Acta Acustica 2020, **4**, 3

https://doi.org/10.1051/aacus/2020001

Available online at:
https://acta-acustica.edpsciences.org

ORIGINAL ARTICLE OPEN ACCESS

Effect of phoneme variations on blind reverberation time estimation

Andrea Andrijašević*

Faculty of Engineering, University of Rijeka, Vukovarska 58, 51000 Rijeka, Croatia

Abstract – This study focuses on an unexplored aspect of the performance of algorithms for blind reverberation time (T) estimation – on the effect that speech signal's phonetic content has on the value of the estimate of T that is obtained from the reverberant version of that signal. To this end, the performance of three algorithms is assessed on a set of logatome recordings artificially reverberated with room impulse responses from four rooms, with their T_{20} value in the [0.18, 0.55] s interval. Analyses of variance showed that the null hypotheses of equal means of estimation errors can be rejected at the significance level of 0.05 for the interaction terms between the factors "vowel", "consonant", and "room", while the results of Tukey's multiple comparison procedure revealed that there are both some similarities in the behaviour of the algorithms and some differences, where the latter are stemming from the differences in the details of algorithms' implementation such as the number of frequency bands and whether T is estimated continuously or only on the selected, the so-called speech decay, segments of the signal.

Keywords: Reverberation time, Estimation, Speech, Logatome, Phoneme, Voiced onset time, ANOVA

1 Introduction

Reverberation time (T) is one of the most important objective measures indicating the severity of reverberation in an enclosure. It represents the value of time it takes for a steady-state sound energy level to gradually decay by 60 dB after an abrupt cessation of the sound source [1]. Since acoustic absorption coefficients of room boundaries, objects, and contained air are all frequency dependent, reverberation time is also frequency dependent and, hence, commonly measured in octave or 1/3 octave bands using existing standardised methods [1–3].

Due to detrimental effects reverberation has on speech signals captured when microphones (e.g., the ones in personal electronic devices such as mobile phones, videoconference systems, and hearing aids) are positioned in the far field of a sound source (e.g., a person speaking), a number of algorithms for reverberation time estimation from the received reverberant speech signals – the so-called *blind algorithms* – have been developed in the last two decades. An estimate of reverberation time obtained with these algorithms can then be used in the process of speech dereverberation for either speech enhancement or speech/speaker recognition purposes [4–6].

One of possibly many classifications of these algorithms, a classification that is based on their complexity, was proposed in a paper by Eaton et al. [7], where the algorithms were grouped into three classes: Analytical with or without Bias Compensation (ABC) class, Single Feature with Mapping (SFM) class, and Machine Learning with Multiple Features (MLMF) class.

Algorithms of the ABC class are based on a stochastic time-domain model of room impulse responses (RIRs), with no information on speech signal characteristics included. One of the first algorithms of this class was developed by Ratnam et al. [8, 9]. The variance of blind T estimates was reduced with the introduction of a sound decay detection step first proposed by Kendrick et al. [10], and later by Löllmann et al. [11] who developed a different variant of the step. Further work presented modifications either to the model or to the preprocessing of the reverberant signal [12–16].

For the second class, SFM, there exists an intermediate feature whose value is estimated and from which, via a polynomial mapping function, a reverberation time estimate is obtained. This feature is commonly a statistic related to the distribution of decay rates obtained in the time–frequency (TF) domain [17–21]. The last class, MLMF, contains classifiers trained on a large speech corpus that was artificially reverberated with a set of room impulse responses [12, 22, 23].

In the same paper, "The ACE Challenge" [7], both the most recent and the most comprehensive comparative study of the performance of blind reverberation time estimation algorithms, the performance of 25 algorithms was assessed on a set of reverberant speech signals. Since in

*Corresponding author: andrea.andrijasevic@riteh.hr

the Challenge the main objective was to assess the influence of the additive noise term on estimation quality, reverberant speech signals had various types and levels of measured noise added to them. The results showed that the machine learning algorithms are being outperformed by the less complex algorithms of the ABC and the SFM class. Interestingly, the results of the ACE Challenge are in agreement with the results of an earlier investigation conducted by Shabtai et al. [24], which indicated that a simpler approach for blind room volume estimation, the one that uses abrupt stops present in the reverberant speech signal, outperforms more complex approaches based on dereverberation or the use of speech recognition features.

Another very comprehensive and equally important study was performed by Kinoshita et al. [25], whose results stress the importance of having an accurate blind estimate of T for the performance of a class of noise robust state-of-the-art speech enhancement algorithms. In their paper [25], the performance of 25 algorithms for speech enhancement (SE) and of 49 algorithms for automatic speech recognition (ASR) was assessed on artificially reverberated speech utterances (T values of 0.3, 0.6, and 0.7 s were considered) as well as speech recordings from a room with T of 0.7 s. For the SE task, the quality of dereverberated speech signals was evaluated using both the objective measures and a listening test. When one-channel algorithms were considered, the second best performing SE algorithm was from the "Statistical RIR modelling" class – a class that needs a blind estimate of reverberation time, which is then used in a statistical, time-domain model of room impulse responses. Finally, the dereverberated speech signal is obtained with spectral subtraction – a technique of low computational complexity that is robust against additive noise.

Also in 2016, the results of a short study inspired by Ratnam's and Kendrick's initial, one speech recording-based, observations that speech signal's offsets introduce errors in the blind T estimation process [8, 12] were presented in a conference paper [26]. Data indicated that there exists a relationship between the values of estimates obtained with one blind T estimation algorithm and the speech signal, asking for a more in-depth and extensive follow-up study to be conducted – a study that could answer the question of whether this was just an isolated case or a general rule. Therefore, the main objective of this study is to explore the outputs of several state-of-the-art blind reverberation time estimation algorithms, and discover how they are related to the characteristics of the speech signal – the type and order of the phonemes and, additionally, speaker's sex. To this end, three of the best performing algorithms from the ACE Challenge were selected for the statistical analysis of their T estimates.

The remainder of this paper is structured as follows: in Section 2, the speech corpus, room impulse responses, and algorithms for blind reverberation time estimation are introduced. This section closes with the details of the T acquisition steps. The results of the analyses of variance and multiple comparison tests of reverberation time estimation errors are presented in Section 3 and discussed in Section 4. Finally, Section 5 summarises the major findings and discusses the objective, speech corpus related, limitations of this study.

2 Methodology

In this section, the speech corpus and room impulse responses are introduced and their characteristics presented. After that, the principles of operation of the three selected state-of-the-art algorithms for blind reverberation time estimation are stated. The section closes with the details of the blind reverberation time estimate acquisition process.

2.1 Speech corpus

The OLdenburg LOgatome (OLLO) corpus version 2.0 [27], a large speech corpus freely available for research purposes that holds recordings of 150 logatomes uttered by 50 speakers of both sexes (25 women), was selected as the most appropriate corpus for this investigation's objectives. It consists of a set of 80 logatomes of the consonant–vowel–consonant (CVC) form and a set of 70 vowel–consonant–vowel (VCV) logatomes, where each of these 150 logatomes was uttered three times by 40 German and 10 French speakers in their normal speaking style. The logatomes that make those two sets are presented in Tables 1 and 2.

In this study, the VCV set of logatomes will stand as a lexical representative for the languages with a predominance of words finishing with an open syllable (such as Italian), while the CVC set will perform an equivalent function for the languages with a predominance of words finishing with a closed syllable, one of them being the English language [28, 29].

2.2 Room impulse responses

Thirty-two room impulse responses were taken from the Aachen Impulse Response (AIR) database v1.4 [30], of which six were measured in the room named "booth", another six in the "office" room, while 10 impulse responses were from the "meeting" room and another 10 from the "lecture" room. In addition to these room impulse responses from the AIR database, three room impulse responses measured in the listening room of the Department of Electroacoustics, at the Faculty of Electrical Engineering and Computing, University of Zagreb (Croatia), were also used. All of these 35 RIRs were obtained without the use of a dummy head and with a sampling rate of 48 kHz. For each RIR, the associated ground truth value of reverberation time T_{20} was obtained from its full-band energy decay curve (EDC) in the $[-5, -25]$ dB interval, and in accordance with the ISO3382 standard [1]. The average full-band T_{20} values for those five rooms were found to be as follows: "booth" – 0.18 s, "meeting" room – 0.28 s, "listening" room – 0.47 s, "office" – 0.55 s, and "lecture" room – 0.82 s.

According to a relatively recent reverberation time measurement study done by Diaz and Pedrero [31], in

Table 1. CVC set of logatomes. For a given row, ten logatomes are constructed using the pair of the 1st and 3rd phoneme combined with one of the vowels given in the 2nd phoneme column.

1st phoneme	2nd phoneme	3rd phoneme
/b/	Vowel group V1: /a/, /ɛ/, /ɪ/, /ɔ/, /ʊ/	/p/
/p/		/p/
/d/		/t/
/t/		/t/
/g/	Vowel group V2: /a:/, /e/, /i/, /o/, /u/	/k/
/k/		/k/
/f/		/f/
/z/		/s/

Table 2. VCV set of logatomes. For each of the 14 consonants in the 2nd phoneme column, five logatomes are constructed using the row pairs of the 1st and 3rd phoneme.

1st phoneme	2nd phoneme	3rd phoneme
/a/	/p/, /t/, /k/, /b/, /d/, /g/, /f/, /s/, /ʃ/, /v/,	/a/
/ɛ/	/m/, /n/, /ʦ/, /l/	/ɛ/
/ɪ/		/ɪ/
/ɔ/		/ɔ/
/ʊ/		/ʊ/

which the authors presented the values of reverberation time for more than 8000 bedrooms and 3000 living rooms, the average value of reverberation time for furbished spaces can be found in the [0.26, 0.77] s interval; the average T value across rooms was smallest in the 4 kHz octave band for the rooms in the 10–20 m^3 range (0.26 s), while the largest average value of T was present for the rooms in the 90–100 m^3 range at the 125 Hz octave band (0.77 s). Thus, it can be concluded that the selected set of measured RIRs with its ground truth values of reverberation time in the [0.18, 0.82] s interval covers well the interval of the most common T values found in the living spaces nowadays.

