TOWARDS A NEW REFERENCE SYSTEM FOR SUBJECTIVE EVALUATION OF CODING TECHNIQUES

T. Etame Etame[1], R. Le Bouquin Jeannès[2,3], C. Quinquis[1], L. Gros[1], G. Faucon[2,3]

[1]France Telecom R&D TECH/SSTP - 2 Av. Pierre Marzin, 22307 Lannion Cedex, France
[2]INSERM, U 642, Rennes, F-35000, France
[3] Université de Rennes 1, LTSI, F-35000, France
LTSI, Campus de Beaulieu, Université de Rennes 1, 35042 Rennes Cedex, France
phone: +33 (0)2 23 23 69 19, fax: +33 (0)2 23 23 69 17
[1] th_etame@yahoo.fr, Catherine.Quinquis@orange-ftgroup.com, Laetitia.Gros@orange-ftgroup.com
[2,3] Regine.Le-Bouquin-Jeannes@univ-rennes1.fr, [2,3] Gerard.Faucon@univ-rennes1.fr
www.ltsi.univ-rennes1.fr; www.francetelecom.com/en_EN/innovation/

ABSTRACT

This paper aims to produce a reference system that can simulate and calibrate degradations of conversational codecs which are currently used on telecommunications networks, for subjective assessment tests of voice quality. At first, 20 wideband codecs are evaluated through subjective tests in order to produce the multidimensional perceptual space underlying the perception of current degradations. Then, a verbalization task helps to identify attribute for each dimension: clear/muffle, background noise, noise on speech and hiss. These dimensions are characterized with correlates such as spectral centroid, energy in the silent part in the high frequency sub-band, ratio of brightness between deterministic part and residual part of the signal and correlation coefficient. This 4-dimension perceptual space is stable for male and female talkers. A new reference system is proposed. The phase of validation leads to a perceptual space with five dimensions including the four previous ones.

1. INTRODUCTION

Together with the evolution of telecommunications systems, the source compression algorithms are more and more complex to cope with limited resources. Foreseen applications for new codecs are not only speech but also mixed content or even music. Mixed content represents advertisement, ring back tones, music on hold and even film trailers. In order to ensure good inter-working between different equipments, source codecs have to be standardized and a part of the standardization exercise is the quality evaluation. The more accurate way to evaluate quality on new algorithms is to use subjective methodology. Those subjective methodologies currently use reference signals to allow comparison of test results across different experiments. For the time being, Modulated Noise Reference Unit (MNRU)[1] is the reference system used together with the Absolute Category Rating (ACR) methodology and the Degradation Category Rating (DCR) [2]. This reference model is directly related to the degradation given by the codecs using pulse code modulation (PCM) like G.711.

Nowadays, the codecs are built on different methods like Code-Excited Linear Prediction (CELP), transform or even hybrid between CELP and transform. Those codecs bring default on the reconstructed signals different from the ones produced by PCM making the MNRU system obsolete.

The paper is organized as follows: section 2 is dedicated to the selection of speech and audio codecs and dissimilarity tests. Section 3 shortly presents the multidimensional scaling and the subjective space obtained with the codecs. Section 4 assesses objectively the characteristics of the subjective space. Section 5 presents the validation steps and Section 6 contains the conclusions of the paper.

2. SELECTION OF SPEECH AND AUDIO CODECS AND DISSIMILARITY TESTS

In the last decade, most of the activities have mainly focussed on Wideband (50-7000 Hz) and preliminary actions have been taken on Superwideband codecs (50-14000 Hz) and Fullband codecs (20-22000 Hz). Consequently, we decided to first focus on the wideband codecs but in order to cover a maximum of existing techniques of compression algorithms, some superwideband or fullband codecs have also been retained.

2.1 Speech and audio codecs

The goal is to include the different coding techniques as much as possible to obtain different types of degradation.

- G.722 is considered at three bitrates: 48 kbit/s, 56kbit/s and 64 kbit/s. This algorithm may be used for 7kHz-bandwidth audio signals. The coding system uses sub-band adaptive differential pulse code modulation. The frequency band is split into two subbands (low and high) and the signals in each subband are encoded using Adaptive Differential PCM.
- G.722.1 is considered at two bitrates: 24 kbit/s and 32 kbit/s. This algorithm may be used for 7kHz-bandwidth audio signals. The coding system uses a Modulated Lapped Transform (MLT) and Huffman Coding. An extension of this algorithm permits 14 kHz audio bandwidth using a 32 kHz audio sample

rate, and is called G.722.1 annex C. This one is considered at the birate of 24 kbit/s.

