Marzena MIĘSIKOWSKA, Evert de RUITER
KIELCE UNIVERSITY OF TECHNOLOGY, ALEJA TYSIĄCLECIA PAŃSTWA POLSKIEGO 7, 24-314 KIELCE, POLAND
PEUTZ ZOETERMEER, POSTBUS 696 2700 AR ZOETERMEER, THE NETHERLANDS

Automatic recognition of voice commands in a car cabin

Dr inż. Marzena MIĘSIKOWSKA

Assistant Professor in the Department of Mechanics, Faculty of Mechatronics and Machine Design, Kielce University of Technology, Poland. Research interests: digital signal processing.

e-mail: marzena@tu.kielce.pl

Dr Evert de RUITER

Part-time consultant at Peutz Zoetermeer, The Netherlands. Research interests: acoustics.

e-mail: e.deruiter@peutz.nl

Abstract

Automatic speech recognition systems are applied in vehicles. It is possible to control a navigation system, an air conditioning system, a media player, and make phone calls using voice commands. The effectiveness of speech recognition systems depends largely on the acoustic conditions in the cabin of a vehicle. In contrast, the recognition accuracy, determines the ability to extend the functionality of the application of speech recognition systems, not only to the basic functions listed above, but also to control the systems that affect the movement of the vehicle. The work shows the preliminary results of research on speech recognition and evaluation of speech intelligibility in the cabin of the vehicle in the presence of noise barriers. These results may be helpful in assessing the speech intelligibility and the results of automatic speech recognition systems in the cabin of the vehicle.

Keywords: in-car speech recognition, acoustics in car cabin, speech intelligibility.

Automatyczne rozpoznawanie komend głosowych w kabinie pojazdu

Streszczenie

Systemy automatycznego rozpoznawania mowy są aplikowane w pojazdach. Za pomocą komend głosowych możemy sterować nawigacją, systemem klimatyzacji, odtwarzaczem multimediów, oraz wykonywać połączenia telefoniczne. Skuteczność systemów rozpoznawania mowy zależna jest w dużej mierze od warunków akustycznych panujących w kabinie pojazdu. Natomiast dokładność rozpoznawania, warunkuje możliwość rozszerzenia funkcjonalności stosowania systemów rozpoznawania mowy nie tylko do podstawowych funkcji wymienionych wyżej, ale także do sterowania układami mającymi wpływ na poruszanie się pojazdu. Praca pokazuje wstępne wyniki badań w zakresie rozpoznawania mowy oraz oceny zrozumiałości mowy w kabinie pojazdu w obecności ekranów akustycznych. Wyniki badań mogą okazać się pomocne w ocenie zrozumiałości mowy i rezultatów automatycznego rozpoznawania mowy w kabinie pojazdu.

Słowa kluczowe: automatyczne rozpoznawanie mowy w kabinie pojazdu, warunki akustyczne w kabinie pojazdu, zrozumiałość mowy.

1. Introduction

Automatic Speech Recognition (ASR) systems are applied in vehicles. It is possible to control a navigation system, an air conditioning system, a media player, and make phone calls using voice commands. The effectiveness of speech recognition systems depends largely on the acoustic conditions in the cabin of a vehicle, especially background levels for its influence on speech intelligibility.

The car interior noise is still problematic and impacts the recognition rates. There have been proposed many solutions to solve the problem of background interior noise. The ASR performance degrades substantially when speech is corrupted by background noise not seen during training. The reason for this is that the observed speech signal does no longer match the distributions derived from the training material. This mismatch between the training and testing conditions is one of the most challenging and important problems in ASR [1, 2]. Many solutions have been proposed to improve the in-car recognition accuracy. The first approach is focused on parameterization methods that are fundamentally resistant to noise or minimize the effect of the noise. The second approach is based on noise reduction by transforming noisy speech into clean speech - the noise is removed/reduced from the representation of speech. The third approach includes methods that are based on adoption of clean models to the noisy recognition environment in order to contaminate the models. The authors of study [3] applied lip detection for audio-visual automatic speech recognition (AVASR) in order to overcome the poor robustness and effectiveness of voice recognition systems in a car environment. Because the implementation of AVASR required algorithms to accurately locate and tract the drivers face and lip area in real-time, it was shown that using the AVICAR in-car database [4], the Viola-Jones approach could be a suitable method.

