Marzena MIĘSIKOWSKA

Kielce University of Technology

doi:10.15199/48.2015.10.35

Discriminant analysis of voice commands in a car cabin

Abstract. *Automatic speech recognition systems are used in vehicles. With this application it is possible to control the navigation system, air conditioning system, media player, and make phone calls by using voice commands. The effectiveness of speech recognition systems depends largely on the acoustic conditions in the cabin of the vehicle. Recognition accuracy determines the ability to extend the functionality of such systems beyond the basic functions listed above. The article shows the preliminary results of research on speech recognition and evaluation of speech intelligibility in the vehicle cabin. The purpose of this article is to present the influence of the background noise levels in a car cabin on speech intelligibility, and to investigate the discriminant analysis as a robust classifier for the speech recognition process.*

Streszczenie. *Automatyczne systemy rozpoznawania mowy są stosowane w pojazdach. Dzięki tej aplikacji możliwe jest sterowanie systemem nawigacji, klimatyzacją, odtwarzaczem multimedialnym i wykonywanie połączeń telefonicznych za pomocą poleceń głosowych. Skuteczność systemów rozpoznawania mowy zależy w dużej mierze od warunków akustycznych w kabinie pojazdu. Dokładność rozpoznawania określa zdolność do rozszerzenia funkcjonalności takich systemów poza podstawowe funkcje wymienione powyżej. W pracy przedstawiono wstępne wyniki badań nad rozpoznawaniem mowy i oceną zrozumiałości mowy w kabinie pojazdu. Celem pracy było przedstawienie wpływu poziomu tła w kabinie samochodu na zrozumiałość mowy i zbadanie analizy dyskryminacyjnej jako klasyfikatora w procesie rozpoznawania mowy. (**Analiza dyskryminacyjna komend głosowych w kabinie pojazdu**).*

Keywords: discriminant analysis, in-car speech recognition, acoustics in car cabin, speech intelligibility.
Słowa kluczowe: analiza dyskryminacyjna, rozpoznawanie mowy, warunki akustyczne w kabinie pojazdu, zrozumiałość mowy.

Introduction

Automatic Speech Recognition (ASR) systems applied in vehicles allow one to control the navigation system, air conditioning system, media player, and make phone calls by using voice commands. The effectiveness of ASR systems depends largely on the acoustic conditions in the cabin of the vehicle, especially on the background noise levels for their influence on speech intelligibility.

The car interior noise level is still problematic and impacts the recognition rates. Many solutions have been proposed to resolve this problem. The ASR performance degrades substantially when a speech is corrupted by the background noise not present during training. The reason for this is that the observed speech signal no longer matches the distributions derived from the training material. This mismatch between 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 a noisy speech into a clean speech - the noise is removed or reduced from the representation of the 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. 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 can be used as a suitable method.

Assessment of speech intelligibility allows to predict speech communication in 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) which offers a method to predict and assess speech intelligibility in cases of direct communication.

Speech intelligibility ratings and speech recognition in a cabin of the vehicle were investigated in study [6]. 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 [6].

The aim of this work was to present the influence of the background noise levels in a car cabin on speech intelligibility and to investigate the discriminant analysis as a robust classifier for the speech recognition process in the cabin of the vehicle.

Methods

Measurements of background levels were made with a Norsonic (Nor) 140 sound analyzer. Measurements were taken in measurement conditions presented in Table 1.

Table 1. Conditions of measurement

Condition	Description
LABORATORY ROOM	Measurements taken in a laboratory room testing conditions.
NO SCREENS, TRAFFIC	Other traffic noise present (traffic) and no noise barriers on both sides of the express road (no screens).
SCREENS, TRAFFIC	Other traffic noise present (traffic) and noise barriers on both sides of the express road (screens).
SCREENS, NO TRAFFIC	No other traffic noise present (no traffic) and noise barriers on both sides of the express road (screens).

Measurements were taken in a hatchback car with three doors and in a laboratory room testing conditions. During measurements in the car, the Nor 140 sound analyzer was situated on the passenger side. The car was moving at 50 km/h. Car windows were in one of the following positions during measurements presented in Table 2.

Table 2. Positions of car windows

Position	Description
LR closed	Both windows (L-Left, R-right) closed from the driver side and the passenger side.
R open	Window from the passenger side (R-Right) open, window from the driver side (L-Left) closed.
LR open	Both windows (L-Left, R-Right) open from the drier side and the passenger side.
L open	Window from the passenger side (R-Right) closed, window from the driver side (L-Left) open.

