scale with four marked degradations. The marked degradations are: "No difference", "Sligthly different", "Some difference", and "Very different". The scale is afterwards converted to figures from 0 to 90.

5. Results

The gross result of the subjective tests is given in a block data matrix $\mathbf{X}' = \left[x'_{i,j,k}\right]$ where $i = 1, \ldots, 21$ represents the rating scales, $j = 1, \ldots, 40$ represents the transmission paths, and $k = 1, \ldots, 40$ represents the subjects. Before further analysis a primary data matrix $\mathbf{X}_{aver} = [x_{i,j}]$ was formed by averaging the subjects $x_{i,j} = E_k\{x_{i,j,k}\}$. Notice that not all subjects rate the test stimuli on all 21 ratings scales.

The interscale variations are calculated by employing a factor analysis model based on the correlation matrix of the averaged scale value datamatrix $\mathbf{X}_{aver}$. As a result of the analysis 90 % of the data variance can be explained by three factors. The first factor explains app. 71 % of the variation, the second factor app. 11 % and the third factor explains nearly 8 %.

The correlation of the scales could be examined by plotting the factor loading elements. The factor loadings for the three most important factors are shown in Figure 2, 3 and 4.

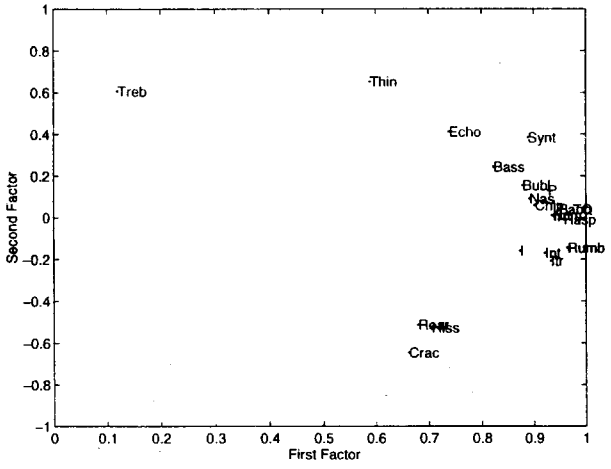

Figure 2: Factor analysis showing the two most dominating factors.

In Figure 2 it is seen that the first factor is relatively dominating and describes most of the information in the different scales. Most of the scales are grouped together on the first factor and it is difficult to identify the scales. In Figure 3 the first and second factor are examined closer by plotting the scales with high values on the first factor.

In general, a common perceptual adjective for the first factor may be "irregularity" and "rasping" in the transmission paths. This incorporates irregularities and rasping in both the speech signal and the background noise. As irregularity and rasping may affect many of the rating scales, it is obvious that most of the scales should correlate highly with the first factor. The factor is most highly correlated with the scales "Rasping", "Babbling", "Irregular", "Humming", "Rumbling", "Intermittent" and "Interrupted".

With respect to the second factor scales like "Thin" and "Treble", and also the scales "Echo" and "Synthetic" are

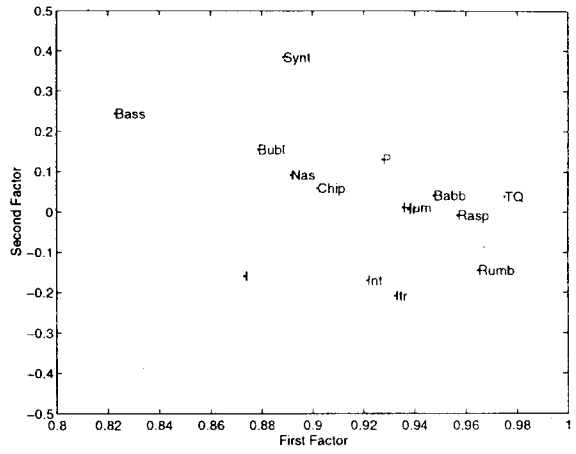

Figure 3: Factor analysis showing the two most dominating factors zoomed on the first factor.

identified on the positive part of the factor. The positive part of the second factor expresses the extent to which the transmission path is thin and contains treble. On the negative part of the second factor the scales "Hissing", "Crackling" and "Roaring" can be seen. These are all associated with the background noise and may describe the perceptual dimension "Hissing/Crackling".

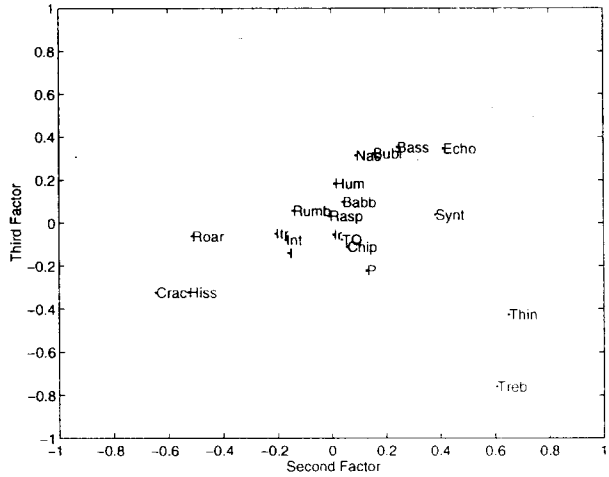

Figure 4: Factor analysis showing the second and third most dominating factors.