Assessment of speech intelligibility allows predicting speech communication under specific conditions. The International Standard specifies the requirements for the performance of speech communication for verbal alert and danger signals, information messages, and speech communication in general [5]. One of the parameters defining speech intelligibility is SIL (en. speech interference level). The SIL offers a method to predict and assess speech intelligibility in cases of direct communication. It is interesting how speech intelligibility rating can refer to communication human-machine with the applied ASR system.

The aim of this work was to present the influence of background levels in a car cabin on speech intelligibility and ASR results and the relation between ASR results and speech intelligibility ratings.

2. Methods

Measurements of background levels were made on June 1, 2014 between 2 pm and 3 pm and on October 6, 2013 between 3 am and 4 am with a Norsonic 140 sound analyzer. The measurements were taken in three measurement variants:

A. with other traffic noise present (traffic) and no noise barriers on both sides of the express road (no screens) (June 1);
B. with other traffic noise present (traffic) and noise barriers on both sides of the express road (screens) (June 1);
C. no other traffic noise present (no traffic) and noise barriers on both sides of the express road (screens) (October 6). Mostly, the noise barriers were a type of green walls barriers. In each variant four measurements were taken in the following order:
    1. LR closed - both windows closed from the driver side and the passenger side;
    2. R open - window from the passenger side open (right open), window from the driver side closed (left closed);

3. LR open - both windows open;
4. L open - window from the passenger side closed (right closed), window from the driver side open (left open).

For example, the A1 symbol means the measurement variant with other traffic noise present (traffic), no noise barriers on both sides of the express road (no screens) and both windows closed (LR closed) during measurements.

The measurements were taken in a hatchback car with three doors. The Norsonic 140 sound analyzer was situated on the passenger side during measurements. The car was moving with 50 km/h.

Speech recordings were taken with a digital OLYMPUS LS-11 recorder in the same variants explained above, after the measurements taken with Nor140 sound analyzer. The recordings consisted of four speech commands: stop, close, open, play. Speech commands were recorded with 44 kHz sampling rate and 16-bit signal resolution.

Speech Interference Level (SIL) parameter was used in this work to predict and assess the speech intelligibility in cases of direct communication [5]. The listener here is the ASR system that is listening to the voice commands of the speaker – the driver. The speech interference level of noise (L_{SIL}) was calculated as the arithmetic mean of the sound-pressure levels in four octave bands with central frequencies 500 Hz, 1 kHz, 2 kHz, 4 kHz. The speech level ($L_{S,A,L}$) was calculated according to vocal effort normal/raised: 60 dB/66 dB and distance to listener: 1 m/2 m [5]. The SIL is given by the difference between $L_{S,A,L}$ and L_{SIL}.

ASR system based on Mel-frequency cepstral coefficients (MFCC) and Hidden Markov Models (HMMs) was used in this experiment to recognize the speech commands recorded by OLYMPUS LS-11 in each measurement variant.

3. Results

In Fig.1 1/3-octave band levels for LR closed in each measurement variant are presented.

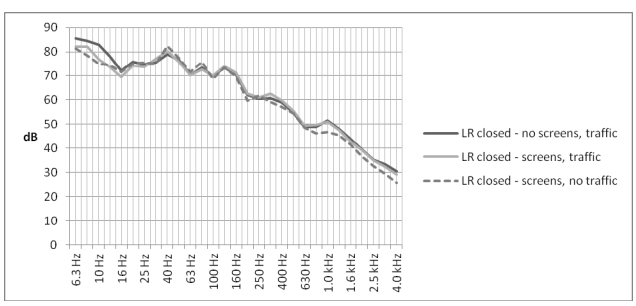

Fig. 1. Background levels in the car cabin for LR closed in three measurement variants
Rys. 1. Tło akustyczne w kabinie pojazdu dla okien zamkniętych w trzech wariantach pomiaru

