Design of reference signals for the quality evaluation of coded speech

T. ETAME[(1)], L. GROS[(1)], C. QUINQUIS[(1)],
R. LE BOUQUIN JEANNES[(2), (3)] and G. FAUCON[(2), (3)]
[(1)] France Telecom R&D TECH/SSTP - 2 Av. Pierre Marzin, 22307 Lannion Cedex, France
tel. +33 (0)2 96 05 98 88 - fax. +33 (0)2 96 05 35 30
(thierry.etameetame, laetitia.gros, catherine.quinquis)@orange-ftgroup.com
[(2)] INSERM, U 642, Rennes, F-35000 France
[(3)] Université de Rennes 1, LTSI, Rennes, F-35000, France
tel. +33 (0)2 23 23 62 20 - fax. +33 (0)2 23 23 69 17
(regine.le-bouquin-jeannes, gerard.faucon)@univ-rennes1.fr

Abstract

This work aims to provide reference signals representative of the new techniques of speech coding, for subjective assessment tests of speech quality. A non-metric weighted multidimensional scaling (MDS) analysis is used for the perceptual analysis of the defects generated by eighteen wideband codecs. The results indicate a four-dimensional space to elucidate the perception of current defects. These dimensions are characterized from a coding point of view.

Keywords

codec, coding defects, MDS, MNRU, perceptive space, reference signals, speech quality, subjective test.

1 Purpose and motivation

The speech quality of digital codecs is often evaluated through subjective tests with naive listeners who are asked to listen to short speech samples processed by the codecs under study and to give their opinion on the speech quality on an appropriate scale (for example, Excellent = 5, Good = 4, Fair = 3, Poor = 2 and Bad = 1) [1]. The Modulated Noise Reference Unit (MNRU) is extensively used in subjective evaluations of digital processes, both in conventional telephone bandwidth and in wideband, to introduce controlled degradations into speech signals [2]. These degradations are characteristic of the quantization noise of the codecs like PCM (Pulse Code Modulation) or ADPCM (Adaptive Differential Pulse Code Modulation). The resulting signals are used as reference conditions so that experiments made in different laboratories or at different times can be sensibly compared [1].

The improvements in audio coding technologies modified the types of distortion and so the MNRU is not representative any more of the current distortions.

At the moment, no study tackles the problem of the subjective test calibration for the quality evaluation of new speech and sound codecs.

Therefore, this work aims to provide reference signals representative of the current coding defects (see Fig. 1). The first step consists of identifying the perceptive characteristics of the defects resulting from the new audio coding techniques by listening experiments. The comparison of these perceptive characteristics with various coding techniques should help to artificially generate these defects and to introduce them into speech signals to get new reference signals.

Codecs under evaluation are presented in section 2. Section 3 presents the MultiDimensional Scaling (MDS) technique used in this study to determine a multidimensional perceptive space which underlies the perception of current impairments. Finally, the listening experiment and its results analyzed by MDS are given in section 4 before concluding.

