# Analysis and Detection of Cognitive Load and Frustration in Drivers' Speech

*Hynek Bořil*[1], *Seyed Omid Sadjadi*[1], *Tristan Kleinschmidt*[2], *John H. L. Hansen*[*1]

[1]Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering,
University of Texas at Dallas, Richardson, Texas, U.S.A.
[2]Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia

## Abstract

Non-driving related cognitive load and variations of emotional state may impact a driver's capability to control a vehicle and introduces driving errors. Availability of reliable cognitive load and emotion detection in drivers would benefit the design of active safety systems and other intelligent in-vehicle interfaces. In this study, speech produced by 68 subjects while driving in urban areas is analyzed. A particular focus is on speech production differences in two secondary cognitive tasks, interactions with a co-driver and calls to automated spoken dialog systems (SDS), and two emotional states during the SDS interactions - neutral/negative. A number of speech parameters are found to vary across the cognitive/emotion classes. Suitability of selected cepstral- and production-based features for automatic cognitive task/emotion classification is investigated. A fusion of GMM/SVM classifiers yields an accuracy of 94.3 % in cognitive task and 81.3 % in emotion classification.

**Index Terms**: Cognitive load, emotions, speech production variations, automatic classification.

## 1. Introduction

Recent advancements in the electronic industry have made access to information and entertainment easier than ever before. While undoubtedly benefiting many areas of our daily lives, there are situations where the presence of electronic gadgets has the opposite effect. In a current study, the Virginia Tech Transportation Institute (VTTI) reports that dialing on a hand-held device whilst driving increases the risk of an accident by a factor of three, and communicating via hands-free set increases the risk by one third [1]. This suggests that performing secondary cognitive tasks while driving may severely impact driving performance. Besides cognitive load, drivers' emotions have also been shown to adversely affect driving performance. In [2], drivers exhibited larger deviations of lane offset and steering wheel angle, and shorter lane crossing times in anger and excitation situations – signs of reduced lane control capability. Availability of an automated system assessing cognitive load and emotional state in drivers would benefit the design of active safety systems and other intelligent in-vehicle interfaces, making them capable of adapting to the driver's current state (e.g. decreasing the frequency of navigation prompts in high cognitive load situations).

Lately, increasing attention has been paid to emotion recognition in the speech community [3–5], gathering research labs in joint evaluation projects such as the INTERSPEECH 2009 Emotion Challenge [6], where natural emotions in children interacting with a pet robot were analyzed. Others have studied the impact of stress in speech (including cognitive task stress) [7, 8], however, a relatively limited body of literature deals with the impact of stress and emotions on drivers.

Driving simulator studies [9, 10] commonly utilize physiological and EEG signals to assess driver emotional states but not driver speech. An exception to this was speech collected in a driving simulator [11, 12], where speech was analyzed under various cognitive tasks and emotional states.

The majority of in-vehicle studies utilize data collected in driving simulators rather than real driving scenarios. This is due to the fact that it is much easier to control emotional/cognitive load scenarios in a lab environment without compromising driver safety. On the other hand, it is questionable how much the observed phenomena correspond to those in real conditions, where driving errors may have severe consequences. In addition, we note that using acted emotions as emotional class templates (as per [11]) may be misleading, as acted emotions often represent exaggerated traits (see discussion in [6]).

The goal of the present study is to analyze speech production variations in real driving conditions. The study extends our initial efforts presented in [13, 14]. In [13], distributions of drivers' speech response delays and proportions of negative emotions with respect to the dialog system failure to recognize the queries, as well as accompanying variations in speech production were analyzed in interactions with two commercial speech dialog systems. In the abstract [14], outcomes of a more comprehensive analysis of speech production were reported for two cognitive tasks and two emotional classes as observed for a group of 42 subjects. A simple classifier employing cepstral features was shown to provide reasonable accuracy in distinguishing two cognitive tasks.

In order to increase the statistical significance of the observed effects of cognitive load and emotions in drivers, this study analyzes speech from a total of 68 subjects (33 females and 35 males). Two cognitive load tasks – communication with a co-driver and interaction with two SDS, and two emotional states (neutral and negative) are studied with respect to speech production changes. Based on the outcomes of the acoustic analysis, suitability of selected speech features for cognitive task/emotional state classification is investigated alongside several common and state-of-the-art cepstral based speech coding schemes. Furthermore, fusion of the acoustic and cepstral domain features is studied and shown to further benefit the classification accuracy.

