Semantic analysis of speech quality in mobile communication: descriptive language development and mapping to acceptability

Ville-Veikko Mattila

Nokia Research Center, Visiokatu 1, P.O. Box 100, 33721 Tampere, Finland

The quality of speech in mobile communication was studied by making use of semantic differentiation and external preference mapping (Mattila, [1], pp. 169-202, 208-248). Semantic differentiation (Osgood, [2]) was used to extract the perceptual characteristics of speech and background noise, originated from speech processing, which can be used to differentiate between processed speech samples. External preference mapping (Carroll, [3]) was then applied to establish the attributes explaining overall acceptability and their relative importance to quality.

Two phonetically rich sentences spoken by a male and a female speaker served as the basic speech samples. These samples were corrupted by a car cabin noise and together with the clean speech samples processed by 41 different processing chains, representative to mobile telecommunication. Additionally, three specific processing chains were only applied to clean speech. These 170 test samples may be divided into seven main processing categories including, e.g., transmissions through real mobile communication systems, speech coding algorithms, speech enhancement algorithms, tandem connections of speech coders, speech and transmission channel coding, etc.

A "listen and describe" technique was used to collect spontaneous verbal descriptions about the test samples from 15 screened and trained subjects (Mattila and Zacharov, [4]). The objective was to identify and verbally describe all the perceptual characteristics of the speech signal and the background noise process. A distinction was made between descriptions originated from processing and the ones that were also present in the original samples so as to separate the perceptual characteristics of the processing chains. In this first collection, a total of about 15100 words were gathered.

Since there may be great differences in the ability to verbalize a perception of an audio characteristic, the subjects were informed about the descriptions given by the other subjects in a replicated run of the collection. This was done to ease the verbalization process and broaden the view to the stimuli. A total of about 21100 words were collected in this second collection.

After the two collections, a preliminary grouping of attributes was carried out by identifying words equal in meaning. Round-table discussions were then held to decide whether a description suggested by one subject was equivalent in meaning to that suggested by someone else, and attempts were made to develop the facility to perceive descriptions once their presence was called to subjects' attention.

21 attributes, eleven for speech signal (*tense/sharp, bright, mechanic, metallic, nasal/whining, muffled, interrupted, rough, scratching* (frequency and intensity), *rustle* and *distant*) and nine for background noise (*humming, creaking, noisy, low vs. high, bubbling,*

hissing, boiling, crackling and fluctuating) were finally considered to represent all the perceptually important aspects of the samples.

The panel discussions were also used to develop a rating scale with associated anchor words for each attribute to guarantee effective use of the scales. Further, the subjects were gathered in groups to select an anchor sample for each attribute from a pre-selected set of samples. These samples served as examples for the specific audio characteristics in question.

Finally, the intensities of the 21 attributes in the 170 test samples were evaluated by 18 screened and trained subjects in an attribute scaling test. Here, each attribute was scored separately so as to minimize cross-correlation between attributes. Each subject gave 3570 judgements and a total of 64260 judgements were collected.

A factorial analysis of variance (ANOVA) was first performed for the whole data to check if the attributes were different. After this, separate ANOVAs were carried out for each attribute to ascertain that the attributes could provide sample differentiation.

Before the semantic differentiation, the test samples were evaluated for absolute overall quality by thirty subjects. 1020 judgements were collected from each subject in six repetitions of the test, resulting in a total of 30600 judgements.

The acceptability and the attribute data were checked in principal component analyses (PCA) for clusters of subjects. As subjects seemed to share a similar view to the overall quality and attributes, both data sets were averaged over the subjects.

Partial least square (PLS) regression (Martens and Næs, [5], pp. 85-165) was used to map the attribute scaling data to the acceptability data. The attractive property of PLS is that it could take into account both data sets in the regression model, extracting relevant factors from an interpretation and a prediction point of view.

Fourteen of the 21 attributes were noticed to be significant by Marten's uncertainty measure (Martens and Martens, [6]) to explain the acceptability. However, eighteen attributes provided the best prediction model but to obtain a simpler model sixteen attributes, achieving roughly the same performance were used to predict acceptability, providing a root mean square error for prediction of about 6 %. Hereby, it can be stated that attributes describing perceptually important characteristics of processed speech and background noise can be used to predict quality and to give a multidimensional view to quality.

References

1. V.-V. Mattila. Perceptual Analysis of Speech Quality in Mobile Communications. PhD thesis, Tampere University of Technology, Tampere, 2001.
2. C. E. Osgood. The nature and measurement of meaning. *Psychology Bulleti*n, 49:197–237, 1952.
3. J. D. Carroll. *Multidimensional Scaling: Theory and Applications in the Behavioral Science*s, volume I, chapter Individual differences and multidimensional scaling, pages 105–155. Seminar Press, New York and London, 1972.
4. 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.
5. H. Martens and T. Næs. *Multivariate Calibratio*n. John Wiley & Sons, 1989.

6. H. Martens and M. Martens. Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food Quality and Preferenc*e, 11:5–16, 2000.