Analysing individual differences in speech quality with internal preference mapping

Ville-Veikko Mattila

Speech and Audio Systems Laboratory, Nokia Research Center, P.O. Box 100, FIN-33721 Tampere, Finland, fax: +358 3 2725899, ville-veikko.mattila@nokia.com

Speech quality in mobile communications was studied by making use of the internal preference mapping method (Mattila, [1], pp. 121–136). The main interest in internal preference mapping was to establish how quality judgements, as highly subjective responses, relate to the properties of the auditory samples under test. Internal preference mapping was then used to derive a preference map of the samples. The method was used to achieve a consensus configuration of the samples based solely on the overall quality data and to identify individual patterns of preferences.
Keywords: speech quality, preference mapping, auditory attributes.

INTERNAL PREFERENCE MAPPING

Internal preference mapping is a technique that can be used to analyse the basic structure of hedonic data, e.g., preference, by identifying patterns of preference. Usually, the method compares a set of competitive products, whose overall acceptability has been scored by a panel of consumers. Internal preference mapping is then used to derive a preference map of the products, effectively describing the main underlying preference dimensions. By this technique, the major sources of variation within preference data are identified and extracted as preference dimensions. Maps of a predetermined dimensionality can be drawn, in which products are presented as points and consumers' main preference criteria as vectors. Theoretically, these maps allow the experimenter to visually identify groups of consumers with different preference patterns.

Internal preference mapping has been widely used in the food (Helgesen et al., [2]) and beverage industry (Costell et al., [3] and in brewing trade (Okada and Miyauchi, [4]).

Background of method

Computationally, internal preference mapping is based on the vector model by Slater [5] and Tucker [6]. This model assumes a set of stimulus points embedded in a multidimensional space, where different subjects are represented by distinct vectors. The vectors indicate the direction of increased preference for individual subjects so that the preference order for a subject is given by the projection of stimuli onto the vector representing that subject. The cosines of the angles the vector forms with the coordinate axes directly measure the relative importances to preference. These importances act like coefficients in a linear combination of dimensions.

An intuitively unattractive property of the vector model is that it assumes preference to change monotonically with all dimensions. In this way, it assumes that if a certain amount of a given thing is good, even more must be even better. In the real world, we know that this is not true in most cases. Rather, for most dimensions there is an optimal value and too much may be as bad as too little.

The idea of an optimal value was realized in Coombs' [7] unidimensional unfolding model. Here, stimuli are presented in one dimension and subjects correspond to different "ideal points" that represent their respective optimal values on that stimulus continuum. The farther a stimulus is from a subject's ideal point, the less the subject will like that stimulus.

Bennett and Hays [8] generalised this model to the multidimensional case. In this multidimensional unfolding model, the stimuli and subjects are both presented as points in the same multidimensional space. The points for subjects represent ideal stimuli or optimal sets of stimulus values for those subjects. Here, the farther a stimulus point is from a subjects's ideal point, the less the subject likes that stimulus. The relative distances were assumed to follow an Euclidean metric.

The simple unfolding model assumes that a difference on a dimension makes as much difference to one subject as to another, as well as assuming that all subjects relate to the same set of dimensions within the space. From a behavioural point of view, these assumptions may, however, impose restrictions and reduce the presence of those individual differences in perception, which a preference analysis seeks to account for.

As presented by Carroll [9], it should be noted that the vector model is a special case of the simple unfolding model. This can be seen when moving the ideal point for

an individual far out along a fixed line from the origin, while holding the stimuli constant. Now, the rank order of the distances from the ideal point approach that of projections of stimuli onto a vector whose direction is the same as that of the fixed line.

Unfolding analysis is similar to multidimensional scaling (MDS) using algorithms like MDSCAL as presented by Kruskal [10], [11]. In [9], Carroll distinguished two different types of analyses that he called internal and external modes of analysis. According to this, an internal analysis was one based entirely on the preference data for a set of subjects without reference to an outside or *a priori* set of stimulus dimensions.

Carroll's internal analysis of paired-comparison preference data, based on the unfolding model, is similar to the methods discussed by Slater [5] and Bechtel [12]. The method is called nonparametric multidimensional analysis of paired comparison data as it does not require explicit parametric assumptions, regarding, e.g., distributions, for justification. A computer program was written by Chang and Carroll [13] for carrying out the analysis. This program, called MDPREF for multidimensional preference analysis, is therefore, a special type of factor analysis of either a derived or given preference score matrix which can also be used to analyse absolute acceptability data. MDPREF has been widely used as a generic term to refer to internal preference mapping without specifying the actual underlying method used to derive the stimulus and subject maps.

