Methods for Assessing Speech Intelligibility and Quality in Cars

Andreas Wendemuth[1], Thomas Starruß[2]

[1] *Institut f. Elektronik, Signalverarbeitung und Kommunikationstechnik, Otto-von-Guericke Universität, 39106 Magdeburg, Deutschland, Email: Andreas.Wendemuth@e-technik.uni-magdeburg.de*

[2] *Ingenieurgesellschaft Auto und Verkehr (IAV), Rockwellstr. 16, 38518 Gifhorn, Deutschland, Email: thomas.starruss@iav.de*

Abstract

Acoustic environments in cars are particularly challenging for automatic processing of speech. Handsfree telephone sets and automatic speech recognition in cars are prominent examples. Established performance measures, e.g. the word recognition rate in speech recognition, are note appropriate in these environments. Instead, intelligibility and quality measures are required which give information about the success of a command or conversation. In this paper, several standardized measures are compared. A test environment is presented and subjective user assessments are reported. It is discussed whether these measures are appropriate for speech quality and speech intelligibility assessment. Criteria are identified for future assessments.

Acoustic Devices in Cars

Since it was required in 2001 by German law that devices like hands-free telephone sets are used for telephoning in cars, they have become standard equipment. Similarly, automatic speech recognition of spoken commands or longer spoken units in cars becomes fashionable: tuning the radio or giving commands to the navigation system are prominent examples.

For optimum performance, for marketing and for comparability reasons, these applications have to be assessed in manufacturing and in usability tests. This poses a serious problem of establishing proper quality and intelligibility measures. Standard measures like the word recognition rate of automatic speech recognition systems do not give information about the success of a desired action, or about the *perceived* quality of speech in telephony. Human user tests carry that information, but they are prohibitive due to their enormous effort and non-standardized character.

Measures for speech intelligibility

Speech *intelligibility* in Germany is defined in the German DIN 9921 as "classification of the part of the spoken speech which was understood". Speech *quality* is speech intelligibility plus factors such as individual perception of disturbing secondary sources, a clear sound, echoes and the required hearing effort.

Technical measures mostly rely on modelling the loudness perceived by man. Typically, the amplitudes of different frequency ranges in speech are weighted to obtain an auditory frequency scale. The perceived loudness is then calculated from the energy in this auditory scale. To obtain a measure of speech intelligibility, two main methods are then applied: analysis of the hall, and of disturbing secondary sources. To incorporate further factors relevant for speech intelligibility, a number of measures have been developed. The AI (Articulation Index) [1] relies on auditorily corrected SNR analysis and is laid down in the norm ANSI S3.5-1969. However, main and secondary signal must be available separately, the signals must be rather stationary over trial time. This is most severe for systems with nonlinear transmission characteristics, as in GSM speech coding, where separation fails. Hence, AI is only partially useful for measuring speech intelligibility in electro-acoustical sound transmission. The SII (Speech Intelligibility Index) is an improvement of AI, laid down in ANSI S3.5-1997. It weights the critical bands differently and individually. As AI, it requires that main and secondary signal must be available separately, leading to the same reduced usefulness. The STI (Speech Transmission Index) is laid down in DIN EN 60268-16:2003. The degree of modulation of an amplitude-modulated test signal or a speech signal [2] is compared between input and output of transmission. A reduced degree of modulation is a measure for information loss and hence for speech intelligibility. Advanced versions take into account masking effects (reduction of auditory sensitivity by lower frequency sound) and absolute auditory perception threshold. STI incorporates noise effects, but fails when vocoders, automatic gain control or noise suppression is active, since this manipulates the modulation severely. RASTI (Room Acoustics Speech Transmission Index) is an improved, simplified version (less frequency bands) of STI which is suitable for assessing direct communication between persons in closed rooms. It suffers from similar drawbacks than STI, in particular the secondary noise must be free from peaks or clearly audible tones. STITEL (Speech Transmission Index for Telecommunication Systems) is also a simplified version of STI, designed for telecommunication systems. The drawbacks are the same as in RASTI.

