

STATISTICAL EFFECTS OF SELECTED NOISE CHARACTERISTICS ON SPEAKER RECOGNITION IN AUTOMOTIVE ENVIRONMENTS – A FIRST ANOVA-BASED INVESTIGATION

Sven Tuchscheerer
Dept. of Computer Science
Otto-von-Guericke-University
Magdeburg
PO Box 4120
39016 Magdeburg, Germany
stuchsch@ovgu.de

Christian Krätzer
Dept. of Computer Science
Otto-von-Guericke-University
Magdeburg
PO Box 4120
39016 Magdeburg, Germany
kraetzer@iti.cs.uni-
magdeburg.de

Tobias Hoppe
Dept. of Computer Science
Otto-von-Guericke-University
Magdeburg
PO Box 4120
39016 Magdeburg, Germany
t.hoppe@iti.cs.uni-
magdeburg.de

Jana Dittmann
Dept. of Computer Science
Otto-von-Guericke-University
Magdeburg
PO Box 4120
39016 Magdeburg, Germany
Jana.dittmann@iti.cs.uni-
magdeburg.de

ABSTRACT

A statistical analysis using the univariate, multifactorial analysis of variance (ANOVA) is used in this paper to investigate the impact of selected noise characteristics (here a 4-factorial design: amplitude, complexity, harmony and fundamental frequency) to speech signals and consecutively to the detection performance in speaker recognition systems (exemplarily used here: the BioSecure reference system ALIZE) in automotive application scenarios. An application scenario specific set of noise signals is recorded and generated and used to evaluate the influence of the noise characteristics. The results show that especially the amplitude and the fundamental frequency show a significant impact (p -values < 0.01), which is completely independent of the features used in the speaker recognition system. The two other characteristics (complexity and harmony) show much less significant impacts (p -values > 0.5).

Index Terms— automotive systems, speaker recognition, noise characteristics, statistical analysis

Copyright held by author(s), AutomotiveUT11, November 29-December 2, 2011, Salzburg, Austria

1. INTRODUCTION

Speaker and speech recognition systems in automotive application scenarios are a current trend to increase the usability in human-machine interaction for this application scenario. Limitations are imposed to the applicability of such speaker and speech recognition systems by the strength and the dynamics of the automotive noise environment. Influences of intra-individual speech modulation and inter-individual differences should not influence the accuracy of speaker recognition since they are designed resilient to individual variances.

To investigate selected characteristics of this noise environment (amplitude, complexity, harmony and fundamental frequency) and their impact, in this paper a first approach based on a 4-factorial design and analysis of variance (ANOVA; [1]) is presented.

The introduced methodology allows an objective investigation of the relationship between the noise characteristics and the performance of these different noise characteristics operationalized and observed by the biometric matching score of an exemplarily chosen speaker recognition system (here ALIZE [4]).

Regarding the results, our investigations, using in the test setup multiple speech samples from 10 speakers as well as 13 typical automotive noise signals (like an accelerating engine, the sound of a wiper, etc.) and three synthetic noise signals, show strong effects for the noise amplitude and the fundamental frequency. Weaker

effects (not statistical significant) are observed for complexity and harmony. The investigation of the ‘‘Lombard effect’’ was not part of the main focus. Nevertheless this effect should be considered in further studies (see section 5 of this paper). For practical applications, these results give a first indication for required filter designs for required audio signal pre-processing in automotive speaker recognition scenarios.

The paper is structured as follows: section 2 introduces our methodology, investigation concept and design. These descriptions are followed by the investigation results in section 3 and finally, a summary and conclusion in section 4.

2. INVESTIGATION METHODOLOGY, CONCEPT AND DESIGN

This section introduces the basic methodology for the investigations performed within this paper, derives a evaluation concept from the presented methodology and outlines a test design and operationalization.