Principal component analysis (PCA) can be used for internal preference mapping (see, e.g., Dillon and Goldstein, [14], pp. 23–52 for a theoretical discussion and Shepherd et al. [15] for an example in practice) on a matrix of data, consisting of samples as objects and subjects as variables. Here, it should be noted that MDS models are based on distances between points whereas PCA models are based on the angles between vectors. Both models generally use Euclidean space but MDS has the advantage in that it is usually easier to interpret distances between points than angles between vectors. The principal components are constructed as linear combinations of the variables that account for a large part of the *total* variation. The aim is to find a low number of dimensions which explain the largest proportion of variation in subjects' scores and, therefore, represent the common conceptual dimensions underlying the data. However, PCA often results in a relatively large number of dimensions, mainly due to the assumption of linear relationships between the variables under study. This, however, may be a severe assumption with regards to perceptual data.

PCA normally uses the covariance matrix rather than the correlation matrix. This means that a subject with small or zero preferences, and consequently a low standard deviation, will not adversely affect the structure of the preference map. However, if it is assumed that the differences of individual variances are artifacts caused by individual ways of using the scale, the PCA can be performed on the basis of a correlation matrix.

Interpretation of PCA is usually done with a biplot representation of both subjects and samples on the first two axes. This biplot is defined as the superimposition of the first two vectors of loadings and the first two principal components on the same plot.

TEST STIMULI

The stimuli under evaluation were speech signals transmitted through various speech processing systems. They were relevant examples of the auditory characteristics typically experienced in mobile communications. All processes were applied to four reference speech samples. Two of the samples were of male and female speakers each speaking different sentences, while the other two were these same speech samples corrupted by car cabin noise to produce an average signal-to-noise ratio (SNR) of about 10 dB. The test stimuli were produced by processing the four speech samples with 44 different processing chains presented in table 1. These processing chains, presented in detail in (Mattila, [1], pp. 71–96), can be categorized into seven classes of processing types:

- transmissions through real mobile communication systems;
- speech coding algorithms;
- speech enhancement algorithms;
- speech coding with speech enhancement;
- tandem connections of speech coders;
- speech coding and discontinuous transmission
- speech and transmission channel coding;
- artificially generated speech distortions.

To simulate the effects of a full transmission path on speech in mobile communication system, the samples were transmitted through real mobile networks. The speech coders were based on waveform coding, vocoding and hybrid coding, and employed different bitrates. Here, the methods in the most important codec standards were present. Also included were some artificially generated speech distortions. These included different types of filtering, quantization, addition of echo and noise, peak and center clipping, etc. Additionally, the clean speech samples were corrupted with car noise at SNR of about 5 dB to study the effect of very high level noise on speech

coding. A total of 170 test stimuli were developed for the analysis.

OVERALL QUALITY JUDGEMENTS

The overall quality data was collected by asking 30 subjects who were screened and trained by the Generalised Listener Selection (GLS) procedure (Mattila, Zacharov, [16]), to rate the overall acceptability of the 170 test samples. Each stimulus was repeated six times to collect a total of 30600 judgements. The subjects had not been trained to evaluate any specific perceptual auditory characteristics and could, therefore, be considered *naïve* in this respect.

Test procedure

A single subject participated in each session and was asked to listen to a test sample at least once after which its overall quality should be rated on a continuous line scale presented in figure 1. The test was performed as a single stimulus absolute quality test with no reference sample within each test item.

Analysis of variance

A factorial analysis of variance (ANOVA) was subsequently applied to the test data to gain a general insight into the variables. A fixed model was used in the ANOVA analysis, i.e., all the factors were considered fixed. The independent variables were *Process* (degrees of freedom (df) = 43), denoting the processing chains, *Bckgrnd* (df = 1), denoting the far-end environments, *Speaker* (df = 1), denoting the test speakers and *Subject* (df = 29), denoting the test subjects, while *Repetitn*, denoting the six repetitions, was used as a co-variate. The ANOVA model included all independent variables and the co-variate as main factors, all two-way interactions between the independent variables and also included the random error. The co-variate was, therefore, not used in interactions. As different playlists were used by all subjects and in all experiments, the effect of the presentation order (block effect) could not be measured. Moreover, because the samples were divided into two adjacent sessions and the samples were different within each session, the number of the session could not be used as a factor to test the influence of the division.

