# Evaluating Speech Recognition Performance in the Car – A Pragmatic Approach

Markus Lieb

*Volkswagen AG, Abt. 1675, D-38436 Wolfsburg, Germany, Email: markus.lieb@volkswagen.de*

## Introduction

During the development of voice-controlled applications, continuous monitoring of the speech recognition performance - from the lab into the final product - is of crucial importance. There is a gap between large-scale off-line automatic speech recognition (ASR) tests, being run in the lab on pre-recorded speech material, and tests with real speech uttered by test persons in the target cars under realistic, e.g. driving, conditions. Both tests are necessary and conclusive, yet not sufficient for objective performance comparisons, such as comparisons of different ASR installations. To evaluate ASR performance in a transferable and repeatable manner for different cars at different points in time without the need to recruit test persons for every new test, a different approach is proposed in this paper. Starting from pre-recorded speech material covering all sorts of variations, such as gender, accent, age, and background noise condition, our objective is a reproducible calibrated playback mechanism in the car. Such controlled playback comprises the ASR performance evaluation in an objective manner and, in particular, allows the comparison of different systems, e.g. in different cars, based on the same speech material. In this paper, we present the experimental set-up for such performance investigations, describe the test procedure, and give insight into results from the field.

## Test Methodologies

Several procedures are available to evaluate voice controlled systems. In the following, we sketch five different categories, describe their use cases, and point out advantages and disadvantages.

### 1. Off-line Testing

When comparing different algorithmic approaches to solve an ASR problem, the way of choice is an off-line evaluation. An automated system simulation is performed using software on fast computers and run over large pre-recorded speech databases. The significance of the results is great due to the typically large amounts of speech data used for such automated algorithm comparisons. Usually, such algorithmic comparisons are practicable only for ASR technology providers having full control over the algorithms and software in use. Further, the many differences between system simulation and a target implementation that influence system performance have to be treated carefully in off-line testing and in interpreting benchmark results. For instance, resource limitations or special timing conditions in a dedicated DSP implementation are crucial points limiting the usefulness of off-line testing for judging real world ASR products.

### 2. User Tests

Opposed to the off-line testing efforts, commonly applied in the scientific context of ASR, is the mean of user tests. The results gained from user tests are used to evaluate the voice-controlled equipment installed in the car operating in real driving conditions. Procedures specify test cases each user has to perform including test sequences to utter, driving speed, window and fan condition etc. Thus, user tests deliver conclusive benchmark results.

The effort required to obtain significant results from user tests is enormous: assembling a large enough number of subjects, the amount of test cases under different driving conditions, and, not least, the size of the application vocabulary to test require large efforts both in time and costs. These efforts are caused by the fact, that the results cannot be reproduced exactly; thus, statistical evaluations are the only basis for conclusive comparable results.

To overcome the effort needed for large-scale user tests, evaluations with pre-recorded speech data as in the off-line testing are required.

### 3. In-situ full-control target system tests

This test scenario is described by simulated sound input into the ASR system while the testing equipment has full control over the recognition process. For example, triggering the recognition process at distinct points in time and checking the recognition results require a control interface, such as a debug API, in order to provide a fully automated test process. A great amount of attention has to be given to the sound input. The common method for preparing the simulated microphone-in data is a two-stage signal processing procedure:

The first step in the process is to mark speech data that has been recorded under quiet-room conditions with no background noise or reverberant room impulse influence, i.e. with close-talk microphone. Since the human speaking style greatly depends on the environmental conditions, the so-called Lombard effect, such recordings are typically conducted while the talkers are presented via headphones with vehicle noise recordings categorised into different noise levels. Thus, sound databases consisting of Lombard speech and labelled according to their signal-to-noise-ratios (SNRs) form the car-independent basis for the actual evaluation.

The first stage for simulating the car-specific environment in the microphone-in signal is to convolute the recordings using an impulse response function reflecting the car-cabin conditions of the car under investigation. Obvious limitations of this approach are that variations of the talker position (e.g. talkers of different sizes) had to be simulated using different response filters, while effects of body movements during an utterance cannot be investigated at all.