- G.722.2 is considered at 8.85, 12.65, 15.85 and 23.85 kbit/s. This algorithm may be used for 7kHz-bandwidth audio signals. The frequency band is split into two sub-bands (low: 50-6400 Hz and high: 6400-7000 Hz) where Algebraic CELP (ACELP) is used in the lower band and the higher band is reconstructed by white noise filtering.
- G.729.1 is considered at bitrates of 14, 20, 24 and 32 kbit/s. This algorithm may be used for 7kHz-bandwidth audio signals. The underlying algorithm is based on a three-stage coding structure: embedded CELP coding of the lower band (50-4000 Hz), parametric coding of the higher band (4000-7000 Hz) by Time-Domain BandWidth Extension (TDBWE), and enhancement of the full band (50-7000 Hz) by a predictive transform coding technique referred to as Time-Domain Aliasing Cancellation (TDAC).
- HE-AAC is considered at bitrates of 16, 24 and 32 kbit/s. This algorithm may be used for 20kHz-bandwidth audio signals. The algorithm consists of MPEG-4 AAC, MPEG-4 SBR. The AAC is a general waveform audio codec using MDCT, SBR is a bandwidth extension technique.
- MP3 is considered at bitrates of 32 and 64 kbit/s. This algorithm may be used for 20kHz-bandwidth audio signals. The algorithm uses,a hybrid filterbank (polyphase + MDCT) and a psychoacoustic model.

2.2 Selection of the codecs

The goal of the selection is to retain about twenty codecs with about the same subjective quality. In order to enlarge the panel of considered codecs, the different codecs described above are considered in one, two or three tandems. One tandem is the fact of encoding the input signal and decoding the bitstream, so that two tandems operation means that the input audio signal is first encoded, the bitstream is decoded and gives a reconstructed audio signal; this reconstructed audio signal is then encoded and the bitstream is decoded giving another reconstructed audio signal. Taking the 19 proposed codecs (including the different bitrates) and adding a condition representing the input signal, we obtain fifty eight conditions.

Four samples are used, two from male talkers and two from female talkers, extracted from France Telecom Database. Each sample consists of two sentences in French language separated by a silence and lasts eight seconds. All samples are at 48 kHz sampling rate. They are presented to the codec at 48, 32 or 16 kHz sampling rate depending on the codec itself. After the processing, all signals are down sampled to 16 kHz and filtered by the P.341 filter to conform to the 50-7000 Hz bandwidth. For MP3 and HE-AAC, the bandwidth reduction is applied after the coding in order to ensure that the considered bandwidth is the same for all conditions.

Those signals are presented to thirty two naive subjects in a test following the ACR methodology.

	Description		Description
+ 1	G722.1C_24kbps_x2	°11	G722_56kbps_x2
+ 2	G722.1C_24kbps_x3	°12	G722_56kbps_x3
+ 3	G722.1_24kbps_x2	*13	G729.1_14kbps_x3
+ 4	G722.1_24kbps_x3	*14	G729.1_20kbps_x3
x 5	G722.2_12.65kbps_x2	*15	G729.1_24kbps_x2
x 6	G722.2_12.65kbps_x3	*16	G729.1_32kbps_x3
x 7	G722.2_15.85kbps_x2	□17	HEAAC_24kbps_x2
x 8	G722.2_8.85kbps_x2	□18	HEAAC_32kbps_x2
° 9	G722_48kbps_x2	□19	MP3_32kbps_x1
°10	G722_48kbps_x3	□20	MP3_32kbps_x2

Table 1. Selected codecs/tandems
(the extensions x1, x2, x3 indicate the number of tandems)

The subjects are asked to score the quality on the following 5-point category scale:

5 - Excellent
4 - Good
3 - Fair
2 - Poor
1 - Bad

The scores are then averaged by condition to give a Mean Opinion Score (MOS).

The MOS values cover a range between 1.33 (for HE-AAC at 16 kbit/s with 3 tandems) and 4.34 (for the direct signal). The selection rule of the codecs/tandems was to retain the ones of medium quality in the range of [2.3; 3.5] and to keep all kinds of coding techniques. The twenty codecs/tandems selected for the dissimilarity test are in Table 1.