2.3 Blind reverberation time estimation algorithms

The first algorithm for blind reverberation time estimation used in this study was developed by Eaton et al. [18]. In this algorithm, the sound decay rates are estimated on the mel-STFT signal representation for each mel-frequency band separately by applying a least-squares linear fit to the log-energy envelopes of the signal. It was shown that the negative-side variance (NSV), defined as the variance of the negative gradients in the distribution of decay rates, is related to reverberation time and is, therefore, used as the input to a polynomial mapping function whose output is the estimate of reverberation time. For this algorithm, the estimated level of signal-to-noise ratio (SNR) defines which decay rates will be chosen to form the aforementioned distribution.

In this investigation, a copy of the Matlab implementation of the algorithm [32], provided on-line by its authors, was used. During the estimation process, the values of the parameters were not changed from the originally set values and were as follows: frame length – 16 ms, overlap – 25%, and the number of mel-frequency bands – 31. Since in this study no noise was added to the reverberant signal, and in order to make this algorithm's results comparable to the results of the remaining two algorithms that do not adapt to the changes in the SNR, the Matlab code was adjusted so that it automatically gives an estimate of T as if the estimated SNR had the highest possible value. This is the mode of operation where all the negative gradients (from all the mel-bands and time frames) form the distribution from which the NSV is calculated.

In the second algorithm, developed by Prego et al. [20, 21], the blind estimation of reverberation time is carried out in the STFT domain for each frequency band independently on detected speech free-decay regions. Estimation of T is performed with a procedure that closely resembles Schroeder's standardized method with the upper decibel limit of the energy decay curve calculated for a detected speech free-decay region set to –5 dB. The lower limit is set by finding the decibel range with the highest value of the regression coefficient when the ranges −60, −40, −20, and −10 dB are considered, and with that order of preference. The final reverberation time estimate is obtained as the output of a linear mapping function, which serves to correct for the differences in the dynamic ranges of the detected free decays, with the median value of all previously calculated sub-band T estimate medians serving as its input.

An implementation of the algorithm in the form of a Matlab script, with no SNR compensation step included, was kindly provided by its authors. During the algorithm evaluation process, the values of the parameters were not changed from the original ones and were as follows: frame length – 50 ms, overlap – 25%, and the number of frequency bands – 1025.

The last algorithm whose sensitivity to speech signal characteristics is assessed, developed by Löllmann et al. [11], is an algorithm in which the estimates of reverberation time are acquired from the time-domain version of the input reverberant speech signal. For this algorithm as well, the detected speech decay sequences serve as an approximation of the true RIR. Since RIRs are modelled with Polack's stochastic model [4], the reverberation time estimate is obtained as the parameter value for which the likelihood of the observed decay sequence is maximal. In order for the speech decays to be detected, the reverberant signal is processed using 2 s long overlapping rectangular windows with 20 ms shift. Each frame is then partitioned into 30 subframes of equal length and their respective energies are calculated, upon which it is checked for a monotonous decrease in energy between at least three adjacent subframes starting from the first. If a monotonous decrease in energy is found, the corresponding subframes are used for the maximum-likelihood estimation of reverberation time.

In this study, an implementation of the algorithm was created in Matlab by following the steps outlined in

reference [11]. In order to reduce the variance of the estimates, only the estimates obtained from the speech decay sequences of dynamic range higher or equal to 20 dB were kept. After quantization of those estimates in steps of 25 ms, a histogram was formed, and the final estimate calculated as the mode of the distribution.

2.4 Data acquisition

During the initial testing phase, it was observed that the two algorithms operating in the time–frequency domain need longer recordings than a speech file that contains the three concatenated repetitions of a logatome belonging to one speaker is long (about 5 s) in order to give a T estimate. Therefore, for women and men separately, the three repetitions of a logatome of 25 speakers were concatenated into a single speech file. The resulting 300 speech files (150 for women and 150 for men) were then artificially reverberated using each of the 35 measured RIRs. Before their convolution with the speech files, the room impulse responses were down-sampled to 16 kHz in order to conform to this lower sampling rate of the speech corpus. Finally, blind estimation of reverberation time was carried out using the algorithms from Section 2.3, after which estimation errors were calculated as the difference between the estimate and the ground truth T_{20} value of the room impulse response.

The inspection of the estimation errors revealed that two of the algorithms performed quite unsatisfactorily for the RIRs from the room with the largest ground truth T_{20} ("lecture"); the first one with a mean error of 0.2 s and a 2.5 ms wide 95% confidence interval, and the second one with a mean error of 0.45 s and a 2 ms wide 95% confidence interval. Given these problems with accuracy, it was decided not to use any of the estimates obtained from the "lecture" room in the following analyses of variance (ANOVAs).

After the data for the "lecture" room were removed, six datasets were created in Matlab – one for each algorithm and logatome set (CVC/VCV) combination, and with the following four independent variables, i.e. factors:

- "speakers' sex" (levels: F, and M),
- "consonant" present in the logatome:
  - CVCs – levels: /p/, /t/, /k/, /b/, /d/, /g/, /f/, and /z/ (i.e., the left consonant of a logatome),
  - VCVs – levels: /p/, /t/, /k/, /b/, /d/, /g/, /f/, /s/, /ʃ/, /v/, /m/, /n/, /ʦ/, and /l/,
- "vowel" present in the logatome:
  - CVCs – levels: /a/, /ɛ/, /ɪ/, /ɔ/, /ʊ/, /a:/, /e/, /i/, /o/, and /u/,
  - VCVs – levels: /a/, /ɛ/, /ɪ/, /ɔ/, and /ʊ/,
- "room" (levels: "booth", "meeting", "listening", and "office"),

and with the dependent, that is, response, variable being the reverberation time estimation error.

Finally, before the analyses of variance were performed, the normality of data was verified for these six datasets using the Jarque-Bera test with the significance level set to 0.05.

3 Analysis of variance

Four-way ANOVA statistical tests with interaction terms, α value set to 0.05 and sum of squares type 3, were performed on the aforementioned six datasets with the factors being "speakers' sex", "consonant", "vowel", and "room".

Although some of the interactions of the factor "speakers' sex" with other factors were statistically significant ($p < 0.05$), Tukey's multiple comparison tests revealed that those interactions were ordinal while the values of the effect size measure η^2 of the interactions of this factor with the remaining three factors were in the [0.07, 2.22]% interval. Furthermore, across the datasets, the differences between the marginal means of the female (F) and male (M) level were in the [10, 30] ms range and with the "speakers' sex" factor value of effect size measure η^2 in the [0.5, 4.58]% interval.

Since those differences were relatively small, both when compared with the order of magnitude of the estimation errors occurring when standardised methods for reverberation time measurement are utilized, and when compared with the differences observed between the marginal means of the levels of other factors, ANOVAs were performed again, but this time without the "speakers' sex" factor. The degrees of freedom (df), the F statistic and p value as well as the measure of effect size η^2 for each factor and their interaction terms are presented in Table 3 for the CVC set and in Table 4 for the VCV set.

3.1 CVC set of logatomes

Figures 1–3 display the marginal means for the three interactions ("consonant" × "room", "vowel" × "room", and "vowel" × "consonant") for the CVC set of logatomes. For Eaton's algorithm (Fig. 1a), for the first three rooms, a pattern is present – the CVCs of the /fVf/ and /zVs/ type give estimates of reverberation time that are significantly higher in value than the ones for the CVCs that contain plosives, which is reflected in the η^2 of 9.58 for the "consonant" factor. Furthermore, it can be observed that the estimates for the plosive CVCs are not significantly different between each other, indicating that it is primarily the manner of articulation of the consonants that defines the value of the estimate. The results for the last room ("office") are somewhat different – indicating the sensitivity of this algorithm to larger values of reverberation time, and are responsible for a $p < 0.001$ for the "consonant" × "room" interaction, with an η^2 of 4.26.

As for the "vowel" × "room" interaction (Fig. 1b), small and generally statistically non-significant upward shifts in the value of the marginal means are present between the adjacent vowels for the first three rooms, while for the "office" room the shifts are larger and are causing a $p < 0.001$ for this interaction term but with a small η^2 of 2.82.

Finally, Figure 1c shows the "vowel" × "consonant" interaction (with a small effect size of $\eta^2 = 2.01$), indicating that when averaging across rooms the distance between the

Table 3. CVC set of logatomes: ANOVA tables for the three algorithms.

Source	*df*	Eaton			Prego			Löllmann		
		F	$p > F$	η^2, %	F	$p > F$	η^2, %	F	$p > F$	η^2, %
Consonant	7	184.14	<0.001	9.58	273.45	<0.001	14.39	146.82	<0.001	11.04
Vowel	9	115.85	<0.001	7.75	50.56	<0.001	3.42	73.31	<0.001	7.08
Room	3	1805.92	<0.001	40.28	1546.02	<0.001	34.86	545.48	<0.001	17.57
Vowel × Consonant	63	4.30	<0.001	2.01	16.77	<0.001	7.94	1.44	0.013	0.98
Consonant × Room	21	27.25	<0.001	4.26	17.01	<0.001	2.68	40.94	<0.001	9.23
Vowel × Room	27	14.05	<0.001	2.82	8.68	<0.001	1.76	20.00	<0.001	5.80
Error	3869									
Total	3999									

Table 4. VCV set of logatomes: ANOVA tables for the three algorithms.

Source	*df*	Eaton			Prego			Löllmann		
		F	$p > F$	η^2, %	F	$p > F$	η^2, %	F	$p > F$	η^2, %
Consonant	13	120.10	<0.001	13.66	73.31	<0.001	3.74	85.04	<0.001	15.08
Vowel	4	448.72	<0.001	15.70	2861.46	<0.001	44.97	11.33	<0.001	0.62
Room	3	660.01	<0.001	17.32	1058.83	<0.001	12.48	466.90	<0.001	19.11
Vowel × Consonant	52	2.97	<0.001	1.35	35.16	<0.001	7.18	1.92	<0.001	1.36
Consonant × Room	39	5.49	<0.001	1.87	1.81	0.002	0.28	18.27	<0.001	9.72
Vowel × Room	12	75.23	<0.001	7.90	66.61	<0.001	3.14	24.51	<0.001	4.01
Error	3376									
Total	3499									

marginal means for the fricative CVCs and the plosive CVCs depends on the central vowel.

