CoLoSS: Cognitive Load Corpus with Speech and Performance Data from a Symbol-Digit Dual-Task

Robert Herms[1], **Maria Wirzberger**[2], **Maximilian Eibl**[1], **Günter Daniel Rey**[2]
[1]Media Informatics, Chemnitz University of Technology, Germany
[2]Psychology of Learning with Digital Media, Chemnitz University of Technology, Germany
{robert.herms, maximilian.eibl}@cs.tu-chemnitz.de

Abstract

In this paper, a new corpus named CoLoSS (Cognitive Load by Speech and performance data in a Symbol-digit dual-task) is presented, which contains speech under cognitive load recorded in a learning task scenario. In order to obtain a reference for cognitive load, a dual-task approach was applied, including a visual-motor primary task that required subjects to learn abstract symbol combinations and an auditory-verbal secondary task to measure the load imposed by the primary task. We report the methodology of collecting the speech recordings, constructing the corpus and describe the properties of the data. Finally, effects of cognitive load on prosodic as well as voice quality features are investigated in conjunction with the corpus. In its current version, the corpus is available to the scientific community, e.g., for exploring the influence of cognitive load on speech or conducting experiments for speech-based cognitive load recognition.

Keywords: speech corpus, speech features, cognitive load, learning performance, dual-task

1. Introduction

The human cognitive system is characterized by capacity limitations in information processing (Plass et al., 2010). They refer to human working memory, which provides temporary storage and manipulation of information (Baddeley, 1992). Cognitive load is generally considered as the load imposed on an individual's working memory by a particular (learning) task (Paas and Van Merriënboer, 1994). According to the cognitive load theory (Sweller et al., 2011), the degree of cognitive load influences the amount and complexity of learned content (Paas et al., 2003).

Speech databases that include audio recordings of speakers under varying levels of cognitive load are rather rare and often created for own research purposes. Different task designs that were developed to investigate the limitations of human working memory in conjunction with speech parameters can be found in the literature; for instance, reading-comprehension (Yin et al., 2007), Stroop interference (Yap et al., 2010), arithmetic abilities (Gorovoy et al., 2010), and driving under cognitive load (Boril et al., 2010). Moreover, the Cognitive load with Speech and EGG (CLSE) database was created by (Yap, 2012), which includes speech recordings of subjects participating in three different tasks: Stroop test with time pressure, Stroop test with dual-task, and reading span task. According to the *INTERSPEECH 2014 Computational Paralinguistics Challenge* (Schuller et al., 2014), the partitioned form of the CLSE database is available for research purposes. Nevertheless, there is still no speech-based corpus available for either the consideration of cognitive load in a learning context or a more sensitive approach to the traditional classification problem.

In this paper, CoLoSS (**C**ognitive **L**oad by **S**peech and performance data in a **S**ymbol-digit dual-task) is introduced—a new corpus that includes speech under cognitive load recorded in a learning task scenario. Compared to existing works in the literature concerning speech-based cognitive load discrimination, the CoLoSS corpus differs in two key aspects: (1) It focuses on cognitive load induced by learning processes. (2) Numeric labels are provided as reference for cognitive load.

The fundamental goal of this work is to encourage scientists in the field of speech technologies to explore the effects of cognitive load (caused by learning) on speech and to provide the basis for regression and/or classification experiments for automatic speech-based cognitive load recognition. The corpus material will be available to the scientific community including audio files, annotations, and labels.

This paper is organized as follows: In the next section we introduce the CoLoSS corpus, including task design, recording conditions, data labelling, and data description. In Section 3, effects of cognitive load on prosodic as well as voice quality features are investigated in conjunction with the introduced corpus. Finally, we conclude this paper in Section 4 and give some future directions.

2. CoLoSS Corpus

The CoLoSS corpus represents a subset of data collected for the experimental study of (Wirzberger et al., 2017b) in which the task design and performance measures were defined. In the following, the task design, cognitive load indicators, recording conditions, and the data of the corpus will be described in detail.

2.1. Task Design

The main goal of this task design was to assess the residual cognitive resources of subjects while they were performing a learning task. For this purpose, a dual-task paradigm was applied: a visual-motor primary task involving the assignment of symbol combinations to a single symbol, while simultaneously memorizing a sequence of five digits from an auditory-verbal secondary task. Symbol assignments of the primary task reflected knowledge schemata that had to be formed across the trials. Inspired by (Wirzberger et al., 2017a), performance measures of the secondary task were considered as reference for cognitive load associated with the primary task.