2.3 Dissimilarity tests

The speech quality is commonly considered as a multi-dimensional phenomenon. In order to take into account the multidimensional nature of speech quality without introducing semantic descriptor, we choose the MultiDimensional Scaling (MDS) technique [3], [4], [5], [6] that consists in studying the perceptual structures which underlie the judgments of similarities given for pairs of stimuli.

The dissimilarity test is run once for a male talker and once for a female talker. The stimuli presented in the dissimilarity test are constituted of pairs of samples. One sample constituted of two sentences separated by a silence is processed through the twenty selected codecs/tandems. The subject is asked to give a score of dissimilarity (0-similar; 100-dissimilar) to each pair of samples; the subject may listen to the pair as many times as necessary. Some null pairs are included in the test to help a post-screening of the subjects in order to reject non reliable subjects.

After running the dissimilarity test, the subject is invited to describe with his own words the impairment he hears when listening to each of the twenty samples.

Twenty nine subjects participated in the test with the male voice and twenty eight subjects in the test with the female voice. After the post screening, twenty five subjects are found reliable for each test.

20 objects 25 subjects	Male talker space		Female talker space	
	Stress	RSQ	Stress	RSQ
2-Dimensional	0.30461	0.51164	0.31862	0.50713
3-Dimensional	0.24347	0.56581	0.25507	0.55138
4-Dimensional	**0.21289**	**0.60833**	**0.22433**	**0.56089**
5-Dimensional	0.18297	0.61067	0.18779	0.56382
6-Dimensional	0.16595	0.60601	0.15889	0.56118

Table 2. Stress and RSQ
for male and female perceptual spaces

For each reliable subject, the dissimilarity scores are translated in distance matrices, the matrices are then used to extract the multidimensional space underlying the perceptual space.

3. MULTIDIMENSIONAL SCALING AND THE SUBJECTIVE SPACE

3.1 Choice of the algorithm

The order induced by the judgments of dissimilarities in subjective listening experiments is more reliable than the numbers given by listeners. In addition, the inter-individual variability between subjects has also to be taken into account. In our study, we chose a non-metric INDividual difference SCALing (INDSCAL MDS from ALSCAL of SPSS-X) which takes into account the characteristics of perceptual evaluations of audio quality.

3.2 Perceptual space

In the analysis, for each dimension in the range of [2;6], stress and squared correlation (RSQ) are estimated as well as the weight of each dimension (see Table 2 for stress and RSQ). Stress values are enhanced as the number of dimensions increases but the RSQ suggests that a 4-dimensional space is the best compromise for the perceptual space. This is true for the perceptual space issued from the male talker as well as the one coming from the female talker. The male talker perceptual space is shown in Figures 1, 2 and 3. The verbalization task, the projection of the objects on the dimensions, and the informal listening test of objects located at the extremities of each dimension suggest the following characterization of the dimensions: clear/muffle for dimension 1, background noise for dimension 2, noise on speech for dimension 3 and hiss for dimension 4.

4. OBJECTIVE CHARACTERISTICS OF THE SUBJECTIVE SPACE AND REFERENCE SIGNALS

4.1 Objective characteristics

4.1.1 "Clear/muffle" attribute

The projection of the objects on dimension 1 shows the CELP codecs/tandems at one end and the others at the other end.

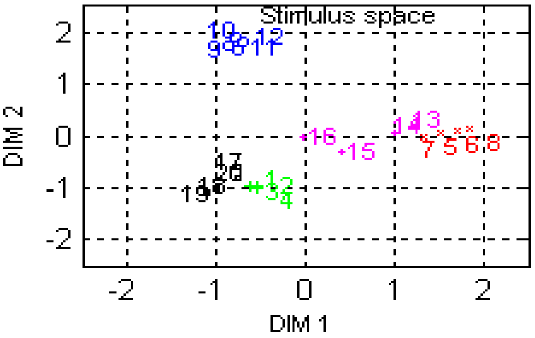

Figure 1 - Perceptual space for male talker (dim. 1 & 2)