Recordings were collected in the following variants presented in Table 3.

Table 3. Variants of measurement (in-car conditions)

Conditions	Car-window positions
NO SCREENS, TRAFFIC	LR closed, R open, LR open, L open
SCREENS, TRAFFIC	LR closed, R open, LR open, L open
SCREENS, NO TRAFFIC	LR closed, R open, LR open, L open

As presented in Table 3, in three different conditions, four options of measurement were taken with different positions of car windows.

Speech recordings were collected with OLYMPUS LS-11 digital recorder in the same variants explained above, after the measurements taken with Nor 140 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 (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}.

Discriminant analysis of speech commands recorded with OLYMPUS LS-11 in each measurement variant was based on Mel-frequency cepstral coefficients (MFCC). Discriminant analysis was performed in STATISTICA Software [7]. Discriminant analysis included discrimination stage and classification stage [8]. In this study, discriminant analysis was based on 12 MFCC features as independent variables and speech commands as grouping variable. After determining variables that discriminate speech commands occurring groups, the classification stage was applied into analysis. Due to four speech command groups (stop, close, open, play), four classification functions were created according to the following formula:

$$(1)\quad K_i(v) = c_{i0} + w_{i1} mfcc_1 + w_{i2} mfcc_2 + \ldots + w_{i12} mfcc_{12}$$

where: the v - command group /stop, close, open, play/, the subscript i denotes the respective group; c_{i0} is a constant for the i'th group, w_{ij} is the weight for the j'th variable in the computation of the classification score for the i'th group; $mfcc_j$ is the observed mel-cepstral value for the respective case.

Results

Table 4 shows the A-weighted sound level and the SIL calculated for each measurement variant [6].

Table 4. A-weighted sound level and the SIL for every measurement variant $L_{S,A,L}$=60 dB [6]

Variant	A-weighted sound level [dB(A)]	L_{SIL} [dB]	SIL [dB]	Intelligibility rating
LABORATORY ROOM TESTING CONDITIONS				
L0	26.7	14.9	45.2	Excellent
NO SCREENS, TRAFFIC				
LR closed	64.3	44.1	15.9	Good
R open	68.6	53.2	6.8	Poor
LR open	72.1	58	2.0	Bad
L open	72.3	58.2	1.8	Bad
SCREENS, TRAFFIC				
LR closed	64.9	43.7	16.3	Good
R open	69.1	54.6	5.5	Poor
LR open	73.4	59.1	0.9	Bad
L open	67.4	50.3	9.7	Poor
SCREENS, NO TRAFFIC				
LR closed	63.7	40.9	19.2	Good
R open	66.3	51.3	8.8	Poor
LR open	66.5	51.3	8.8	Poor
L open	65.1	48.2	11.8	Fair

As presented in Table 4, the A-weighted sound level measured in the car cabin was between 63.7 dB(A) and 73.4 dB(A), and changeable. The A-weighted sound level was mostly influenced by other traffic. The highest ratings were obtained for variants with the presence of other traffic. The traffic determined also the speech intelligibility ratings. The best intelligibility ratings were obtained for the last variant – screens and no traffic. The worst intelligibility ratings were obtained for variant with traffic and no screens. In general, for both windows closed, the intelligibility rating was Good. For R open, the intelligibility rating was 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. For laboratory room testing conditions, the sound level was equal to 26.7 dB(A). The intelligibility rating for such conditions was excellent.

After investigation of the background levels and speech intelligibility ratings in car cabin and in laboratory room testing conditions, it was proceeded discriminant function analysis for speech recordings. Discriminant analysis of speech commands spoken in the laboratory room showed significant main effects for 12 MFCC and 100 % of classification – every speech command was successfully classified to its classification function (see Table 5).

Table 5. Classification results – laboratory room conditions

Speech command	close	open	play	stop
close	100 %	-	-	-
open	-	100 %	-	-
play	-	-	100 %	-
stop	-	-	-	100 %
Mean value:	100 %			

A dash "-" in column means that no case has been classified to this command group.

Discriminant analysis performed for speech commands spoken in the car cabin showed significant main effects for 12 MFCC used in the model (Wilks'-Lambda: 0.040, approximation, $p < 0.0001$). Three discriminant functions (Root1, Root2, and Root3) based on 12 MFCC entry variables were created. Chi-square tests with successive roots removed performed in canonical stage are presented in Table 6.