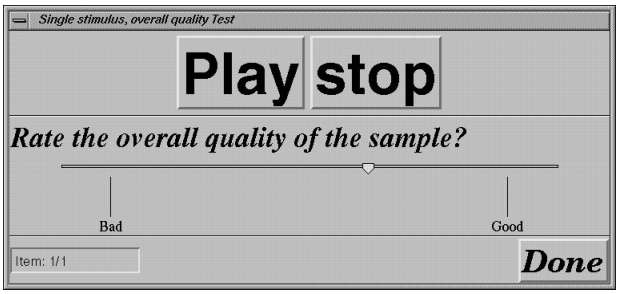

FIGURE 1. The user interface for the single stimulus absolute quality test

The ANOVA assumption of a normal sampling distribution at each independent variable could not be stated in Kolmogorov-Smirnov tests. However, the departure from normality was not severe as, e.g., the absolute standard errors for skewness and kurtosis of the *Process* were less than four. Moreover, the residuals were normally distributed and the main assumption of the equality of the residual error variance was fulfilled. In this way, the ANOVA analysis could be performed.

The results of the ANOVA with the main effects and the two-way interactions are shown in table 2. All the main effects and interactions except the interaction *Bckgrnd*Speaker* were significant ($p \leq 5\%$). By far the highest F-ratio was obtained by *Bckgrnd*, indicating the clear difference between clean speech and noisy speech. The second highest F-ration was measured by *Process*, expressing that the processing chains were judged to be clearly different in overall quality. This was an important result as the aim was to develop and select processing chains providing a wide range of subjective acceptability. Note that the co-variate *Repetitn* is significant and the reported mean values are, therefore, corrected for this influence. Here, the significance was assumed to be due to learning effect, i.e., the subjects were gradually learning to use the scale consistently as they were familiarized with the test samples.

The highest F-ratio for interactions was resulted by *Process*Bckgrnd* denoting that the perceptual characteristics produced by some processing chains are dominating the role of background noise on judgements. Moreover, the presence of noise, the perceptual properties created when processing noisy speech and noise alone influenced subjects' judgements rather differently as indicated by the *Bckgrnd*Subject* interaction, scoring the second highest F-ratio. On the other hand, although the *Process*Subject* interaction was significant, its relatively low F-ratio denotes that processing chains were evaluated in a roughly similar manner.

As shown at the bottom of table 2, $R^2 = 0.852$, meaning that the model explains about 85 % of the variability of the overall quality judgements. As the adjusted R^2 obtained only slightly smaller value than R^2, the model

Table 1. Summary of the processing chains

Nr.	Processing	Symb.	Nr.	Processing	Symb.
1	Land line phone to mobile phone transmission, GSM-EFR	LME	23	GSM-EFR and transmission channel errors (*C/I* = 7 dB)	ES2
2	Land line phone to mobile phone transmission, GSM-HR	LMH	24	GSM-EFR and transmission channel errors (*C/I* = 4 dB)	ES3
3	Mobile phone to mobile phone transmission, GSM-EFR	MME	25	GSM-EFR and transmission channel errors (*C/I* = 3-15 dB)	DH3
4	Mobile phone to mobile phone transmission, GSM-HR	MMH	26	GSM-HR and transmission channel errors (*C/I* = 7 dB)	HS2
5	Land line phone to mobile phone transmission, GSM-EFR, HATS	VTE	27	GSM-EFR, muting of lost frames	EMT
6	Land line phone to mobile phone transmission, GSM-HR, HATS	VTH	28	Transducer	TRD
7	PCM, ITU G.711	PCM	29	Wideband additive noise	GNS
8	ADPCM, ITU G.726	ADP	30	Narrowband additive noise	NBN
9	LD-CELP, ITU G.728	LDC	31	Addition of echo	ECH
10	GSM-FR, ETSI GSM 06.10	FR	32	Clipping	CLP
11	GSM-HR, ETSI GSM 06.20	HR	33	Center clipping	CCL
12	GSM-EFR, ETSI GSM 06.60	EFR	34	Highpass filtering	HP
13	EVRC, TIA IS-127	EVR	35	Lowpass filtering	LP
14	TETRA, TETRA 06.20	TRA	36	Bandpass filtering	BP
15	PDC-HR, RCR	PDC	37	Angular pole distortion	APD
16	U-ALWE (noise suppression)	UAW	38	Radial pole distortion	RPD
17	UDRC (dynamic range control)	DRC	39	Pole distortion	POD
18	U-ALWE and GSM-EFR	EAW	40	Narrowband frequency distortion	NBD
19	U-ALWE and GSM-HR	HAW	41	No Processing	ORG
20	GSM-EFR and GSM-EFR (tandem connection)	ETD	42	No Processing (SNR = 5 dB)	OR5
21	GSM-HR and GSM-HR (tandem connection)	HTD	43	GSM-HR, ETSI GSM 06.20 (SNR = 5 dB)	EF5
22	GSM-EFR with DTX	EDX	44	GSM-EFR, ETSI GSM 06.60 (SNR = 5 dB)	HR5