The current (and often successful) trend in the main stream emotion recognition is to extract typically hundreds to thousands of speech parameters and use automatic feature selection strategies to obtain semi-optimal task-oriented feature sets. In contrast to this, given the limited knowledge of the impact of real in-vehicle environments on speakers, the goal of our study is to analyze and understand the variation in primary speech production parameters. In addition, we show that a careful selection and combination of a very limited number of 'elementary' acoustic features results in promising classification performance.

The remainder of the paper is organized as follows. First, the data corpus used in the study is described. Second, the procedure and results of the acoustic analysis of speech are summarized.

[*] This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen. Approved for public release; distribution unlimited.

Third, performance of selected acoustic and cepstral speech representations is evaluated in the cognitive task/emotion state classification. The final section summarizes the outcomes of the study.

## 2. Corpus Description

The analyses and experiments presented in this paper are conducted on a set of 68 drivers (33 females and 35 males) from the UTDrive database [15] collected in real conditions while driving in urban areas. The session routes comprise a mixture of secondary, service, and main roads in residential and business districts in Richardson, TX. The data were collected in a Toyota RAV4 vehicle equipped with a set of microphones, CCD cameras for monitoring the driver and the road scene, optical distance sensor, GPS, CAN-Bus OBD II port for collecting vehicle speed, steering wheel angle, gas and brake inputs from driver, and gas/brake pedal pressure sensors. In this study, a speech signal from the microphone mounted above the windshield is utilized.

Each driving session includes a mixture of several secondary tasks that the driver is asked to perform while driving such as sign reading, operating a radio and AC, talking to a co-driver, and calling two commercial automated dialog systems – American Airlines for online flight departure/arrival information, and Tell ME for general information including weather, sports scores, movie theaters, etc. Our focus in this study is exclusively on the driver's interactions with a co-driver and the calls to the automated SDS.

Table 1: Definition of Emotion Classes.

| | |
|---|---|
| **Neutral (Non-Negative)** | Neutral, Confident, Happy, Humored, Uninterested |
| **Negative** | Hesitancy, Confusion, Frustration, Anger, Increased Vocal Effort, Decreased Speaking Rate, Altered Pitch |

While the real level of cognitive load in drivers is unknown, it can be argued that calls to the dialog system are likely to induce higher load than co-driver interactions. The cell-phone interaction with the dialog system can be broken down into a set of subtasks: holding and dialing, interaction, and processing. Due to the frequent errors of the automatic speech recognition engine in the dialog system, the driver is frequently asked to confirm and repeat queries. This induces further load, where the subject tries to figure out how to become more intelligible to the system. On the other hand, the interactions with the co-driver are of a relaxed nature and the discussed topics do not require any extensive focus (discussing weather, etc.). Considering this and following [11], we use *cause-type* annotation of the cognitive load scenarios and map cognitive load labels to the tasks – co-driver interactions (low cognitive load) vs. dialog system interactions (high cognitive load).

In addition, it was observed that frequent requests of the dialog system for query repetitions are likely to induce negative emotions in drivers [13]. To further study this phenomenon, in Table 1 we define two broad emotional classes. Subjectively perceived emotions in drivers' speech were manually labeled by an expert annotator. The proportion of negative interactions with the increasing number of repetition request is shown in Fig. 1. In the remainder of this paper, the speech variations in the two cognitive task and emotional state scenarios are analyzed and classified.

## 3. Speech Production Analysis

Past investigations have observed variations in a number of speech parameters across stress and emotion classes [3, 7, 16]. In this section, the following speech production factors are analyzed:

- mean utterance fundamental frequency $F_0$

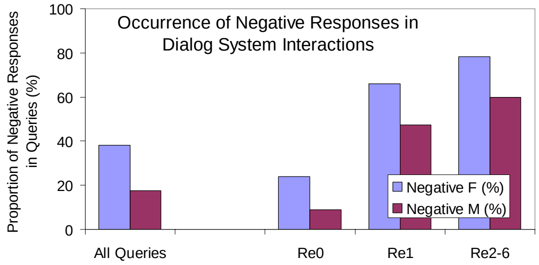

Figure 1: Proportion of negative interactions with dialog systems. F and M denote female and male subjects respectively; Re0 – no repetition, Re1 – $1^{st}$ repetition, Re2-6 – $2^{nd}$–$6^{th}$ repetitions.