The "consonant" × "room" interaction for Prego's algorithm is presented in Figure 2a. The marginal means for the /fVf/ and /zVs/ type of CVCs are significantly higher than the ones for the logatomes that contain plosives, as reflected in a $p < 0.001$ for the "consonant" term with $\eta^2 = 14.39$. Furthermore, the distance between the means depends on the value of reverberation time, and the marginal means for the CVC pairs that differ from each other only by the voicing of the first plosive (e.g., /dVt/ and /tVt/) are not statistically different.

From Figure 2b, it can be observed that the "vowel" × "room" interaction is similar to the one presented in Figure 1b – the marginal means increase from /a/ to /ʊ/, and from /a:/ to /u/. Figure 2c shows the "vowel" × "consonant" interaction, revealing a very consistent pattern across consonants for the vowels /a/, /ɛ/, /ɪ/, /a:/, /e/, and /i/, and with $\eta^2 = 7.94$ for this interaction term.

The panel showing "consonant" × "room" interactions for Löllmann's algorithm (Fig. 3a) indicates that the difference in the marginal means between the /fVf/ and /zVs/ CVCs and the CVC logatomes that contain plosives is significant and depends on the value of reverberation time (with $\eta^2 = 9.23$ for the interaction term and $\eta^2 = 11.04$ for the "consonant" term). Again, the place of articulation of a plosive is not a predictor of the value of the estimate, while the influence of vowels changes with the change in the true value of reverberation time (Fig. 3b), reflected in a $p < 0.001$ for the "vowel" × "room" interaction with $\eta^2 = 5.80$.

3.2 VCV set of logatomes

Figures 4–6 display the marginal means for the three interactions ("consonant" × "room", "vowel" × "room", and "vowel" × "consonant") for the VCV set of logatomes. Figure 4 gives results for Eaton's algorithm, and shows that for all four rooms the estimates obtained from the VCVs with an unvoiced plosive or the affricate /ts/ are significantly lower in value than the estimates obtained from the nasals /m/ and /n/, voiced fricative /v/ and liquid /l/, reflected in an η^2 of 13.66 for the "consonant" term and a very small η^2 value of 1.87 for the interaction with "room". The marginal means for the voiced plosives and unvoiced fricatives can be found in between.

As for the "vowel" × "room" interaction, Figure 4b shows that the T estimates are lower for the vowels /a/ and /ɛ/ than those for the remaining three vowels, and with the value of the difference between the vowels /a/ and /ʊ/ depending on the factor "room", where $\eta^2 = 7.90$ for the interaction term and $\eta^2 = 15.70$ for the "vowel" term.

Finally, the panel 4c indicates that the influence of consonants in the VCV type of logatomes is very consistent across vowels (and, therefore, the interaction term has a very small effect size $\eta^2 = 1.35$) – the VCV logatomes that contain either an unvoiced plosive or the affricate /ts/ produce significantly lower estimates of reverberation time than the VCVs with a nasal /m/ or /n/, voiced fricative /v/ or the liquid /l/, while the estimates for the remaining central consonants from Table 2 can be found in between.

The results of the multiple comparison test for Prego's algorithm are presented in Figure 5. For this algorithm,

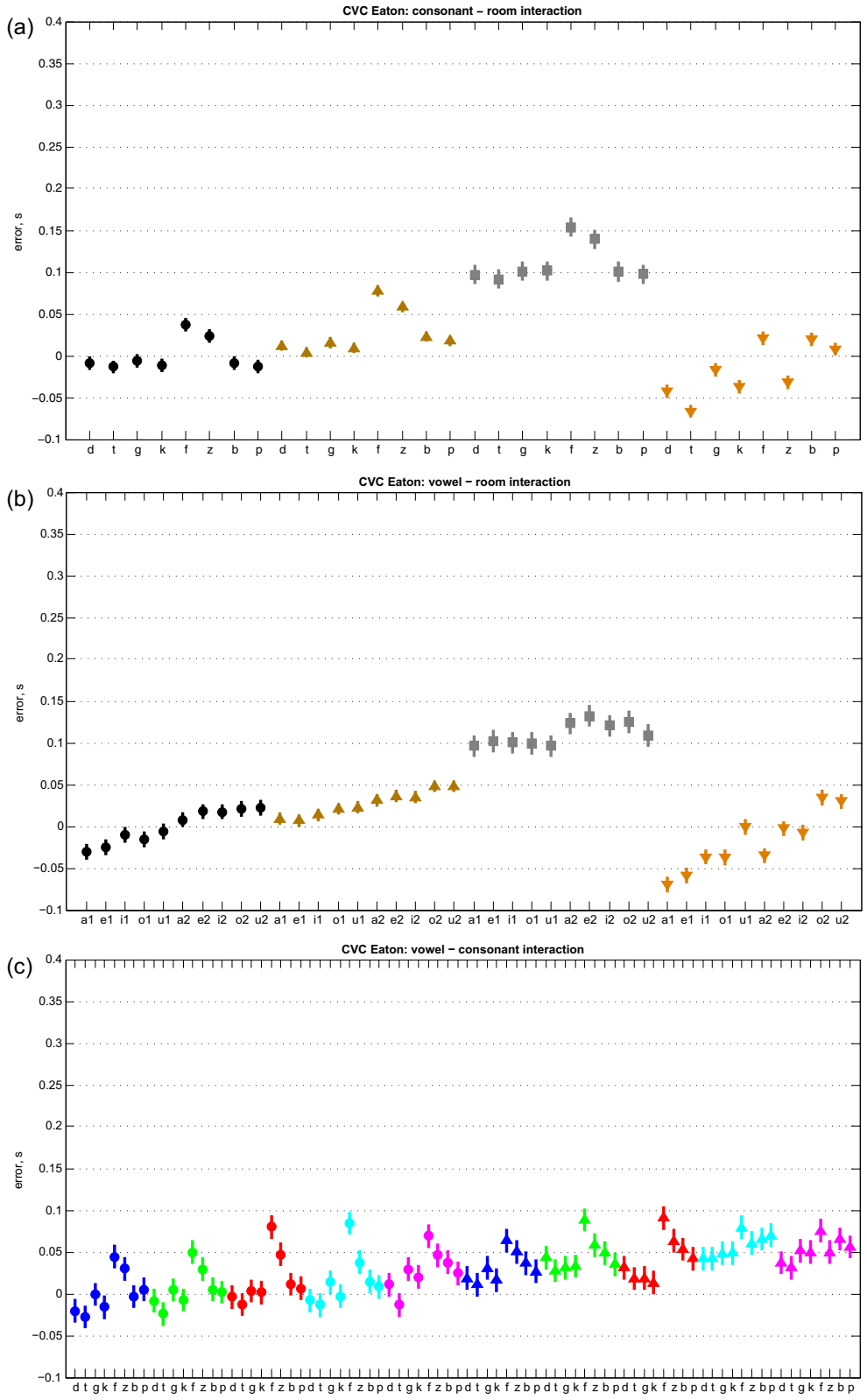

Figure 1. CVC logatomes: marginal means for Eaton's algorithm. (a) "Consonant" × "room" interaction, (b) "vowel" × "room" interaction, (c) "vowel" × "consonant" interaction. *X*-axis details: For the panels (a) and (c) left consonants of the CVCs are denoted on the *x*-axis. For the panel (b) vowel group V1 (/a/, /ɛ/, /ɪ/, /ɔ/, /ʊ/) as "a1"–"u1" and vowel group V2 (/a:/, /e/, /i/, /o/, /u/) as "a2"–"u2". Symbols for the marginal means: For the panels (a) and (b) "booth" – circle, "meeting" – upward triangle, "listening" – square, "office" – downward triangle. For the panel (c) vowel group V1 (/a/, /ɛ/, /ɪ/, /ɔ/, /ʊ/) – circle and vowel group V2 (/a:/, /e/, /i/, /o/, /u/) – upward triangle.