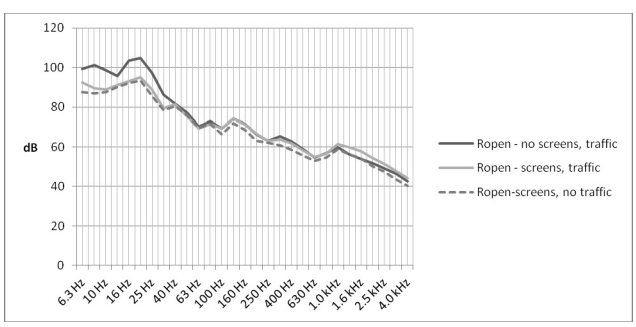

Fig. 2. Background levels in the car cabin for R open in three measurement variants
Rys. 2. Tło akustyczne w kabinie pojazdu dla prawego okna otwartego w trzech wariantach pomiaru

The background level with both windows closed is not dependent on the presence of other traffic. Probably it is mainly the vehicle noise.

In Fig. 2 1/3-octave band levels for R open in each measurement variant are shown.

The right window open shows a little difference with/without other traffic - so this is mainly the influence of the vehicle; the contribution of reflection via a noise barrier is not clear.

In Fig. 3 1/3-octave band levels for LR open in each measurement variant are presented.

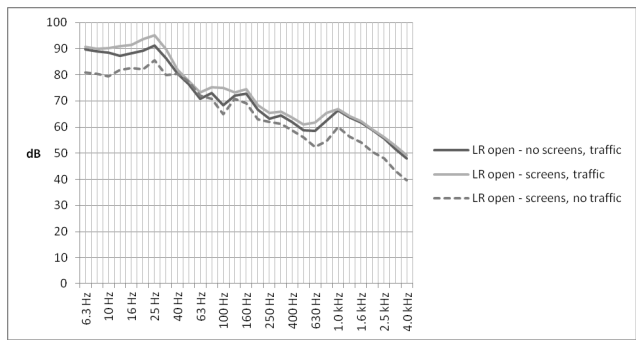

Fig. 3. Background levels in the car cabin for LR open in three measurement variants
Rys. 3. Tło akustyczne w kabinie pojazdu dla okien zamkniętych w trzech wariantach pomiaru

The background level for both windows open is more dependent on the presence of other traffic.

In Fig. 4 1/3-octave band levels for L open in each measurement variant are shown.

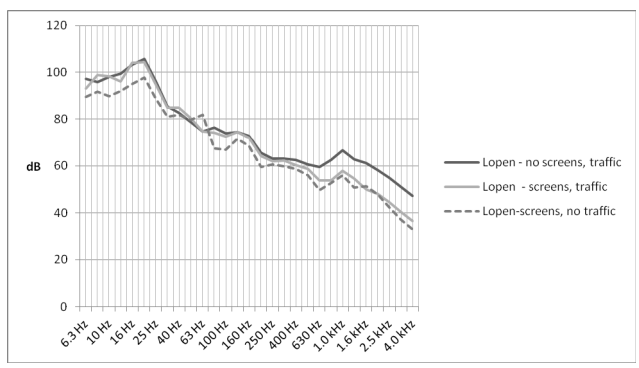

Fig. 4. Background levels in the car cabin for L open in three measurement variants
Rys. 4. Tło akustyczne w kabinie pojazdu dla lewego okna otwartego w trzech wariantach pomiaru

He left window open shows the influence of other traffic; probably a little influence of reflections.

Table 1 presents the A-weighted sound level and the SIL calculated for each measurement variant.

The worst intelligibility rating was obtained for the variant with traffic and no screens. The traffic determined the intelligibility rating. For both windows closed, the intelligibility rating is Good. For R open, the intelligibility rating is Poor. For both windows open, the intelligibility rating was Bad, except when there was no traffic (Poor). For L open, the intelligibility rating was strongly influenced by the presence of other traffic, from Bad, Poor, Fair. The best intelligibility ratings were obtained for the last variant – screens and no traffic.

Tab. 1. A-weighted sound level and the SIL for each measurement variant $L_{S,A,L}$=60 dB
Tab. 1. Poziom dźwięku A i parametr SIL dla każdego wariantu pomiaru $L_{S,A,L}$=60 dB

Variant	A-weighted sound level dB(A)	L_{SIL} dB	SIL dB	Intelligibility rating
NO SCREENS, TRAFFIC				
A1	64,3	44,1	16,0	Good
A2	68,6	53,2	6,8	Poor
A3	72,1	58	2,0	Bad
A4	72,3	58,2	1,8	Bad
SCREENS, TRAFFIC				
B1	64,9	43,7	16,3	Good
B2	69,1	54,6	5,5	Poor
B3	73,4	59,1	0,9	Bad
B4	67,4	50,3	9,7	Poor
SCREENS, NO TRAFFIC				
C1	63,7	40,9	19,2	Good
C2	66,3	51,3	8,8	Poor
C3	66,5	51,3	8,8	Poor
C4	65,1	48,2	11,8	Fair

Table 2 presents the ASR results of four commands stop, close, open, play for every measurement variant.

