IDENTIFICATION OF PERCEPTIVE DIMENSIONS OF SPEECH AND AUDIO CODECS SUBJECTIVE QUALITY

Yves Zango [1,2,3], Régine Le Bouquin Jeannès [2,3], Nathalie Costet [2,3] and Catherine Quinquis [1]

[1] Orange Labs - Lannion, 2 Av. Pierre Marzin, 22307 Lannion Cedex, France
[2] INSERM, U 642, Rennes, F-35000 France
[3] Université de Rennes 1, LTSI, Rennes, F-35000, France
{yves.zango, catherine.quinquis}@orange-ftgroup.com, {regine.le-bouquin-jeannes, nathalie.costet}@univ-rennes1.fr

ABSTRACT

New generations of speech and audio codecs have several complex types of impairment. To assess their quality, telecommunication laboratories perform subjective and/or objective assessments. Anchor signals are required to get reliable subjective assessment and comparable results of studies performed at different moments or results of studies from different laboratories. The design of these signals needs a description of the impairments useful to find out their correlated physical characteristics. Assuming that codecs quality is multidimensional, the codecs were projected into a perceptive space whose principal axes represented their main impairments. We carried out a verbalization task to label these main dimensions. A Multiple Factor Analysis highlighted a four-dimensional perceptive space. Two of these dimensions were modelled and validated. The proposed reference signals allowed covering the defaults of two different coding techniques.

INTRODUCTION

The quality of speech and audio codecs is important in the high competitive world of telecommunications where voice communication remains one of the most used services. Improvement of codecs quality is assessed by subjective and/or objective tests. Due to the subjective nature of voice, subjective assessment remains the most reliable. However, codec ratings can differ significantly from one subject to another one. Moreover, results of assessment tests run at different periods and/or in different laboratories are difficult to compare, when these tests are not based on a reference system. Law and Seymour [1] proposed a reference system, the MNRU (Modulated Noise Unit Reference) for the earlier codecs, the log-PCM waveform coding systems. The MNRU system assumes that speech and audio codecs quality depends only on the Signal-to-Noise Ratio, waveform codecs having just quantization noise as impairment. In the 80's the MNRU has been standardized by ITU-T (Standardization section of the International Telecommunication Union) in [2].

However, recent codecs use new compression techniques and have other types of impairment. Therefore, the MNRU system is no longer adapted to their assessment. The purpose of our study was to elaborate a new reference system intended to replace the MNRU. In this phase, the work was limited to wideband codec applied on clean speech, focusing on intrinsic quality of codecs and transmission impairments were not taken into account. The different steps of this work were first to find the main dimensions of the perceptive space in which new generation codecs can be projected. Then, we ran a verbalization task to label these main dimensions. Identifying these dimensions helped in finding the physical characteristics to which they were correlated, the goal being to create a new reference system linked to these dimensions.

In this paper, after describing the MNRU system in section 1, we present in section 2 the main results of a previous study which highlighted that codecs quality can be projected in a four-dimensional space. In section 3, we describe the reference signals designed to model two dimensions of this perceptive space. Section 4 is devoted to the description of the validation phase. The statistical analysis is developed in section 5 and results are discussed in section 6 before concluding.

1. REFERENCE SIGNALS AND MNRU

Since speech and audio signal quality is intrinsically subjective, its assessment depends on the "rater" involved. Let S be an original speech or audio signal and $\hat{S}$ the signal distorted by a system under assessment. Reference signals correspond to the original signal S transformed by different functions such as they reproduce as close as possible the distorted signal in the perceptive domain. These reference signals are useful to:

- help subjects in rating tasks,
- allow comparison of results obtained by a laboratory across time,
- allow comparison of results across laboratories,
- allow validating objective assessment models.

At the beginning of speech coding, the only impairment of codecs was quantization noise and the MNRU was designed to reproduce it. Let S be the original signal, N a Gaussian white noise and Q the ratio, in dB, of speech power to modulated noise power. The MNRU system generates an output signal Y defined by:

$$Y = \left(S + 10^{-Q/20} S \cdot N\right) * H \ .$$

The symbol $*$ represents the convolution and H is the impulse response of a filter whose bandwidth depends on the system under study. Let $\Gamma = (c_1, c_2, \cdots, c_n)$ be a set of n codecs whose quality must be evaluated. The j MNRU signals are introduced in Γ to get a new set

$\tilde{\Gamma} = \left(c_1, c_2, \cdots, c_n, MNRU_{Q_1}, \cdots, MNRU_{Q_j}\right)$. The $MNRU_{Q_k}$ is the reference signal obtained by the MNRU process depending on the $SNR\ Q_k$. The new set is easier to assess than the initial set because the $MNRU_{Q_k}, k = 1, 2, \cdots, j$, constitute a reference scale that helps the subject in rating the other signals. However, the recent codecs have other types of impairment and the one-dimensional character of codec quality is no longer true.