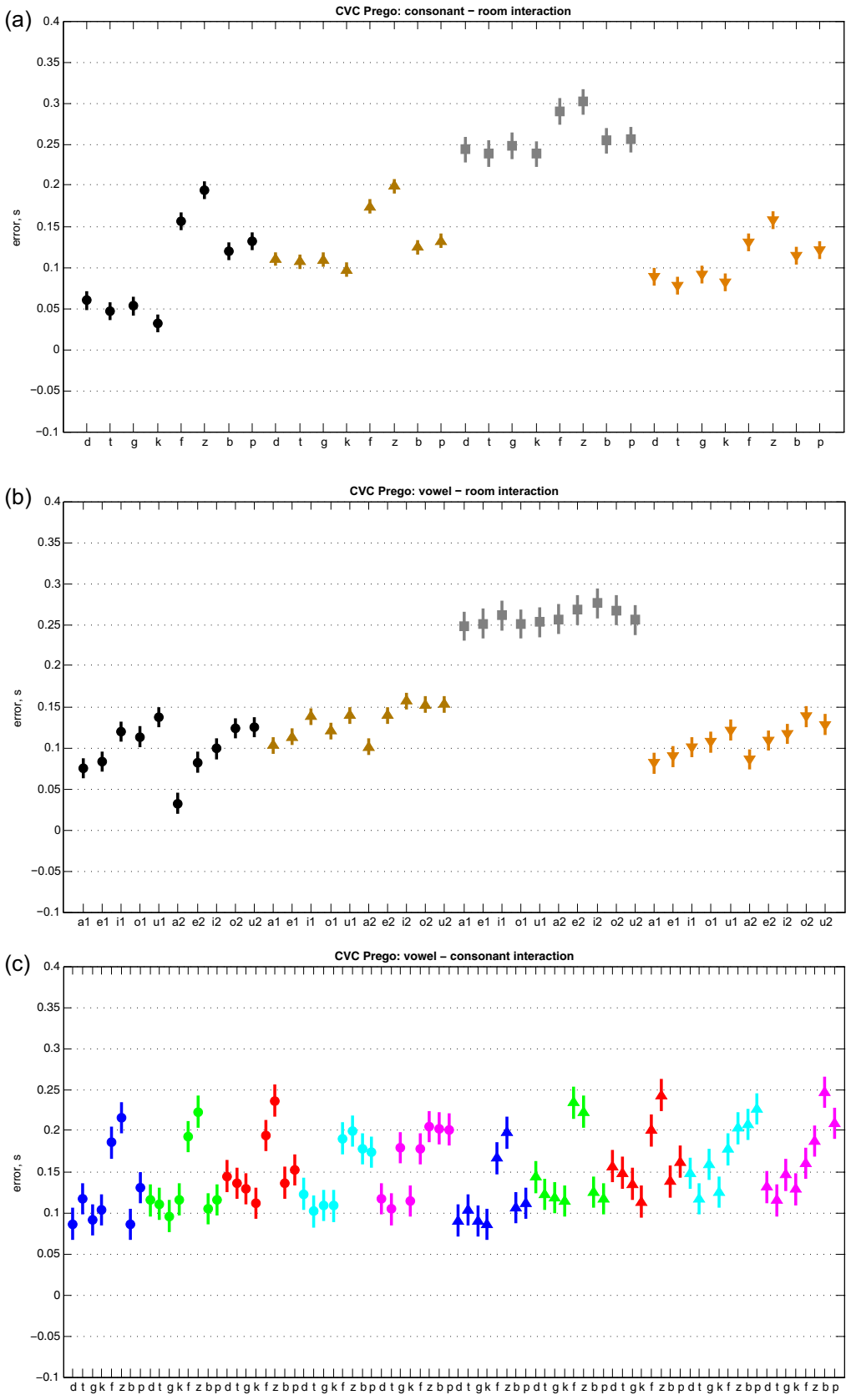

Figure 2. CVC logatomes: marginal means for Prego's algorithm. (a) "Consonant" × "room" interaction, (b) "vowel" × "room" interaction, (c) "vowel" × "consonant" interaction. For more details see Figure 1.

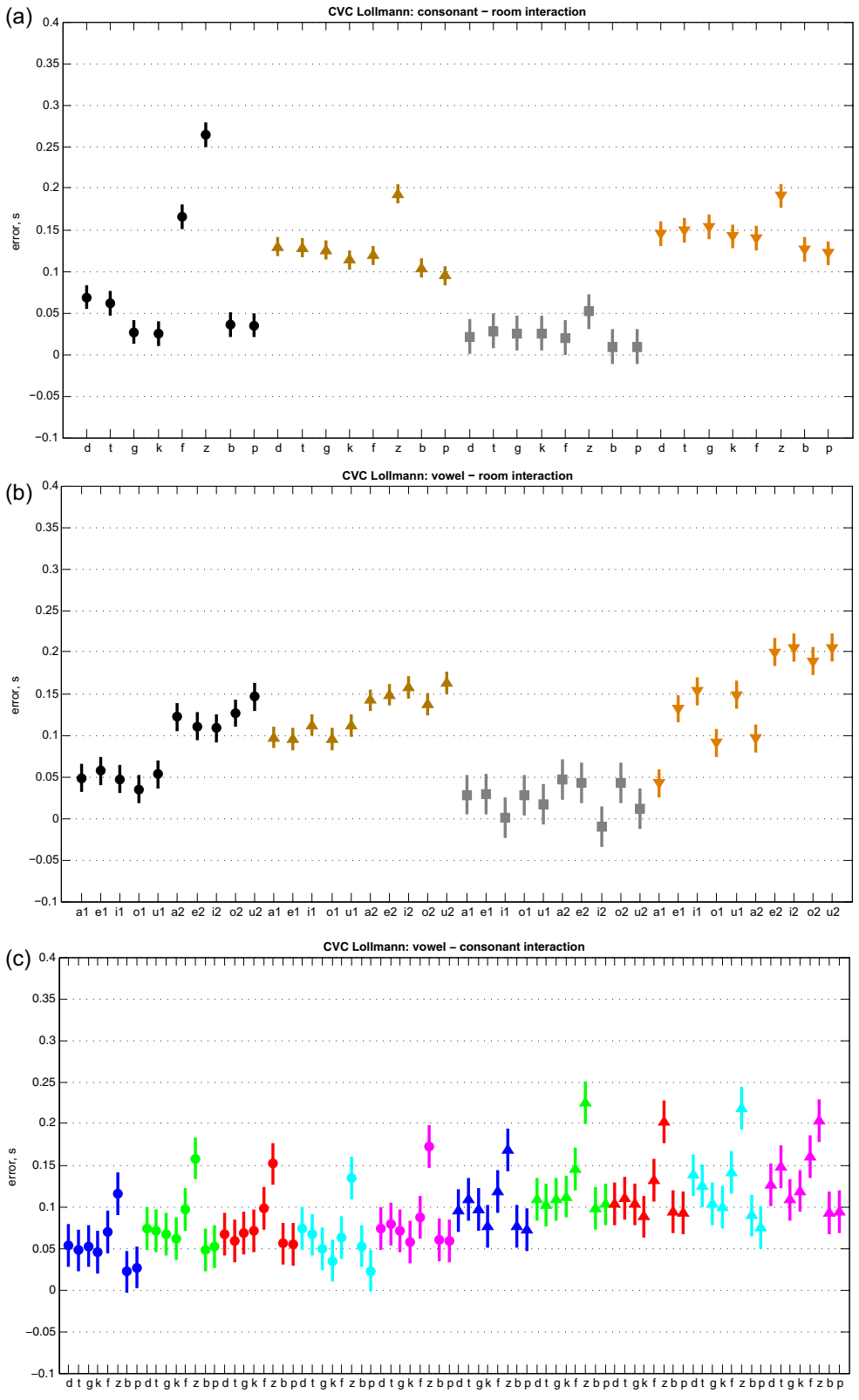

Figure 3. CVC logatomes: marginal means for Löllmann's algorithm. (a) "Consonant" × "room" interaction, (b) "vowel" × "room" interaction, (c) "vowel" × "consonant" interaction. For more details see Figure 1.

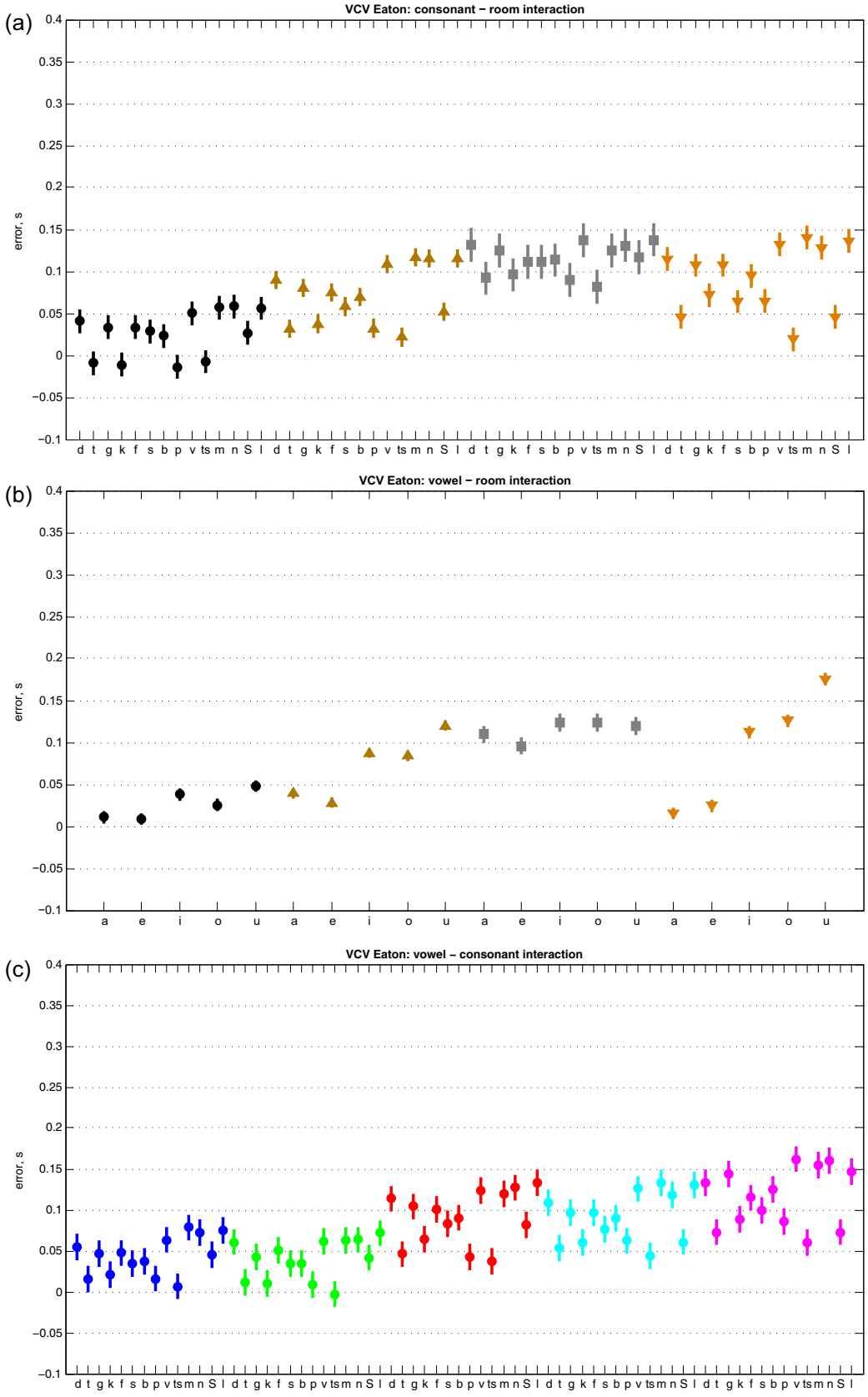

Figure 4. VCV logatomes: marginal means for Eaton's algorithm. (a) "Consonant" × "room" interaction, (b) "vowel" × "room" interaction, (c) "vowel" × "consonant" interaction. *X*-axis details: For the panels (a) and (c) central consonants of the VCVs are denoted on the *x*-axis. For the panel (b) vowels (/a/, /ɛ/, /ɪ/, /ɔ/, /ʊ/) as "a"–"u". Symbols for the marginal means: For the panels (a) and (b) "booth" – circle, "meeting" – upward triangle, "listening" – square, "office" – downward triangle. For the panel (c) vowels (/a/, /ɛ/, /ɪ/, /ɔ/, /ʊ/).