2.1. Methodology

The methodology applied here, for investigating the impact of noise characteristics on the noises masking performance in automotive speaker recognition scenarios, is a statistical analysis using the univariate, multifactorial analysis of variance (ANOVA; [1]) method. The input for this variance analysis is the matching score of a biometric speaker verification system on complete noiseless and noisy signals. As long as the chosen speaker verification system is kept constant and shows significant performance (i.e. a n accuracy better than the probability of guessing correctly), this variance testing with differently parameterized noise characteristics will tell which influence the considered noise characteristics have on the signal. Hereby it is assumed that those noise characteristics, which are additively modelled onto the speech signals, would be able to completely alter the decision of any speaker verification system, since a too strong parameterization of the noise completely replaces the original distribution of the features computed from speech samples by the speaker verification system with the noise influence. Furthermore it is assumed here that especially in the considered automotive context the noise influence can range from zero to complete masking/cancellation of the speech signal to be used for the speaker verification.

In summary, our methodology allows an investigation of the relationship between the independent variables (noise characteristics) and dependent variables (masking performance of these different noise characteristics operationalized and observed by the matching score of a speaker recognition system – here: ALIZE [4]).

2.2. Test concept

The test concept for this paper has to consider four different parts. The first one is the used speaker recognition system, the second is the used noise model (including the used characteristics of noise signals), the third is the concept for the test material generation and the fourth is the method applied for the statistical analysis.

As an exemplary **speaker verification system** we use the open source software platform ‘‘ALIZE’’, developed by the Laboratoire Informatique d’Avignon ([4], [5]). It is applied to WAV-format audio files (44.1kHz, 16 Bit).

The feature extractor component of ALIZE, which extracts 16 Mel Frequency Cepstral Coefficients (MFCC), 16 first order derivatives of the MFCCs, the energy and the delta energy, is based on the the filter-bank based cepstral analysis library `sfbcep` from SPro [2].

This biometric speaker recognition system fulfils the two requirements imposed by the methodology described above in section 2.1: on one hand it performs with significant results (as shown e.g. in [5] or [6]) and on the other hand it can be used with a fixed parameter setting (here the default setting for the extractor: analysis frame length = 20ms, interval between two consecutive frames = 10ms, weighting window = Hamming, Mel frequency warping = on, add log-energy to the feature vector = on, pre-emphasis coefficient = 0, lower frequency bound = 300 Hz and upper frequency bound = 8000 Hz). Nevertheless it has to be mentioned again here that the actual performance of the used biometric system is of no direct concern for the introduced statistical analysis on the noise characteristics influence, as long as these two requirements are fulfilled and no limiting filters for the speech signals are applied.

For the automotive context considered here an additive **noise model** appears to be the most suitable approach. In contrast to its alternatives (multiplicative and context sensitive noise modelling) it allows for a simple handling of the most significant noise component in automotive scenarios, which is the superposition of the users speech commands with environmental noise. Therefore the noise model used here is described in equation (1) with S being the speakers commands, N being the noise and S' being the result of the superposition of both.

$$S' = S + N \quad (1)$$

The noise component N is in this paper modelled as the product between the moderating amplitude A and the output of a noise generator function parameterized using the three variables complexity c , harmony h and fundamental frequency b :

$$N = A \cdot \text{GenNoise}(c, h, b) \quad (2)$$

The variables A , c , h and b are the noise characteristics considered within this paper. These exemplary selected, physical **characteristics of noise signals** the following four are chosen and operationalized as follows:

- **Amplitude A :** The amplitude of an audio signal is correlated to its perceived volume. Therefore, the amplitude of a noise signal can be expected as a major influence factor on speaker recognition. Its effect could be considered as being equivalent to the loudness of a noise disturbing a human conversation.
- **Complexity c :** The complexity of a composed sound signal correlates to the number of distinct, relevant frequencies (that can e.g. be determined by an FFT transformation followed by a threshold filtering). Different potential approaches to determine the complexity of a sound signal have been analysed in preliminary tests. These approaches included counting the number of distinct frequencies after a Fourier transformation as well as determining achievable data compression rates. However, these approaches did not yield satisfying results, i.e. they did not lead to a sufficiently discriminating distribution. Finally, two complexity classes are defined here by the distinction of

recorded, natural sounds with high complexity (*hc*) and synthetic tones (as sinus-formed signals) with low complexity (*lc*).