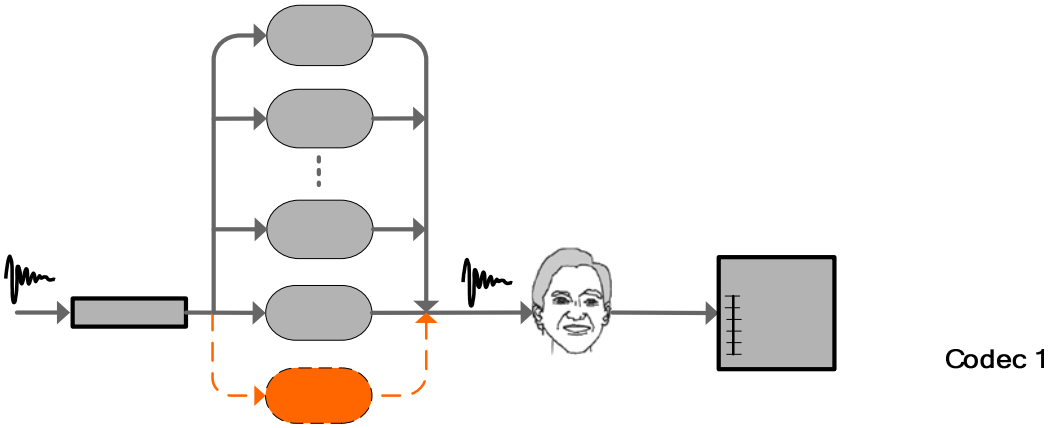

Fig. 1 Subjective evaluation of codec quality

2 Technical specification of codecs

For this study, we considered a maximum of codecs presenting varied coding techniques in order to have a maximum of possible defects. As it will be detailed in paragraph 4.1, we selected wideband and full band codecs which are similar in terms of global speech quality. Each codec output signal is filtered in order to limit its bandwidth to wideband, so that the subjective judgments do not relate too much to the bandwidth of the codecs. Table 1 briefly presents the technical specifications of wideband (16 kHz sampling rate input signal), super-wideband (32 kHz sampling rate input signal) and full band (48 kHz sampling rate input signal) codecs of which some will be considered in this study.

3 The multidimensional scaling technique (MDS)

The quality of speech is commonly considered as a multidimensional phenomenon [3]. In order to take into account the multidimensional nature of speech quality, one can use the method of semantic descriptors in which listeners are asked to judge the sound quality on different scales using adjective descriptors [4]. But people usually do not have a sufficient vocabulary to describe most of the auditory sensations. In addition, the choice of descriptors is biased by the experimenter.

The multidimensional scaling technique was selected as the analysis method because there is no presumption on dimensions contrary to the methods using semantic descriptors. However, MDS is based on the hypothesis of continuous dimensions. The multidimensional scaling is a technique that consists of studying the perceptive structures which underlie the judgments of similarities (or preference) given for pairs of stimuli, by translating them into a matrix of distance. This matrix is used to project the whole of the stimuli or objects, in a multidimensional space according to a mathematical model. Therefore, the similarity judgments allow the experimenter to find a mathematical solution for perception of auditory characteristics based on experimental quantitative data without having to rely on subjective and often conceptually incomplete lists of verbal descriptors [5].

The typical multidimensional scaling technique determines the order of objects in an n-dimensional object space that minimizes the disparity between the Euclidean distances given in the dissimilarity matrix and Euclidean distances in the object space, in the least squares sense. This is the basic concept of the metric multidimensional scaling technique. There is only a single dissimilarity matrix and the values in this one are derived by a linear transformation of the distances between objects.

It is difficult for a listener to use numbers consistently throughout the course of a listening test [6]. Therefore, it is assumed that the order induced by the judgments of dissimilarities in human listening experiments is more reliable. The nonmetric multidimensional scaling ([7] [8]) takes into account this point by computing a dissimilarity matrix with values which are a rank ordering of the distance between objects.

In addition, it is assumed that subjects differently weight the perceptive dimensions when they elaborate an overall judgment. This inter-individual variability is taken into account in the weighted multidimensional scaling also referred to as INDSCAL (INDividual differences SCALing) ([9] [10]). Results obtained from weighted multidimensional scaling include not only a stimulus space but also a subject space that shows the weight given to each dimension by each subject.

In our study, we chose a nonmetric INDSCAL MDS which takes into account the characteristics of perceptive evaluations of audio quality.