could be considered to be representing the subjective data well.

Table 2. ANOVA results for the single stimulus absolute quality test data

Tests of Between-Subjects Effects
Dependent Variable: Q

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Observed Power
Corrected Model	14440705.246	1464	9863.870	114.304	0.000	1.000
Intercept	12947378.324	1	12947378.324	150035.826	0.000	1.000
REPETITN	39937.673	1	39937.673	462.803	0.000	1.000
PROCESS	10649624.180	43	247665.679	2869.981	0.000	1.000
BCKGRND	806048.780	1	806048.780	9340.593	0.000	1.000
SUBJECT	645841.451	29	22270.395	258.072	0.000	1.000
SPEAKER	1766.629	1	1766.629	20.472	0.000	0.995
PROCESS * BCKGRND	456378.071	40	11409.452	132.214	0.000	1.000
PROCESS * SUBJECT	1485998.982	1247	1191.659	13.809	0.000	1.000
PROCESS * SPEAKER	136669.829	43	3178.368	36.831	0.000	1.000
BCKGRND * SUBJECT	122391.565	29	4220.399	48.907	0.000	1.000
BCKGRND * SPEAKER	1.506	1	1.506	0.017	0.895	0.052
SUBJECT * SPEAKER	50086.337	29	1727.115	20.014	0.000	1.000
Error	2514211.957	29135	86.295			
Total	90682334.147	30600				
Corrected Total	16954917.202	30599				

a Computed using alpha = .05

b R Squared = .852 (Adjusted R Squared = .844)

Principal component analysis of acceptability data

Internal preference mapping analysis was carried out by PCA on the acceptability data. PCA was selected as the analysis method because MDPREF would have required a full paired-comparison test to be done. With 170 test samples, the test would have required a collection of $170 \times 169 = 28730$ judgements from each subject, which was not considered feasible due to listener fatigue. Furthermore, the size of the test could not have been reduced enough with an incomplete block design.

The acceptability data was averaged over the test speakers, though the voice types were noticed to influence judgements in the ANOVA. As the associated F-ratio of the *Speaker* main effect was, however, about twelve times smaller than the second smallest F-ratio of the *Subject* main effect and about 140 times smaller than the third smallest F-ratio of the *Process* main effect, the averaging was justified. The data matrix for the Q-type

PCA consisted, therefore, of 85 samples as objects, including 41 processing chains for clean speech and 44 processing chains for noisy speech, and 30 subjects as variables. Therefore, the ratio of objects to variables was about 2.8.

An initial PCA analysis was run to find the right dimensionality using no rotation of the component matrix. The covariance matrix of the subjects was used as the basis for the analysis because the correlation matrix would have removed differences attributable to both the mean and the dispersion of the individuals. The first three components accounted for about 88.8 %, 3.2 % and 2.4 %, respectively, of the total variation in the data. Thus, cumulatively, about 92.0 % of the total variance of the subjects could be explained by the first two components, as the third component and the higher components were expected to model only noise in the data. The use of two components was also supported by the scree plot, presented in figure 2, where a clear knee could be stated in the eigenvalue graph at component two.

Bartlett's test of sphericity (see, e.g., Basilevsky, [17], pp. 185–194) resulted in the approximate Chi-Square of about 6426.4 (df = 435) that was non-significant, indicating that a correlation exists between the subject variables that can be utilised in PCA.

A PCA was then rerun on the data by setting the dimensionality to two components. The component matrix was rotated using the Varimax method that seeks to rotate the components so that the variation of the squared loadings for a given component is made large. Generally, the method tries to make large loadings larger and small loadings smaller, making the interpretation of the components easier. The use of Varimax is recommended, e.g., by Kline ([18], pp. 76–77). After the rotation, the proportions of the total variance accounted for by the first two components remained the same, i.e., 88.8 % and 3.2 %.

