ECHOCOMPENSATION AND NOISE SUPPRESSION FOR SPEECH RECOGNITION APPLICATIONS

Dr. Walter Stammler, Matthias Schulz and Frank Scheppach
Daimler-Benz Aerospace AG
Wörthstr. 85
D-89077 Ulm, Germany

Summary

This contribution deals with the role and the performance of echocompensation and noise suppression, when used in combination with speech recognition systems. For two applications of interest (speech control in car or via telephone) there are quite significant differences to classical echocompensation and noise suppression for telephone conferences. It will be pointed out, how the systems are structured, what performance can be achieved and how realtime solutions are looking like.

Introduction

Daimler Benz Aerospace (DASA) and Daimler Benz Research have developed various speech recognition systems for "command and control applications", such as

- interactive voice response (IVR) for access of information - or ordering systems via public telephone
- voice control of audio, radio telephone, air conditioning etc. in a vehicle.

The recognition algorithms are characterized as phoneme based, speaker-independent, connected word input with typical vocabularies of less than a few hundred words. The systems are aiming at high performance under adverse conditions. In these systems echocompensation as well as noise suppression algorithms contribute to robustness, intuitive handling, comfort and ergonomy.

Interactive voice response with echocompensation

In the case of IVR, echo compensation takes care of line echos ocurring at two wire/four wire transition points (fig.1). These line echos are resulting from speech output of the dialog system (SODS) and are observable at the speech input of the recognizer. Near-end echoes have their origin in hybrids of the board that connects the speech dialog system (SDS) with the analog telephone network. Far end echos are ocurring at the user`s telephone. In the case of digital coupling (e.g. in the ISDN) near end echos are not relevant, whereas far echos from the digital-analog transition can return to the SODS nearly without attenuation. All these line echos cited are causing erroneous recognition results or superponing actual voice commands of the user. In the past, the recognizer was not in operation as long as the SODS was active. Thus, however, the user could not interrupt long explanations of the dialog system ("barge in") or give an immediate answer / command ("talk thru"). Here the "talk thru" solution will be considered only, since it reduces dialog duration and cuts down telephone cost for the user.

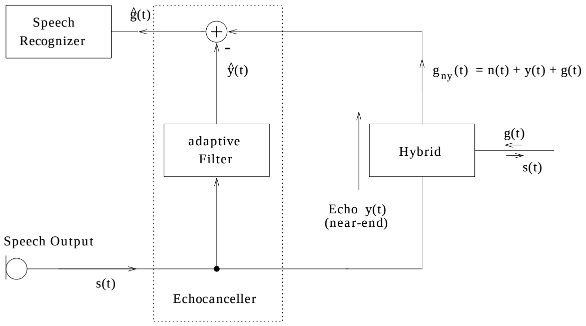

Fig. 1. Speech Dialog System

To get an overview of typical channels to be adapted, several telephone connections were measured. Fig. 2 shows two examples of echo-path impuls responses. With dominant near-end echos, the delay of the far-end echos corresponds to the growing distance between the user at the far-end and the speech dialog system (SDS). The maximum time delay requires a FIR-compensation filter with N =256 taps for a sampling period of T=125µs.

For compensation of line echos, the goal is to achieve a high echo return loss enhancement (ERLE) and a fast filter adjustment to allow a "talk thru" right at the beginning of the SODS.