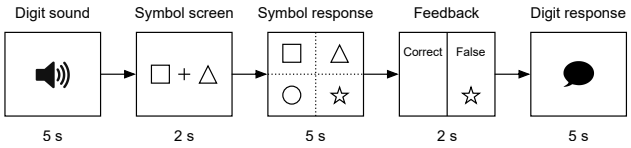

Figure 1: Schematic representation of the dual-task method applied for the CoLoSS corpus. Primary task: step two to three; secondary task: step one and five.

The task comprised a set of 64 trials and was presented to the participants using a desktop computer. Participants were guided in each trial by different screens as depicted in Figure 1. Each trial consisted of the following five steps: (1) Digit sound: a random sequence of five digits in the range of 1 to 9 (in random order) was generated by a text-to-speech system (in German); (2) Symbol screen: one out of four randomly chosen combinations of abstract geometrical symbols was displayed where the order of the symbols must be considered; (3) Symbol response: one out of four possible symbols in a randomly arranged 2×2 grid had to be selected via mouse click; (4) Feedback: feedback was obtained, accompanied by the correct symbol in case of false responses to foster correct schema acquisition; (5) Digit response: the verbal recall of the five digit sequence of step 1 in correct order was requested.

Additionally, task difficulties varied between subjects, but not within the task, by the number of symbols displayed on the screen in step 2. At this point, a distinction was made between an easy and a difficult condition by two and three symbols, respectively.

With reference to the cognitive load theory (Sweller et al., 2011), this task design is associated to various assumptions: Intrinsic cognitive load is represented in the described framework by the number of symbols used to form the combinations. Extraneous cognitive load is represented by the embedded secondary task requirements. Finally, the overall cognitive load, including germane load, is reflected in performance measures of the secondary task.

2.2. Chosen Performance Measures

In order to obtain sensitive measures of the subjects' performance concerning the primary and secondary task, an efficiency score was computed using the likelihood model approach after (Hoffman and Schraw, 2010). The calculation based upon the ratio between performance and effort, whereby performance is represented by the accuracy of problem solving and effort is represented by the time required.

For primary task efficiency, performance was obtained by symbol response correctness while reaction time needed to select a symbol constitutes the effort component. Note, reaction time was related to the visual stimulus regarding the appearance of the 2×2 grid in the *symbol response* stage.

For secondary task efficiency, the performance component was defined by the word accuracy of subjects' responses regarding the five-digit sequence in the *digit response* stage. More precisely, substituted, inserted and deleted words were considered to calculate the word error rate (WER)—the common evaluation measure for automatic speech recognition systems. The word accuracy was then computed by WA = 1 − WER. Since the number of words in the reference added up to five and negative accuracy values were set to zero, the following values could be obtained by parameter WA: 0, 0.2, 0.4, 0.6, 0.8, and 1.0. The effort component of the secondary task was determined by the verbal response duration, i.e., the time starting from the presentation of the visual stimulus (speech bubble) to the end of the last uttered digit. In addition to the actual utterance duration of the subject, the verbal response duration includes indeed the onset latency, i.e., the reaction time from the stimulus to the onset of the first uttered digit. This time span reflects complex cognitive processing for mentally representing the message, selecting words, and retrieving syntactic and phonetic properties; moreover, motor processing for articulation is required.

2.3. Recording and Postprocessing

In total, 123 German students from the Chemnitz University of Technology (Germany) participated in the task. Speech was recorded using a mono clip-on microphone at a sampling frequency of 48 kHz and a 24 bit resolution via a mobile recording device (Roland R-88). Each recording session refered to a particular subject who performed the learning task across 64 trials. A recording session lasted about 20 minutes. Afterwards, the audio segments of the uttered five-digit sequence within the *digit response* stage were extracted using time-codes (5 seconds + 0.5 seconds tolerance) from the task log-data.

The data of 28 subjects had to be excluded for different reasons (lack of sufficient working memory capacity, lack of confirmation for data sharing, lack in language proficiency, or violation of instructions). Furthermore, the speech corpus was restricted by excluding audio segments due to manifold reasons: a segment contains only silence; a segment does not include at least one digit; a segment contains disturbing noise while speaking, for example, caused by unintended gesticulation. In order to provide enough data per subject for various investigations, only subjects with at least 75% of valid audio segments were included in the corpus.