2. PREVIOUS STUDY

In a previous study [3], a dissimilarity test was run on a selected group of recent codecs in order to get dissimilarity matrices which represent the perceptive pairwise distances between these codecs.

2.1 Dissimilarity test

As the study focused on new generation of codecs, we selected a set of 19 wideband and super wideband codecs [3]. The transmission impairments were not considered. In order to take into account several types of compression techniques, they were chosen among different families of codecs. In order to highlight the impairments of the codecs, we applied two or three times a tandeming technique. From the 58 stimuli (19×3 + the original signal), we retained 20 codecs/tandems which were approximately in the same range of quality (around the middle of the Mean Opinion Score scale). The codecs/tandems finally retained (index 1 to 20) are presented in Table 1 and their characteristics are given in Table 2. Stimuli were obtained by processing the original signal by the 20 codecs/tandems. The original signal was a double sentence uttered by a male speaker and separated by a short silence (6-second total duration). We carried out a listening test where subjects were asked to rate the distance they perceived by pairwise comparison. A dimensional reduction technique allowed obtaining a four-dimensional space [3]. The next step of the study consisted in labelling these dimensions through a verbalization task.

Index	Description	Index	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
5	G722.2_12.65kbps_x2	15	G729.1_24kbps_x2
6	G722.2_12.65kbps_x3	16	G729.1_32kbps_x3
7	G722.2_15.85kbps_x2	17	HEAAC_24kbps_x2
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 – Codecs/tandems under assessment (x2 and x3 mean respectively that tandem speech coding is applied two and three times to the considered codec)

Codecs	Technical characteristics
G722.1C [5]	Modulated Lapped Transform (MLT)
G722.2 [6]	Algebraic Code Excited Linear Prediction (ACELP)
G722 [7]	Waveform codec
G729.1 [8]	Hybrid codec
HEAAC	Modified Discrete Cosine Transform (MDCT)
MP3	

Table 2 – Technical description of codecs under assessment

2.2 Dimensions of the perceptive space

Listeners were asked to describe with their own vocabulary the impairments they perceived on the 20 codecs/tandems selected. The analysis of this verbalization test highlighted that the two preponderant dimensions were labelled by the attributes "muffled" and "background noise" [3]. The remaining two dimensions were more difficult to label so that no reference signals were derived.

3. DERIVED REFERENCE SIGNALS

The first experimentation showed that codec quality could be described in a four-dimensional space. The two first dimensions were characterized and two reference signals were proposed.

3.1 First reference signal

A muffled sound is a sound whose bandwidth is limited towards the high frequencies. Thus, the function designed for the first dimension is a low-pass filter whose cut-off frequency is at least 3400 Hz (the upper limit of the narrowband). The reference signal of the first dimension was obtained by applying the above filter with different cut-off frequencies to the original signal.

3.2 Second reference signal

Listening to the stimuli presenting the most "background noise" (stimuli 9, 10, 11 and 12) [4], listeners noticed that this noise was always present in the silence. Consequently, the reference signal was obtained by adding a Gaussian white noise to the original signal. The SNR was controlled by the gain of an amplifier.

4. VALIDATION PHASE

After designing the reference signals we ran a new dissimilarity test followed by a new verbalization task to validate them.

4.1 Dissimilarity test

The set of stimuli of this test was composed of the previous 20 codecs/tandems, three reference signals for the first dimension and three reference signals for the second dimension. The first dimension reference signals were the original signal filtered by a low-pass filter whose cut-off frequency was respectively 3500 Hz (index 21), 4500 Hz (index 22) and 5500 Hz (index 23). The second dimension reference signals were the original signal corrupted by an additive white Gaussian noise such as the SNR was respectively 35 dB (index 24), 45 dB (index 25) and 55 dB (index 26).

We recruited 30 subjects to participle to the dissimilarity test. During the test, the listeners were asked to give a score from 0 to 100 that reflected the perceptive distance between two stimuli. A null score indicated that stimuli were strictly similar whereas a score equal to 100 meant they were really different. Two of the listeners were felt unreliable (they indicated high values for null pairs) and were discarded from this study. After the dissimilarity test, the subjects were asked to qualify with "attributes" the impairments they perceived on the 26 stimuli.