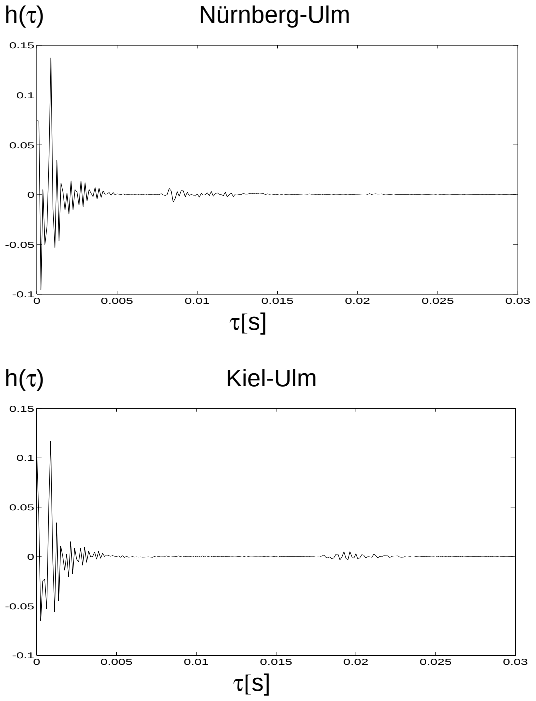

Fig. 2. Electrical echo path impulse responses of two telephone far-connections

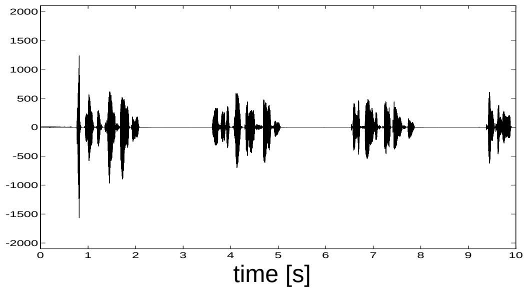

Fig. 3. The user's speech input

In practical applications, four to eight telephone channels (users) need to be handled by one floating point signalprocessor, employing echo compensation as well as speech recognition. To guarantee high computational efficiency, a stochastic gradient algorithm realized in the frequency domain was implemented by using FFT processing. Its weight vector update equation is

$$\underline{W}_{i+1} = \underline{W}_i + 2\underline{\mu}_i \otimes \underline{S}_i^* \otimes \underline{E}_i \qquad (1)$$

where $\underline{\mu}_i$ denotes the frequency dependant adaption step size multiplicated element-by-element (operator $\otimes$) with $\underline{S}_i^*$ denoting the complex conjugate transformed SODS-block of N samples and $\underline{E}_i$ is given by the transformed error signal-block $\underline{g}_{ny}(i) - \hat{\underline{y}}(i)$ of the i - th block.

The user's speech input to simulate the echo-canceller under real conditions of double talk is shown in fig. 3.

To observe the current system identification, the learnig curve D [dB] is calculated as

$$D(i) = 10 \log \frac{\left\| \underline{h}(i) - \underline{w}(i) \right\|^2}{\left\| \underline{h}(i) \right\|^2}, \mathrm{time} = iNT \qquad (2)$$

where $\underline{h}(i)$ describes the impulse response of the unknown channel and $\underline{w}(i)$ is the coefficient vector adjusted, the double bars denoting the Euclidean norm. Stability and robustness of the canceller under worst-case conditions depend on the filter adjustment control with consideration of the nonstationary and spectral caracteristics of speech signals. This problem is handled by an adaption step size driven by a combination of coherence estimation at the speech input and a multistage double talk detector depicted in fig. 4. Delayed flanks at the end of double talk g(t) make the canceller more insensitive towards disturbance n(t) in these critical moments.

To increase convergence, the frequency specific step-size factors are normalized to their averaged spectral density. To ensure filter stability, only the spectral factors, that climbed over a minimum power level of the reference signal (SODS) are used. In times of SODS-power under this level, the coefficient adjustment is frozen. This is also illustrated in form of the learning curve in fig. 4 where the plateaus of D(i) represent the timeduration of speech output energy not exceeding the minimum power level.

The energy contour of the returned echo in this example is plotted in fig. 5.

The achieved total ERLE, estimated by

$$\mathrm{ERLE}(i) = 10 \log \frac{\sum_{j=0}^{N-1} g_{ny}^{2}(j)}{\sum_{j=0}^{N-1} \left(g_{ny}(j) - \hat{y}(j) \right)^2}, \mathrm{time} = iNT \qquad (3)$$