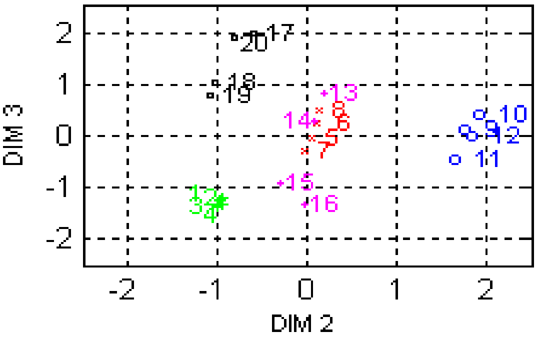

Figure 2 - Perceptual space for male talker (dim. 2 & 3)

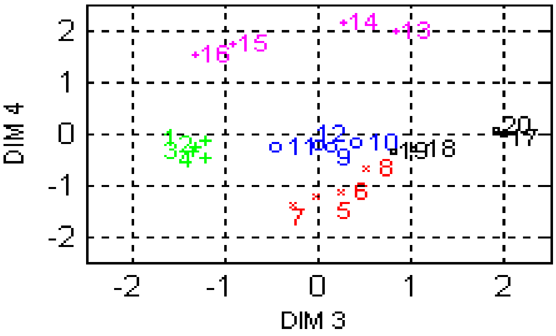

Figure 3 - Perceptual space for male talker (dim. 3 & 4)

Correlation values	DIM 1	DIM 2	DIM 3	DIM 4
SC (male talker)	-0.93*	0.01	-0.07	0.06
SC (female talker)	-0.91*	0.17	0.23	-0.35

Table 3. Correlations of SC (spectral centroid)
for all dimensions
(the asterisk means that the correlation value is significant)

CELP codecs have difficulty to reproduce the highest frequencies of the spectrum and this problem is even enhanced with the G.722.2 family where the coding scheme works on sub-bands and the higher sub-band is reproduced via spectral duplication. During the verbalization task, subjects tend to characterize G.722.2 codec with the attribute muffle. Spectral Centroid (SC) is an indicator of brightness [7]; we compute SC and the correlation coefficient for all dimensions. The results are given in Table 3. It appears that the Pearson correlation is significant only for dimension 1 for both cases, male perceptual space and female perceptual space.

Correlation values	DIM 1	DIM 2	DIM 3	DIM 4
Av. energy high band (male)	-0.47	0.85*	-0.05	-0.15
Av. energy high band (female)	-0.50	0.84*	-0.02	-0.14

Table 4. Correlations of energy (high band) for all dimensions (the asterisk means that the correlation value is significant)

Correlation values	DIM 1	DIM 2	DIM 3	DIM 4
RSC (male)	-0.11	-0.54	-0.61*	-0.27
RSC (female)	-0.40	-0.54	-0.61*	-0.32

Table 5. Correlations of RSC for all dimensions (the asterisk means that the correlation value is significant)

Correlation values	DIM 1	DIM 2	DIM 3	DIM 4
Rmax (male)	-0.19	0.01	0.08	-0.90*
Rmax (female)	-0.16	0.11	-0.00	-0.56*

Table 6. Correlations of Rmax for all dimensions (the asterisk means that the correlation value is significant)

4.1.2 *"Background noise" attribute*

The projection on the second dimension shows the G.722 family as representative of this dimension and the subjects characterize its impairment as background noise. A measure of the averaged energy in the silent part in the high and low frequency sub-bands may be a good objective measurement to characterize dimension 2.

Pearson correlation of the averaged energy in the silent part on the high frequency sub-band for the four dimensions is shown in Table 4. The correlation with dimension 2 is very high for both perceptual spaces (male and female talkers).

4.1.3 *"Noise on speech" attribute*

The third dimension is characterized by the subjects as noise on speech. We propose to use the ratio of brightness (RSC) between deterministic part and residual part of the signal as an objective measurement to characterize dimension 3. Pearson correlation of this measure for the four dimensions is shown in Table 5. The correlation with dimension 3 is higher than for the other dimensions for the perceptual spaces of both male and female talkers.

4.1.4 *"Hiss" attribute*

Pearson correlation of the averaged energy in the silent part on the low frequency sub-band and dimension 4 is -0.87 for both male and female perceptual spaces.

When analyzing the perceptual space it appears that G.729.1 family breaks away from the others along dimension 4 and this correlation may be related to the silence cleaning introduced in the G.729.1. Another impairment brought by G.729.1 is a sort of hiss. Hiss and reverberation are linked to variation in spectrum, so that we compute the maxima of correlation (Rmax) between the original signal and its reconstructed version (see Table 6).

