SNR-based Evaluation of Speech Recognition Products in Real-Life Car Environments

Sascha Hohenner[1], Klaus Lukas[2]

[1] *Siemens AG, Corporate Technology, 81730 Munich, Germany, Email: sascha.hohenner@siemens.com*
[2] *Siemens AG, Corporate Technology,81730 Munich, Germany, Email: lukas@siemens.com*

Abstract

The evaluation of speech recognition systems is an essential issue for the customer acceptance of a product. However, evaluation results of individual customer-specific tests are hardly comparable due to different test setups, while database tests often do not reflect the real-life environment for the final product.

Therefore, we will present a well-defined procedure for an independent evaluation of speech recognizer products in real-life car-environments. The procedure has been exemplified and validated in comprehensive in-car tests by deploying the Siemens speech recognizer VSR Very Smart Recognizer® [1] and allows a comprehensive, objective and comparable assessment of recognizer performances under real-life conditions. Thus it may serve as a basis for future standardized evaluation procedures for automotive speech recognition products.

Common Evaluation Procedures

The common evaluation procedures for speech recognition products can roughly be subdivided into three different approaches: customer-specific tests, database tests, and a hybrid approach of both.

With customer-specific tests the fulfillment of different customer requirements can be verified very well. However, there is nearly no possibility to compare evaluation results from different customer-specific test setups.

With tests on common databases it is possible to produce comparable test results. However, such database tests often do not reflect the real-life environment for the final product: for example, the audio path of the final product is not taken into account, and the recordings were often performed in cars different to the target car. Finally, common databases often do not contain all commands of the final product, and therefore allow to verify only a subset of the overall commands.

Another common approach can be seen as a hybrid of a customer-specific test and common database tests: clean speech from databases is mixed with noise recorded with the final target in the final environment. However, this approach has also some disadvantages. First, as for database tests already stated, often not all commands of the final product are included in a database. And second, the mixing of clean speech and noise is not the same as speech recorded in real noise, as e.g. due to the Lombard-effect speech characteristics often change under noise [2].

Proposed SNR-Approach

In order to overcome the constraints of the previously described approaches, we propose in the following an objective and practical evaluation procedure especially for automotive environments based on SNR (Signal-To-Noise Ratio) values. For our evaluation procedure, recordings are taken on a normalized roundtrip with typical traffic and road situations. For each utterance a specific signal-to-noise ratio is calculated to assign the utterance to an SNR-bin of the main car noise range. In a defined evaluation procedure, the SNR-bins are compiled into normalized SNR recognition curves. This SNR-based approach has the advantage of better comparability in opposition to conventional tests with fixed driving speed (e.g. 0/50/130 km/h), as environmental conditions like weather, tires or road type are implicitly considered. Furthermore an SNR-based approach takes the speaker loudness correctly into account.

Data Recording

To provide a comparable and comprehensive assessment of the recognizer performance, recordings from 12 test speakers (6 male and 6 female from target group) are taken on a normalized roundtrip. All recordings are performed with the final target. The audio signals are recorded as provided to the recognizer engine as well as the recognition results. The normalized roundtrip should contain most of normal traffic situations and road types like town traffic, country roads and highways. To diversify the in-car situation, recordings are taken with opened as well as with closed windows and comprise different settings of the air conditioning. All test speakers get a list of the same test utterances. This list should contain all commands, that have to be tested, and every command should occur in the same quantity (preferably at least five times each).

Evaluation

After the calculation of the signal-to-noise ratio with a well-defined algorithm [3] for every recorded utterance, all recordings are grouped into SNR-bins from 2 to 16 with a step of 2, where every SNR-bin *X* includes all recordings with a signal-to-noise ration higher or equal *X–2* and lower *X+2*. The idea of this grouping is that every utterance within the same SNR-bin has the same "level of challenge" for the recognizer, as the challenge notably depends on the ratio between the intensity of speech signal and environmental noise. Furthermore the grouping summarizes the utterances in few SNR-bins improving the statistical relevance of the particular clusters.

After this grouping, the recognition rate for every SNR-bin is calculated. To avoid a strong influence of very good or very bad recognized speakers on the recognition rate, the best and the worst speaker are removed from all SNR-bins. For this purpose, an individual recognition rate R_n over all SNR-bins S is calculated for every speaker n as follows:

$$R_n = \frac{\sum_{S=\min}^{\max} [R_n(S) - R_{mean}(S)] * utt_n(S)}{\sum_{S=\min}^{\max} utt_n(S)} \quad (1)$$

Where $R_{mean}(S)$ is the mean recognition rate over all speakers for the corresponding SNR S, and $utt_n(S)$ is the number of utterances of speaker n in the SNR-bin S. With this approach the decision on best and worst speaker is based on their relative performance to the other speakers per SNR-bin rather than on an overall recognition rate, which heavily depends on the particular noise conditions per speaker.