- **Harmony *h*:** A sound signal is harmonic, if it has a constant cycle period and amplitude. Typically, Natural noise is disharmonic – but it may contain harmonic sequences (i.e. fragments with constant period and amplitude). To classify a noise signal with respect to harmony, the harmonic sequences within a certain time interval have been determined. The length of the harmonic sequences is put into relation to the total length of the sample to build a total harmony measure *h* for each sample, using the following formula:

$$h \equiv \frac{\text{time}(\text{harmonic_sequences})}{\text{time}(\text{total})} \quad (3)$$

- **Fundamental frequency *b*:** describes the lowest significant frequency in a complex, harmonic signal. The determined harmony properties of the samples are additionally used to derive their fundamental frequency. Determining the fundamental frequency of noise samples is only possible on fragments which are close to harmonic. It can yield to varying results not only between different harmonic fragments but also within each single one. As a simple measure to address this, the average fundamental frequency over all harmonic sequences is calculated. In addition, the average frequency is used as a reference (assuming it is identical or related to the fundamental frequency). Again a classification into two classes (high (*hb*) and low (*lb*)) is performed based on a threshold for the fundamental frequency.

For **test material generation**, as a basis for the corresponding investigations a context relevant test setup using equally numbered adult male and female speakers speaking automotive relevant voice commands has to be generated. Furthermore the introduced noise characteristics *A*, *c*, *b* and *h* have to be investigated in automotive-relevant ranges. To identify the effect of these characteristics (independent variables) on the speaker recognition performance (depend variable), the respective noise signals are superimposed as described in equation (1) into a set of reference voice samples and a cross comparison with the speaker recognition accuracy is performed. Considering the possible operationalizations of the ANOVA approach, we use a full factorial design to compute the effects for each factor (independent variable).

For the **statistical analysis** we compute descriptive statistics (means and two sided confidence intervals) and inference statistics tests. Based on the results from the univariate, multifactorial variance-analyses (ANOVA) effect sizes are computed. An effect size is a measure of the strength of the relationship between two variables in a statistical population. We used “Cohen’s *d*” as it is defined in [3] as the difference between two means divided by the standard deviation for the data. Furthermore we use the standard metrics *F* (from F-Test), *p* (or *p*-value, as lower the better indicator for the significance level) and the *statistical power* as they are defined in [1]. For the computation of these metrics SPSS [8] v.18 is used.

2.3. Test design and operationalization

As mentioned above in the evaluation concept, a **test-set** of automotive relevant speaker commands are recorded for equally numbered adult male and female speakers (here five speakers per gender). These voice commands consist of the phrase “please start the engine” spoken five times by each person as well as the spoken numbers from “1” to “10”. In Table 1 the used **parameterizations for the noise** characteristics are summarised.

Table 1: Noise characteristics parameterizations

Char.	Evaluated values
Noise/Speech-Ratio (Amplitude)	(0/1) and (0,3/1) and (0,66/1) and (1/1)
<i>c</i>	Natural sounds with high complexity (<i>hc</i>) and synthetic tones (as sinus-formed signals) with low complexity (<i>lc</i>) – a closer description of the used instantiations of the two complexity classes is found in Table 2.
<i>h</i>	A two-level classification is chosen (i.e.: rather harmonic (<i>hh</i>) / disharmonic (<i>lh</i>) samples) based on a threshold of 20% for <i>h</i>
<i>b</i>	The threshold to distinguish between the two defined classes high (<i>hb</i>) and low (<i>lb</i>) is set here to 700 Hz.

As output of the noise modelling function *GenNoise()* (see equation (2)) 16 different, application scenario relevant noise samples are modelled (see Table 2). The major part of these test samples is taken from a free online library [7]. They contain recordings of driving and braking, rain and wind as well sounds of a tram, a redirector and a rain wiper, as well as three synthetic sounds.