Table 6. Chi-Square Tests with Successive Roots Removed – car cabin

Roots Removed	Canonical R	Wilks' - Lambda	p-value
0	0.903	0.040	0.000001
1	0.799	0.217	0.000003
2	0.633	0.599	0.015079

As presented in Table 6, chi-square tests of canonical stage showed significance of all created discriminant functions used in the model (R = 0.903, Wilks'-Lambda = 0.040, $p < 0.000001$). Removing of the first discriminant function showed high canonical value R between groups and discriminant functions (R = 0.799, Wilks'-Lambda = 0.217). In general, the more removed functions the less discrimination between groups (R = 0.633, Wilks'-Lambda = 0.599).

After deriving discriminant functions and determining variables, 12 MFCC features that discriminate most between speech groups, it was proceeded classification stage. The coefficients of classification functions obtained for speech commands are presented in Table 7.

Table 7. The coefficients of classification functions – car cabin

c_i	$K_1(close)$	$K_2(open)$	$K_3(play)$	$K_4(stop)$
c_{i0}	127.71	130.03	127.77	131.21
w_{i1}	19.50	26.18	16.57	26.11
w_{i2}	216.11	185.66	198.60	206.16
w_{i3}	-79.44	-75.19	-60.00	-83.35
w_{i4}	234.22	231.20	228.88	242.68
w_{i5}	73.76	38.32	51.66	60.75
w_{i6}	300.58	312.61	287.62	329.05
w_{i7}	155.20	168.28	158.78	152.78
w_{i8}	538.36	544.54	529.82	559.90
w_{i9}	-36.84	-64.57	-24.79	-54.13
w_{i10}	259.12	293.38	245.97	283.32
w_{i11}	18.88	-23.58	-7.64	-6.26
w_{i12}	-882.55	-844.69	-828.71	-912.11

Results of classification using classification functions $K_i(v)$ for speech command groups are presented in Table 8.

Table 8. Classification results – car cabin

Speech command	close	open	play	stop
close	100 %	-	-	-
open	-	92 %	8 %	-
play	-	-	100 %	-
stop	16 %	-	-	84 %
Mean value:	94 %			

Mean value (94 %) was calculated as a mean value of the best results of classification obtained for each command group. The highest score was obtained for /close/ and /play/ command groups (100 %). Every case in /close/ and /play/ command group was classified in 100 %. The lowest result of classification (84 %) was obtained for the /stop/ command – the 16 % of cases in /stop/ command group were classified as /close/ command. The 8 % of cases in /open/ command group were classified as /play/ command.

Conclusions

Background levels presented in this article 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 measurement variant.

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

For such acoustic conditions, discriminant analysis applied as classifier to the recognition of voice commands in the car cabin resulted in 94 % of classification. For laboratory conditions, the classification stage of discriminant analysis resulted in 100 % of classification. The efficiency of classifier was 6 % less in the car cabin due to higher background levels that influenced speech recordings. The value of 94 % in the car cabin is a good recognition rate for such changeable conditions. The preliminary results showed that discriminant analysis can be considered as robust classifier for recognition process in the cabin of the vehicle, but it requires further study with more advanced and extended database of speech commands.

REFERENCES

[1] Gong Y., Speech recognition in noisy environments: a survey. *Speech Communication*, 16 (3) (1995), 261–291

[2] Cavalcante B.A., Shinoda K., Furui S., Robust Speech Recognition in the Car Environment. LTC 2009, LNAI 6562 (2011), 24–34

[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, *Proc. of Interspeech 2004*, Jeju Island, Korea

[5] ISO/IEC 9921 – Assessment of Speech Intelligibility

[6] Mięsikowska M., Evert de Ruiter, Automatic Recognition of voice commands in a car cabin, *Pomiary, Automatyka, Kontrola*, 60 (2014), nr. 8, 652-654

[7] Discriminant analysis – STATSoft Electronic Documentation: http://www.statsoft.com/textbook/discriminant-function-analysis/

[8] The R Project for Statistical Computing: http://www.r-project.org/

***Author**: Marzena Mięsikowska (Ph.D.), Kielce University of Technology, Faculty of Mechatronics and Mechanical Engineering, Aleja Tysiąclecia Państwa Polskiego 7, 25-314 Kielce, E-mail: marzena@tu.kielce.pl.*