- first four formant center frequencies in voiced speech segments $F_{1-4}$
- spectral slope
- duration of voiced segments
- spectral center of gravity (SCG)
- spectral energy spread (SES).

SES is defined as a frequency interval of one standard deviation from SCG and is expected to be sensitive to changes in the energy concentration across the frequency axis. $F_0$ and $F_{1-4}$ were extracted using the open source tool WaveSurfer. To reduce the effect of segmental and contextual variation [17] on the analyzed factors, their values are averaged on the utterance level.

Since many speech parameters have distinctive nominal values in males and females, aggregating them may mask any observable effects of cognitive load and emotions. To account for this and improve the statistical significance of the following analyses, each speech parameter was first normalized by a fixed mean with respect to gender and then aggregated. The fixed mean for each parameter was calculated as the average of the respective male and female means in the co-driver interactions, in order to preserve the notion of the original sample amplitudes. For example, for observed female mean $F_0 = 208\,\text{Hz}$ and male mean $F_0 = 136\,\text{Hz}$, both male and female $F_0$ distributions were centered to $F_0 = 172\,\text{Hz}$.

The task/emotion dependent mean values for $F_0$, SCG, and voiced segment durations are shown in Fig. 2, where the vertical bars represent 95 % confidence intervals. Similar trends can be observed for all three parameters: there is an increase in $F_0$, SCG, and voiced segment duration when switching from co-driver interactions to the dialog system task. Similarly, a parameter increase can be seen from neutral to negative emotions, and mostly also from no-repetition (Re0) to first repetition (Re1), and $2^{nd}$–$6^{th}$ repetition. Note that two variants of SCG were analyzed – 'SCG-All' refers to the spectral center of gravity extracted from all utterance segments while 'SCG-Voiced' was extracted just from voiced parts of speech. For the remainder of this analysis, only 'SCG-All' is analyzed and denoted as SCG.

The different rates of error bar overlaps give a notion of the significance of the effects of cognitive task and emotions on the speech parameters. When inspecting spectral slope plots, considerable overlaps of the 95% error bars were observed across all classes, suggesting that this parameter is not sensitive to any of the analyzed effects. This is probably due to the relatively high energy noise content in the low frequency portion of spectra. For this reason, spectral slope was not considered in the rest of the study.

To get a more concrete idea about the significance of the feature interactions, paired t-tests were performed on the samples from co-driver vs. dialog system classes, and neutral vs. negative