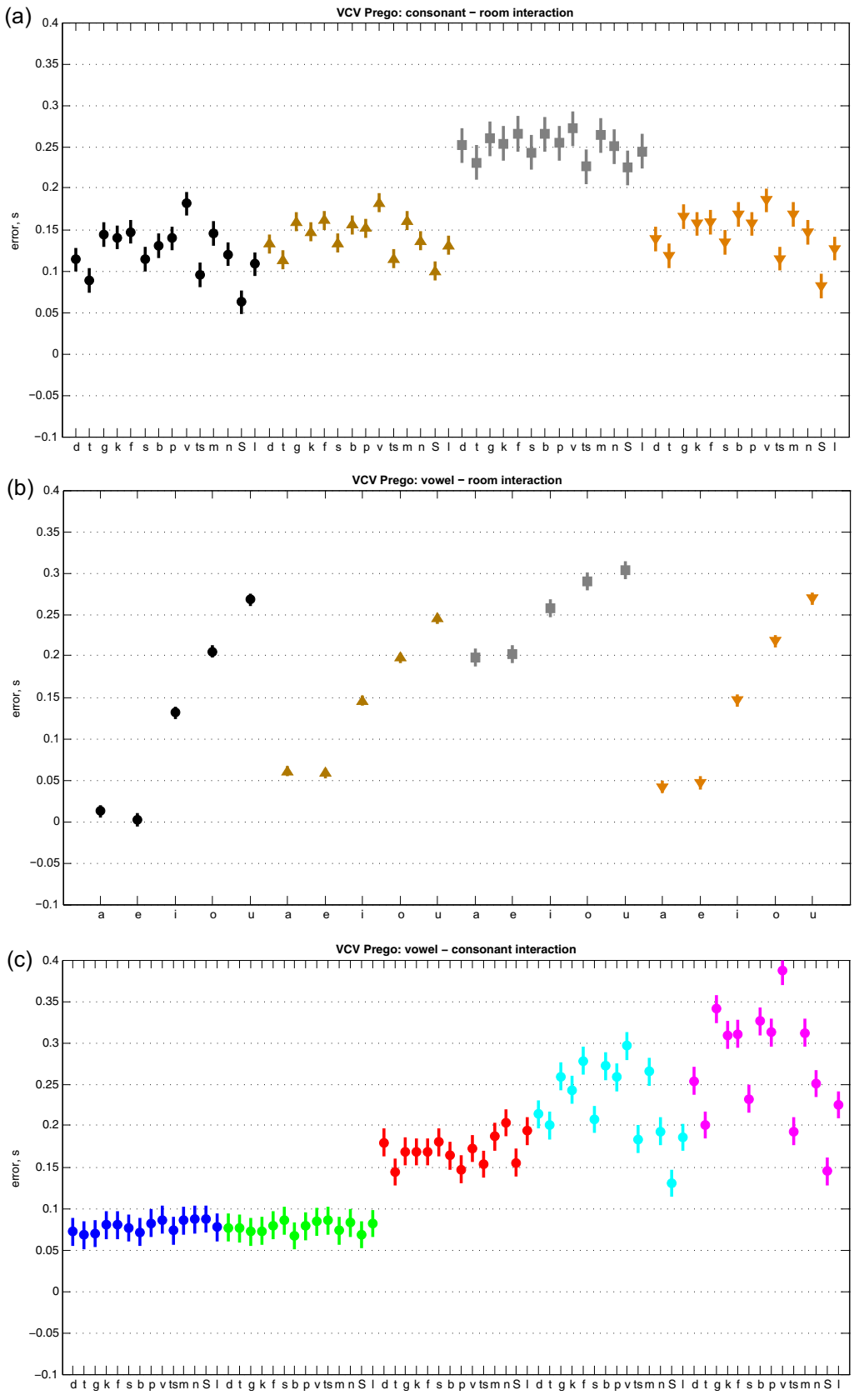

Figure 5. VCV logatomes: marginal means for Prego's algorithm. (a) "Consonant" × "room" interaction, (b) "vowel" × "room" interaction, (c) "vowel" × "consonant" interaction. For more details see Figure 4.

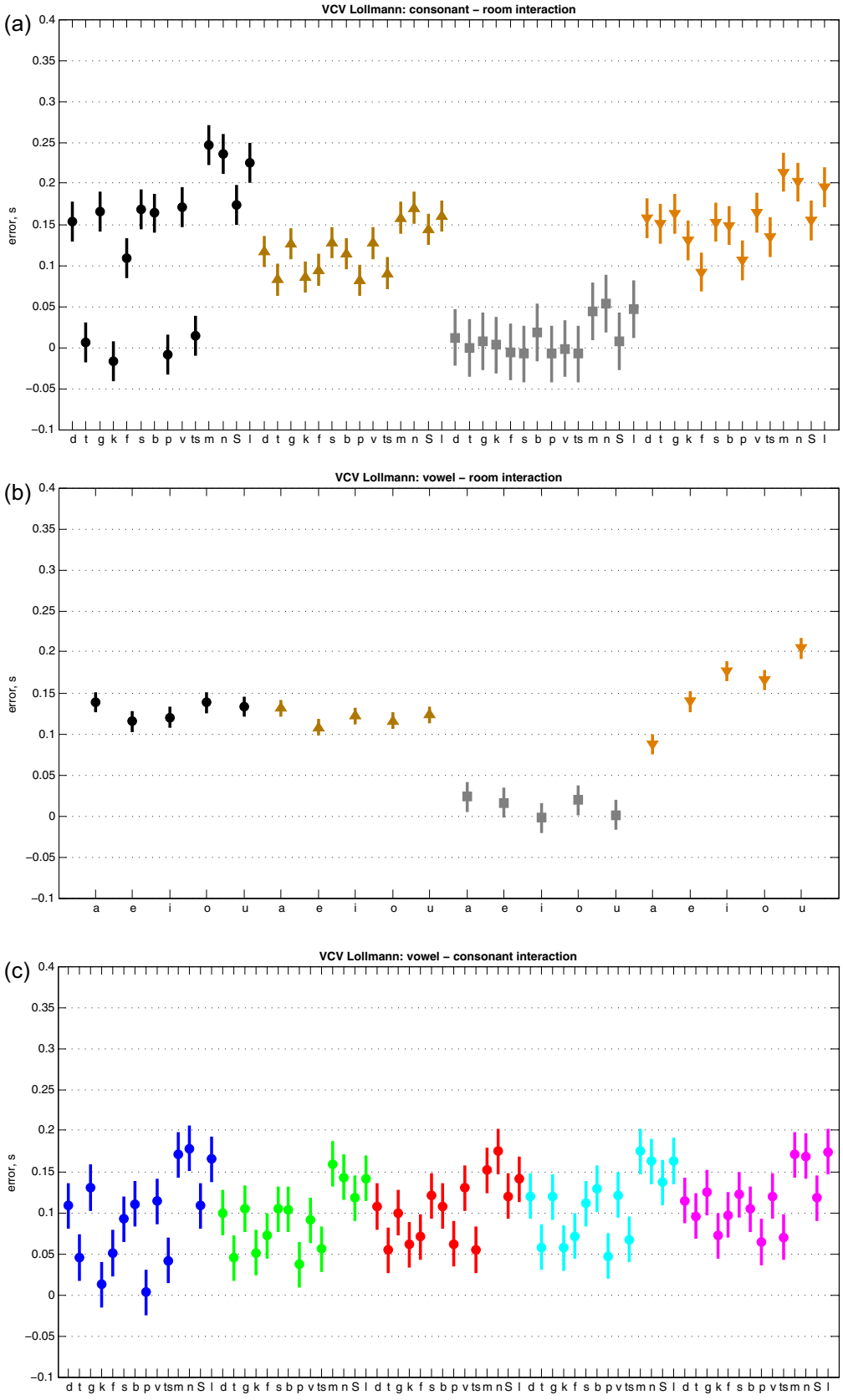

Figure 6. VCV logatomes: marginal means for Löllmann's algorithm. (a) "Consonant" × "room" interaction, (b) "vowel" × "room" interaction, (c) "vowel" × "consonant" interaction. For more details see Figure 4.

the behaviour across consonants is different than the one observed in Figure 4a but, more importantly, it is very consistent across rooms, which is reflected in a negligible η^2 of 0.28 for the interaction term and a larger η^2 of 3.74 for the "consonant" term.

Figure 5b reveals the strong influence that vowels have on this algorithm's output; the differences in the values of marginal means between the VCVs that contain an /a/ or an /ʊ/ are from 0.1 s to 0.25 s, depending on the room in question, where the effect of interaction is small ($\eta^2 = 3.14$), while the effect of "vowel" term is very large ($\eta^2 = 44.97$).

Figure 5c presents the "vowel" × "consonant" interactions ($\eta^2 = 7.18$), where a pattern is visible – in the case the outer vowel is an /a/ or an /ɛ/, the value of the reverberation time estimate is essentially the same regardless of the central consonant while, when going from the vowel /ɪ/, to /ɔ/, and finally /ʊ/, the influence of the type of consonant embedded in the VCV on the value of the estimate increases drastically. The change in the value of the estimate across vowels for a fixed consonant is largest for the voiced fricative /v/ and smallest for the unvoiced fricative /ʃ/.

Finally, Figure 6 presents interactions between the factors for Löllmann's algorithm. The influence of consonants strongly depends on the "room" factor, as shown in the panel 6a and an η^2 of 9.72. For the first room ("booth"), the inter-consonant differences are on the order of hundreds of milliseconds – the estimates of reverberation time for the VCVs that contain an unvoiced plosive or an affricate are approximately 0.25 s lower than the ones for the VCV logatomes that contain a nasal or a liquid. As the true value of reverberation time for "booth" is 0.18 s, this observed difference is almost 140% of that value. For the other two rooms from the Aachen database the influence of the central consonant is also quite strong and statistically significant between some consonant pairs. The same cannot be said for the "listening" room where the values for the nasals and the liquid are indeed again higher but the difference is not statistically significant.