It appears that the Pearson correlation is significant only for dimension 4 for both cases (male perceptual space and female perceptual space).

	Description		Description
+ 1	G722.1C_24kbps_x3	□11	MP3_32kbps_x1
+ 2	G722.1_24kbps_x3	^12	**refDIM1_fc3500**
x 3	G722.2_15.85kbps_x2	^13	**refDIM1_fc4500**
x 4	G722.2_8.85kbps_x2	^14	**refDIM2_Gain40**
° 5	G722_48kbps_x2	^15	**refDIM2_Gain45**
° 6	G722_56kbps_x3	^16	**refDIM3_Mnru30**
* 7	G729.1_14kbps_x3	^17	**refDIM3_Mnru35**
* 8	G729.1_20kbps_x3	^18	**refDIM3_Mnru40**
* 9	G729.1_32kbps_x3	^19	**refDIM4_decorHF_x1**
□10	HEAAC_24kbps_x2	^20	**refDIM4_decorHF_x2**

Table 7. Selected objects for validation phase (the extensions x1, x2, x3 indicate the number of tandems)

20 objects 17 subjects	Male talker space	
	Stress	RSQ
2-Dimensional	0.35178	0.41078
3-Dimensional	0.26628	0.44901
4-Dimensional	0.21528	0.46155
5-Dimensional	**0.18584**	**0.49011**
6-Dimensional	0.16763	0.49807

Table 8. Stress and RSQ for validation

4.2 Reference signals

These statistical analyses allow us to produce some reference signals. For dimension 1, which is characterized by the muffle/clear attribute, a reference system may be produced by filtering the input signal using a low-pass filter. Dimension 2 being characterized by background noise, we propose a reference signal by adding white noise to the original signal with different signal-to-noise ratios. Since dimension 3 is characterized by noise on speech, we propose to use the old reference system MNRU. For dimension 4, we introduce a phase discontinuity between the low and high sub-bands. To this end, we first filter the signal to separate high and low sub-bands and the high sub-band is filtered by a decorrelation filter before the reconstruction of the signal. Decorrelation is achieved by filtering the high sub-band signal with an all-pass filter having random phase response as indicated by Kendall in [8].

5. VALIDATION PHASE

The validation phase uses the same process as the ones that created the perceptual space. First, an ACR test is run to select twenty objects to be used in the dissimilarity test. The dissimilarity test provides matrices used in the MDS analysis to create the final perceptual space.

In the ACR test, thirty one of the conditions of the previous ACR test are kept and reference signals are introduced. Seven band limited signals are introduced as reference for dimension 1, nine created signals for dimension 2 and eight signals for both dimension 3 and dimension 4. Twenty four naive subjects participated in this ACR test.

As previously the selection is done in the medium quality range and with the limitation to keep codecs/tandems from each family of coding technique and reference for each dimension. The selected objects are shown in Table 7.

A dissimilarity test is run with seventeen subjects and the dissimilarity scores are translated in distance matrices, the matrices are then used to feed an INDSCAL MDS. The results of the MDS in terms of stress and RSQ are given in Table 8. The RSQ suggests a 5-dimensional space but it appears more difficult to characterize the five dimensions of this perceptual space. The male talker space is shown in Figures 4, 5, 6 and 7.

Dimension 1 may be characterized by the "hiss" attribute but no reference signal is characterizing this dimension. Dimension 2 is related with "muffle/clear" attribute and the reference signals 12 and 13 are at one end of this dimension. Dimension 3 is characterized by noise on speech and reference signals 16 and 17 are on one end of this dimension. Dimension 4 is characterized by background noise since reference signals 14 and 15 as well as the representative signals of G.722 family are on one end of this dimension. It seems more difficult to characterize Dimension 5.

6. CONCLUSION

The paper presents the multidimensional perceptual space underlying the perception of current degradations provided by the different coding techniques. This perceptual space is stable across gender. A set of reference signals is proposed. The stress and RSQ of the validation step leads to a 5-dimensional perceptual space. Nevertheless, the projection of the objects in the 5-dimension space does not allow a complete explanation of the dimensions. In fact, four of them may be characterized using the attributes of the previous phase, but the fifth dimension remains hard to characterize.

The reference signals have an influence on the resulting perceptual space and defining better suited reference signals may lead to a different perceptual space easier to interpret. For example, a reference system for the dimension characterized by the "hiss" attribute might be defined using different decorrelation filters and might lead to a stable 4-dimension perceptual space.

