Article

Characterization of the Intelligibility of Vowel–Consonant–Vowel (VCV) Recordings in Five Languages for Application in Speech-in-Noise Screening in Multilingual Settings

Giulia Rocco [1], Giuliano Bernardi [2], Randall Ali [2], Toon van Waterschoot [2], Edoardo Maria Polo [3], Riccardo Barbieri [1] and Alessia Paglialonga [4,*]

1   Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, 20133 Milan, Italy
2   Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, 3001 Leuven, Belgium
3   Dipartimento di Ingegneria Informatica Automatica e Gestionale Antonio Ruberti (DIAG), Università La Sapienza di Roma, 00185 Rome, Italy
4   Cnr-Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni (CNR-IEIIT), 20133 Milan, Italy
*   Correspondence: alessia.paglialonga@cnr.it

Abstract: The purpose of this study is to characterize the intelligibility of a corpus of Vowel–Consonant–Vowel (VCV) stimuli recorded in five languages (English, French, German, Italian and Portuguese) in order to identify a subset of stimuli for screening individuals of unknown language during speech-in-noise tests. The intelligibility of VCV stimuli was estimated by combining the psychometric functions derived from the Short-Time Objective Intelligibility (STOI) measure with those derived from listening tests. To compensate for the potential increase in speech recognition effort in non-native listeners, stimuli were selected based on three criteria: (i) higher intelligibility; (ii) lower variability of intelligibility; and (iii) shallower psychometric function. The observed intelligibility estimates show that the three criteria for application in multilingual settings were fulfilled by the set of VCVs in English (average intelligibility from 1% to 8% higher; SRT from 4.01 to 2.04 dB SNR lower; average variability up to four times lower; slope from 0.35 to 0.68%/dB SNR lower). Further research is needed to characterize the intelligibility of these stimuli in a large sample of non-native listeners with varying degrees of hearing loss and to determine the possible effects of hearing loss and native language on VCV recognition.

Keywords: intervocalic consonants; hearing screening; Short-Time Objective Intelligibility (STOI); speech intelligibility; speech-in-noise testing

1. Introduction

Despite the significant incidence of hearing loss in older adults and the severe consequences of this condition on communication and quality of life, age-related hearing loss is still largely underdiagnosed [1]. The majority of individuals with hearing loss tend to live with their hearing problems for several years before seeking help until they experience communication problems and disability due to the slow progression of hearing loss [2,3]. One way to help prevent or limit the progression of disabling hearing loss is to raise awareness about the importance of early detection in adults and promote adult hearing screening [4]. In the field of adult hearing screening, there is increasing interest in easy-to-use, user-operated tools that are able to detect hearing difficulties and can be administered at a distance [5–8]. The inability to understand speech in noise is one of the most common complaints of older adults with hearing loss, even at early hearing impairment stages [9]. As such, speech-in-noise tests may reflect real-life communication performance, and they may be helpful in promoting awareness, early detection of hearing loss, and supporting

patient motivation to seek care [10]. In particular, hearing tests delivered remotely may help increase access and facilitate regular monitoring.

In comparison to pure-tone audiometry, speech-in-noise tests can be performed in noisy environments outside the clinic, and they can be self-administered in the absence of calibrated equipment and with no supervision by hearing healthcare providers [9–13]. The increasing use of speech-in-noise tests raised awareness of the limitations of diagnostic validity in specific testing conditions and applications. For example, appropriately serving underrepresented populations with special attention to racial, ethnic, and linguistic minorities [14]. It is worth noting that several mobile health apps have been recently introduced to support hearing screening in adults, as well as self-testing for education purposes or for hearing aid fitting [15]. However, hearing screening at a distance may open new challenges in hearing health care. For example, remote testing using pure-tone audiometry has limitations in terms of output levels calibration and sensitivity to environmental noise [16]. Speech-in-noise testing can help overcome this limitation, but language-specific test stimuli may lead to issues with non-native listeners [17].

Examples of speech-in-noise tests viable for adult hearing screening include, but are not limited to: the Speech Perception Test [18] and the Earcheck [19,20], based on consonant–vowel–consonant (CVC) words, the digits-in-noise test, based on digit triplets [13], and the Speech Understanding in Noise test, based on nonsense vowel–consonant–vowel (VCV) stimuli [12,21,22]. All the above-mentioned tests were originally developed in a single language and then adapted to different languages, (i.e., by translation of stimuli and by using recordings from native language speakers).

1.1. Speech-in-Noise Screening Test Design

To account for common situations in which the language is not known (e.g., screening in multilingual settings, screening at a distance), we recently developed a novel, automated speech-in-noise test for hearing screening in multilingual settings. Specifically, we used a set of nonsense VCV stimuli viable for use across listeners of various languages presented in a multiple-choice format using an adaptive three-alternative forced-choice (3AFC) task [5,23–26]. Several ways to communicate the individual responses to stimuli in psychophysical tasks have been used in the literature, e.g., automatization of typed and spoken responses, for example, using automated speech recognition algorithms [27]. However, using automated approaches to record individual responses has limitations, as individuals, particularly the older ones, may have difficulties repeating the stimuli due to higher cognitive load and potential unease and anxiety [28]. In this study, a forced-choice task is adopted for several reasons. First of all, multiple-choice tasks are easy to implement in a user-operated and automated procedure, which could be a simple graphic user interface in which the subject selects the desired response (e.g., [5,12]). Moreover, the simple interaction required (i.e., clicking on a screen) can help reduce the negative influence of cognitive decline, slowed temporal processing and added memory demands on speech recognition performance in older adults [28]. In this study, a 3AFC task was used as it is shown to perform better than the popular and widespread two-alternative forced-choice task in terms of threshold estimation stability under different types of adaptive procedures [29].