Table 1 - Technical Specifications of some codecs

Codec	Bit rates (kbit/s)	Technical Specification
ITU-T G722 wideband speech codec	64, 56, or 48	Sub-Band Adaptive Differential Pulse Code Modulation (SB-ADPCM) with 4,5,6 bits/sample (lower band) and 2 bits/sample (higher band)
ITU-T G722.1 wideband speech and audio codec Low Complexity Transform Codec (LCTC)	24 or 32	• Uses a transform-coding scheme called Modulated Lapped Transform (MLT) with Scalar Quantized Vector Huffman Coding (SQVH) and bit allocation by categorisation • 20 ms frame size + 20 ms look-ahead
ITU-T G722.1 C Super-wideband speech and audio codec	24, 32 or 48	Low-complexity extension mode to G.722.1 which permits 14-kHz audio bandwidth using a 32-kHz audio sample rate
ITU-T G722.2 or 3GPP AMR-WB Adaptive Multi-Rate WideBand codec	6.6 to 23.85	The codec is based on Algebraic Code-Excited Linear Predictive (ACELP) coding model (50-6400 Hz) and bandwidth extension (BWE, 6.4 to 7 kHz)
ITU-T G729.1 or G729EV (Embedded Variable bit-rate) scalable wideband speech and audio codec	14-32	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 (TD-BWE) • enhancement of the full band (50-7000Hz) by a predictive transform coding technique referred to as Time-Domain Aliasing Cancellation (TDAC)
High Efficiency-AAC (aacPlus or Enhanced aacPlus) Full band audio codec, MPEG-4 standard for low bit rate coding	10-32 (mono) 16-48 (stereo), stimulus is a stereo-to-mono down-mixing signal)	• HE-AAC uses Spectral Band Replication (SBR) for regeneration of high-frequency signal components and Parametric Stereo (PS) to achieve further reduction in bitrate • SBR consists of a 64-QMF (Quadrature Mirror Filter) analysis filterbank
MPEG-1 layer III (ISO/IEC 11172-3) Full band audio codec Constant Bitrate encoding with Standard MP3, try by Cool Edit Pro (technology licensed from Coding Technologies, Fraunhofer IIS and Thomson multimedia)		The MP3 format uses a hybrid filterbank: • polyphase filterbank (32 subbands) 18-spectral point (long block) or 6-spectral point (short block) MDCT with scalar Huffman coding and rate loop. It is based on perceptual Audio Coding (psychoacoustic model)

4 Experiments

4.1 Selection of codecs

In a MDS analysis, it is desirable to include as many stimuli as practically possible in an experiment. This is because the number of dimensions which can be extracted increases with the number of stimuli. Ideally, Kruskal recommends nine stimuli for two dimensions, thirteen for three and seventeen for four. These recommendations are, however, for a single matrix of data, and when more than about ten matrices are to be analyzed, the recommendations can be weakened [5]. Therefore, around twenty codecs with more than ten subjects seem appropriate for a perceptive space of four or five dimensions, which is a reasonable number of dimensions from a perceptive point of view.

A preliminary ACR (Absolute Category Rating) test (P.800 [1]) was run to select around twenty codecs with similar speech quality (MOS (mean opinion score) around 3) so that the judgments of dissimilarity do not relate to global quality, but only on the type of defect.

In order to have codecs with different coding techniques but with similar quality, tandem speech coding was applied to the nineteen following codecs: G722_64kbps, G722_56kbps, G722_48kbps, G722.2_8.85kbps, G722.2_12.65kbps, G722.2_15.85kbps, G722.2_23.85kbps, G722.1_24kbps, G722.1_32kbps,

G729.1_14kbps, G729.1_20kbps, G729.1_24kbps, HEaac_16 kbps, HEaac – 24k, HE aac_32 kbps, G722.1 C_24 kbps, MP3_32 kbps, MP3_64kbps. Tandem speech coding, where two codecs operate on a signal in cascade, as in the case of mobile communication, can amplify the distortion generated if the two codecs are identical. Our study will consider the cases where one (_x1), two (_x2), or three (_x3) codec(s) are present.

The samples used in the ACR test consisted of pairs of sentences spoken by two males and two female talkers, two samples per talker, processed by the fifty-eight (19 codecs x 3 + original signal) resulting tandem codecs, including original source. Thirty-two subjects participated in this ACR listening test. The ACR MOS values for the 58 conditions are reproduced in Appendix A. We consider the codecs having scores between 2 and 3.5 included. To increase perceived degradations, we prefer tandems (_x2, _x3) to the others. We choose codecs presenting varied coding techniques in order to have a maximum of possible defects. We selected 18 tandem codecs for the test of dissimilarity. These ones are given in Table 2.