4.2 Verbalization task

After completion of the dissimilarity test, the 28 subjects retained were asked to describe the impairments they perceived on the stimuli first by using words from a panel of attributes provided (pre-defined list) and also via their own vocabulary.

Label	Attribute	Label	Attribute
RV	Robot voice	BR	Breath
MF	Muffled	DS	Distorted speech
SC	Scratching	CR	Crackling
HS	Hissing	EC	Echo
BN	Background noise	EV	Energy variation
MN	Modulated noise		

Table 3 – Retained attributes for the verbalization task

5. STATISTICAL ANALYSIS

Multiple dissimilarity matrices are usually analyzed with a three-way MultiDimensional Scaling (MDS). In our study, we used a three-way factor analysis technique, the Multiple Factor Analysis (MFA) presented in [9]. MFA aims at analyzing a set of objects described by different groups of variables. In our case, we applied MFA to the dissimilarity matrices.

5.1 Preprocessing of dissimilarity matrices

The 28 dissimilarity matrices of the 28 listeners who compared 26 codecs (26×26 matrices) were first transformed into Euclidean distance matrices $E_k, k \in \{1,\cdots,28\}$. Then, a Principal Component Analysis (PCA) was applied to each Euclidean distance matrix in order to obtain a representation of the 26 codecs in a p_k dimensional and orthogonal space, with p_k the dimension of the space for listener k. Let us note X_k the matrix representing the 26 codecs (lines) in the p_k factorial space (columns) for listener k. The 28 X_k were concatenated and submitted to MFA.

5.2 MFA algorithm

Let X be the concatenation of $X_k, k \in \{1,\cdots,28\}$, X is a $26\times p$ matrix where $\sum_{k=1}^{28} p_k = p$.

MFA can be viewed as a double PCA which processes in two steps:

1. A PCA is performed on each X_k. The first eigenvalue λ_k is extracted and used to compute the weighted matrix $Z_k = X_k / \sqrt{\lambda_k}$

2. A global PCA is performed on $Z = [Z_1,\cdots,Z_k,\cdots,Z_{28}]$.

Eigenvalues of this global PCA allows determining the optimal number of dimensions.

5.3 Integration of verbalization data in MFA

The MFA allows analyzing simultaneously quantitative and qualitative variables. For each listener, we derived from the verbalization task a matrix indicating which attributes he/she cited to qualify a codec. The rows of the verbalization matrices represented the codecs and the "attributes" were represented by the columns. For each stimulus, we put "1" if the subject quoted the attribute and "0" otherwise. The 28 individual matrices were summed up into one single matrix containing for each codec (row) the number of times each attribute (column) was cited by the 28 listeners. Attributes or words with too few citations were merged with attributes having similar meanings. Finally, eleven attributes were kept (Table 3). This matrix was then transformed into the percentage 26×11 matrix V, indicating for each codec the percentage of citation of each attribute among all attributes cited for this codec. It is the quantitative representation of the verbalization matrix. Then the median value of the citation percentage of each attribute was computed and attributes were binary coded as follows:

Let m_j be the median value of citation percentage for attribute j. The binary verbalization matrix W was defined as:

$$W(i,j) = \begin{cases} \text{attribute "}j\text{"} & \text{if } V(i,j) > m_j \\ \text{" " (nothing)} & \text{otherwise} \end{cases}.$$

The matrix finally analyzed by the MFA was the concatenation of X, V and W. All elements of X were considered as active in the analysis, whereas the elements of V and W were considered as supplementary, i.e. they did not participate in determining the perceptive space, but were used to qualify a posteriori the factors obtained.