The VCV stimuli used in this study could yield optimal screening for a number of reasons. Firstly, decreased VCV recognition may indicate decreased ability to recognize consonants and fast transients, which are among the first clues of the presence of age-related high-frequency hearing loss [11]. Moreover, nonsense syllable recognition tests such as VCV tests present good test–retest reliability and precision under multiple experimental conditions, such as stimulus randomization and uncontrolled background noise [30]. VCVs are also associated with smaller inter-stimulus variability of intelligibility within the set compared to, e.g., CVC words or sentences and may be more easily equalized before the test [26]. Additionally, VCV recognition is largely independent of lexicon and semantics, thus limiting the influence of auditory-cognitive interactions and the involvement of higher-level processing centers in the recognition task [31]. By using VCV stimuli,

multiple-choice recognition tasks can be implemented in a simple user-operated automated procedure (e.g., by using a graphical user interface on a touch screen). Last but not least, the intelligibility of VCVs is less influenced by the subjects' education, literacy, and native language compared to words or sentences [32]. This issue is particularly relevant in today's globalized world with multilingual societies, especially in remote screening settings where the languages spoken by participants are not necessarily known. In any case, to date, there is still a lack of validated, automated speech-in-noise screening tests viable for use at a distance in non-native listeners.

1.2. Study Rationale and Specific Aims

The aim of this study is to introduce a methodology for characterizing the intelligibility of a corpus of spoken VCVs in different languages in order to identify the most suitable subset of stimuli for application in listeners of an unknown language. The proposed approach is based on the integration of objective, computational estimates of VCV intelligibility derived from the Short-Time Objective Intelligibility (STOI) measure [33] and subjective, experimental estimates of VCV intelligibility derived from listening tests [21,22].

The rationale is to estimate the psychometric curves of VCV recorded from speakers of different languages and compare them based on the following criteria, designed to maximize the ability of the speech material to compensate for the possible additional effort of non-native listeners:

1. High average intelligibility, especially at low signal-to-noise ratios (SNR). This criterion is justified by the fact that the higher the theoretical intelligibility of a stimulus, the lower the effort associated with the recognition task, particularly in non-native listeners (e.g., [34]);
2. Low variability of intelligibility as a function of noise, especially at low SNR. This criterion is justified by the fact that the lower the variability of intelligibility with varying realizations of noise, the lower the aleatoric uncertainty introduced by noise on the measured speech recognition performance. Stated differently, if the variability of intelligibility is low, differences in measured speech recognition are more likely to reflect intra-individual differences in performance rather than random fluctuations of performance due to noise; and
3. Shallow slope of the psychometric function. This criterion is justified by the fact that the shallower the slope of the intelligibility functions, the wider the dynamic range of SNRs that can be used by the speech-in-noise test, therefore enabling higher flexibility to adapt the test to the individual performance of non-native listeners.

2. Materials and Methods

A summary of the study methodology is reported in Figure 1. First, the dataset for each language is built, by combining VCV recordings in different languages with varying levels of noise, as described in Section 2.1. Then, the intelligibility of noisy signals is estimated by using the objective STOI measure and by conducting subjective listening tests, as described in Section 2.2. Finally, the objective and subjective estimates are compared between VCVs in the different languages by extracting different performance measures, as reported in Section 2.3.

2.1. Dataset Creation: VCV Recordings and Noise

In this study, sets of spoken VCVs in English, French, German, Italian and Portuguese were compared, i.e., five languages among the top 25 languages with 50 or more million first-language speakers each [35]. The list of stimuli included 12 consonants common across the five languages (/b/, /d/, /f/, /g/, /k/, /l/, /m/, /n/, /p/, /r/, /s/, /t/) in the context of the vowel /a/ (e.g., aba, ada). The VCVs were single exemplars spoken in a sound-treated room by a professional language native male speaker who pronounced the VCVs with no prosodic accent and with constant pitch across the list. Stimuli were recorded in a professional recording studio by using a TLM 103 microphone (Neumann,

Berlin, Germany), SSL S4000 64 channels mixer, HD192 A/D converters (44.1 kHz, 16 bit) (Motu, Cambridge, MA, USA), and a GENELEC 1025A control room monitor. The level of the recordings was equalized within and across the sets to meet the equal speech level requirement as in the ISO 8253-3:2012 standard [36] and to guarantee equal average levels of the sets of speech materials.