Table 2 – Codecs used for the test of dissimilarity

Codec	Description
+ O1	G722.1C_24kbps_x2
+ O2	G722.1C_24kbps_x3
+ O3	G722.1_24kbps_x2
+ O4	G722.1_24kbps_x3
x O5	G722.2_12.65kbps_x2
x O6	G722.2_12.65kbps_x3
x O7	G722.2_15.85kbps_x2
x O8	G722.2_8.85kbps_x2
° O9	G722_48kbps_x2
° O10	G722_48kbps_x3
° O11	G722_56kbps_x2
° O12	G722_56kbps_x3
* O13	G729.1_14kbps_3
* O14	G729.1_20kbps_x3
* O15	G729.1_24kbps_x2
* O16	G729.1_32kbps_x3
> O17	MP3_32kbps_x1
> O18	MP3_32kbps_x2

4.2 Test procedure

A 6-sec speech sample uttered by one male (in french, "La vanille est la reine des arômes. Fragile, il ne résiste pas à l'air glacé."), originally a full band signal sampled at 48 kHz, was downsampled as input signal for wideband and super-wideband codecs. It was then filtered with a 50 Hz-7 kHz band-pass filter for wideband codecs and a 50 Hz-14 kHz band-pass filter for super-wideband versions. The resulting stimuli were processed by the eighteen selected codecs with output limited to wideband. Finally, these coded versions were upsampled to 48 kHz for compatibility with stimulus presentation equipment.

The tests were performed on the headphone STAX Signature SR-404 (open model) and its amplifier SRM-006t. The stimuli were stored on a Windows 2k workstation. The digital sound was played through the PC board Digigram VX 222 and converted using a 24 bits DAC (3Dlab DAC 2000). Stimuli were presented diotically at a comfortable listening level to the subject sat inside an industrial acoustics company soundproof booth.

All in all, 171 pairs (153 + 18 null pairs) were presented in random order to subjects, with a different randomization for each subject. For each pair, the subject was asked to evaluate perceptual distance between coded speech samples. The similarity between the samples is rated on a continuous line scale varying between 0 (similar) and 100 (different). For each pair, the two coded versions of the speech sample were presented so as the subject could freely switch from one version to another to better detect differences. A single subject participated in each session and was asked to listen to the test sample pair at least once, after which the overall similarity was to be rated on the scale. Session results were typically collected in two sessions around a hundred trials each in two different days. Each session took typically 90 minutes and included a ten-pair preliminary session. Fourteen subjects participated in the experiment.

4.3 Results

The analyses were carried out by the SPSS (Statistical Package for Social Sciences) INDSCAL algorithm that is based on the ALSCAL (Alternating Least Square SCALing) algorithm.

The stimulus space derived for the eighteen speech samples used with the male talker is shown in Fig. 2. The coordinates are given in Appendix B.

Several criteria, including variance accounted for, residual stress, and average weight given by the listeners to each dimension, indicate that four is an appropriate number of dimensions (stress = 0.18, RSQ (squared correlation) = 69%).

The results reveal a regrouping of the codecs according to technical specifications.