In order to determine the verbal response duration (Section 2.2), audio segments were annotated by two student assistants using time markers in the software Audacity (Team, 2012). This process involved to omit any sound including uttered content after the end of the last uttered digit. Afterwards, all duration values were double checked by another student assistant.

The secondary task efficiency, introduced in Section 2.2, constitutes a promising cognitive load indicator and, consequently, provides the basis for data labelling (Section 2.4). Since the audio data of the corpus were partly contaminated with information on the verbal response duration and thus partly with the secondary task efficiency, audio segments were further processed by trimming (see Figure 2). In more detail, energy threshold based audio activity detection was applied on the speech signal to obtain the onset of the first activity and the end of the last activity. The audio activity refers to any sound which can be caused by speech, breathing, filled pauses, lip-smacking, and so forth. Subsequently,

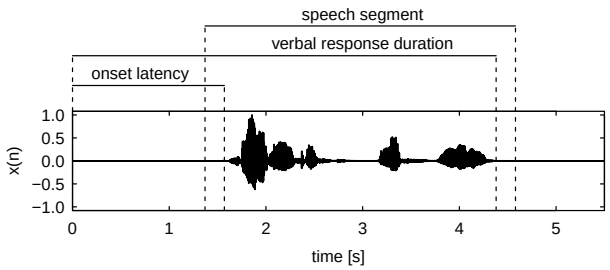

Figure 2: Onset latency, verbal response duration and the resulting segment of a recorded speech signal.

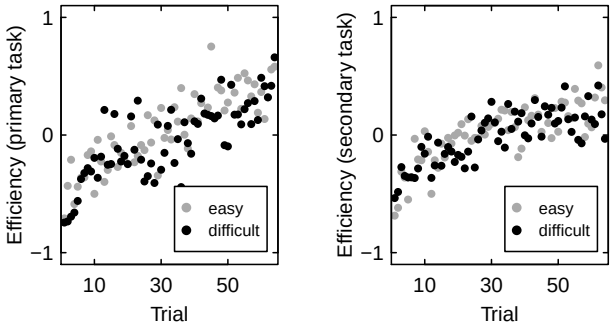

Figure 3: Standardized primary and secondary task efficiency over trials, averaged across subjects for each task difficulty.

a tolerance of 200 ms was added to the segmental boundaries of the activity detector to ensure that information on speakers' activity was not lost by detection errors. If the length of the tolerance exceeded the limits of the original audio segment, as much silence as needed was added to fill the 200 ms at the beginning and/or the end. The resulting segments were then transcoded to 16 kHz with a 16 bit resolution in mono WAV, which constitutes the audio format of the corpus.

2.4. Cognitive Load Labels

As pointed out in Section 2.1, performance measures of the secondary task can be used as a reliable and valid reference for cognitive load associated with the primary task. Furthermore, the study of Wirzberger et al. (Wirzberger et al., 2017b)—the basis for the CoLoSS corpus—backed up the hypothesis that as learning progresses with the sequence of trials, the subjects' efficiency increases concerning the primary as well as secondary task. Hence, the variable of interest for data labelling comprises the secondary task efficiency, which considers performance (word accuracy) as well as effort (time required). Again, this label assignment is linked to the following assumed relationship: Secondary task efficiency reflects the amount of the speaker's cognitive resources devoted to performing the secondary task. The higher the load imposed by cognitive learning processes in terms of the primary task, the lower the efficiency score of the secondary task.

A second variant of cognitive load labels was realized by performing a discretization of the numeric values. For instance, in this way, classification models can be trained for the automatic assessment of cognitive load as an alternative approach to the regression problem. The labels were transformed into nominal values with three distinct cognitive load levels by equal-width binning. Note, this method is an experimental approach for another representation of the original labels. The appropriate separation of numeric indicators into cognitive load classes constitutes an open issue for future work.