The second stage of the simulation is an additive feeding with appropriate background noise. Here, combining the SNR-labelled speech recordings with the corresponding noise files is crucial.

Thus, given Lombard speech databases, in order to conduct tests in a new car, two things have to be prepared: the impulse response function(s) as well as speed dependent background-noise recordings.

Given these prerequisites together with a fully controlled target ASR system being tested, an automated test procedure makes reliable performance measurements on large databases possible.

## 4. Target system tests with controlled sound-in

Full control access to devices being tested is only possible in rare situations. Particularly, it is not possible when it comes to benchmarking among different suppliers' prototype devices or among different OEMs' product solutions.

Thus, the former fully automated testing procedure is often stripped down to a controlled sound-in test, where the microphone-in signal is generated in the same way as described above, yet the testing process itself is conducted through human interaction. Operating the voice-controlled device, starting the sound feeding, and taking note of the device-recognition results become manual tasks. Due to the manual interaction, exact reproducibility and repeatability are inherently not assured by this approach. An especially delicate task is activating the speech file at an appropriate time after a corresponding system prompt, while the background noise should be fed independently from such activation of the speech signal.

## 5. Target system tests with simulated talker

The controlled sound-in approaches outlined earlier have the drawback of requiring impulse response function(s) specific to a given car as well as noise recordings specific to the car and microphone or to the microphone array. Furthermore, physical microphone-in channel access is required for the system being tested.

For evaluation situations where these requirements are not feasible or not available, a simulated talker approach is proposed. The microphone-in signal is not produced by the signal processing scheme of car cabin filtering and noise addition but by an artificial talker. Mobile playback equipment comprising an ITU-specified [3] artificial head with mouth simulation allows realistic sound propagation of the speech signals inside the car cabin. The background noise condition can be achieved by two means: The first possibility is to use playback equipment separately from the artificial talker. This equipment is fed with background noise files recorded in the given car with high quality recording tools, not via car built in microphone(s), as with the above approaches. To assure reliable and comparable performance measurements, a correctly calibrated sound level of the artificial talker and of the noise playback is required. The requirements are similar to the measurement specifications fixed in the VDA standard for hands-free telephony systems [1].

The other possibility offered by an artificial talker approach is an evaluation under real driving conditions. With the restriction that the head- and torso-simulator can only be seated on the co-driver's position during such driving, this approach doesn't require recording car-specific background noise files at all. PC-based software helps our test personal operate the test equipment as well as  collect the reactions and to generate test reports.

Manual operation of the voice-controlled system being tested as well as manual interaction to start the artificial talker's test utterances lead to the same inexact reproducibility of the experiments described before in the controlled sound-in approach. If real driving replaces the car-specific noise files that were recorded and played back, inexact reproducibility becomes an even greater issue .

However, the major advantage of this simulated talker approach is the mobility, the quick transfer of the measurement tools from one car to another, and its applicability from one device to another with little effort of preparation.

An even more simplified approach did not deliver  sufficient results: Playback equipment without artificial mouth characteristics was used to play back both noise and speech sound recordings together or even noisy speech sound files. This oversimplified measurement approach, interesting from an effort point of view, has not led to reliable assessments and has been replaced by the artificial talker approach with a separate source of background noise.

# Conclusion

The range of different approaches for evaluating recognition performance of in-car voice-controlled devices is broad. Several procedures differing in requirements, flexibility, and effort have been collected and categorized. A set-up with artificial talker, separate background noise playback, and manual test operation is a good compromise for a flexible, quick, yet still conclusive evaluation.

Thanks to the companies Scansoft Inc., TEMIC SDS GmbH and Headacoustics GmbH for their support to.

## References

[1] "VDA specification for hands-free terminal testing"

[2] NIST "Speech recognition scoring package (SCORE)", www.nist.gov/speech/tools/

[3] ITU Recommendation P.581, "Use of head and torso simulator (HATS) for hands-free terminal testing"