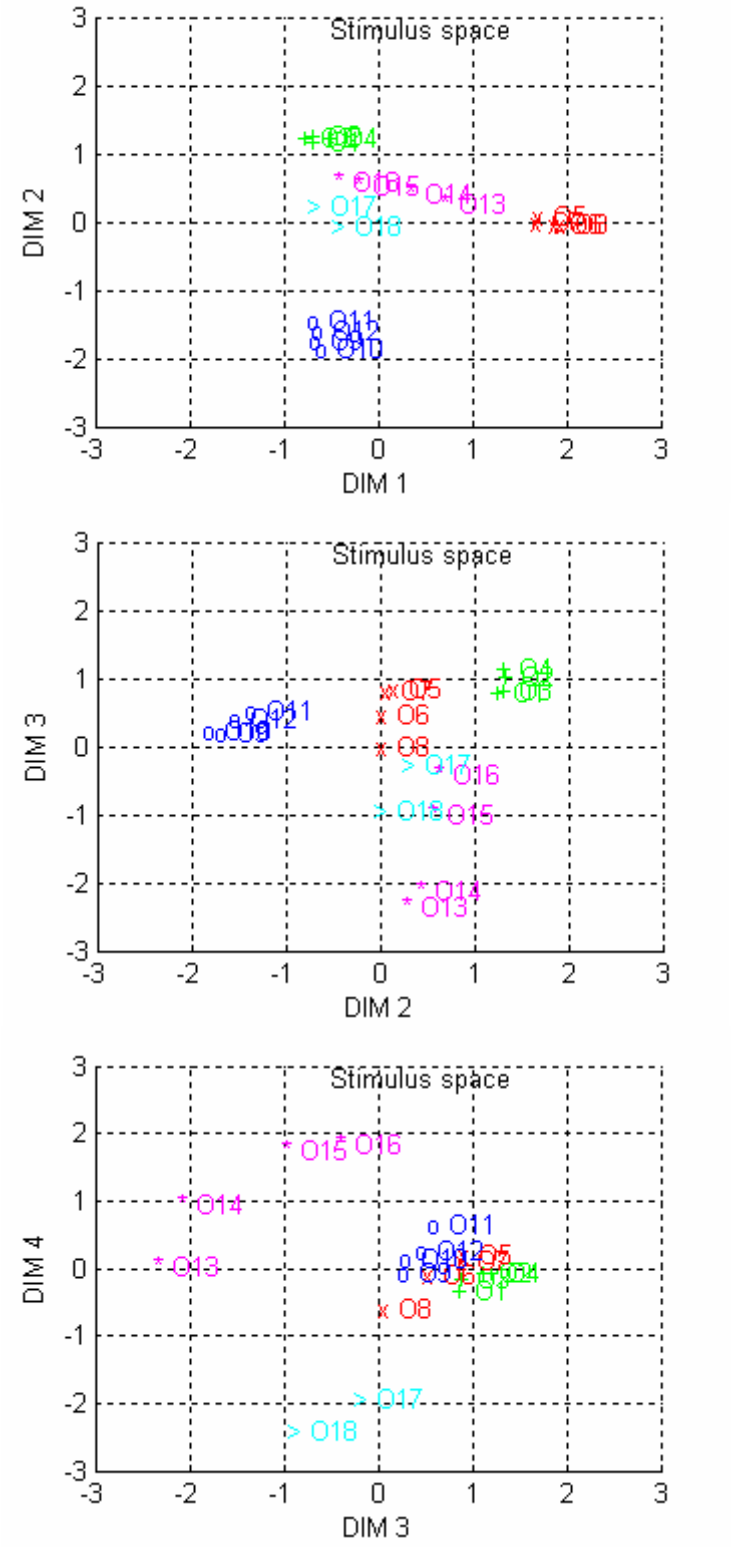

Fig. 2 Plot of object space

Dimension 1 accounts for 23.22 % of the total explained variance. When listening to different objects along this dimension (O1-G722.1C_24kbps_x2, O15-G729.1_24kbps_x2, O7-G722.2_15.85kbps_x2), it seems that this dimension is characterized by a "clear/muffled" attribute. The negative end of the dimension is covered by codecs of MLT transform using categories for bit allocation [O1-G722.1C_24kbps_x2, O2-G722.1C_24kbps_x3, O3-G722.1C_24kbps_x2, O4-G722.1C_24kbps_x3] that preserve the naturalness of the speech signal. On the other hand, the positive end of the dimension is covered by ACELP codecs [O5-G722.2_12.65kbps_x2, O6-G722.2_12.65kbps_x3, O7-G722.2_15.85kbps_x2, O8-G722.2_8.85kbps_x2].

Dimension 2 contributes to 23.18 % of the total explained variance. As shown in Fig. 2, there was a clear grouping of samples at the negative end of the dimension 2. The objects in this group are ADPCM codecs [O9-G722_48kbps_x2, O10-G722_48kbps_x3, O11-G722_56kbps_x2, O12-G722_56kbps_x3], which clearly present a background noise contrary to the other codecs. Therefore, dimension 2 would be defined by background noise.

Dimension 3 contributes to 12.58 % of the total explained variance. The negative end of the dimension is represented by hybrid codecs [O13-G729.1_14kbps_x3, O14-G729.1_20kbps_x3, O15-G729.1_24kbps_x2, O16-G729.1_32kbps_x3] and codecs of transform using psycho-acoustics models [O17-MP3_32kbps_x1, O18-MP3_32kbps_x2] that present pre-echo in the signal. So, the attribute that can characterize dimension 3 is pre-echo.