Table 2: Modelled noise samples

ID	Description	ID	Description
1	Brake and Roll	9	Wind 1
2	Drive through 1	10	Wind 2
3	Drive through 2	11	Oil pan vibration (damped)
4	Rain	12	Oil pan vibr. (not damped)
5	Rainwiper	13	Tram turn
6	Redirector & Acceleration 1	14	440 Hz Sine
7	Redirector & Acceleration 2	15	880 Hz Sine
8	Redirector & Acceleration 3	16	1320 Hz Sine

The normalization of all sounds to a common level (sampling rates (44.1 kHz), amplitudes (89db) and the assignment of each noise signal from the test set to distinct classes w.r.t. the noise characteristics is performed using the tools Audacity ([9]) and Praat ([10]). In

Table 3 the noise samples are classified with respect to their corresponding noise characteristics.

Table 3: Noise sample classification w.r.t. the noise characteristics

sample char.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
<i>c</i>	<i>hc</i>	<i>hc</i>	<i>hc</i>	<i>hc</i>	<i>hc</i>	<i>hc</i>	<i>hc</i>	<i>hc</i>	<i>hc</i>	<i>hc</i>	<i>lc</i>	<i>lc</i>	<i>hc</i>	<i>lc</i>	<i>lc</i>	<i>lc</i>
<i>h</i>	<i>hh</i>	<i>lh</i>	<i>lh</i>	<i>lh</i>	<i>hh</i>	<i>hh</i>	<i>lh</i>	<i>hh</i>	<i>lh</i>	<i>lh</i>	<i>hh</i>	<i>hh</i>	<i>hh</i>	<i>hh</i>	<i>hh</i>	<i>hh</i>
<i>b</i>	<i>lb</i>	<i>hb</i>	<i>hb</i>	<i>hb</i>	<i>hb</i>	<i>lb</i>	<i>lb</i>	<i>lb</i>	<i>hb</i>	<i>lb</i>	<i>lb</i>	<i>lb</i>	<i>hb</i>	<i>lb</i>	<i>hb</i>	<i>hb</i>

We further classify the 16 noise samples into 6 categories using equipartition (limiting the number of samples in a category and exclude redundant samples). We could not find any sounds for the “low complexity / low harmony” categories. Probably, these two properties can be expected to be mutually exclusive. Table 4 presents the output of this equipartition as the final test-design matrix.

Table 4: Test-design matrix

		<i>b</i> < 700Hz	<i>b</i> > 700 Hz
Low <i>c</i>	<i>h</i> < 20% (low)	no samples identified	no samples identified
	<i>h</i> > 20% (high)	Category 1 Sounds: 11; 12; 14	Category 2 Sounds: 15; 16
High <i>c</i>	<i>h</i> < 20% (low)	Category 3 Sounds: 7; 10	Category 4 Sounds: 2; 3
	<i>h</i> > 20% (high)	Category 5 Sounds: 1; 8	Category 6 Sounds: 5; 13

This test design yields a 4-factorial design for the introduced methodology and design. The four dimensions (here the noise characteristics) have been chosen because they are relevant in common automotive use-cases, where they are expected to have an influence on the performance of a speaker recognition system. However, due to run-time complexity reasons in the context of this work, further potentially suitable dimensions (e.g. frequency priority or frequency spectrum) could not be included and remain subject for future work.

3. TEST RESULTS

As introduced in section 2.1 we perform here an ANOVA variance analysis to test the influence of amplitude, complexity, harmony and fundamental frequency on the average matching score of the exemplarily chosen speaker recognition system.

For the amplitude we see a significant effect ($F=1765$, $p=0.000$, *statistical power*=1.0). A clear, nearly linear influence of this noise characteristic is recognisable: the louder the noise signal, the worse the resulting matching scores. Unfortunately, the analyses of the influence of the complexity *c* (Fig. 1) and harmony *h* (Fig. 2) do not lead to such clear results. Noise samples classified as complex (*hc*) can simultaneously be found with comparatively high and low impact on the resulting matching scores. The effect of *c* on the matching score is not significant ($p>0.5$). The same is

true for the harmony characteristics *h*. The mean matching score is the average of the matching scores from the tested samples (e.g. related to a specific factor).