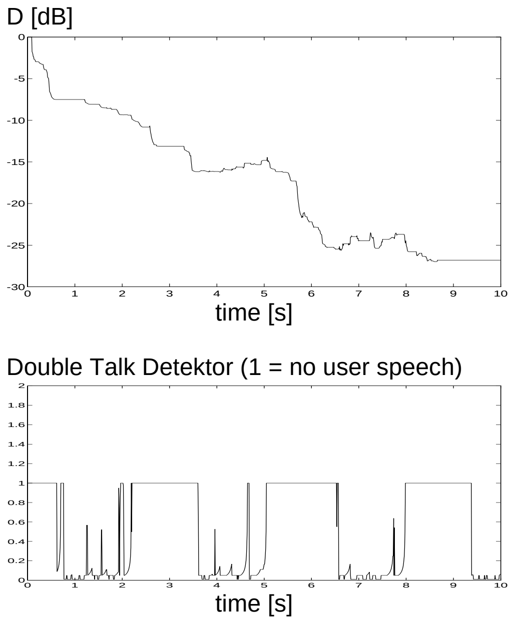

Fig. 4. Learning curve D(t) and
Multistage Double Talk Detection

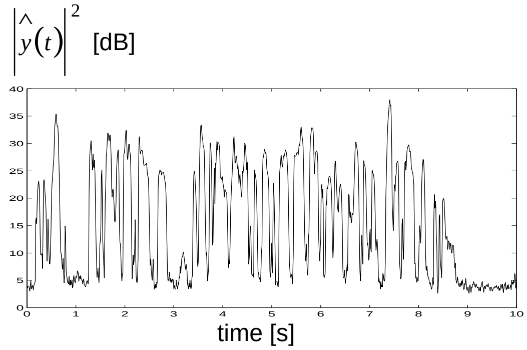

Fig. 5. Energy contour of echos occuring at the input
of the SDS

for each block during the active SODS depicts fig. 6. It is obvious that in times of double talk, the ERLE decreases because of the user's speech contributing to the echo signal as well as the disturbance n(t) always existing at the input of the SODS.

It should be mentioned that in the application of speech recognition, the ERLE must be adequate over all frequency bands to prevent erroneous feature extraction based on filterbank analysis. Another specificum of this application is that the echo path hardly changes its characteristics during a telephone conversation. Thus the weights may be frozen, when the ERLE has reached its maximum.

Noise suppression in this application primarily copes with ambient noise from the far-end user. In most cases though explicit noise reduction during the preprocessing phase can be abandoned, if the training of the recognizer is based on noise corrupted data registered on actual telephone channels.

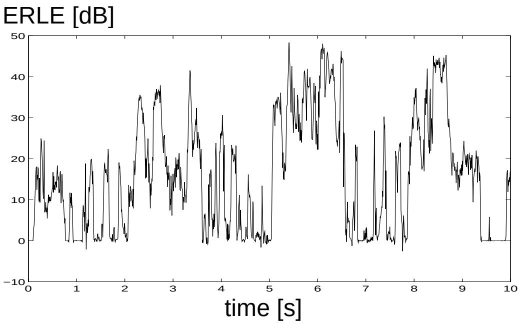

Fig. 6. Averaged echo return loss enhancement
during active SODS and Double Talk

Speech dialog system in the vehicle

The second example of speech recognition combined with echocompensation and noise suppression is the automotive application. Voice control of car telephones, audio and navigation equipment provides more comfort as well as more safety (permitting the driver to keep his eyes on the road and his hands on the steering wheel) and in the long run it will help reducing the immense number of switches, displays, knobs etc. Noise reduction is essential for the recognizer to cope with a variety of disturbances reaching from wheel-noise, wind-noise, engine-sound to noise of blinkers, rain drops and wind-shield wipers. Echo compensation permits to talk to the recognizer, even though the radio/CD/Cassette is active. The electrical input of the loudspeaker serves as the input to the echo compensator. In case of stereo signals two echocompensators have to be employed. A complete structure of the DASA-automotive speech dialog system is shown in figures 7 and 8.

Typical echos in a large vehicle reach a length of 10... 40 msec. Measurements of Dr.Linhard from Daimler Benz Research indicate, that stereo-echocompensators, based on FLMS or LMS-algorithms achieve a signal to noise ratio enhancement of approximately 20dB under realistic test conditions in a vehicle. With increasing speed, this value will