Tab. 2. Speech recognition results for three measurement variants
Tab. 2. Wyniki rozpoznawania mowy wyznaczone dla trzech wariantów pomiaru

COMMAND R-recognized, NR – not recognized				
Variant	**STOP**	**CLOSE**	**OPEN**	**PLAY**
NO SCREENS, TRAFFIC – 31%				
A1	NR	**R**	NR	NR
A2	**R**	NR	**R**	NR
A3	**R**	NR	NR	NR
A4	NR	NR	NR	**R**
SCREENS, TRAFFIC – 18%				
B1	NR	NR	NR	**R**
B2	**R**	NR	NR	NR
B3	NR	NR	NR	NR
B4	**R**	NR	NR	NR
SCREENS, NO TRAFFIC – 25%				
C1	**R**	NR	NR	**R**
C2	NR	NR	NR	**R**
C3	**R**	NR	NR	NR
C4	NR	NR	NR	NR

The recognition results for the variant with other traffic noise present and no screens on both sides of the express road are better than those for other variants – 5 recognized commands on 16 (31%). For the second variant – screens and traffic present – 3 recognized commands on 16 (18%). For the last variant – screens and no traffic present – 4 recognized commands on 16 (25%). The recognition process is less efficient that 31%. In the conditions, where the speech intelligibility rating was Good – for both windows closed, the ASR system resulted for most cases in one recognized command. In the conditions, where the speech intelligibility rating was Bad or Fair, the ASR system resulted in one or none recognized command. For Poor intelligibility rating, the ASR system resulted in one/two recognized command(s).

4. Conclusions

The background levels presented in this work were dependent on the presence of other traffic – the more traffic, the highest sound levels. The A-weighted sound levels were between 63,7 dB(A) and 73,4 dB(A), and changeable in the measurement variant.

The background levels influenced the speech intelligibility ratings. The speech intelligibility ratings were accordingly: Good for LR closed; Fair, Poor, Bad for L open; Poor for R open; Poor, Bad for LR open.

The speech intelligibility ratings were consistent with the ASR results for Bad and Poor ratings - none/one/two recognized commands. For Good ratings of speech intelligibility, the ASR results were opposite to good results and were rather fair – the recognition results less than 50% - one or two recognized commands on four expressed commands. For Fair intelligibility rating, the ASR results were none. The speech intelligibility ratings were not consistent with the ASR results for Good and Fair intelligibility ratings in this experiment. The ASR system resulted in low recognition rates, especially in the presence of screens. The speech intelligibility is better when the screens are present, but the speech recognition in such conditions resulted in bad rates in this experiment.

The future work will include investigation on how well speech intelligibility ratings correlates with ASR results, when the listener is the ASR system and the investigation on how well discriminant function analysis deals as a robust classifier in the speech recognition process.

5. References

[1] Gong Y.: Speech recognition in noisy environments: a survey. Speech Communication, 16 (3), pp. 261–291, 1995.
[2] Cavalcante B.A, Schinoda K., Furui S.: Robust Speech Recognition in the Car Environment. LTC 2009, LNAI 6562, pp. 24–34, 2011.
[3] Navarathna R, Lucey P, Dean D, Fookes C, Sridharan S. Lip detection for audio-visual speech recognition in-car environment, 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010).
[4] Lee B., Hasegawa-Johnson M., Goudeseune C., Kamdar S., Borys S., Liu M. and Huang T.: Avicar: Audio-visual speech corpus in a car environment. In Proc. Interspeech 2004, Jeju Island, Korea.
[5] ISO/IEC 9921 – Assessment of Speech Intelligibility

otrzymano / received: 15.05.2014
przyjęto do druku / accepted: 01.07.2014 *artykuł recenzowany / revised paper*