The scores of the sample objects in the rotated space are shown in figure 3. Here, the samples are named as in table 1. The scores are standardised with a mean of zero and a standard deviation of one. The scores were pretty well distributed with a circular shape, indicating that the components were approximately normally distributed. Normality is not a necessary assumption for PCA, but outliers can distort results. As only scores outside about ± 2.5 could be considered extremes with the sample sizes used here, no unusual samples were identified.

The loadings of the subject variables are also shown in figure 3 as vectors drawn from the origin. These loadings

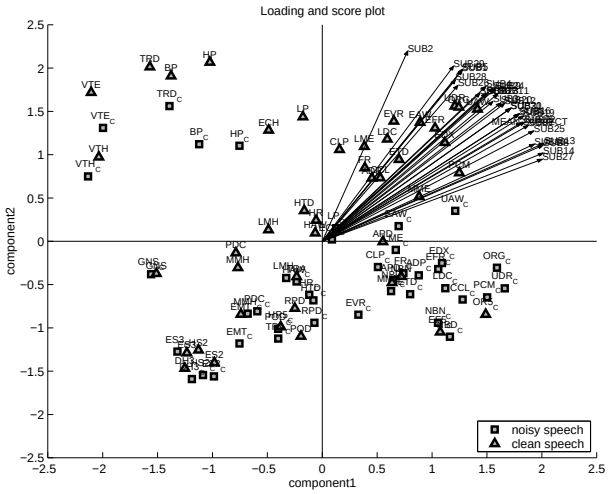

FIGURE 3. Biplot of the sample scores and subject loadings

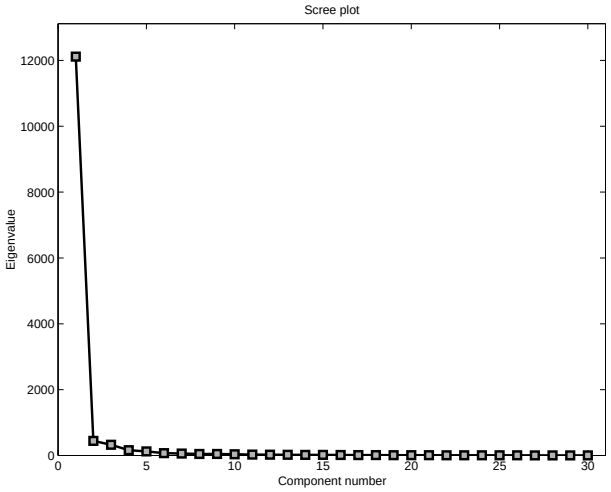

FIGURE 2. Scree plot for PCA of the acceptability data

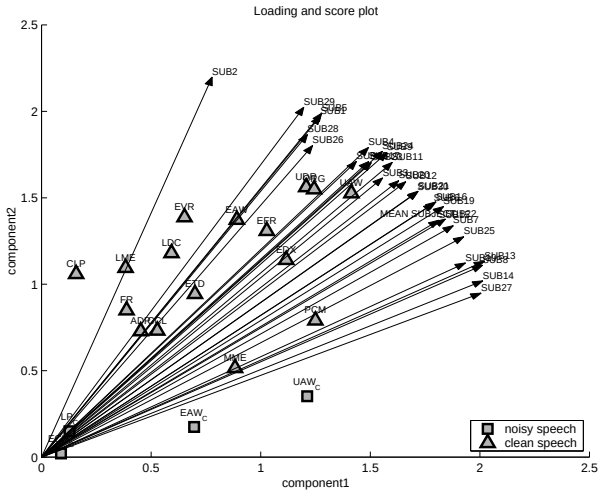

FIGURE 4. Biplot of the sample scores and subject loadings in the first quarter

are the correlations of the variables with the two components. The vectors indicate the directions of increased preference for the subjects. As the samples (scores) were mean-centered, the vectors indicate preference that is greater than the mean preference. This is how subject vectors are usually presented in internal preference mapping. It can be noticed that the subjects had a very similar opinion about the overall quality of the samples and only subject 2 had a slightly different view to the quality as shown in figure 4. Subject 2 is clearly more sensitive to the presence of background noise than the other subjects as shown, e.g., by the processing chains OR5, representing speech with very high noise level, NBD, representing narrow band frequency distortion, and NBN, representing narrow band additive noise. Moreover, high bitrate coding of noisy speech, e.g., PCM, ADP, LDC and EFR, was evaluated to be of less overall quality by subject 2 than by the other subjects. On the other hand, subject 2 considered the effect of high- and bandpass filtering on speech quality clearly less important than the other subjects as shown by the processing chains HP and BP. In clean speech, both filtering types made the clean speech sound brighter and, in addition, removed low-frequency background noise in noisy speech.