The panel 6b indicates that for the smaller values of reverberation time (the first three rooms), the influence of vowels is not significant while, for the fourth room, a pattern similar to the one previously observed for Eaton's algorithm emerges, causing a $p < 0.001$ with $\eta^2 = 4.01$ for this interaction term. Unlike the results for the previous two algorithms, where a relatively high level of consistency of both the vowel and consonant behaviour was observed across rooms, the panel 6c is not so informative, since it hides the fact that the influence of consonants depends quite strongly on the "room" factor, that is, the true value of reverberation time.

4 Discussion

In this Section, the most likely sources for the patterns of reverberation time estimation errors observed in the previous section are now proposed. They are based on the results of the analyses of spectral characteristics of the phonemes used, ratios of spectra of the adjacent speech sounds as well as the durations of these phonemes.

4.1 CVC set of logatomes

For all three blind reverberation time estimation algorithms, the following regularities could be observed in Figures 1–3:

- when averaging the estimates across vowels (i.e., the "consonant" × "room" interaction), the CVCs that contained fricatives produced a higher estimate of T than the CVCs that contained plosives,
- when averaging the estimates across consonants (i.e., the "vowel" × "room" interaction):
  - for the TF algorithms: the CVCs that contained the vowel /a/ gave lower T estimates than the CVCs that contained the vowel /ʊ/, and the same could be observed for the /a:/ and /u/ vowel pair.
  - for the time domain algorithm: for the first two rooms the vowels from the V1 and V2 vowel groups formed two separate clusters of marginal means.

The first point can be explained by the fact that vowels are a class of speech sounds up to 40 dB higher in energy level than the consonants surrounding them, thus forming an "energy arc" [33]. Consequently, the blind estimates of T will be primarily obtained from the vowel decay segments of a reverberant speech signal – occurring when the vowel is either the final phoneme in a word or when it is followed by a consonant.

Given that, it is reasonable to assume that the differences in the T estimates between the fricative CVCs and the plosive CVCs can be attributed to the influence of the consonant that follows the central vowel where, on the one hand, an abrupt closure of the vocal tract at the beginning of the closed phase of a plosive [28] causes an almost instantaneous stop of the preceding vowel while, on the other hand, fricatives have relatively time–stable frequency profiles which are disrupting the energy decay of the vowel in the higher frequency bands, consequently inducing larger values of the reverberation time estimates.

To give an appropriate explanation for the second point, two figures (Figs. 7 and 8), must be first introduced. Figure 7 shows the average RMS-normalised spectra for the 10 vowels that were obtained from the CVC logatome recordings using the time stamps of the beginning and the end of the central vowel given in the speech segmentation text files – an integral part of the OLLO corpus. The second figure, Figure 8, presents another important vowel feature – its duration, which was also obtained with the help of those speech segmentation text files. From Figure 7, it can be observed that the very likely cause for the increase in the values of the marginal means from /a/ to /ʊ/ and from /a:/ to /u/ can be found in the decrease of the overall energy above 1 kHz that the vowels have – starting from the vowel /a/ to /ʊ/ (Fig. 7a) and from the vowel /a:/ to /u/ (Fig. 7b), which, in the end, makes the decays of the vowels /a/ and /a:/ of "better quality" than the ones of the vowels /ʊ/ and /u/.

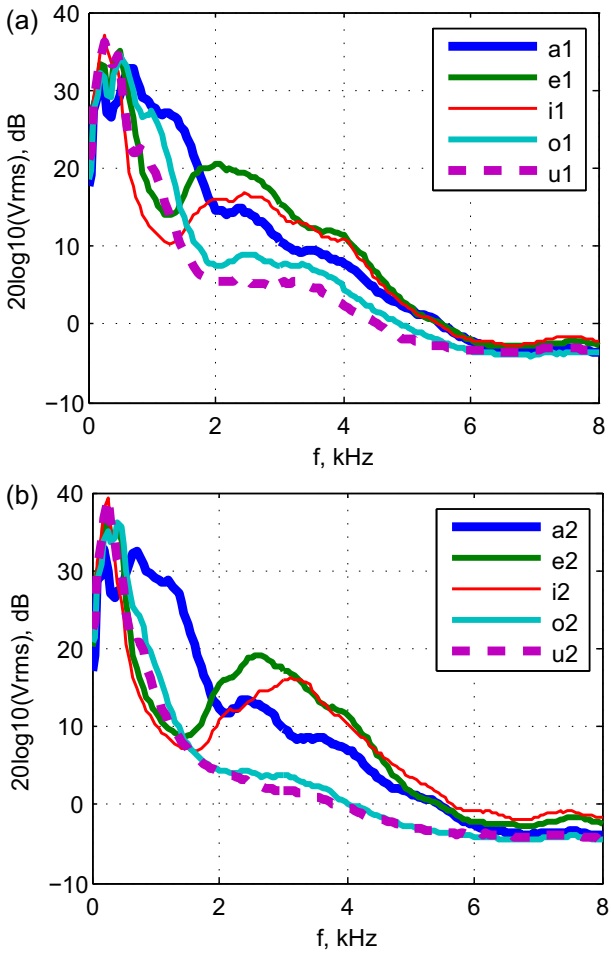

Figure 7. CVC logatomes: average RMS normalised spectra of central vowels. (a) Vowel group V1, (b) vowel group V2.

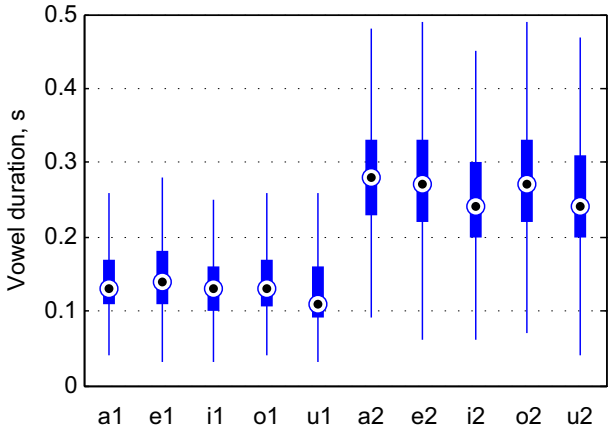

Figure 8. CVC logatomes: duration of central vowels.

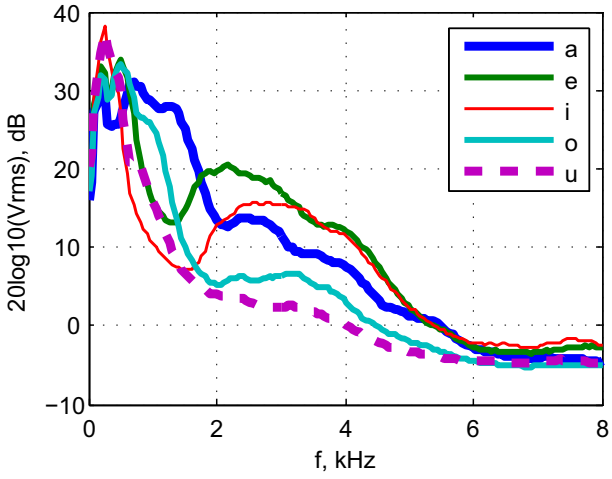

Figure 9. VCV logatomes: average RMS normalised spectra of left vowels.

Finally, the vowel group V1 and V2 clusters of the marginal means observed for the first two rooms for Löllmann's algorithm (Fig. 3b) as well as for Eaton's algorithm (Fig. 1b) could be explained by the difference in the average duration of the vowels belonging to those two groups, as shown in Figure 8.

4.2 VCV set of logatomes

4.2.1 VCV–CVC comparison

As explained in the previous sub-section, for the CVC logatomes the estimates of T were obtained primarily from the vowel–consonant (V–C) transition segment of the recordings. The VCV logatomes, on the other hand, contain two vowels and, consequently, the recordings will have two decays from which the final T estimates will be obtained: the V–C transition and the V-silence transition for the right VCV vowel.

Figure 9 shows the average RMS-normalised spectra of the left vowel obtained from the recordings of the VCV logatomes. It can be noticed that the spectra look, as expected, almost exactly the same as the ones in Figure 7a for the V1 group of CVC vowels.

The differences in the energy profiles of vowels are again very visible; /a/ and /ɛ/ have substantially more energy than /ɔ/ and /ʊ/ in the region above 1.5 kHz, and /ɪ/ is similar to /ɛ/ but with a lower inter-formant minimum at about 1.6 kHz and a lower formant peak in the 3 kHz area.

Figure 10 presents the average ratios of the left vowel spectrum and the spectrum of the central consonant (VC ratios) calculated from the same logatome recording. These VC ratios are calculated and presented only for the central consonants that have relatively stable time–frequency characteristics and not for the plosives and affricate /ts/, which are the classes of transitory speech sounds. It can be observed that these average VC ratios decrease as one goes from /a/ to /ʊ/, indicating that there will be more (on average) frequency bands with free-decays of higher decibel range for the vowels /a/ and /ɛ/ than for the vowels /ɔ/ and /ʊ/ during the vowel–consonant transition.