Index	RV	MF	SC	HS	BN	MN	BR	DS	CR	EC	EV
1	0.37	0.03	0.03	0	0.11	0.03	0.16	0.11	0.03	0.16	0
2	0.26	0.03	0	0.03	0.13	0.08	0.08	0.15	0.03	0.18	0.05
3	0.29	0.05	0	0.02	0.1	0.02	0.05	0.07	0.1	0.27	0.02
4	0.38	0.06	0	0.03	0.03	0.03	0.06	0.06	0.09	0.21	0.06
5	0.16	0.38	0	0	0.11	0.02	0.04	0.09	0.04	0.07	0.09
6	0.13	0.23	0	0	0.08	0	0.08	0.13	0.06	0.15	0.15
7	0.14	0.51	0	0	0.09	0	0.09	0.06	0.03	0.03	0.06
8	0.21	0.3	0	0.02	0.02	0.02	0.05	0.05	0.09	0.12	0.12
9	0.1	0.08	0.06	0.02	0.21	0	0.14	0.04	0.33	0.04	0
10	0.04	0.07	0.02	0.04	0.18	0.02	0.12	0.02	0.37	0.12	0.02
11	0.02	0.02	0	0.07	0.28	0	0.13	0.04	0.35	0.07	0.02
12	0	0.05	0	0.02	0.31	0.02	0.21	0	0.31	0.05	0.02
13	0.21	0.25	0.03	0	0.11	0.05	0.03	0.1	0.1	0.05	0.08
14	0.12	0.17	0	0	0.14	0.03	0.05	0.19	0.09	0.1	0.12
15	0.13	0.17	0	0	0.15	0.09	0.02	0.21	0.09	0.09	0.06
16	0.17	0.06	0	0.02	0.17	0.02	0.06	0.14	0.14	0.1	0.14
17	0.3	0.02	0	0	0.04	0	0.09	0.13	0.17	0.23	0.02
18	0.18	0.04	0	0.04	0.11	0	0.13	0.09	0.22	0.18	0.02
19	0.24	0.07	0	0	0	0.02	0.07	0.07	0.16	0.36	0.02
20	0.44	0.02	0	0.02	0.02	0.02	0.04	0.04	0.14	0.22	0.04
21	0.07	0.58	0.03	0	0.07	0	0.07	0	0.07	0.1	0.03
22	0.09	0.53	0	0	0.09	0	0.03	0.06	0.03	0.09	0.06
23	0.04	0.48	0	0	0.12	0.04	0.04	0.04	0.04	0.12	0.08
24	0	0.02	0	0.02	0.42	0.02	0.17	0	0.29	0.02	0.02
25	0.02	0.04	0	0.09	0.28	0	0.22	0.02	0.26	0.04	0.02
26	0.05	0	0.03	0.03	0.43	0.03	0.24	0	0.16	0	0.03
Mean	**0.16**	**0.16**	**0.01**	**0.02**	**0.15**	**0.02**	**0.09**	**0.07**	**0.14**	**0.12**	**0.05**

Table 4 – Quantitative verbalization matrix V (the last value of each column represents the mean of all other values of the column)

6. RESULTS

6.1 Verbalization task

A first analysis of Table 4 suggests the following characteristics for the families of codecs: the MLT family (stimuli 1, 2, 3 and 4) was qualified by RV and EC. CELP codecs (stimuli 5, 6, 7 and 8) were qualified by MF, the waveform codecs (stimuli 9, 10, 11 and 12) were characterized by BN, BR and CR, and the hybrid family (stimuli 13, 14, 15 and 16) was essentially characterized by MF and DS. The transform codecs based on MDCT (stimuli 17, 18, 19 and 20) were characterized by RV, CR and EC. Moreover, the reference signals 21, 22 and 23 constructed as MF were actually characterized by MF and the reference signals 24, 25 and 26 constructed to represent BN were characterized with BN, BR and CR. The created reference signals were able to reproduce the following defaults: BN, BR, CR, and also MF. Comparatively, defaults characterizing the transform codecs such as RV and EC are still to be elaborated as well as the default DS present in hybrid codecs (Table 4).

6.2 Dimensions from the dissimilarity test

The number of dimensions of the perceptive space was determined from the eigenvalues plot. Figure 1 displays an elbow between the 4th and 5th dimensions. After dimension 5, the variation in eigenvalues was very low. Furthermore, the 3rd and 4th eigenvalues were approximately the same (respectively 10.55 and 10.35). We set to four the number of dimensions of the perceptive quality space. This result reinforces the four-dimensional space found in the previous study [3]. This indicates that the reference signals inserted in the test did not modify the perceptive space.