The individual loadings were averaged to produce the mean direction of preference for subjects. This mean subject vector was extended to cover all samples as shown in figure 5. Here, the dotted line that is orthogonal to the mean subject vector, indicates the mean preference dividing, to the right hand side of which the samples are above the mean preference and to the left hand side below the mean preference.

As suggested in the ANOVA, the processed clean speech samples were generally clearly preferred to the corresponding processed noisy speech samples. However, for example processes ES2, ES3, DH3 and HS2, representing speech coding with erroneous transmission channels, have been judged approximately equal in acceptability regardless of the far-end environment. This means that the associated perceptual characteristic common to these samples has been dominating the judgements. As presented in multidimensional scaling and semantic differentiation analyses in (Mattila, [1]) this characteristic was related to the abrupt muting of lost frames (interruptions). Other examples include processes RPD and POD, representing pole distortions in a LPC model, these distortions being perceived as a bubbling sound alternating with the speech signal. Poor general naturalness was clearly dominating the judgements with some processing chains. For example, processes VTE and VTH, representing mobile phone to mobile phone transmissions with GSM-EFR and GSM-HR speech coding, including especially transducer and leakage effects, caused the speech specifically sound tense, mechanic, rough, rustling, scratching and crackling. In addition, processing chains, representing different types of noise distortion, were evaluated similarly in clean and noisy speech.

In figure 6, a clear trend of increasing difference in acceptability between processed clean speech and noisy speech samples can be noticed with increasing MOS values. This effect is also suggested by the relatively high F-ratio of the *Process*Bckgrnd* interaction in the ANOVA.

Above the mean preference, there is a clear difference between the far-end environments in the judgements of

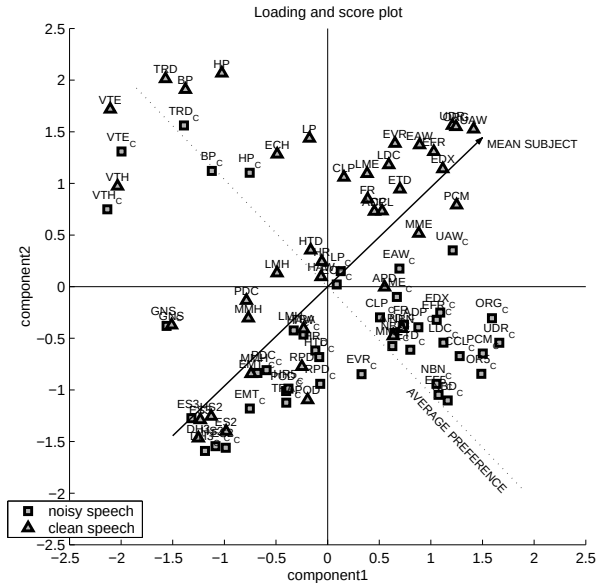

FIGURE 5. Biplot of the sample scores and the mean subject loading

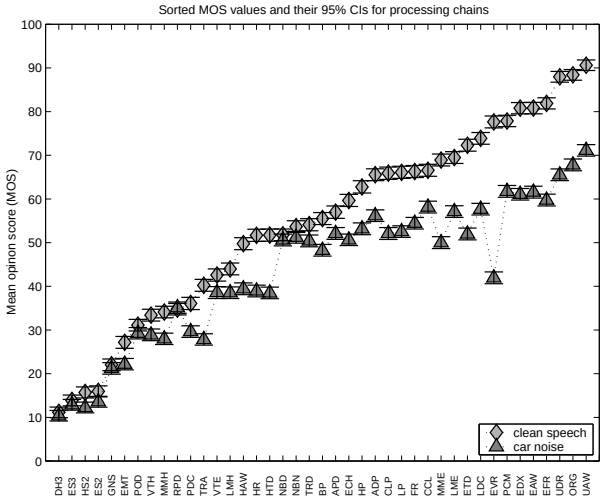

FIGURE 6. MOS values and their 95 % confidence intervals (CIs) for processing chains (sorted with respect to MOS values)

the processing chains. The ten most preferred samples represent processes applied to clean speech. The three most preferred samples include the processes UAW, representing acoustic background noise suppression, UDR, representing dynamic range control and ORG, representing the original, unprocessed clean speech sample. The UAW processing was slightly preferred to ORG as the just audible electrical noise in clean speech samples was suppressed by about 10 dB, making it inaudible. On the other hand, the UDR processing should not have any effect on clean speech signals.