Measures for speech quality

All measures for speech *quality* base on loudness computed from the original and the disturbed (output) signal in various frequency bands. From the loudness patterns, perceptive patterns are computed and weighted. The weight factors are derived from subjective tests, such that objective and subjective measurements are equalized. The methods we present now differ in preprocessing and weighting. PSQM (Perceptual Speech Quality Measurement) is laid down in ITU-T P.861, where the parameters were optimized on the basis of auditive tests

according to ITU-T P.800 which recommends the framework for subjective speech quality tests. The perceptive patterns aim at modelling cognitive features, such as loudness, secondary sources, asymmetries and speech pauses. PSQM is ideally suited to assess speech codecs. Variable and short local delays are a problem, and filter characteristics are not properly detected. This is especially harmful if vocoder technologies are used. PESQ (Perception Evaluation of Speech Quality) was developed jointly by KPN (Dutch Telecom) and British Telecom. It amends PSQM by a compensating transfer function for the transmission, a different time alignment and methods for compensating variable delays. It is laid down in ITU-T P.862 which replaced PSQM in 2000. This results in a so-called MOS (mean opinion score) which measures the disturbances. TOSQA (Telecommunication Objective Speech Quality Assessment) was independently developed by Deutsche Telekom to overcome the drawbacks of PSQM. Its main focus is the temporal synchronization of the signals by correlation analysis. Frequency shifts due to the transmission can be compensated for. TOSQA has been improved by TASQ (Telecom analysis of speech quality). Here, the MOS are not limited from below. This allows analysis of very low quality speech. TASQ also indicates the likely reason for the disturbance, since various potential disturbing scenarios were trained and can be recovered.
In general, PESQ and TOSQA(+TASQ) are both suitable for the assessment of speech quality. However, they were developed for telecommunications applications and may not be suitable in free-speech scenarios under adverse car noise. The reason is that the technical processing is geared towards comparing clean and disturbed signals which should be not too different from each other, as it is the case in office environments but not in car environments. The new ITU-T P.862.3 which was approved in November 2005 (status in March 2006: pre-published) gives further application guidelines.

Test environment

We describe here a typical test environment which is used for intelligibility and quality tests. Car noise samples are recorded in a medium-class limousine at 50 km/h on poor road, and at 100 km/h and 130 km/h on high-quality roads. Speech material from 3 male and 3 female speakers were recorded under laboratory conditions, where the car noise was displayed to the speakers. The volume (audio pressure) was recorded as well. Later, the speech was displayed by loudspeakers at head position in the car where the volume was measured and the same pressure was realized than in the recordings. The speech was then recorded again at the electrical interface between microphone and telephone driver unit. This enabled us to simulate different telephone drivers later in the lab. We used drivers with galvanic (systems 2,3,4) and bluetooth (1,5,6) connection to a (mobile) phone, where internal signal processing was either disturbed (1), activated with standard (2) and high (3) noise reduction, or de-activated (4). We further tested the GSM module provided by the telephone driver unit, which enabled car-specific noise suppression (5,6) and additional dynamic compression (5). Numbers correspond to table 1.
3 male and 3 female subjects had to assess the displayed speech *intelligibility* after the telephone line on a scale of $0 \cdots 1$. The well-established rhyme test after Sotscheck was used for this task. The only assessment which, according to ISO TR 4870:1991, corresponds to insufficient intelligibility, was obtained for the disturbed bluetooth (1) interface at 50 km/h on poor road. All other measurements showed fair, none showed excellent intelligibility. As a result, all test situations were burdened by increased hearing effort, however speech intelligibility was still sufficient for the following *quality* test. This was performed by subjects according to ITU-T P.800 with pairs of undisturbed and disturbed sentences recommended in DIN 45621/2 on a scale 1-5, which can directly be related to MOS (ITU-T P.800), giving values of 5 (excellent) - 1 (bad). Table 1 displays the subjective quality values at the three road conditions. Subjects downvalue test conditions 5 and 6 with active GSM noise suppression, in particular dynamic compression (5) leads to very low quality. The subjective quality decreases as well for disturbed bluetooth (1) and de-activated signal processing in galvanic systems (4). Even though the test with 50 km/h was performed on poor road, the results are still better than with the tests with 100 km/h and 130 km/h which were performed on high-quality roads. At 50 km/h, the quality was roughly assessed as "fair" or better, which degraded to just slightly better than "poor" at 130 km/h.

System	1	2	3	4	5	6
50 km/h	2.87	3.71	3.76	3.58	2.73	3.12
100 km/h	2.37	3.19	3.26	3.11	2.49	2.76
130 km/h	2.35	2.51	2.51	2.33	1.94	2.11

Table 1: Subjective MOS values under different test conditions (description see text)

Outlook

The presented objective measurements, if applicable at all, ignore the asymmetry in the assessment of background noise or secondary sounds. Objective measurement indices like PESQ and TOSQA+TASQ were optimized for (galvanic) telecommunication and therefore interprets noise as a quality malfunction of the transmission. Subjects interpret this as an inevitable feature of environment of a driving car and ignore it mostly, as long as speech is intelligible at all. Future automatic systems must compensate for this asymmetry.

References

[1] French, N.R., Steinberg, J.C. (1947). Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Am. 19, 90-119.

[2] Steeneken, H.J.M., Houtgast, T. (1985). A tool for evaluating auditoria. Bruël & Kjaer Technical Review (1985) 13-39.