The VCV recordings were then additively combined with filtered speech-shaped noise. The speech-shaped noise (SSN) was generated by filtering a Gaussian white noise with a filter representing the international long-term average speech spectrum [37] and, subsequently, with a low pass filter (cut-off 1.4 kHz, roll-off slope 100 dB/octave). Adding this kind of noise is a common viable solution used in speech-in-noise tests for screening in multilingual settings, as it allows to differentiate between normal hearing and mild hearing-impaired subjects more accurately than other processed noise (e.g., high pass, modulated, etc.) [19,20]. To further limit the possible influence of ambient noise levels, a noise floor was added, consisting of the same SSN described above and attenuated by 15 dB [19,20], thus the final delivered stimulus resulting as the addition of a VCV, the filtered SSN and the filtered SSN attenuated by 15 dB.
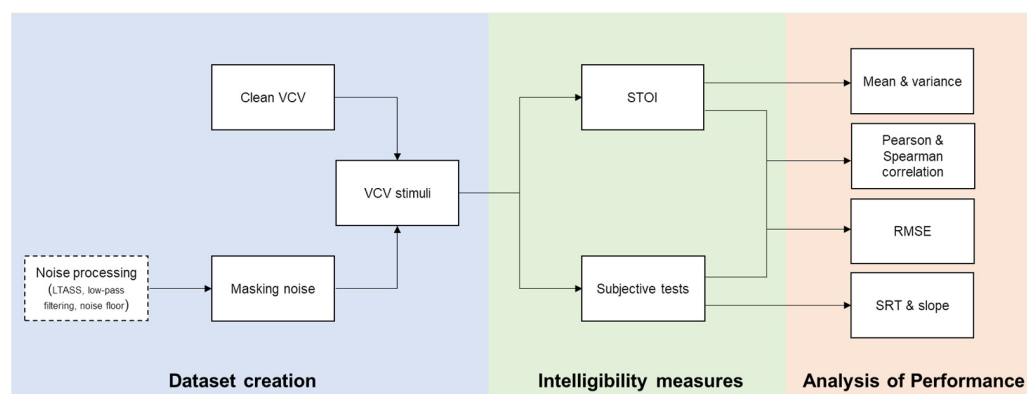

Figure 1. Outline of the study methodology.

2.2. Intelligibility Measures

In order to assess specific sets of spoken VCVs in the five languages in light of the three aforementioned criteria (in Section 1.2) for use in multilingual settings, the intelligibility of spoken VCVs in speech-shaped noise was estimated as a function of the SNR by using the STOI measure [33], an objective measure of speech intelligibility that has been shown to generate accurate predictions in various languages [33,38,39]. Even though the recognition of intervocalic consonants, such as the VCV stimuli used here, is less influenced by the subject's native language than that of words and sentences, there is an intrinsic linguistic perceptual factor that does not allow subjects to determine whether a language is more intelligible than another one [40]. An objective intelligibility measure, such as the STOI used here, can help address differences in intelligibility between speech materials in different languages. The STOI is an intrusive measure of speech intelligibility based on a correlation coefficient between the temporal envelopes of the time-aligned reference (i.e., the VCV spoken stimulus, with an average duration of about 1 s) and the processed speech signal (VCV plus noise) in short-time overlapped segments of 386 ms [33]. To evaluate the influence of noise on the STOI estimates, the mean STOI was computed in this study by averaging over 100 simulations with different noise realizations. For each VCV recording, the STOI for the 3AFC task was computed as a function of the SNR in the range -12 to $+6$ dB in 2 dB steps to cover most of the dynamic range of stimuli, as shown by our earlier experimental data [12,21,22].

The accuracy of the STOI estimates of VCV intelligibility in the five languages was validated using psychometric estimates from previously conducted experimental subjective tests [15,16] in a 3AFC speech recognition task. The three alternatives followed a

maximal opposition criterion: the two wrong alternatives differ from the spoken stimulus in manner, voicing, and place of articulation (for example, ata, ava, ama) (details are reported in: [12,21,22]) and subjects had to select their response among the three alternatives displayed on a touch-sensitive screen. Subjective data were collected from a sample of young, normal-hearing native listeners for each language. The sample size ranged from 20 to 25 participants across the five languages. The 12 VCVs used in this study were presented in each of the five languages, except ama and ana in Portuguese, which were not used in the original experiment. The sets of VCVs in each language were presented monaurally, one ear at a time, at SNRs ranging from −10 to +6 dB in 2 dB steps. At each SNR, VCVs were presented 20 times in a randomized sequence of 2160 stimuli per ear (12 VCVs × 20 presentations × 9 SNRs) for an approximate session duration of 45 min, and the intelligibility of each VCV as a function of SNR was estimated as the average percentage of correct responses at each SNR.

2.3. Analysis of Performance

To extract the speech reception threshold (SRT) and slope for each VCV, the average psychometric functions of VCVs in the five languages were estimated by fitting the mean STOI values (pooled, for each language, across VCVs) with a cumulative normal model (sigmoid function) [41]. The SRTs (dB) were estimated as the SNR values where the psychometric functions reached 79.4%, as per the definition of SRT in a 3AFC test [29]. The slopes (%/dB SNR) of the psychometric functions were estimated by using the following formula, which returns the value of the maximum slope on the cumulative normal function, i.e., at the point of inflection [42]:

$$slope = \frac{y_N - y_1}{\sqrt{2\pi}}\frac{1}{\sigma} \qquad (1)$$

where y_1 and y_N are the minimum and the maximum value of the psychometric function, respectively, and σ is the standard deviation of the cumulative normal function used to fit data.

To assess the quality of the estimates of intelligibility provided by the STOI measure, the Pearson product moment correlation coefficient, the Spearman correlation coefficient, and the root mean square error (RMSE) between mean STOI and mean VCV intelligibility from speech recognition tests were computed, as suggested by [43–45]. The mean and variance of STOI as a function of SNR and the estimated SRT and slope were compared across languages to characterize the intelligibility of VCVs in the five languages and to identify the set of stimuli that fulfilled the three criteria listed in Section 1.2.