2.5. Data Description

Statistics of the constructed corpus are given in Table 1. It includes 70 native speakers of German, whereby 18 are male and 52 are female (9 male and 26 female per task difficulty). Due to the exclusion of some speech files (cf. Section 2.3), the number of instances varies across subjects (min = 48, max = 64, μ = 58.23, σ = 18.41). In the following, the corpus material is given at a glance:

- Audio files (WAV, mono, 16 kHz, 16 bit) containing German speech (digits) from secondary task trials
- Subject id, trial id, and information about the gender.
- Primary task condition assignment (easy and difficult)
- Primary task performance measures (symbol response correctness, reaction time, efficiency)
- Secondary task performance measures (word accuracy, verbal response duration, efficiency)
- Cognitive load labels (secondary task efficiency as numeric and nominal values)

The progression in primary task and secondary task efficiency over trials, averaged across subjects, is illustrated in Figure 3. For a deeper analysis, linear mixed-effects models with z-standardized predictors were used to consider individual effects (condition, interaction between trial and condition, and efficiency across trials per subject). Results confirm an increasing performance across the primary task, β = .273, p < .001, RMSE = 1.003, R^2 = .178, as well as the secondary task, β = .186, p < .001, RMSE = 0.983, R^2 = .478. Neither for the primary task nor the secondary task significant differences between conditions were observed. The RMSE was obtained from a leave-one-subject-out cross validation approach, whereas the R^2 resulted from a Pseudo-R^2 procedure, taking into account random effects in linear mixed-effect models.

Since the efficiency score of the secondary task was suggested for data labeling, the underlying parameters are described in more detail: Regarding the word accuracy (WA), in almost all cases, the response of the five-digit sequence was error free (WA = 1) with a frequency of 3,927 whereas the lowest frequency of 2 occurrs at WA = 0. Such confirms that the secondary task is rather simple so that it does not tend to distract subjects from working on the primary task. Considering all WA values, a mean of 0.94 and standard deviation of 0.15 is obtained. For the verbal response duration, the skewness is 0.81 and the kurtosis is 1.41 indicating that the distribution (min = 0.83, max = 5.5,

	Condition		
Description	easy	difficult	all
Number of subjects	35 (9 M, 26 F)	35 (9 M, 26 F)	70 (18 M, 52 F)
Number of instances	1,993	2,083	4,076
Average number of instances per subject	56.94	59.51	58.23
Average duration per instance [s]	2.68	2.65	2.66
Total duration [hh:mm]	01:29	01:32	03:01

Table 1: Data description of the CoLoSS corpus.

$\mu = 2.85$, $\sigma = 0.68$) is slightly skewed to the right with heavier tails and a sharper peaks than the normal distribution. Similar characteristics are given by the distribution of the efficiency scores (min = 0, max = 0.86, $\mu = 0.35$, $\sigma = 0.11$) where the skewness is 0.21 and the kurtosis is 1.15.

As described in Section 2.4, a discretized version of the numeric labels is included in the corpus. With respect to the assumptions concerning secondary task efficiency (Eff_{ST}) and by involving all conditions, the following three cognitive load (CL) classes were obtained:

$$\text{CL}(\text{Eff}_{\text{ST}}) = \begin{cases} L_1 & \text{for} \quad 0.58 < \text{Eff}_{\text{ST}} \leq 0.86 \\ L_2 & \text{for} \quad 0.29 < \text{Eff}_{\text{ST}} < 0.58 \\ L_3 & \text{for} \quad 0 \leq \text{Eff}_{\text{ST}} < 0.29 \end{cases}$$

where L_1, L_2, and L_3 represent the low, medium, and high cognitive load level, respectively. Note—for clarity reasons—the shown ranges of Eff_{ST} values are rounded to two decimals; the actual values are more accurate. The resulting distribution among classes is highly unbalanced (L_1: 109, L_2: 3.051, and L_3: 916). Therefore, it is strongly recommended to apply resampling techniques before classification models are trained.

3. Effects of Cognitive Load on Speech

Effects of cognitive load on speech were investigated by analyzing means and 95% confidence intervals of six different parameters under three different cognitive load levels (L_1, L_2, and L_3; cf. Section 2.5). Significance across these levels was tested using post-hoc pairwise t-tests with Bonferroni-Holm correction (Holm, 1979), following analyses of variance (ANOVAs) with a significance-level of $\alpha = .05$.