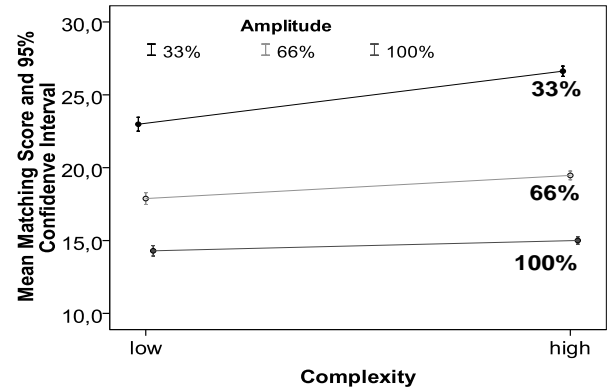


Fig. 1: Influence of the complexity *c* on speaker recognition

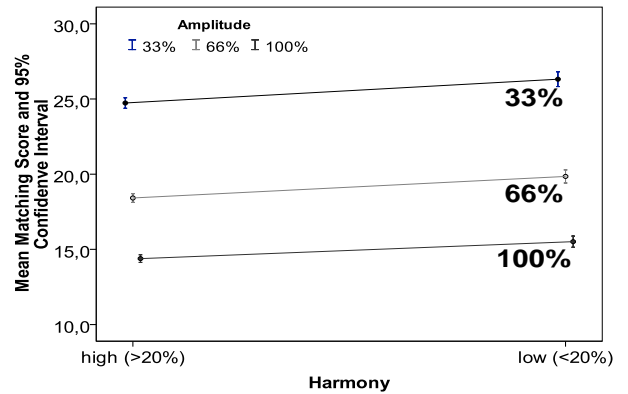


Fig. 2: Influence of the harmony *h* on speaker recognition

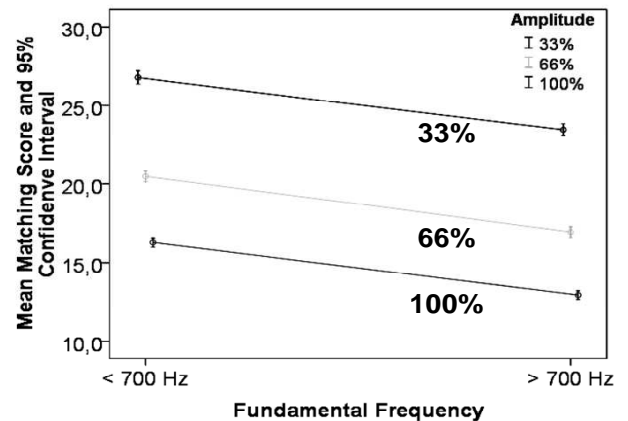


Fig. 3: Influence of the fundamental freq. *b* on speaker rec.

Finally, the evaluation of the influence of the fundamental frequency *b* (see Fig. 3) shows the following effect on matching-scores: Noise signals with a relatively low *b* (*lb*) have less influence on the matching score than signals which are classified *hb*. The mean degradation of matching scores caused by noise signals with a higher fundamental frequency (*hb*) is with ($p<0.01$) reasonably significant. For each pair-wise comparison of

categories (see table 4) and noise characteristics, the effect sizes are computed. The following comparisons show the largest effects: Amplitude (0,3/1 vs. 1/1) with $d= 0.834$, Category 3 vs. Category 6 with $d= 0.778$ and Category 4 vs. Category 5 with $d= 0.772$.

Summarising the pair-wise comparisons we can say that they are pointing out an interactive effect between harmony and fundamental frequency. If higher fundamental frequencies and lower harmonies appear together in a noise, this results in a high reduction of the matching score.