3. Results

3.1. Accuracy of STOI Estimates

Tables 1 and 2 show the Pearson and the Spearman correlation coefficients, respectively, computed between mean STOI and mean estimated intelligibility for all the VCV recordings in the corpus. The correlation coefficients for the stimuli ama and ana in Portuguese were not available (n.a.) because these recordings were not used in the original experiment, as reported in Section 2.2. The correlation coefficients for the stimulus asa in Italian and German were n.a. because the distributions of subjective intelligibility were highly skewed towards saturation in the tested range of SNR. In general, a high linear correlation is observed, above 0.8 in most cases, confirming the accuracy of STOI estimates of VCV intelligibility. VCV recordings in English and Portuguese yield the highest values of correlation (e.g., Pearson correlation mean: 0.87 and 0.85; range: 0.74–0.98 across 12 VCVs and 0.75–0.94 across 10 VCVs, respectively). The Spearman correlation coefficients are slightly higher than the Pearson ones, suggesting a monotonous relationship between the STOI and subjective measures that is not necessarily linear.

Table 1. Pearson correlation coefficient between STOI and subjective measures.

VCVs	English	French	German	Italian	Portuguese
aba	0.90	0.83	0.92	0.80	0.94
ada	0.91	0.66	0.88	0.94	0.83
afa	0.96	0.95	0.90	0.96	0.89
aga	0.88	0.55	0.89	0.87	0.85
aka	0.83	0.72	0.91	0.99	0.91
ala	0.98	0.83	0.80	0.70	0.85
ama	0.74	0.56	0.63	0.69	n.a.
ana	0.75	0.59	0.67	0.88	n.a.
apa	0.94	0.87	0.98	0.83	0.86
ara	0.91	0.78	0.79	0.86	0.78
asa	0.92	0.55	n.a.	n.a.	0.75
ata	0.78	0.79	0.96	0.97	0.85
average	0.87	0.72	0.72	0.86	0.85
s.d.	0.08	0.14	0.45	0.10	0.06
range	0.74 ÷ 0.98	0.55 ÷ 0.95	0.65 ÷ 0.98	0.69 ÷ 0.99	0.75 ÷ 0.94

Table 2. Spearman correlation coefficient between STOI and subjective measures.

VCVs	English	French	German	Italian	Portuguese
aba	0.90	0.95	0.93	0.76	0.98
ada	0.88	0.76	0.93	0.98	0.93
afa	0.84	0.97	0.85	0.98	0.97
aga	0.74	0.18	0.95	0.87	0.90
aka	0.81	0.58	0.92	0.99	0.76
ala	0.96	0.88	0.72	0.77	0.71
ama	0.51	0.41	0.28	0.73	n.a.
ana	0.53	0.48	0.34	0.89	n.a.
apa	0.84	0.93	0.97	0.81	0.94
ara	0.95	0.65	0.65	0.97	0.96
asa	0.80	0.01	n.a.	n.a.	0.57
ata	0.78	0.91	0.98	1	0.97
average	0.79	0.64	0.77	0.88	0.87
s.d.	0.14	0.32	0.25	0.10	0.14
range	0.51 ÷ 0.96	0.01 ÷ 0.97	0.28 ÷ 0.98	0.73 ÷ 1	0.57 ÷ 0.98

Table 3 shows the RMSE computed between the STOI estimates and subjective estimates. The observed values of RMSE suggest VCVs in English are associated with the lowest error and, thus, the best prediction performance of the STOI. The set of VCVs in Portuguese yielded similar prediction errors as the one in English, whereas the poorest performance was observed for VCVs in Italian.

Table 3. RMSE between STOI and subjective measures.

	English	French	German	Italian	Portuguese
RMSE%	10.45	11.30	11.53	20.18	10.48

3.2. Characterization of VCV Intelligibility

Figure 2 shows the mean value of STOI (panel a) and the variance of STOI (panel b) as a function of the SNR for the five sets of VCVs used here. As the STOI is used here as an estimate of intelligibility, the mean and variance are reported as percent values rather than in the range 0–1. Figure 2a shows that there are differences in mean STOI across the five languages, with VCVs in English being, on average, those with the highest estimated intelligibility (highest mean STOI values throughout the SNR range), followed by VCVs in Italian, whereas VCVs in Portuguese were, on average, those with lowest estimated

intelligibility (lowest mean STOI values). English VCVs led, on average, to estimates of intelligibility from about 1% (at +6 dB SNR) to about 8% (at −12 dB SNR) higher compared to VCVs in Italian, which is the second language in terms of intelligibility. Compared to Portuguese, the improvement in intelligibility ranged from about 4% (at +6 dB SNR) to about 13% (at −12 dB SNR). On average, for negative SNRs, the intelligibility of English VCVs was from about 9% to about 10% higher compared to the average of the other languages.
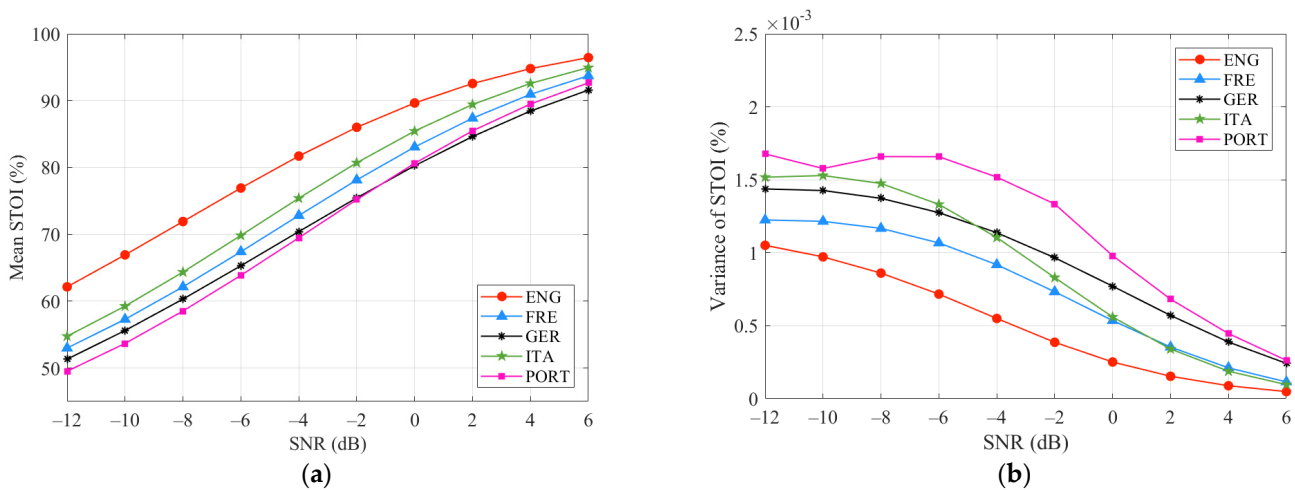