MAPPING ATTRIBUTES TO A PREFERENCE MAP

As the internal preference mapping is only based on the preference data, it cannot be used to understand the reasons for preference on its own. Although it is the external preference mapping that is dedicated to relating sensory properties to preference scores, some scientists have also used internal preference mapping for that purpose. Here, the sample coordinates in the internal map are correlated to the direct attribute scaling data of each sensory attribute (Schlich, [19]). If, e.g., two principal components (PCs) are providing a sufficient explanation of preference in a PCA, the attributes can be presented in the score plot together with the studied samples by plotting an (x,y) coordinate for each attribute, as two correlation values are obtained for each attribute, one with the first preference dimension and one with the second preference dimension (McEwan, [20], p. 99).

It has been argued that the performance of this approach, based solely on bivariate correlations, in explaining preference is far worse than the performance of external preference mapping and that the approach is used to overcome technical difficulties attached to external preference mapping models (Schlich, [19]).

In the following, the auditory attributes referred to have been developed in semantic differentiation analysis in (Mattila, [1], pp. 169–202) to describe auditory characteristics that can be used in analytic evaluation of speech quality in mobile communications.

Correlation analysis of preference dimensions and attributes

The Q-type PCA analysis of the overall acceptability data was carried out as above with the exception that, now, the samples were not averaged over the test speakers. The data matrix for PCA included 170 samples as rows and 30 subjects as columns. The speakers were separated as, especially, attributes *dark–bright, rough, scratching* and *rustle*, in general, obtained different values with male and female speakers as presented in (Mattila, [1], p. 199).

The first three principal components accounted for about 87.0 %, 3.3 % and 2.4 %, respectively, of the total variation in the data. Hence, cumulatively about 90.0 % of the total variance of the subjects could be explained by the first two components, as the third component and the higher components were expected to model only noise in the data. The use of two components was also supported by the scree plot, where a clear knee could be stated in the eigenvalue graph at component two.

Bivariate Pearson's correlations, measuring the strength of the linear relationship between variable values, were now computed between the sample coordinates in the resulted score plot and the averaged direct attribute ratings. The correlation values that are significant at the 0.05 level, are shown in figure 7. It can be noticed that the correlation of the attributes to the first two principal components, *PC*1 and *PC*2, was, in general, very low. There were no attributes, at which both of the PCs would have reached significance at the 0.05 level. On the other hand, the *dark–bright, smooth–interrupted, creaking* and *low–high* attributes obtained significance on *PC*2 at the 0.05 level, whereas the *hissing* attribute on *PC*1, and the *humming, noisy, bubbling* and *boiling* attributes on *PC*2 were significant at the 0.01 level. However, the correlations were very low, ranging in absolute value from 0.155 to 0.350.

Due to the absence of significant correlation pairs and due to low correlation values, it appeared that, based on bivariate correlation, the attributes could not be used to interpret the internal preference mapping solution. The coordinates of the test samples on the first two principal components (component scores) had weak linear relationships to the averaged direct ratings of the attributes.

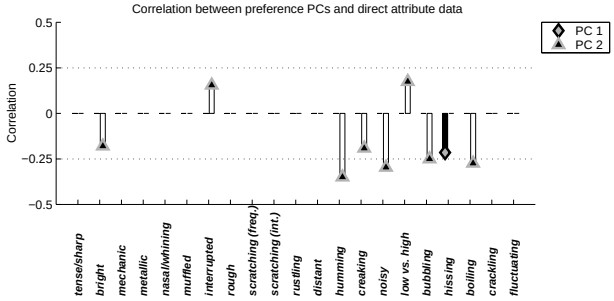

FIGURE 7. Correlations between the first two principal components and direct attribute ratings (only statistically significant correlations shown)

The disadvantage of making use of bivariate correlations is that they remove differences attributable to both the mean and the dispersion of the component scores and the direct attribute ratings. Thus, a principal component and an attribute exhibiting the same rank order of samples and relative spacing will correlate highly regardless of any differences in mean levels and dispersion.