After this first step of validation, the reference system should be validated for different types of input signal, for example noisy speech and music. Then, the case of other bandwidths should be addressed.

REFERENCES

[1] ITU-T Recommendation P.810 "Modulated Noise Reference Unit"

[2] ITU-T Recommendation P.800 "Methods for subjective determination of transmission quality"

[3] M. Wältermann, K. Scholz, A. Raake, U. Heute and S. Möller, Underlying quality dimensions of modern telephone connections, Interspeech, paper 1089-Wed3FoP.11, 2006.

[4] V.V. Mattila, Ideal point modelling of the quality of noisy speech in mobile communications based on multidimensional scaling, AES 114th convention, Amsterdam, The Netherlands, 2003.

[5] T. Etame Etame, G. Faucon, R. Le Bouquin Jeannès, L. Gros and C. Quinquis, Characterization of the multidimensional perceptive space for current speech and sound codecs, AES 124th convention, Amsterdam, The Netherlands, May 17-20, 2008.

[6] J.B. Kruskal, Nonmetric multidimensional scaling: a numerical method, Psychometrika, Vol. 29, pp. 115-129, 1964.

[7] K. Scholz, M. Wältermann, L. Huo, A. Raake, S. Möller, and U. Heute, Estimation of the quality dimension "directness/frequency content" for the instrumental assessment of speech quality, Interspeech, paper 1219-Wed1A3O.6, 2006.

[8] G. S. Kendall, The decorrelation of audio signals and its impact on spatial imagery, Computer Music Journal, 19(4), pp 71-87, 1995.

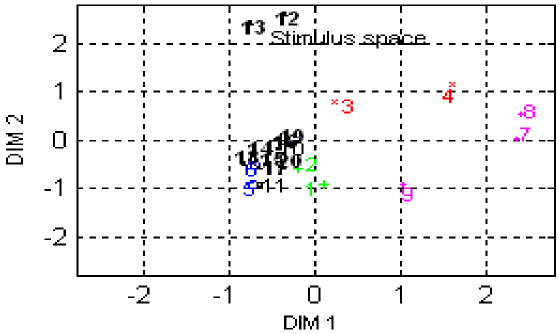

Figure 4 - Perceptual space (dim. 1 & 2)

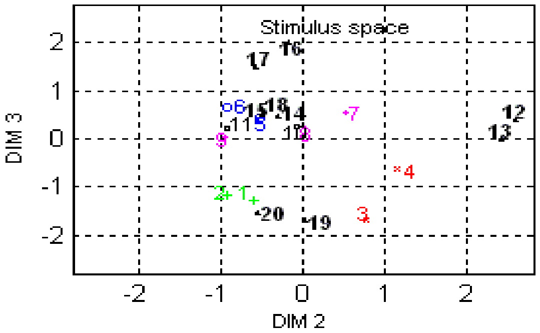

Figure 5 - Perceptual space (dim. 2 & 3)

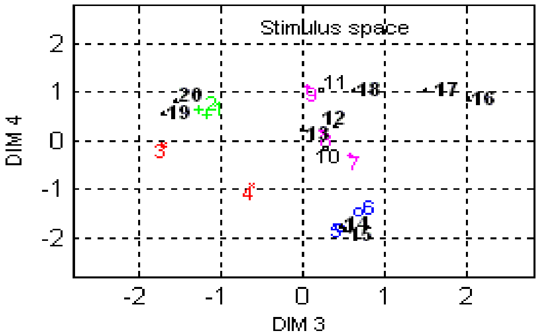

Figure 6 - Perceptual space (dim. 3 & 4)

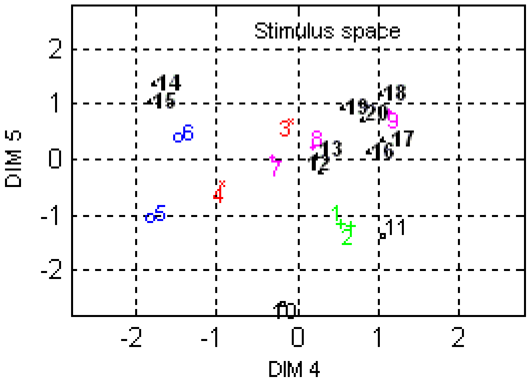

Figure 7 - Perceptual space (dim. 4 & 5)