3.1. Feature Extraction

In this section, six common speech-related parameters, that were investigated in conjunction with the CoLoSS corpus, are introduced. Two phoneme-based as well as two acoustic prosodic features and two voice quality features were extracted:

- *Articulation rate*: This rate describes the tempo in speech using the total number of syllables divided by the total duration of the utterance excluding silent pause duration.
- *Silent pause duration*: The total duration of silent pauses within an utterance is determined; it can be an indicator for disfluency in speech.
- *Intensity*: This parameter was computed by the root mean square energy of a signal and can be understood as the acoustic equivalent to the perceptual quantity loudness.
- F_0: The fundamental frequency F_0 represents the frequency of the vocal fold vibration and can be regarded as the acoustic equivalent to the perceptual unit pitch.
- *Jitter* and *shimmer*: Both parameters are the most common descriptors that characterize the voice quality. While jitter is defined as the period-to-period variation in vocal fold frequency, shimmer refers to the period-to-period variation in the amplitude of a voice.

The intensity, F_0, jitter, and shimmer were determined using the analysis tool Praat (Boersma and others, 2002). Afterwards, means of feature contours were computed for each instance. For computing the articulation rate as well as silent pause duration for each instance, the phoneme-based feature extraction system, introduced in (Herms, 2016), was applied. In order to consider phonemes of German language in the phoneme-based feature extractor, an acoustic model including 43 phonemes was trained on the basis of the *German open source corpus for distant speech recognition* (Radeck-Arneth et al., 2015). The resulting acoustic model consists of context-dependent triphone Hidden-Markov-Models with 32 Gaussians per state.

In order to remove the inter-speaker variability, Z-score normalization was applied within each speaker context. For this purpose, the features were adjusted with a mean of 0 and a standard deviation of 1 across instances for each subject separately.

3.2. Results

Figure 4 shows the means and 95% confidence intervals of prosodic and voice quality features under different cognitive load (CL) levels.

Results concerning the mean values of the articulation rate show a significant separation between low and medium ($p < .001$) as well as low and high cognitive load ($p < .001$). Between medium and high load, a difference cannot be observed for this parameter ($p > .05$); confidence intervals overlap. For silent pause duration, there is a statistically significant increase from low to high ($p < .001$) and medium to high cognitive load ($p < .001$). Nevertheless, an overall linear trend cannot be derived across cognitive load levels due to a slight drop from low to medium load.

Regarding the distribution of the intensity, no significant differences between groups can be observed (all $p > .05$).

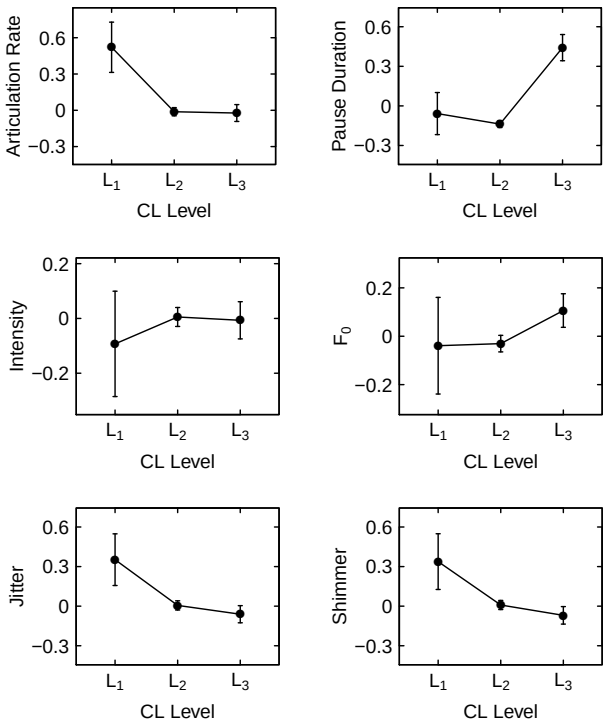

Figure 4: Means and 95% confidence intervals of speech parameters under different cognitive load (CL) levels.

From the visual impression in Figure 4, there are only marginal differences between the mean values of intensity as cognitive load increases and, moreover, confidence intervals overlap completely across cognitive load levels. The mean values of F_0 exhibit a statistically significant difference between medium and high cognitive load ($p < .001$). On the other hand, confidence intervals of F_0 overlap between low and medium as well as low and high cognitive load.