Figure 11 presents the corresponding average VC ratios obtained from the CVC recordings for the consonants /f/ and /s/. It is visible that the VC ratios for the same

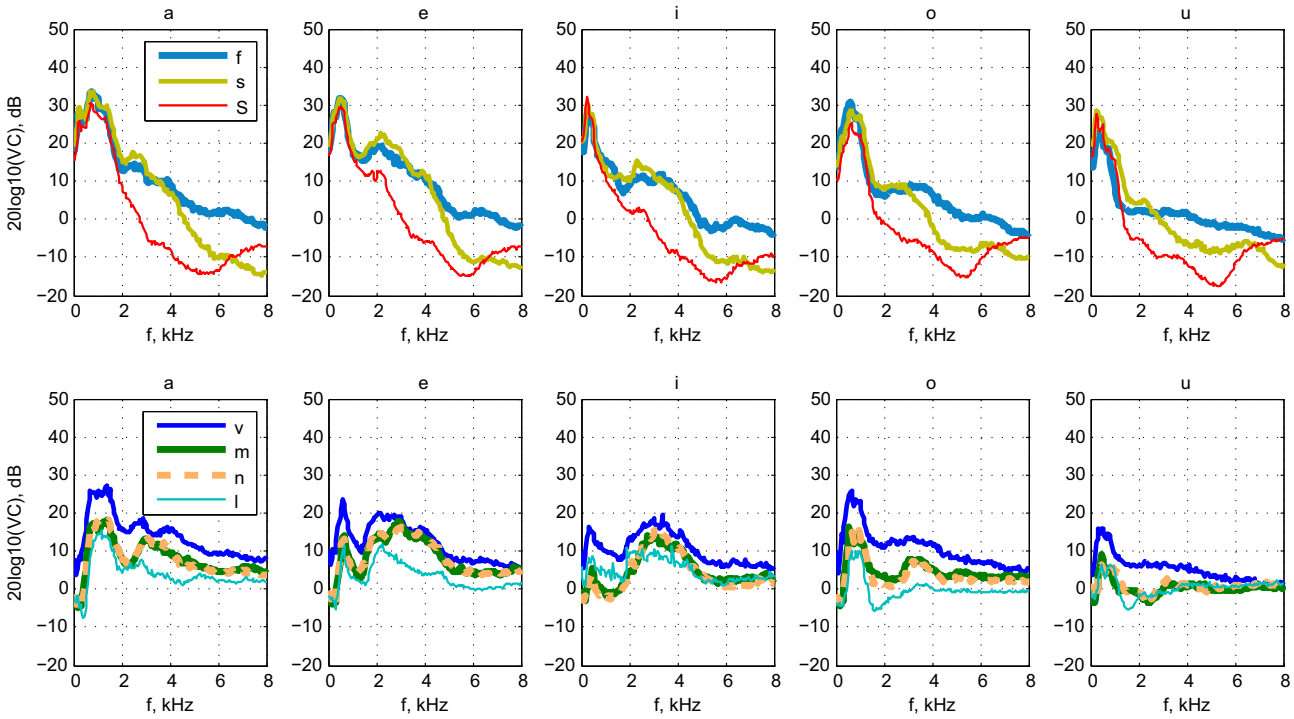

Figure 10. VCV logatomes: average vowel–consonant spectra ratio.

vowel–consonant pairs of the VCV and CVC recordings have the same frequency profiles. The only difference is that the VC ratios are a few decibels higher for the CVC recordings, meaning that the vowel in a CVC was stressed more by the speakers than the first vowel of a VCV.

The similarity of both the average spectrum profiles of the vowels (Figs. 7 and 9) and the VC ratios (Figs. 10 and 11) enables for the comparison of the results obtained for the CVC logatomes and the ones for the VCV logatomes that contain the consonants /p/, /t/, /k/, /f/, and /s/ to be performed. By comparing the results presented in Figures 1a and 4a for Eaton's algorithm, it can be concluded that the V-silence decay provides new negative gradients for this algorithm, so that the final estimate of T for the VCVs that contain an /f/ or /s/ is closer in value to the estimates obtained with unvoiced plosives than it was the case for the CVC logatomes. This effect, caused by the presence of a new, V-silence, decay is even more pronounced for Prego's (Figs. 2a and 5a) and Löllmann's algorithm (Figs. 3a and 6a).

4.2.2 Influence of the central C

As described in Section 2.3, the first algorithm, Eaton's, continuously calculates the decay rates on the time–mel-frequency decomposed reverberant speech signal, and from whose distribution it estimates the full-band value of reverberation time. Since it uses a window of fixed length to estimate those decay rates, it is sensitive to the duration of the gap between two high energy speech sounds, such as the two vowels in a VCV logatome. Furthermore, the two main differences between the unvoiced and voiced plosive pairs are: (i) the existence of a voiced bar for the latter, and (ii) longer voiced onset time (VOT) for the former, where the VOT is defined as the difference between the time of the plosive burst and the onset of voicing of the following vowel [28]. Figure 12 shows the durations of central consonants of the VCV recordings obtained using the time stamps of the beginning and the end of the consonant given in the speech segmentation text files. It can be observed that, in line with previous research [28], the durations of the unvoiced plosives, unvoiced fricatives and the affricate /ts/, a phoneme that is composed of those two speech classes, are longer than those for the voiced phonemes. Given these results and the results of the "consonant" × "room" interactions (Fig. 4a), it can be concluded that this algorithm's output is sensitive to both the inter-vowel gap duration and the frequency distribution and energy level of the consonant, where the latter is responsible for the difference between the results for the voiced plosives and the results for the voiced fricative /v/, the nasals and /l/.

The second, Prego's, algorithm differs from Eaton's primarily in two things: (i) in the much larger number of frequency bands used for the time–frequency reverberant speech signal decomposition, and (ii) that it searches for the free-decay regions and calculates their corresponding EDCs, from which an estimate of reverberation time is obtained. Since the minimum EDC dB range this algorithm finds acceptable for T calculation is 10 dB and given the average VC ratios (Fig. 10), it can be now understood why this algorithm shows strong dependence on the vowel used in a VCV – in contrast to the stability of the reverberation time estimates for /a/ and /ɛ/ across consonants, for the vowels /ɔ/ and /ʊ/ there simply is not much high

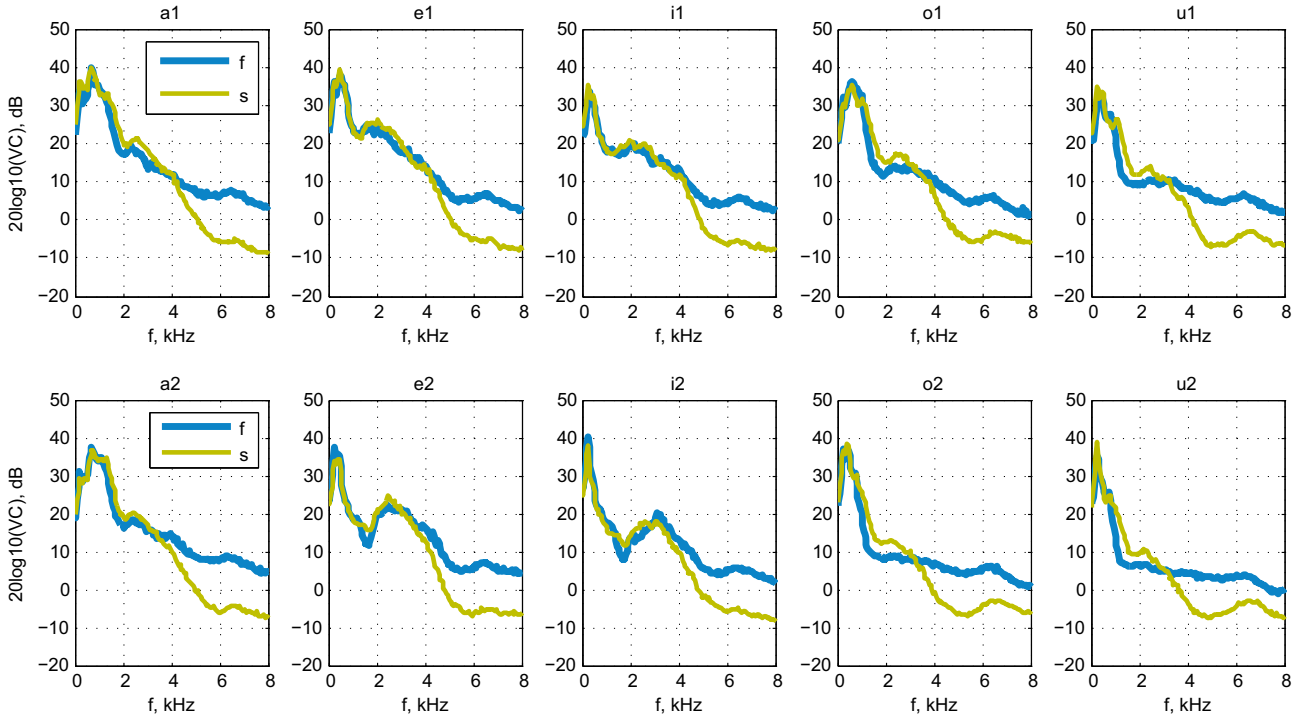

Figure 11. CVC logatomes: average vowel–consonant spectra ratio.

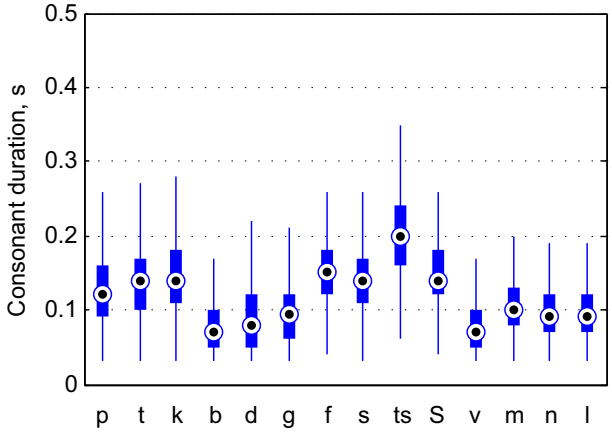

Figure 12. VCV logatomes: duration of central consonants.

frequency energy and, consequently, the interaction between the vowel and the consonant that follows becomes much more pronounced for a large subset of frequency bands.

Lastly, Löllmann's algorithm differs from the previous two in that it operates on the signal's time series, while its speech decay detection procedure is similar in its essence to the one used in Prego's algorithm. Due to the first point, the quality of a vowel decay will depend on the overall energy of the consonant that follows. This is in agreement with the results for "booth" (Fig. 6b), where the unvoiced plosives produce the smallest values of T estimate, voiced plosives, and fricatives larger, while the nasals and /l/ the largest ones. The difference between the unvoiced and voiced plosives most probably stems from the presence of the energy of the voiced bar for the latter. At the same time, the results for the affricate /ts/, somewhat perplexing due to the existence of the high energy /s/ part, must be also carefully considered – they are the same as for the plosive /t/, meaning that for the T value of 0.18 s the plosive part of this unvoiced affricate is long enough to enable good reverberation time estimation on the V – /t/ decay segment. For this algorithm, as T increases, the interactions change as well, and for the larger T's ("office"), vowels become the second source of estimate variability.