CONCLUSION

It was shown that internal preference mapping could be used to visualize preference patterns. However, it appeared that the method could not be used to establish and interpret auditory attributes behind the acceptability judgements. Extensive methods are needed to develop auditory attributes that can be used to distinguish between auditory stimuli. In addition, complex analysis methods are needed to map the auditory attributes to overall quality judgements, i.e., to establish the interrelationship between auditory attributes and preference. Both multidimensional scaling and semantic differentiation can be used to develop auditory attributes relevant for speech quality as presented in (Mattila, [1], pp. 137–168 and pp. 169–202; Mattila, [21]; Mattila, [22]). Furthermore, external preference mapping and partial least-squares regression with a simple vector model for preference were shown in (Mattila, [1], pp. 203–248; Mattila, [21]; Mattila, [22]) to be successful methods and techniques to map auditory attributes to overall speech quality with a moderate prediction error.

REFERENCES

1. V.-V. Mattila. *Perceptual Analysis of Speech Quality in Mobile Communications*. PhD thesis, Tampere University of Technology, Tampere, 2001.
2. H. Helgesen, R. Soleim, and T. Næs. Consumer preference mapping of dry fermented lamb sausages. *Food Quality and Preference*, 8:97–109, 1969.
3. E. Costell, M. Vicenta Pastor, L. Izquierdo, and L. Durán. Relationship between acceptability and sensory attributes of peach nectars using internal preference mapping. *European Food Research and Technology*, 3:199–204, 2000.
4. A. Okada and A. Miyauchi. Predicting the amount of purchase by a procedure using multidimensional scaling: An application to scanner data on beer. In P*roceedings of the 21st annual conference of the Gesellschaft für Klassifikationen*, University of Potsdam, 1997.
5. P. Slater. Inconsistencies in a schedule of paired comparisons. *Biometrika*, 48:303–312, 1960.
6. L. R. Tucker. *Psychological Scaling: Theory and Applications*, chapter 13: Intra-individual and inter-individual multidimensionality, pages 155–167. John Wiley & Sons, New York, 1960.
7. C. H. Coombs. Psychological scaling without a unit of measurement. *Psychological Review*, 57:148–158, 1950.
8. J. F. Bennet and W. L. Hays. Multidimensional unfolding: determining the dimensionality of ranked preference data. *Psychometrika*, 25:27–43, 1960.
9. J. D. Carroll. *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, volume I, chapter Individual differences and multidimensional scaling, pages 105–155. Seminar Press, New York and London, 1972.
10. J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.
11. J. B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29:115–129, 1964.
12. G. Bechtel. Individual differencies in the linear multidimensional scaling of choices. *Presented at the meeting of the Psychometric Society, Princeton, New Jersey*, 1969.
13. J.-J. Chang and J. D. Carroll. How to use MDPREF, a computer program for multidimensional analysis of preference data. Bell Telephone Laboratories, 1968.
14. W. R. Dillon and M. Goldstein. *Multivariate Analysis*. John Wiley & Sons, 1984.
15. R. Shepherd, N. M. Griffiths, and K. Smith. The relationship between consumer preferences and trained panel responses. *Journal of Sensory Studies*, 3:19–35, 1988.
16. V.-V. Mattila and N. Zacharov. Generalized listener selection (GLS) procedure. In *Proceedings of the Audio Engineering Society; 110*th *International Convention*. Audio Engineering Society, 2001.
17. A. Basilevsky. *Statistical Factor Analysis and Related Methods*. John Wiley & Sons, 1994.
18. P. Kline. *An Easy Guide to Factor Analysis*. Routledge, 1994.
19. P. Schlich. Preference mapping: Rerating consumer preferences to sensory or instrumental measurements. *in Bioflavour 95, Paris*, 48:135–150, 1995.
20. J. A. McEwan. *Multivariate Analysis of Data in Sensory Science*, chapter Preference mapping for product optimization, pages 71–102. Elsevier, 1996.
21. V.-V. Mattila. Multidimensional scaling of speech quality in mobile communications. In *Proceedings of the International Congress on Acoustics; 17*th *International Congress on Acoustics*. International Comission for Acoustics, 2001.
22. V.-V. Mattila. Descriptive analysis of speech quality in mobile communications: Descriptive language development and external preference mapping. In *Proceedings of the Audio Engineering Society; 111*th *International Convention*. Audio Engineering Society, 2001.