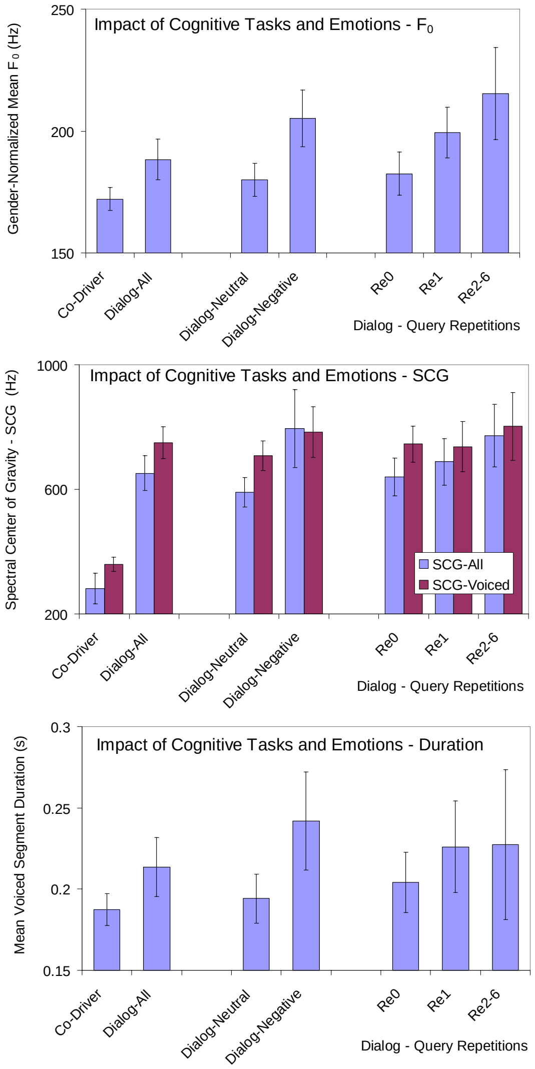

Figure 2: Impact of cognitive tasks and emotions on speech production.

classes. Data from 58 subjects were evaluated in the co-driver vs. dialog system paired tests, since data from one of the two domains were not available for the remaining 10 subjects. The paired test set was reduced to 30 subjects in the case of emotion analysis, since many subjects did not exhibit negative emotions. Finally, one-way analysis of variance (ANOVA) together with Levene's test of homogenity of variance and Fisher's least significant difference (LSD) post-hoc test were conducted on the dialog system repeated queries.

- *Co-driver vs. dialog system* interactions: $F_0$ – significant increase ($t\,(57) = -3.820, p < 0.001$); $F_1$ – significant increase ($t\,(57) = -7.282, p < 0.001$); $F_{2,3}$ – not significant effects ($p > 0.1$); $F_4$ – significant increase ($t\,(57) = -3.339, p = 0.001$); $SCG$ – significant increase ($t\,(57) = -9.803, p < 0.001$); $SES$ – significant increase ($t\,(57) = -10.487, p < 0.001$); $duration$ – significant increase ($t\,(57) = -2.726, p = 0.008$).
- *Neutral vs. negative* interactions: $F_0$ – significant increase ($t\,(29) = -2.472, p = 0.02$); $F_{1-4}$ – no significant effects ($p > 0.187$); $SCG$ – significant increase ($t\,(29) = -3.008, p = 0.005$); $SES$ – no significant effects ($p = 0.541$); $duration$ – significant increase ($t\,(29) = -3.004, p = 0.005$).
- *Effect of query repetitions* interactions: $F_{1-4}$ – not significant effects ($p > 0.833$); $F_0$ – significant interaction ($p = 0.002$), post-hoc test – significant increases from Re0 to Re1 (p = 0.029) and from Re0 to Re2-6 ($p < 0.001$) significant, Re1 vs. Re6 – no significant interaction (p = 0.076); $SCG$ – no significant effects ($p > 0.05$); $SES$ – no significant effects ($p = 0.165$); $duration$ – no significant effects ($p = 0.406$).

## 4. Cognitive Task/Emotion Classification

In this section, the effectiveness of selected acoustic features and cepstral representations in cognitive task and emotion classification is evaluated. In the cognitive task classification, the goal is to identify whether the driver's utterance comes from the co-driver or automated dialog system interaction, while in the emotion classification, neutral and negative emotional states are to be distinguished on the utterance level. For the cognitive task and emotion classification respectively, a Gaussian Mixture (GMM) based maximum likelihood classifier was trained. Data from 40 sessions (20 per gender) were used to train the GMM's; the reminder (28 sessions) were used for speaker/gender-independent open test set evaluation. For each setup, a binary decision threshold providing an equal error rate – EER (balanced error of the class assignments) was found in an iterative procedure on the test set.

From the cepstral domain, common Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP), and Expolog cepstral coefficients (Expolog) [18] are compared. Unlike MFCC and PLP, Expolog was designed with focus on stressed speech recognition. Two variants of MFCC and PLP, where discrete cosine transform cepstrum was replaced by linear prediction cepstrum (MFCC-LPC) and vice-versa (PLP-DCT) were also considered. All cepstral representations were evaluated with and without applying a full-wave spectral subtraction algorithm.