(**a**) (**b**)

Figure 2. Analysis of STOI values (percent values) as a function of the SNR (dB) observed in the five sets of VCVs, averaged within each set: (**a**) mean values of STOI; (**b**) standard deviation of STOI.

Figure 2b shows that the variance of STOI is low, on average, for all five sets of recordings (i.e., below 0.2% on average), suggesting that, in general, the influence of different noise realizations on STOI estimates of VCV intelligibility is limited in all the tested languages. VCVs in English showed the lowest variance and were, therefore, the least influenced by noise, whereas VCVs in Portuguese showed the highest variance of STOI, i.e., from about two to four times that of VCVs in English.

Table 4 reports the average, standard deviation, and range of the observed SRT and slope for VCVs in the five languages studied here. On average, the lowest (better) SRTs were obtained with the VCVs in English (i.e., −4.86 dB SNR), whereas the highest (worse) SRTs were obtained with VCVs in Portuguese (i.e., −0.85 dB SNR). Similarly, the slowest slopes were observed, on average, with VCVs in English (i.e., 2.58%/dB SNR), whereas the fastest slopes were observed with VCVs in Portuguese (i.e., 3.26%/dB SNR).

Table 4. Average, standard deviation, and range of the SRT (dB) and slope (%/dB SNR) of the psychometric functions.

	English		French		German		Italian		Portuguese	
	SRT	**Slope**	**SRT**	**Slope**	**SRT**	**Slope**	**SRT**	**Slope**	**SRT**	**Slope**
avg	−4.86	2.58	−1.83	3.05	−0.89	2.93	−2.82	3.06	−0.85	3.26
s.d.	1.74	0.47	2.18	0.47	3.10	0.56	2.42	0.71	3.31	0.63
range	−6.79÷−0.63	2.07÷3.52	−6.94÷2.34	2.05÷3.79	−6.96÷2.64	2.15÷3.82	−5.64÷3.10	2.19÷4.18	−4.36÷7.07	2.22÷4.30

The Kolmogorov–Smirnov test was conducted to check whether data were normally distributed. As the assumption of normality was not confirmed, the non-parametric Kruskal-Wallis test was performed to assess possible between-language differences in STOI mean and variance, SRTs, and slope. Statistical analysis confirmed that the mean STOI values for English VCVs (Figure 2a) are significantly higher than for VCVs in French, German, and Portuguese in the range from −8 to −2 dB SNR ($p < 0.05$), whereas the observed differences with VCVs in Italian were not statistically significant. Regarding the

variable of STOI (Figure 2b), statistical analysis showed that the observed differences were not statistically significant ($p > 0.05$). Statistical analysis of the distributions of SRTs shown in Table 4 revealed that the estimated SRTs were significantly lower with VCVs in English compared to those in French, German, and Portuguese ($p < 0.05$), whereas the observed difference between the average SRTs for VCVs in English and Italian was not statistically significant. Regarding the distributions of slope in Table 4, the differences in average values were small, and the range of slope values was similar across languages. Statistical analysis showed that the observed differences are not statistically significant ($p = 0.098$), although there is a trend for English to provide lower slope values compared to the other languages.

In summary, Figure 2 and Table 4 shows that among the five languages here studied, the set of spoken VCVs in English fulfilled the three criteria for application in multilingual settings, i.e., (1) high mean STOI (Figure 2a), (2) low variance of STOI (Figure 2b), and (3) shallow slope and low SRT (Table 4). As such, these results suggested that the subset of VCV recordings in English may be the best candidate language among the five languages here considered for the implementation of speech-in-noise tests to be administered to listeners whose native language is not known in advance.

4. Discussion

The results from the analysis of the VCV intelligibility using the STOI measure and validation with subjective listening experiments demonstrate that there are inherent differences in the intelligibility of spoken VCVs across the five different sets of recordings—English, French, German, Italian, and Portuguese. (Figure 2, Table 4). The estimated intelligibility of spoken VCVs recorded from the English speaker was, on average higher than those in other languages, especially at low SNRs (on average, from about 9% to about 10% higher at negative SNRs compared to the average of the other languages). This relatively higher intelligibility, especially at low SNRs, may be useful when testing non-native listeners as processing speech in a non-native language requires more effort, especially in adverse conditions, and the higher intelligibility of English VCVs at low SNRs may, at least in part, compensate for the increased demand.