In Figure 4 the second and third factor are shown. The third factor correlates positively with the scales "Nasal", "Bobbling", "Bass" and "Echo" and could therefore describe the degree of bass in the transmission paths. On the negative part of the third factor scales belonging to the perceptual dimension already defined are seen.

In total, four perceptual dimensions were identified by the factor analysis. These are:

1. Speech and background: Irregularity/Rasping
2. Speech: Thin/Treble
3. Background: Hissing/Crackling
4. Speech and background: Bass/Nasal

As can be seen, the first and fourth perceptual dimensions incorporate both the speech signal and the background noise. The second dimension incorporates only the speech signal while the third dimension only takes the background into account.

6. Relation Between the Perceptual Dimension and the Total Quality

In order to examine the correlation between the individual perceptual dimensions and the total quality, a linear relation is assumed.

$$TQ_j = b_0 + \sum_{i=1}^{4} b_i S_{i,j} \quad (1)$$

The $S_{i,j}$ $i = 1, \ldots, 4$ coefficients refer to the average values of rating scales associated with the four defined perceptual dimensions and the b_i $i = 1, \ldots, 4$ coefficients are the linear regression constants to be estimated. TQ_j is the total quality for test stimulus j.

Using the average values for the rating scales: Irregularity/Rasping, Thin/Treble, Hissing/Crackling and Bass/Nasal as representatives of the four perceptual dimensions, the following linear model could be established:

$$TQ_j = 8.95 + 0.38S_{1,j} + 0.20S_{2,j} + 0.22S_{3,j} + 0.32S_{4,j} \quad (2)$$

Based on this prediction model a correlation of 96 % to the total quality is achived. In Figure 5 the relation between the measured and the predicted total quality is shown.

7. Summary of Results

In this paper a three layer model has been proposed for assessing the speech quality of compounded digital telecommunication systems. By using such model not only the total speech quality could be evaluated, but the cause for a possible degradation could also be observed and used in network planning. In this paper four perceptual dimensions have been identified as sufficient in evaluating the speech quality in compounded networks. The dimensions are Irregularity/Rasping, Thin/Treble, Hissing/Crackling, Bass/Nasal. Based on a linear prediction model a correlation of 96 % to the total quality could be achived by using these dimensions.

8. References

[1] John G. Beerends and Jan A. Stemerdink. A perceptual audio quality measure based on a psychoacoustic sound representation. *J. Audio Eng. Soc.*, 42(3), 1994 March.

[2] CCITT. *CCITT Recommendation G.711, 'Pulse code modulation (PCM) of voice frequencies'*, 1988.

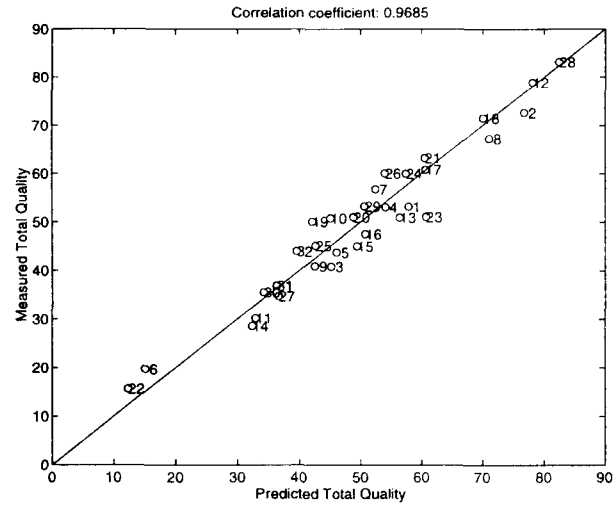

Figure 5: Relation between the measured and the predicted total quality. The numbers 1-16 refer to the transmission paths with the male speaker and the numbers 17-32 refers to the transmission paths with the female speaker

[3] CCITT. *CCITT Recommendation G.726, '40-, 32-, 24-, and 16-kbit/s adaptive differential pulse code modulation'*, Geneva, December 1990.

[4] CCITT. *CCITT Recommendation G.728, 'Coding of speech at 16 kbit/s using low-delay code excited linear prediction'*, Geneva, September 1992.

[5] ETSI. *Recommendation ETS 300 581-2, 'European digital cellular Telecommunications system - Half rate speech transcoding'*. ETSI.

[6] ETSI. *Recommendation ETS 300 580-2, 'European digital cellular Telecommunications system - Full rate speech transcoding (GSM 06.10)'*. ETSI, September 1994.

[7] ITU-T. *ITU-T Recommendation G.729, 'Coded Exited Linear Prediction'*, 1996.

[8] ITU-T. *Draft revised recommendation P.83 'Subjective performance assessment of telephone-band and wideband digital codecs'*. ITU-T., April 1995.

[9] T. E. Tremain J. P. Campbell and V. C. Welch. *The DoD 4.8 kbps standard (proposed federal standard 1016).* in Advances in Speech Coding (B.S Atal, V. Cuperman, and A. Gersho, eds), pp. 121-134, Kluwer Academic Publishers, 1991.

[10] W. D. Voiers. Diagnostic acceptability measure for speech communication systems. *Proc. IEEE ICASSP*, pages 204-207, 1977.