After removing all utterances from the best and worst speakers, the final recognition rate for every SNR-bin is calculated, receiving SNR-curves like shown in Fig. 1.

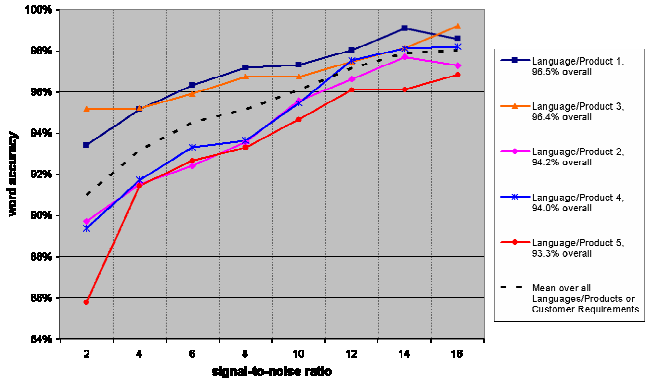

Figure 1: SNR-curves for different languages/products

These SNR-curves as shown in Fig. 1 give a comparative overview about the performance of a recognition product under different aspects. First, the distribution of a SNR-curve shows the characteristics of the recognizer under different noise levels. Second, all recognition curves (e.g. for different languages or products) are directly comparable, as every SNR-curve has been created under the same conditions and subdivided into the SNR-bins by the same algorithm. For example, if two recognizers with a similar performance have been recorded under different weather conditions (e.g. sunny vs. rainy), the overall recognition rate will normally differ due to the different noise levels. However, with our approach, SNR-curve of both recognizers can be compared directly, as every SNR-bin reflects a similar noise level for the corresponding utterances.

The SNR-curves can finally be translated into a normalized SNR-curve as shown in Fig. 2, taking the mean over all curves within every SNR-bin (or e.g. customer requirements) as baseline. With such a normalized representation it is now very easy to compare the recognition performances visually.

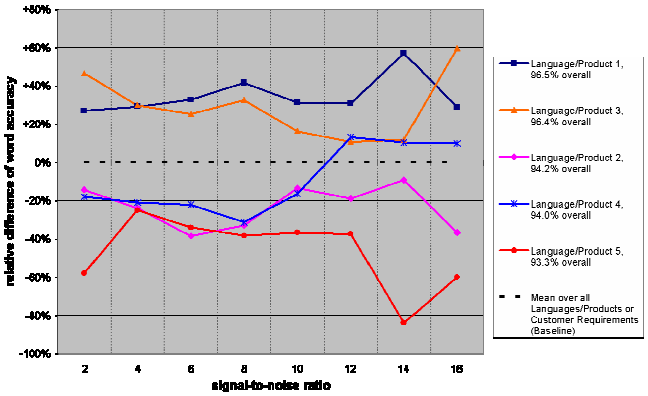

Figure 2: Normalized SNR-curves for different languages/products

Best Practice

We experienced some points to be considered to get a good coverage of the utterances over the whole SNR-range and to avoid unnatural accentuation of the test utterances. First of all, the sequence of the test utterances should be varied a little bit for every speaker to avoid dependences between certain utterances in the sequence and certain traffic situations, as often the same circuit will be driven. Second, the sequence of utterances should not contain a series of same commands. Otherwise the test speakers sometimes start to play with the accentuation of this command. The same applies for unnatural digit-sequences (e.g. like 0102030405). Finally, the round trip should contain as much different driving situations as possible to get a good distribution of the utterances over the whole SNR-range for all speakers and therefore to get a comprehensive assessment of the recognizers performance.

Conclusion

The proposed evaluation procedure has been exhaustively field-tested in comprehensive in-car tests by deploying the Siemens speech recognizer VSR Very Smart Recognizer® in various conditions. With the SNR-based evaluation approach the disadvantages of the conventional evaluation procedures have been accomplished. As it allows a comprehensive, objective and comparable assessment of recognizer performances under real-life conditions, it provides a sound basis for future standardized evaluation procedures for automotive speech recognition products.

References

[1] Varga, S. et al.: ASR in mobile phones – an industrial approach. Transactions on Speech and Signal Processing, Vol. 10, 2002, 562-569

[2] Junqua, J.C.: The Lombard reflex and its role on human listeners and automatic speech recognizers. J. Acoustic. Soc. Amer., Vol. 93, 1993, 510-524

[3] Höge, H. and Andrassy, B.: Human and machine recognition as a function of SNR. LREC 2006 ELRA, Genoa, Italy, 2006, 2060-2063