In general, the STOI estimates of intelligibility are highly correlated with subjective intelligibility. However, it should be noted that the quantitative relationship between changes in STOI values and changes in subjective perception is not unique and may depend on the speech material used [46,47]. In other words, the minimum amount of increase (or decrease) in STOI that corresponds to a noticeable variation in speech perception is not known. However, in previous studies, changes in STOI values as small as 1% were considered meaningful [48]. As such, the differences in mean STOI observed suggest that there may be differences in VCV intelligibility across subsets, especially between VCVs in English versus those in French, German, and Portuguese. The lower intelligibility observed for stimuli in Portuguese may be due to particular language characteristics, as already pointed out in previous studies regarding, for example, nasal consonants [22,49,50].

Regarding the variance of STOI estimates, our results show that higher mean STOI is associated with lower variance throughout the SNR range (on average, from about 28% lower (at −12 dB SNR) to about 73% lower (at +6 dB SNR) for English compared to the average variance of the other languages). Low variance, i.e., limited variability of intelligibility with varying noise (as estimated in this study by the STOI variance), can be useful in a screening test to ensure better repeatability of results, especially at low SNR. Future studies will need to measure the actual repeatability of SRT estimates for spoken VCVs recorded from speakers of various languages in native and non-native listeners.

Our STOI analysis, together with the estimated SRTs and slopes of the psychometric curves, overall points at English as a good candidate language for the implementation of VCV-based speech-in-noise tests for screening to be delivered in multilingual settings. Indeed, psychometric functions of English VCVs show, on average, high intelligibility, low variability of intelligibility, and larger dynamics.

Of note, the STOI was considered in this study because it has been shown to result in better correlation with speech intelligibility in listening tests compared to the above-mentioned models, as they rely on the global statistics over longer segments than the 386 ms segments used by the STOI [33]. In addition, the use of STOI was recommended with respect to other intrusive measures for intelligibility prediction for users with assistive listening devices, and it was tested on three subjectively rated speech data sets covering reverberation-alone, noise-alone, and reverberation-plus-noise degradation conditions, as well as degradations resultant from nonlinear frequency compression and different speech enhancement strategies [43]. It will be important in future studies to address recently proposed intrusive measures, particularly the extended STOI [51] that can be applied to a wider range of input signals compared to STOI, as it is based on fewer assumptions.

The strength of the approach presented here relies on the combination of computational estimates (STOI) and experimental estimates from subjective listening tests (SRT) to characterize the intelligibility of a speech corpus, including stimuli in different languages. The experimental data support and validate the results obtained using the STOI model; however, performing listening tests undeniably demands a significant amount of time and resources. Therefore, considering the demonstrated high correlation between STOI estimates and experimental estimates tests in multiple experimental settings [43,52], the STOI measure remains a good surrogate for measuring human speech intelligibility whenever subjective tests are not feasible. It might be argued that, as an intrusive measure, STOI has the major shortcoming of requiring a clean reference signal which is not always available, thus it may limit its applicability in real-world scenarios. However, there is increasing use and development of STOI-based non-intrusive measures [53] which can be a viable solution if the clean reference speech cannot be accessed. Some of these measures may be used to expand the characterization of VCV stimuli here presented (reviewed in [53]), for example, the Coherence Speech-Intelligibility Index [54], the Normalized Covariance-Based Speech Transmission Index [55], and the Frequency-Weighted Segmental SNR [56].

The suitability of VCV recordings in English confirms evidence from previous studies where a significant degree of language independence in English VCV perception was shown, and acoustic and auditory factors have been shown to have a dominant influence in determining perceptual responses even across groups with different native languages (e.g., [32]). Moreover, the use of English VCVs within a speech-in-noise test to be used in multilingual settings is also supported by the widespread diffusion of English as a second language and by the widespread exposure of individuals to English (e.g., through the web and social media). Preliminary analysis of speech recognition performance, as measured using an adaptive 3AFC task, reveals that the SRTs measured in 12 native and 12 non-native English listeners with similar pure-tone average thresholds (difference $\leq$ 5 dB HL; mean difference = 1.67 dB HL $\pm$ 1.44) and similar age (difference $\leq$ 5 years; mean difference = 2 years $\pm$ 1.6) were similar (i.e., -9.1 vs. -9.2 dB SNR) (details are reported in [5]). The preliminary experimental evidence in non-native listeners reported in [5] confirms that the speech recognition performance of non-native listeners may be similar to that of native listeners when VCV recordings in English are used. It is important to note that such a screening procedure, based on speech stimuli viable for use in listeners of an unknown language, can substantially increase access to screening by eliminating possible barriers related to language, opening the delivery of screening tests at a distance able to identify undetected hearing loss with high accuracy [5,24]. For example, previous studies have considered an application of the adaptive 3AFC procedure to a population of 350 participants, including native and nonnative unscreened adults showing that the speech-in-noise test can accurately identify hearing loss of mild and moderate degree with accuracy up to 0.87 and 0.90, respectively [5,23–26,57]. The observed accuracy is higher than that observed for most of the available validated, language-specific speech-in-noise screening tests. For example, with the U.S. version of the digits-in-noise test, the observed sensitivity for mild hearing loss was equal to 0.80, and the specificity was equal to 0.83 [58]. With the original digits in the noise test in Dutch, the sensitivity for mild hearing loss was equal to 0.75,

and the specificity was 0.91 [13]. The Speech Understanding in Noise (SUN) test in Italian, administered sequentially in both ears, showed sensitivity equal to about 0.85 for moderate hearing loss and specificity equal to 0.85 [8,12].