Finally, dimension 4 opposes hybrid codecs [O13-G729.1_14kbps_x3, O14-G729.1_20kbps_x3, O15-G729.1_24kbps_x2, O16-G729.1_32kbps_x3], to the codecs of transform using psycho-acoustics models [O17-MP3_32kbps_x1, O18-MP3_32kbps_x2] which let perceive a quantization noise, which is more important than in the other codecs. Therefore, dimension 4 would be defined by noisiness.

5 Conclusion

The aim of this study was to link the current coding technologies with their potential perceptive defects, in order to be able to artificially and quantitatively generate these defects, and so to create a reference quality system for subjective tests.

We presented a brief review of multidimensional scaling techniques, including INDSCAL, weighted multidimensional scaling technique that can be applied to pair dissimilarity judgments to determine a multidimensional perceptive space in which auditory stimuli can be represented.

We next presented results from a listening experiment in which subjects gave dissimilarity judgments on pairs of speech samples coded with eighteen wideband tandem codecs. The resulting dissimilarity matrices were processed by INDSCAL to generate stimulus and subject spaces. Coded speech samples for one male talker are therefore represented in a 4-dimensional perceptive space characterized by different attributes such as background noise, naturalness, pre-echo, noisiness, and interpreted with technical characteristics of codecs.

This result is consistent with our objective to link these perceptive characteristics to various coding techniques. That should help to artificially generate these defects and to introduce them into speech signals.

Other listening experiments carried out with female talker are in progress to allow us to justify that a four-dimensional space is appropriate to characterize the perception of current defects.

6 References

[1] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality". International Telecommunications Union, 08/96.

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

[3] T. Letowski, "Timbre, tone color, and sound quality: concepts and definitions," Archives of Acoustics, 17(1):17–30, 1992.

[4] C. Osgood, "The nature and measurement of meaning," Psychological Bulletin, 49 (197-237), 1952.

[5] 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, March 22-25. 2003.

[6] J.L. Hall, "Application of multidimensional scaling to subjective evaluation of coded speech," J. Acoust. Soc. Am. 110 (4), Oct. 2001.

[7] J.B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," Psychometrika, Vol. 29, pp. 1-27, 1964.

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

[9] J.D. Carroll, "Individual differences and multidimensional scaling," R.N. Shepard, A.K. Romney & S.B. Nerlove (Eds.) Multidimensional Scaling: Theory and Applications in the Behavioral Sciences (Vol.1,

pp.105-155). New York and London: Seminar Press, 1972.

[10] J.D. Carroll and J.J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition," Psychometrika, 35,283-319, pp. 283-319, 1970.

Appendix A: ACR MOS values (average, standard deviation and confidence interval) for the 58 conditions