In case of the voice quality features jitter and shimmer, a monotonically decreasing trend can be observed for both parameters as the level of cognitive load increases. More precisely, the reduction of jitter and shimmer from low to medium cognitive load exhibit a statistically significant difference (both $p < .001$), whereas from medium to high cognitive load, a significant difference was obtained only for the parameter shimmer ($p < .001$). The results of both voice quality features indicate that speech includes less rough or hoarse characteristics as cognitive load increases.

4. Conclusion

We presented a new corpus named CoLoSS, which contains speech under cognitive load recorded in a learning task scenario. We used a dual-task approach to determine subjects' residual cognitive resources reflecting the degree of cognitive load. It comprises a visual-motor primary task that required subjects to learn abstract symbol combinations and an auditory-verbal secondary task to measure the load imposed by the primary task. This paper reports the methodology of collecting the speech recordings, constructing the corpus and gives a description of the data. Furthermore, effects of cognitive load on prosodic as well as voice quality features have been investigated in conjunction with the speech data of corpus.

For future work, we plan to conduct experiments aimed at the automatic speech-based recognition of cognitive load using both the numeric as well as the class labels of the corpus.

5. Acknowledgements

The presented research was conducted within the Research Training Group "CrossWorlds - Connecting virtual and real social worlds" (GRK 1780/1). The authors gratefully acknowledge funding by the German Research Foundation (DFG).

6. Bibliographical References

Baddeley, A. (1992). Working memory. *Science*, 255(5044):556.

Boersma, P. P. G. et al. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5.

Boril, H., Omid Sadjadi, S., Kleinschmidt, T., and Hansen, J. H. (2010). Analysis and detection of cognitive load and frustration in drivers' speech. *Proceedings of INTERSPEECH 2010*, pages 502–505.

Gorovoy, K., Tung, J., and Poupart, P. (2010). Automatic speech feature extraction for cognitive load classification. In *Conference of the Canadian Medical and Biological Engineering Society (CMBEC)*.

Herms, R. (2016). Prediction of deception and sincerity from speech using automatic phone recognition-based features. In *INTERSPEECH*, pages 2036–2040.

Hoffman, B. and Schraw, G. (2010). Conceptions of efficiency: Applications in learning and problem solving. *Educational Psychologist*, 45(1):1–14.

Holm, S. (1979). A simple sequentially rejective multiple test procedure.

Paas, F. G. and Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational psychology review*, 6(4):351–371.

Paas, F., Tuovinen, J. E., Tabbers, H., and Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, 38(1):63–71.

Plass, J. L., Moreno, R., and Brünken, R. (2010). *Cognitive load theory*. Cambridge University Press.

Radeck-Arneth, S., Milde, B., Lange, A., Gouvêa, E., Radomski, S., Mühlhäuser, M., and Biemann, C. (2015). Open source german distant speech recognition: Corpus and acoustic model. In *International Conference on Text, Speech, and Dialogue*, pages 480–488. Springer.

Schuller, B. W., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., and Zhang, Y. (2014). The interspeech 2014 computational paralinguistics challenge: cognitive & physical load. In *INTERSPEECH*, pages 427–431.

Sweller, J., Ayres, P., and Kalyuga, S. (2011). *Cognitive load theory*, volume 1. Springer.

Team, A. (2012). Audacity (version 2.0. 2). Retrieved from https://sourceforge.net/projects/ audacity/.

Wirzberger, M., Bijarsari, S. E., and Rey, G. D. (2017a). Embedded interruptions and task complexity influence schema-related cognitive load progression in an abstract learning task. *Acta Psychologica*, 179:30–41.

Wirzberger, M., Herms, R., Bijarsari, S. E., Rey, G. D., and Eibl, M. (2017b). Influences of cognitive load on learning performance, speech and physiological parameters in a dual-task setting. In *Abstracts of the 20th Conference of the European Society for Cognitive Psychology*, page 161.

Yap, T. F., Epps, J., Ambikairajah, E., and Choi, E. H. (2010). An investigation of formant frequencies for cognitive load classification. In *Eleventh Annual Conference of the International Speech Communication Association*.

Yap, T. F. (2012). *Speech production under cognitive load: Effects and classification*. Ph.D. thesis, University of New South Wales, Sydney, Australia.

Yin, B., Ruiz, N., Chen, F., and Khawaja, M. A. (2007). Automatic cognitive load detection from speech features. In *Proceedings of the 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces*, pages 249–255. ACM.