5 Conclusion

The results presented in this study demonstrate that, in addition to their well explored sensitivity to background noise, the values of T estimates of the state-of-the-art algorithms for blind reverberation time estimation are strongly influenced by the phonemes present in the speech material. This influence that phonemes have on the value of reverberation time estimates is statistically significant not just for the algorithm based on a stochastic time-domain model of room impulse responses which uses no information on speech signal characteristics, but also for the more complex algorithms, whose values of mapping parameters have been adjusted during the training phase in order to compensate for the speech signal characteristics.

Since reverberation time rarely exceeds the value of one second in dwellings [1, 31], very large relative (i.e., percentage) errors in blind T estimation can occur in the electronic

devices that process captured speech signals – both for the languages with a predominance of open syllables and the ones with the predominance of closed syllables, such as the omnipresent English language. In addition to that, many of the differences between the marginal means for different vowel–consonant combinations were much larger than 42 ms, which is the value of the subjective difference limen – a measure describing the T resolution of the human auditory system, and defined as the minimal change in the value of T that human subjects can register in listening tests [34].

Despite the corpus related limitations of this study, primarily due to a small number of consonants present in the CVC set, the results implicitly indicate ways in which the algorithms can be adapted to become more speech signal-robust; for the time–frequency algorithms, the frequency bands could be selected so that the ones where the vowels /ɔ/ and /ʊ/ have relatively low energy level are not used for the decay rates/EDC estimation, while for the time-domain operating algorithm, since the values of its estimates are primarily consonant-driven for smaller values of reverberation time, the V–C transition of only one of the many classes of consonants should be utilized.

Acknowledgments

The author would like to thank the authors of the three algorithms used in this study – Mr. Eaton for making his blind reverberation time estimation algorithm (Ref. [32]) publicly available, Mr. Prego for kindly providing his algorithm described in reference [20], and Mr. Löllmann for the very detailed description of his algorithm as presented in reference [11].

The author would also like to express her gratitude both to the Institute for Communication Systems at RWTH Aachen University and to the Department of Medical Physics and Acoustics at Carl von Ossietzky University in Oldenburg for creating the AIR database and the OLLO corpus, respectively, and sharing them selflessly with other researchers.

It was these blind T estimation algorithms and the aforementioned two data sets that inspired the author to draw the first sketches of this study, and were so essential for its realisation – thank you all.

Lastly, Ms. Andrijašević would like to express her sincere appreciation to Prof. Juraj Šimunić, Prof. Viktor Sučić, Prof. Ivan Štajduhar, Prof. Ivan Dražić, Prof. Neven Bulić, and assistant Karlo Radman from the Faculty of Engineering in Rijeka, as well as to Prof. Kristian Jambrošić from the Faculty of Engineering and Computing in Zagreb, and to her Ph.D. mentor, the late Prof. Hrvoje Domitrović, for their support, encouragement and insightful comments.

Conflict of interest

Author declared no conflict of interests.

References

1. Acoustics – Measurement of the reverberation time of rooms with reference to other acoustical parameters, ISO 3382-2: 2008/AC – Reverberation time in ordinary rooms. International Organization for Standardization, Geneva, 2009.
2. F.A. Everest: Reverberation, in Master Handbook of Acoustics, Chap. 11, New York, NY, McGraw-Hill. 2009, pp. 151–179.
3. H. Kuttruff: Measuring techniques in room acoustics, in Room Acoustics, Chap. 8, Oxford, Spon Press. 2009, pp. 251–293.
4. P.A. Naylor, N.D. Gaubitch, Editors: Speech dereverberation using statistical reverberation models, in Speech Dereverberation, Chap. 3, Berlin, Springer. 2010, pp. 57–94.
5. I.J. Tashev: De-reverberation, in Sound Capture and Processing: Practical Approaches, Chap. 8, NY, USA, Wiley. 2009, pp. 341–358.
6. H.W. Löllmann, P. Vary: Low delay noise reduction and dereverberation for hearing aids. EURASIP Journal on Advances in Signal Processing 2009 (2009) 1–9.
7. J. Eaton, N.D. Gaubitch, A.H. Moore, P.A. Naylor: Estimation of room acoustic parameters: The ACE Challenge. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24 (2016) 1681–1693.
8. R. Ratnam, D.L. Jones, B.C. Wheeler, W.D. O'Brien, C.R. Lansing, A.S. Feng: Blind estimation of reverberation time. Journal of the Acoustical Society of America 114 (2003) 2877–2892.
9. R. Ratnam, D.L. Jones, W.D. O'Brien: Fast algorithms for blind estimation of reverberation time. IEEE Signal Processing Letters 11 (2004) 537–540.
10. P. Kendrick, F.F. Li, T.J. Cox: Blind estimation of reverberation parameters for non-diffuse rooms. Journal of the Acoustical Society of America 93 (2007) 760–770.
11. H.W. Löllmann, E. Yilmaz, M. Jeub, P. Vary: An improved algorithm for blind reverberation time estimation, in Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC), Israel, Tel Aviv. 2010, pp. 1–4.
12. P. Kendrick: Blind estimation of room acoustic parameters from speech and music signals. PhD dissertation, University of Salford, UK, 2009.
13. P. Kendrick, T.J. Cox, F. Li, Y. Zhang, J. Chambers: Monaural room acoustic parameters from music and speech. Journal of the Acoustical Society of America 124 (2008) 278–287.
14. T. Jan, W. Wang: Blind reverberation time estimation based on Laplace distribution, in Proc. 20th European Signal Processing Conference (EUSIPCO 2012), Bucharest, Romania. 2012, pp. 2050–2054.
15. A. Keshavarz, S. Mosayyebpuor, M. Biguesh, T.A. Gulliver, M. Esmaeili: Speech-model based accurate blind reverberation time estimation using an LPC filter. IEEE/ACM Transactions on Audio, Speech, and Language Processing 20 (2012) 1884–1893.
16. C. Schuldt, P. Handel: Blind low-complexity estimation of reverberation time, in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New York, USA. 2013, pp. 1–4.
17. J.Y.C. Wen, E.A.P. Habets, P.A. Naylor: Blind estimation of reverberation time based on the distribution of signal decay rates, in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, USA. 2008, pp. 329–332.
18. J. Eaton, N.D. Gaubitch, P.A. Naylor: Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost, in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada. 2013, pp. 161–165.

19. N. Lopez, Y. Grenier, G. Richard, I. Bourmeyster: Low variance blind estimation of the reverberation time, in Proc. International Workshop on Acoustic Signal Enhancement (IWAENC), Aachen, Germany. 2012, pp. 1–4.
20. T.M. Prego, A.A. Lima, S.L. Netto: A blind algorithm for reverberation-time estimation using subband decomposition of speech signals. Journal of the Acoustical Society of America 131 (2012) 2811–2816.
21. T.M. de Prego, A.A. de Lima, R. Zambrano-Lopez, S.L. Netto: Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition, in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA. 2015.
22. T.H. Falk, W.Y. Chan: Temporal dynamics for blind measurement of room acoustical parameters. IEEE Transactions on Instrumentation and Measurement 59 (2010) 978–989.
23. F. Xiong, S. Goetze, B.T. Meyer: Blind estimation of reverberation time based on spectro-temporal modulation filtering, in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada. 2013, pp. 443–447.
24. N.R. Shabtai, Y. Zigel, B. Rafaeli: Towards room-volume classification from reverberant speech using room-volume feature extraction and room-acoustics parameters. Acta Acustica United with Acustica 99 (2013) 658–669.
25. K. Kinoshita, M. Delcroix, S. Gannot, E.A.P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Mass, T. Nakatani, B. Raj, A. Sehr, T. Yoshioka: A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. EURASIP Journal on Advances in Signal Processing 2016 (2016) 1–19.
26. A. Andrijašević, H. Domitrović: Effects of word phonetic contents and speaking style on blind estimation of reverberation time, in Proc. of Alps Adria Acoustics Association Congress on Sound and Vibration, Ljubljana, Slovenia. 2016, pp. 201–208.
27. B.T. Meyer, T. Jürgens, T. Wesker, T. Brand, B. Kollmeier: Human phoneme recognition depending on speech-intrinsic variability. Journal of the Acoustical Society of America 128 (2010) 3126–3141.
28. T.F. Quatieri: Production and classification of speech sounds, in Discrete-Time Speech Signal Processing: Principles and Practice, 1st edn., Chap. 3, NJ, USA, Prentice Hall. 2001, pp. 55–110.
29. T.D. Rossing, Editor: The human voice in speech and singing, in Springer Handbook of Acoustics, Chap. 16, Berlin, Springer. 2007, pp. 669–712.
30. M. Jeub, M. Schäfer, P. Vary: A binaural room impulse response database for the evaluation of dereverberation algorithms, in Proceedings of the International Conference on Digital Signal Processing, Santorini, Greece. 2009, pp. 1–4.
31. C. Diaz, A. Pedrero: The reverberation time of furnished rooms in dwellings. Applied Acoustics 66 (2005) 945–956.
32. Available online: http://www.commsp.ee.ic.ac.uk/~sap/projects/blindestimation-of-acoustic-parameters-from-speech/blind-t60-estimator/. Last viewed on: 16th May 2018.
33. J. Schnupp, I. Nelken, A. King: Hearing speech. Auditory Neuroscience, Chap. 4, Cambridge, MA, MIT Press. 2011, pp. 139–176.
34. T.I. Niaounakis, W.J. Davies: Perception of reverberation time in small listening rooms. Journal of the Audio Engineering Society 50 (2002) 343–350.

Cite this article as: Andrijašević A. 2020. Effect of phoneme variations on blind reverberation time estimation. Acta Acustica, **4**, 3.