The results for the cepstral representations are summarized in Table 2; 'Def' and 'NS' denote the default front-end setup without and with noise subtraction respectively. The results are reported as equal accuracy rates (EAR), calculated as $100 - EER$ (%). In both cognitive and emotion classification tasks, the best performance is provided by the Expolog-based classifier (32-mixture GMM's). In the case of emotion classification, standard PLP provides similar performance to Expolog. The overall performance suggests that cognitive task classification is somewhat easier than emotion recognition. This corresponds well with the intuition obtained in the previous section, where more speech parameters were found to significantly vary in the cognitive tasks compared to the emotion states. In addition, a higher $p$-value was found for two of the three parameters that displayed significant changes in emotion states compared to $p$-values found in cognitive tasks.

Second, performance of acoustic features in the classification tasks was evaluated. Based on the observations presented in the previous section, $F_0$, $F_1$, SCG, SES, and voiced duration ('Dur') were used as cognitive task classification features. In a similar setup like for the cepstral features, separate classifiers were first trained for each individual feature type, and also a classifier combining all of them into a single vector was trained. The performance in the cognitive task is shown in Table 3. The best performance can be seen for SCG-based classifier (single Gaussian GMM's). Surprisingly, combining all feature types into a single classification vector results in a slightly reduced performance. Overall, the selected acoustic features provided lower cognitive class classification accuracy than cepstral-based classifiers, however, the performance is still considered promising.

$F_0$, SCG, and duration were employed in the acoustic feature-based emotion classification (results shown in Table 4). Here, the

Table 2: Cog. task/emotion class. – cepstral features; EAR (%).

| | Cog. Task | | Emotion Task | |
|---|---|---|---|---|
| Front-End | Def | NS | Def | NS |
| MFCC | 92.2 | 93.7 | 66.9 | 68.8 |
| MFCC-LPC | 93.0 | 92.5 | 66.9 | 66.4 |
| PLP | 92.4 | 93.4 | **69.3** | 66.9 |
| PLP-DCT | 93.3 | 93.3 | 66.4 | 64.8 |
| Expolog | 91.5 | **94.0** | **69.3** | 64.8 |

Table 3: Cog. task classification – acoustic features; EAR (%).

| $F_0$ | $F_1$ | SCG | SES | Dur | Together |
|---|---|---|---|---|---|
| 58.3 | 73.7 | **90.0** | 83.0 | 65.5 | 88.2 |

best performance was reached when combining all three features into one vector and the equal accuracy rate obtained here (using single Gaussian GMM's) is considerably higher than the cepstral-based classifiers.

Finally, a support vector machine (SVM) based classifier employing RBF kernel is used to perform feature fusion. Since the cepstral features are extracted on a frame level while acoustic features are extracted on the suprasegmental (utterance) level, instead of combining them directly into a supervector for SVM, the cepstral features are replaced by utterance-level scores from the GMM classifier trained on cepstral features. For each utterance, two scores from class-dependent GMM's are obtained. For example, given an MFCC classifier in the emotion classification task, the overall likelihoods that the utterance is generated by neutral and negative models are calculated and used as cepstral-based score features in the SVM system. Similarly, a variation of the acoustic features represented by scores from the corresponding classifier is evaluated. Among several feature combinations, the SVM fusion of SCG acoustic features, SCG likelihood scores from a GMM classifier, and Expolog likelihood scores from a GMM classifier provided a slight performance improvement in the cognitive task, yielding 94.3% accuracy (compared to 94.0% accuracy of SCG-based classifier), and the fusion of $F_0$, SCG, duration, and their scores from the respective GMM classifiers yielded an accuracy of 81.3 % in the emotion classification (compared to 78.1 % obtained from the acoustic based classifier).

## 5. Conclusion

In this study, speech produced by 68 subjects while driving in realistic scenarios was analyzed. A particular focus was on the impact of two types of secondary cognitive tasks that were considered to represent different levels of cognitive load, and two emotional states, on speech production parameters. A number of parameters were found to vary significantly with the cognitive task and emotional state. Based on the outcomes of this analysis, features likely to provide discrimination between cognitive tasks and emotional states were selected and compared alongside cepstral-domain speech representations in automatic classification tasks. The best performance was reached by fusing acoustic features, their corresponding GMM scores, and cepstral features in the cognitive task classification (94.3 % accuracy). For emotion classification, the best result was obtained by SVM-based fusion of selected acoustic features and their respective GMM scores (81.3%).