4. SUMMARY

Concluding the work done in this paper, the result is a first approach based on a 4-factorial design to achieve an approximation of the influences of selected noise characteristics on speaker recognition. We see different levels of influence in respect to the tested noise characteristics. The strongest effects are observed for the noise amplitude and its fundamental frequency. Weaker effects (not statistical significant) are observed for complexity and harmony.

For a practice-oriented application scenario like a speaker recognition system in automotive systems the results imply the following: the noise amplitude and its fundamental frequency have a strong influence on relevant audio content in such environments. This influence is completely independent of the features used in the speaker recognition system and its significance is documented by the computed effect sizes.

Active noise cancellation techniques, like active noise control or active noise reduction, would have no positive influence in the application scenario since the cancellation would also distort the spoken content. An approach that could include our results would be the intended use of FFT-transformations, separating sine-frequencies by detecting typical noise frequency-patterns and frequency changes in comparison to the FFT patterns of the speakers voice and of course a specific sound insulation by car design. For practical applications, these results give a first indication for required filter designs for required audio signal pre-processing in automotive speaker recognition scenarios.

5. OUTLOOK

With respect to the small sample group we found a statistic power above 90%. From theoretical statistics point of view, this is enough. Nevertheless, to generalize the results, especially for multiple languages and speech-frequency-ranges, that differ more than 1 SD (standard deviation) from standard normal distribution the sample group should be increased in further studies.

The investigation of the "Lombard effect" was not part of the main focus. Natural sounds were investigated – and, derived from that, the impact on speaker recognition accuracy was computed using an exemplary speaker recognition system. The recording of the speech samples was realized in an absolutely quiet environment (a soundproof anechoic chamber), so the Lombard effect did not take place in our design. We further normalized all speech samples to the same amplitude level. If we would not do that, the effects we measured would have needed to be corrected. With respect to this aspect, we plan further studies, using a structural equation model, sorting out the covariances to compute the causal connection between the factors and the recognition performance.

Beside vehicles, our findings could also be applied in other domains as, for example, noisy production environments. As well as in cars, authentication processes, interaction processes or safety processes could be realized hands-free. Therefore it will be extremely necessary to design those systems in a robust way, which means regarding the environmental typical noises, inter- and intra-individual variance.

6. ACKNOWLEDGEMENTS

The work described in this paper has been supported in part by the European Commission in the context of the programme COMO - Competence in Mobility (EU/EFRE) under Contract No. C(2007)5254. The information in this document is provided as is, and no guarantee or warranty is given or implied that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.



European Commission
European Regional Development Fund
INVESTING IN YOUR FUTURE

Sven Tuchscheerer is funded by the German Ministry of Education and Science (BMBF), project 01IM10002A. Parts of the presented work, in particular application to production environment, is part of the ViERforES project.

7. REFERENCES

- [1] D.A. Freedman, "Statistics," W.W. Norton & Company, New York, NY, 2008.
- [2] G. Gravier, "Quick reference guide – sfbcep," 2004; Retrieved Dec. 2010 from <http://www.irisa.fr/metiss/guig/spro/>
- [3] P. Sedlmeier, "Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen," *Methods of Psychological Research*, 1(4), pp.41 – 63, 1996.
- [4] ALIZE, "Open Tool for Speaker Recognition" [Online]. Available: <http://www.lia.univ-avignon.fr/heberges/ALIZE/>
- [5] J.-F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a free toolkit for speaker recognition," In Proc. ICASSP, pp. 737-740, 2005.
- [6] D. Fauve, D. Matrouf, J.-F. Scheffer, J.-F. Bonastre, J. Mason, "State-of-the-Art Performance in Text-Independent Speaker Verification Through Open-Source Software," *IEEE Trans. on Audio, Speech and Lang. Proc.*, Vol. 15(7), pp. 1960-1968, 2007.
- [7] Thefreesoundproject, <http://www.freesound.org>, 2010.
- [8] IBM: SPSS 18 incl. AMOS for Windows. (Available from <http://www.spss.com/software/statistics/>). 2009.
- [9] <http://audacity.sourceforge.net/>
- [10] <http://www.fon.hum.uva.nl/praat/>