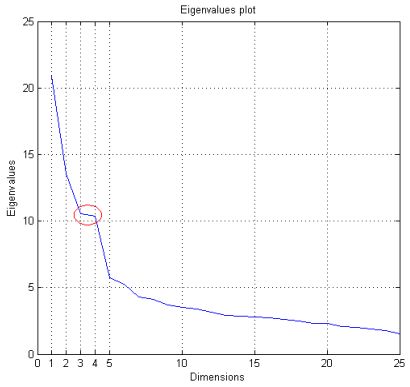

Figure 1 – Eigenvalues plot

6.3 Plausibility of dimensions

6.3.1 First dimension

The first dimension represented 17.6 % of the total explained variance. It separated the muffled codecs from the others. As shown in Figure 2, the reference signals designed to model the MF characteristic were grouped with the most muffled codecs [4] in plane (Dim 1, Dim 2) of the perceptive space. The Table 4 indicates clearly that the stimuli 5, 6, 7, 8, 13, 14, 15, and 16 had the highest occurrence percentages of attribute MF. Moreover, Pearson's correlations between the dimensions and the attributes represented in Table 5 display high negative values for the attribute MF (-0.8). The most important contributions of the codecs in the first dimension were those of stimuli 13, 21, 22, 14, 8 and 6 (respectively equal to 13.8%, 10.6%, 7.7%, and around 6% for the last three). These contributions were higher than 4% (the contribution value of the stimuli if their contributions were all equal). These results confirm the MF attribute for dimension 1.

6.3.2 Second dimension

The second dimension contributed for 11.4% of the total explained variance. Figure 2 shows that the group of stimuli 9, 10, 12, and the reference signals of 24, 25 and 26 were the only stimuli having positive coordinates on this dimension. The rank order of these stimuli along this second dimension followed their noise level. As seen in Table 4, these codecs had a high frequency of nomination for attribute BN and low frequencies for attributes DS, RV and EC. The stimuli 1 and 2 were located at the extreme opposite side and they contributed to 6% to this dimension. These results were consistent with those presented in Table 4: stimuli 1 and 2 had high frequencies of nomination for DS, RV and EC. A high negative correlation was observed between the second dimension and the attributes RV, EC and DS and a positive correlation was observed with BN. Since the second dimension separated the stimuli characterized by the BN attribute from the others, this dimension represented the opposition between attributes BN and RV/EC. The background noise reference signals 24, 25 and 26 might have a too high signal-to-noise ratio so that the resulting perceptive space was impacted.

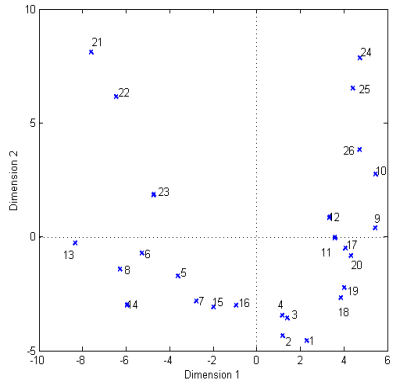

Figure 2 – Attributes and stimuli plot in (Dim 1, Dim 2) plane

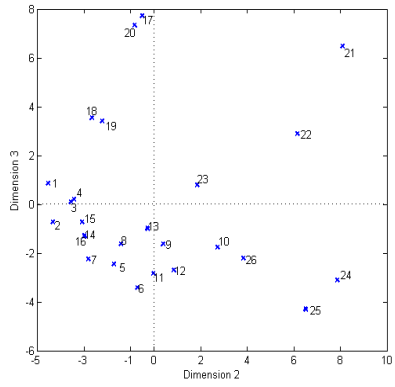

Figure 3 – Attributes and stimuli plot in (Dim 2, Dim 3) plane

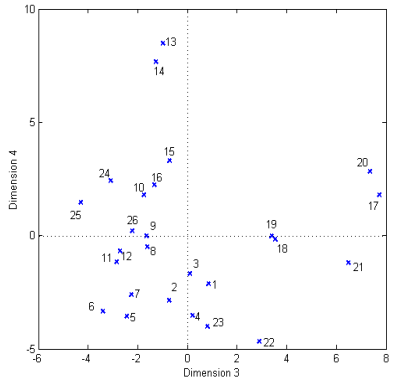

Figure 4 – Attributes and stimuli plot in (Dim 3, Dim 4) plane

Attributes	Labels	Dim1	Dim2	Dim3	Dim4
Robot voice	RV	0.06	-0.64	0.52	-0.02
Muffled	MF	-0.8	0.27	0.11	-0.31
Scratching	SC	0.05	0.22	0.01	0.17
Hissing	HS	0.55	0.23	-0.36	0.01
Background noise	BN	0.41	0.48	-0.58	0.16
Modulated noise	MN	-0.15	-0.37	-0.15	0.25
Breath	BR	0.64	0.37	-0.37	-0.02
Distorted speech	DS	-0.3	-0.66	0.05	0.26
Crackling	CR	0.69	0.34	-0.24	0.26
Echo	EC	0.18	-0.46	0.59	-0.15
Energy variation	EV	-0.72	-0.24	-0.25	0.08