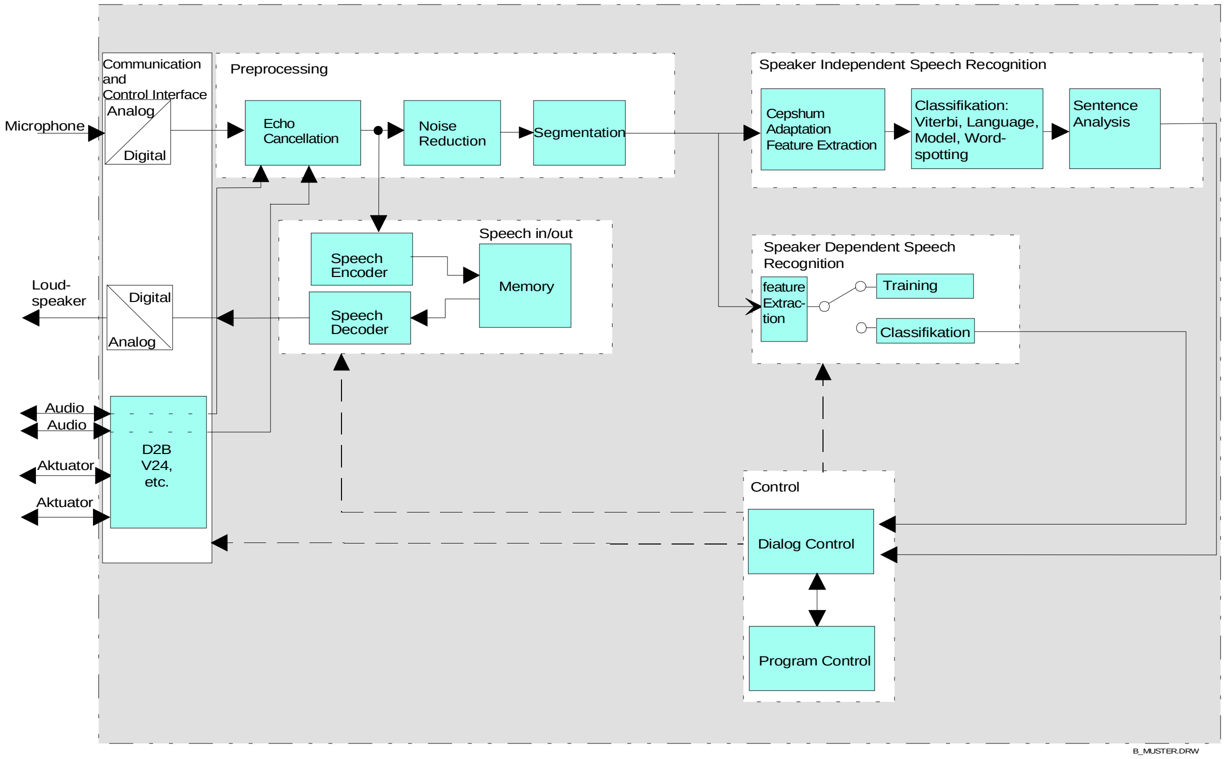

Figure 7: Structure of speech dialog system for auto motive application

even be lower. Listening tests have shown, that in the case of stereo radio the rest echo of a music signal remains audible, but due to ambient noise and subsequent noise reduction it hardly deteriorates recognition results.

For noise reduction a spectral subtraction method is applied together with a nonlinear filter to reduce musical tones. Optimizations of the recognition and noise reduction process finally lead to approximately identical recognition results up to 100 km/h.

The structure realized for a vehicle application is shown in fig. 7. Besides echo-compensation and noise control a combination of speakerindependent recognizer (for command- or number sequences) and speakerdependent recognizer (for individual names of a telephone list), of speech codec and dialog- as well as interface-software is implemented on a commercially available signal processor. Processing power required ranges from 15 MIPs to 50 MIPs depending on the size of vocabulary (up to 300 words) and the necessity of echocompensation . The processor provides 24Bit fixed point arithmetic, the program code consists of approximately 40k LoC.

References:

Linhard,K.:Frequenzbereichsverfahren zur Echokompensation bei Störgeräuschen und Gegensprechen, 8.Aachener Kolloquium „Signaltheorie- Mobile Kommunikation“, März 1994

Scheppach,F: Untersuchungen zur Unterdrückung der dominanten Störungen durch Echos..., Diploma-Thesis, University of Ulm, Department of Informationtechniques, 1995

Structure of D2B-optical System

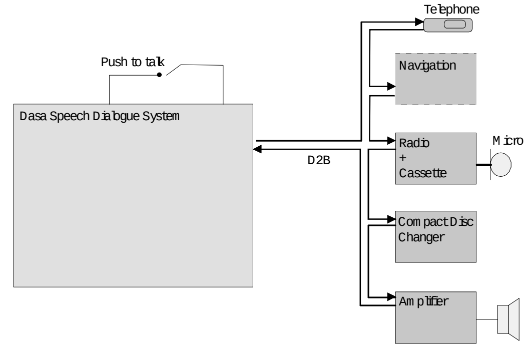

Fig. 8: Standard bus combination of speech dialog system and automotive components.