Table 4: Emotion classification – acoustic features; EAR (%).

| F0 | SSG | Dur | $F_0$+SCG | $F_0$+SCG+Dur |
|---|---|---|---|---|
| 73.2 | 62.2 | 57.0 | 75.7 | **78.1** |

## 6. REFERENCES

[1] B. Magladry and D. Bruce, *In-Vehicle Corpus and Signal Processing for Driver Behavior*. USA: Springer, 2008, ch. Improved Vehicle Safety and How Technology Will Get US There, Hopefully). K. Takeda, H. Erdogan, J. H. L. Hansen, H. Abut (Eds.), pp. 1–8.

[2] H. Cai, Y. Lin, and R. R. Mourant, "Study on driver emotion in driver-vehicle-environment systems using multiple networked driving simulators," in *Proc. Driving Simulation Conference – North America 2009*, Iowa City, Iowa, 2007.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Sig. Proc. Mag.*, vol. 18, no. 1, pp. 32–80, 2001.

[4] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. on Speech & Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[5] W. Kim and J. H. L. Hansen, "Angry emotion detection from real-life conversational speech by leveraging content structure," in *IEEE ICASSP'10*, Dallas, TX, 2010, pp. 5166–5169.

[6] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *INTERSPEECH'09*, Brighton, pp. 312–315.

[7] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Comm.*, vol. 20, no. 1-2, pp. 151–173, 1996.

[8] B. Yin, N. Ruiz, F. Chen, and M. A. Khawaja, "Automatic cognitive load detection from speech features," in *OZCHI '07: Proc. of the 19th Australasian conference on Computer-Human Interaction*. New York, NY, USA: ACM, 2007, pp. 249–255.

[9] J. Wang and Y. Gong, "Normalizing multi-subject variation for drivers' emotion recognition," in *Proc. IEEE Int. Conf. on Multimedia and Expo 2009*, NY, USA, 2009, pp. 354–357.

[10] L. J. M. Rothkrantz, R. Horlings, and Z. Dharmawan, "Recognition of emotional states of car drivers by EEG analysis," *Neural Network World; International Journal on Neural and Mass - Parallel Computing and Information Systems*, vol. 19, no. 1, pp. 119–128, 2009.

[11] R. Fernandez and R. W. Picard, "Modeling drivers' speech under stress," *Speech Communication*, vol. 40, no. 1-2, pp. 145–159, 2003.

[12] C. M. Jones and I.-M. Jonsson, *Universal Access in Human-Computer Interaction. Ambient Interaction*. Berlin/Heidelberg: Springer, 2007, ch. Performance Analysis of Acoustic Emotion Recognition for In-Car Conversational Interfaces. W.B. Kleijn and K.K. Paliwal (Eds.), pp. 411–420.

[13] T. Kleinschmidt, P. Boyraz, H. Bořil, S. Sridharan, and J. H. L. Hansen, "Assessment of speech dialog systems using multi-modal cognitive load analysis and driving performance metrics," in *IEEE International Conference on Vehicular Electronics and Safety ICVES'09*, Pune, India, November 2009, pp. 167–172.

[14] H. Bořil, T. Kleinschmidt, P. Boyraz, and J. H. L. Hansen, "Impact of cognitive load and frustration on drivers' speech." *The Journal of the Acoust. Soc. of America*, vol. 127, no. 3, pp. 1996–1996, 2010.

[15] P. Angkititrakul, M. Petracca, A. Sathyanarayana, and J. Hansen, "UTDrive: Driver behavior and speech interactive systems for in-vehicle environments," in *Proc. of the IEEE Intelligent Vehicle Symposium*, June 2007, pp. 566–569.

[16] Z. Callejas and R. López-Cózar, "Influence of contextual information in emotion annotation for spoken dialogue systems," *Speech Communication*, vol. 50, no. 5, pp. 416 – 433, 2008.

[17] J. Volín and D. Studenovský, "Normalization of Czech vowels from continuous read texts," in *Proc. of 16th Int. Congress of Phonetic Sciences (ICPhS XVI)*, Saarbrücken, Germany, 2007, pp. 185–190.

[18] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Trans. on Speech & Audio Proc.*, vol. 8, no. 4, pp. 429–442, 2000.