Further research will be needed to more precisely characterize the measured speech recognition performance of non-native listeners as a function of their hearing sensitivity using VCV recordings of varying languages and using different experimental procedures (e.g., both fixed-level and adaptive procedures). Knowledge of the psychometric performance of non-native listeners might also help improve the adaptive procedure as the presentation levels may be adapted to the estimated individual intelligibility rather than on hypothetical psychometric models. Moreover, future experiments will be important to investigate the influence of the selection criteria used here, particularly the requirement of the shallow slope, on test sensitivity and efficiency and to assess whether the intelligibility of intervocalic consonants is not significantly influenced by the listeners' native language for the sake of hearing screening.

5. Conclusions

This study proposes an analysis of estimated VCV intelligibility using computational estimates such as the STOI measure in combination with experimental estimates from listening tests to identify specific sets of stimuli that may be applied in speech-in-noise testing in multilingual settings. Results show that the three criteria used for the testing procedure—high intelligibility, low variability of intelligibility, shallow slope and low SRT–are fulfilled by a subset of VCV stimuli in English when a 3AFC task is used (i.e., average intelligibility from 1% to 8% higher than the other languages; SRT from 4.01 to 2.04 dB SNR lower; average variability up to 4 times lower; slope from 0.35 to 0.68%/dB SNR lower). The criteria used here can help identify stimuli that may compensate for the additional speech recognition effort that non-native listeners may occur into. The results of this study can be used as a basis for developing speech-in-noise tests that can be administered to individuals of varying languages. Moreover, using an adaptive procedure using the VCVs in English in a 3AFC task, high accuracy for the identification of mild and moderate hearing loss has been documented. Preliminary findings from listening tests in individuals with varying degrees of hearing loss suggest a limited effect of native language on speech recognition performance using the selected subset of VCV recordings. Despite the encouraging results, this study has limitations, particularly in relation to the use of a single computational measure (the STOI) and in relation to the use of single VCV exemplars in each language. Widening the scope of the obtained findings, further research is needed to characterize the intelligibility of these stimuli using different computational measures and a more robust experimental validation in a large sample of non-native listeners with varying degrees of hearing loss to determine the possible effects of hearing loss and native language on VCV recognition.

Author Contributions: Conceptualization, G.R., T.v.W., R.B. and A.P.; methodology, G.R., T.v.W. and A.P.; software, G.R., G.B., R.A. and E.M.P.; validation, G.R., E.M.P. and A.P.; formal analysis, G.R., E.M.P. and A.P.; investigation, E.M.P. and A.P.; resources, T.v.W., A.P. and R.B.; data curation, G.R. and A.P.; writing—original draft preparation, G.R., E.M.P. and A.P.; writing—review and editing, G.R., G.B., R.A., T.v.W., E.M.P., R.B. and A.P.; visualization, G.R.; supervision, T.v.W., R.B. and A.P.; project administration, T.v.W., R.B. and A.P.; funding acquisition, T.v.W. and A.P. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available upon reasonable request to the Authors.

Conflicts of Interest: The authors declare no conflict of interests.