Conditions	Aver.	St Dev.	Conf. Int.
c01 Original source	4.34	0.66	0.11
c56 MP3_64 kbps_x1	4.20	0.77	0.13
c38 G729.1_32 kbps_x1	4.09	0.77	0.13
c02 G722_64 kbps_x1	4.07	0.80	0.14
c57 MP3_64 kbps_x2	4.05	0.86	0.15
c20 G722.2_23.85 kbps_x1	3.99	0.83	0.14
c03 G722_64 kbps_x2	3.87	0.75	0.13
c50 HE aac_32 kbps_x1	3.86	1.01	0.17
c26 G722.1_32 kbps_x1	3.82	0.82	0.14
c27 G722.1_32 kbps_x2	3.81	0.86	0.15
c21 G722.2_23.85 kbps_x2	3.79	0.87	0.15
c35 G729.1_24 kbps_x1	3.79	0.90	0.16
c05 G722_56 kbps_x1	3.78	0.78	0.14
c58 MP3_64 kbps_x3	3.75	0.96	0.17
c17 G722.2_15.85 kbps_x1	3.73	0.83	0.14
c28 G722.1_32 kbps_x3	3.72	0.82	0.14
c39 G729.1_32 kbps_x2	3.72	0.79	0.14
c04 G722_64 kbps_x3	3.70	0.81	0.14
c22 G722.2_23.85 kbps_x3	3.63	0.87	0.15
c32 G729.1_20 kbps_x1	3.60	0.88	0.15
c14 G722.2_12.65 kbps_x1	3.59	0.84	0.15
c06 G722_56 kbps_x2	3.58	0.87	0.15
c24 G722.1_24 kbps_x2	3.58	0.97	0.17
c23 G722.1_24 kbps_x1	3.55	0.90	0.16
c18 G722.2_15.85 kbps_x2	3.52	0.91	0.16
c40 G729.1_32 kbps_x3	3.52	0.86	0.15
c53 MP3_32 kbps_x1	3.51	1.10	0.19
c29 G729.1_14 kbps_x1	3.48	0.96	0.17
c36 G729.1_24 kbps_x2	3.45	0.89	0.16
c41 G722.1 C_24 kbps_x1	3.38	1.02	0.18
c42 G722.1 C_24 kbps_x2	3.34	0.99	0.17
c25 G722.1_24 kbps_x3	3.24	1.04	0.18
c43 G722.1 C_24 kbps_x3	3.24	0.99	0.17
c08 G722_48 kbps_x1	3.22	0.90	0.16
c15 G722.2_12.65 kbps_x2	3.21	0.94	0.16
c07 G722_56 kbps_x3	3.20	0.90	0.16
c51 HE aac_32 kbps_x2	3.16	1.08	0.19
c11 G722.2_8.85kbps_x1	3.16	0.98	0.17
c33 G729.1_20 kbps_x2	3.14	1.00	0.17
c47 HE aac_24 kbps_x1	3.09	1.16	0.20
c19 G722.2_15.85 kbps_x3	3.00	1.01	0.18
c37 G729.1_24 kbps_x3	2.88	1.03	0.18
c52 HE aac_32 kbps_x3	2.84	1.16	0.20
c30 G729.1_14 kbps_x2	2.80	1.00	0.17
c16 G722.2_12.65 kbps_x3	2.73	1.03	0.18
c09 G722_48 kbps_x2	2.52	1.01	0.18
c34 G729.1_20 kbps_x3	2.47	1.01	0.18
c54 MP3_32 kbps_x2	2.35	1.17	0.20
c12 G722.2_8.85kbps_x2	2.28	1.06	0.18
c44 HE aac_16 kbps_x1	2.20	1.10	0.19
c48 HE aac_24 kbps_x2	2.19	1.06	0.18
c10 G722_48 kbps_x3	2.15	1.06	0.18
c31 G729.1_14 kbps_x3	2.04	1.01	0.17
c13 G722.2_8.85kbps_x3	1.72	0.86	0.15
c49 HE aac_24 kbps_x3	1.70	0.87	0.15
c55 MP3_32 kbps_x3	1.69	1.02	0.18
c45 HE aac_16 kbps_x2	1.57	0.84	0.15
c46 HE aac_16 kbps_x3	1.33	0.68	0.12

Appendix B: Coordinates of stimulus spaces (eighteen speech samples)

Codec	Dim. 1	Dim. 2	Dim. 3	Dim. 4
G722.1C_24kbps_x2	-0.78	1.17	0.78	-0.34
G722.1C_24kbps_x3	-0.77	1.24	1.03	-0.09
G722.1_24kbps_x2	-0.86	1.21	0.80	-0.18
G722.1_24kbps_x3	-0.61	1.22	1.13	-0.08
G722.2_12.65kbps_x2	1.62	0.06	0.83	0.19
G722.2_12.65kbps_x3	1.80	-0.05	0.46	-0.10
G722.2_15.85kbps_x2	1.62	-0.02	0.82	0.12
G722.2_8.85kbps_x2	1.87	-0.05	-0.01	-0.62
G722_48kbps_x2	-0.74	-1.75	0.20	-0.08
G722_48kbps_x3	-0.67	-1.87	0.22	0.13
G722_56kbps_x2	-0.76	-1.44	0.53	0.63
G722_56kbps_x3	-0.72	-1.61	0.39	0.24
G729.1_14kbps_3	0.66	0.25	-2.37	0.02
G729.1_20kbps_x3	0.31	0.38	-2.12	0.94
G729.1_24kbps_x2	-0.25	0.51	-1.00	1.74
G729.1_32kbps_x3	-0.47	0.58	-0.43	1.84
MP3_32kbps_x1	-0.76	0.23	-0.27	-1.95
MP3_32kbps_x2	-0.49	-0.07	-0.97	-2.42