Table 5 – Correlations between attributes and dimensions

6.3.3 Third dimension

The third dimension represented 8.9% of the total variance. The attributes EC and RV had high positive correlations (respectively 0.59 and 0.52) with this dimension whereas the attribute BN had a high negative correlation (-0.58) (see Table 5). Dimension 3 was strongly characterized by the extreme position of the stimuli 17 and 20 and also by the position of the stimuli 18 and 19. Table 4 shows that these stimuli were mostly characterized by a high frequency of nomination for the attributes EC and RV and a low frequency for the attribute BN. The relative position of stimuli 17 and 18 on this dimension was explained by the lower bitrate of stimulus 17. Similarly, the relative position of stimuli 19 and 20 was explained by the higher number of tandeming for stimulus 20 compared to stimulus 19. Stimulus 20 was a twice tandeming of a MP3 codec at a rate of 32 kbps whereas stimulus 19 corresponded to this codec itself. From these analyses, we labelled the third dimension with the EC/RV attribute.

6.3.4 Fourth dimension

The fourth dimension contributed to 8.7% of the total explained variance. The correlation matrix (Table 5) shows that the best attributes that described the fourth dimension were CR, DS and MN. Now, the stimuli 13, 14 which were the most contributory to the fourth dimension (respectively 27% and 22%), and to a lesser extent stimuli 15 and 16, were also characterized by the attribute DS as shown in Table 4. Therefore, we labelled the fourth dimension with the DS attribute.

7. CONCLUSION

In this study we presented the analysis of the perceptive quality of codecs using MFA. After running out dissimilarity tests, we found that the perceptive quality of new generation of codecs could be described in a four-dimensional space. Thanks to the verbalization task, we tried to associate labels to each stimulus, characterizing the codecs with attributes. Each family of coding techniques can be characterized by a set of attributes.

The MLT family (stimuli 1, 2, 3 and 4) was characterized by "robot voice" and "echo" attributes, CELP codecs (stimuli 5, 6, 7 and 8) by "muffled" attribute, the waveform codecs (stimuli 9, 10, 11 and 12) by "background noise", "breath" and "crackling" attributes. The hybrid family (stimuli 13, 14, 15 and16) was labelled by "muffled" and "distorted speech" attributes. The transform codecs based on MDCT (stimuli 17, 18, 19 and 20) were characterized by "robot voice", "crackling" and "echo" attributes.

From now on, reference signals representing the "muffled" attribute (stimuli 21, 22 and 23) and reference signals representing "background noise", "breath" and "crackling" (stimuli 24, 25 and 26) allow covering the defaults of two families of coding techniques, CELP codecs and waveform codecs. Reference signals revealing the defaults "distorted speech" and "robot voice/echo" remain now to be generated in order to complete the set of reference signals.

REFERENCES

[1] H. B. Law, R. A. Seymour, "A reference distortion system using modulated noise," IEEE, pp. 484-485, November 1962.

[2] ITU-T Recommendation P.810, "Modulated Noise Reference Unit (MNRU)". International Telecommunications Union, 02/96.

[3] T. Etamé, G. Faucon, R. Le Bouquin Jeannès, L. Gros and C. Quinquis, "Characterization of the multidimensional perceptive space for current speech and audio," *AES 124th convention,* The Netherlands, May 17-20, 2008.

[4] T. Etamé, R. Le Bouquin Jeannès, C. Quinquis, L. Gros, G. Faucon, "Towards a new reference impairment system in the subjective evaluation of speech codecs," *IEEE*, Issue 99, October 2010.

[5] ITU-T Recommendation G.722.1, "Low-complexity coding at 24 and 32 Kbit/s for hands-free operation in systems with low frame loss," 2005.

[6] ITU-T Recommendation G.722.2, "Wideband coding of speech at around 16 Kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)," 2003.

[7] ITU-T Recommendation G.722, "7 kHz audio-coding within 64 Kbit/s," 1988.

[8] ITU-T Recommendation G.729.1, "G.729-based embedded variable bit-rate coder: An 8-32 Kbit/s scalable wideband coder bitstream interoperable with G.729," 2006.

[9] B. Escofier, J. Pagès, "Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data," *Computational Statistics & Data Analysis*, vol.52, pp. 3255-3268, February 2008.