References

1. Mick, P.; Pichora-Fuller, M.K. Is Hearing Loss Associated with Poorer Health in Older Adults Who Might Benefit from Hearing Screening? *Ear Hear.* **2016**, *37*, e194–e201. [CrossRef] [PubMed]
2. Davis, A.; Smith, P.; Ferguson, M.; Stephens, D.; Gianopoulos, I. Acceptability, Benefit and Costs of Early Screening for Hearing Disability: A Study of Potential Screening Tests and Models. *Health Technol. Assess.* **2007**, *11*, A1–A32. [CrossRef]
3. Davis, A.; Smith, P. Adult Hearing Screening: Health Policy Issues—What Happens Next? *Am. J. Audiol.* **2013**, *22*, 167–170. [CrossRef] [PubMed]
4. Dubno, J.R.; Majumder, P.; Bettger, J.P.; Dolor, R.J.; Eifert, V.; Francis, H.W.; Pieper, C.F.; Schulz, K.A.; Silberberg, M.; Smith, S.L.; et al. A Pragmatic Clinical Trial of Hearing Screening in Primary Care Clinics: Cost-Effectiveness of Hearing Screening. *Cost. Eff. Resour. Alloc.* **2022**, *20*, 26. [CrossRef] [PubMed]
5. Paglialonga, A.; Polo, E.M.; Zanet, M.; Rocco, G.; van Waterschoot, T.; Barbieri, R. An Automated Speech-in-Noise Test for Remote Testing: Development and Preliminary Evaluation. *Am. J. Audiol.* **2020**, *29*, 564–576. [CrossRef] [PubMed]
6. Bastianelli, M.; Mark, A.E.; McAfee, A.; Schramm, D.; Lefrançois, R.; Bromwich, M. Adult Validation of a Self-Administered Tablet Audiometer. *J. Otolaryngol.—Head Neck Surg.* **2019**, *48*, 59. [CrossRef]
7. Liu, C.-F.; Collins, M.P.; Souza, P.E.; Yueh, B. Long-term cost-effectiveness of screening strategies for hearing loss. *J. Rehabil. Res. Dev.* **2011**, *48*, 235–243. [CrossRef] [PubMed]
8. Grandori, F.; Tognola, G.; Paglialonga, A. A Novel Approach for Adult Hearing Screening. *Audiol. Neurotol.* **2012**, *17*, 3–25. [CrossRef]
9. Humes, L.E. Understanding the Speech-Understanding Problems of Older Adults. *Am. J. Audiol.* **2013**, *22*, 303–305. [CrossRef]
10. Folmer, R.L.; Saunders, G.H.; Vachhani, J.J.; Margolis, R.H.; Saly, G.; Yueh, B.; McArdle, R.A.; Feth, L.L.; Roup, C.M.; Feeney, M.P. Hearing Health Care Utilization Following Automated Hearing Screening. *J. Am. Acad. Audiol.* **2021**, *32*, 235–245. [CrossRef]
11. Killion, M.C.; Niquette, P.A. What Can the Pure-Tone Audiogram Tell Us about a Patient's SNR Loss? *Hear. J.* **2000**, *53*, 46–48. [CrossRef]
12. Paglialonga, A.; Tognola, G.; Grandori, F. A User-Operated Test of Suprathreshold Acuity in Noise for Adult Hearing Screening: The SUN (Speech Understanding in Noise) Test. *Comp. Biol. Med.* **2014**, *52*, 66–72. [CrossRef]
13. Smits, C.; Houtgast, T. Results from the Dutch Speech-in-Noise Screening Test by Telephone. *Ear Hear.* **2005**, *26*, 89–95. [CrossRef] [PubMed]
14. Sanchez, V.A.; Arnold, M.L.; Moore, D.R.; Clavier, O.; Abrams, H.B. Speech-in-Noise Testing: Innovative Applications for Pediatric Patients, Underrepresented Populations, Fitness for Duty, Clinical Trials, and Remote Services. *J. Acoust. Soc. Am.* **2022**, *152*, 2336–2356. [CrossRef]
15. Paglialonga, A.; Cleveland Nielsen, A.; Ingo, E.; Barr, C.; Laplante-Lévesque, A. eHealth and the Hearing Aid Adult Patient Journey: A State-of-the-Art Review. *BioMed Eng. OnLine* **2018**, *17*, 101. [CrossRef] [PubMed]
16. Liu, H.; Du, B.; Liu, B.; Fu, X.; Wang, Y. Clinical Comparison of Two Automated Audiometry Procedures. *Front. Neurosci.* **2022**, *16*, 1011016. [CrossRef]
17. Warzybok, A.; Brand, T.; Wagener, K.C.; Kollmeier, B. How Much Does Language Proficiency by Non-Native Listeners Influence Speech Audiometric Tests in Noise? *Int. J. Audiol.* **2015**, *54*, 88–99. [CrossRef]
18. Blamey, P.J.; Blamey, J.K.; Saunders, E. Effectiveness of a Teleaudiology Approach to Hearing Aid Fitting. *J. Telemed. Telecare* **2015**, *21*, 474–478. [CrossRef]
19. Leensen, M.C.J.; de Laat, J.A.P.M.; Dreschler, W.A. Speech-in-Noise Screening Tests by Internet, Part 1: Test Evaluation for Noise-Induced Hearing Loss Identification. *Int. J. Audiol.* **2011**, *50*, 823–834. [CrossRef]
20. Leensen, M.C.J.; de Laat, J.A.P.M.; Snik, A.F.M.; Dreschler, W.A. Speech-in-Noise Screening Tests by Internet, Part 2: Improving Test Sensitivity for Noise-Induced Hearing Loss. *Int. J. Audiol.* **2011**, *50*, 835–848. [CrossRef] [PubMed]
21. Paglialonga, A.; Grandori, F.; Tognola, G. Using the Speech Understanding in Noise (SUN) Test for Adult Hearing Screening. *Am. J. Audiol.* **2013**, *22*, 171–174. [CrossRef]
22. Vaez, N.; Desgualdo-Pereira, L.; Paglialonga, A. Development of a Test of Suprathreshold Acuity in Noise in Brazilian Portuguese: A New Method for Hearing Screening and Surveillance. *BioMed Res. Intern.* **2014**, *2014*, 652838. [CrossRef] [PubMed]
23. Zanet, M.; Polo, E.M.; Rocco, G.; Paglialonga, A.; Barbieri, R. Development and Preliminary Evaluation of a Novel Adaptive Staircase Procedure for Automated Speech-in-Noise Testing. In Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC 2019), Berlin, Germany, 23–27 July 2019; pp. 6991–6994. [CrossRef]

24. Zanet, M.; Polo, E.M.; Lenatti, M.; van Waterschoot, T.; Mongelli, M.; Barbieri, R.; Paglialonga, A. Evaluation of a Novel Speech-in-Noise Test for Hearing Screening: Classification Performance and Transducers' Characteristics. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 4300–4307. [CrossRef]

25. Lenatti, M.; Moreno-Sánchez, P.A.; Polo, E.M.; Mollura, M.; Barbieri, R.; Paglialonga, A. Evaluation of Machine Learning Algorithms and Explainability Techniques to Detect Hearing Loss from a Speech-in-Noise Screening Test. *Am. J. Audiol.* **2022**, *31*, 961–979. [CrossRef] [PubMed]

26. Polo, E.M.; Lenatti, M.; Zanet, M.; Barbieri, R.; Paglialonga, A. Development and Characterization of a Novel Adaptive Staircase for Speech Recognition Testing. Unpublished work.

27. Bruns, T.; Ooster, J.; Stennes, M.; Rennies, J. Automated Speech Audiometry for Integrated Voice Over Internet Protocol Communication Services. *Am. J. Audiol.* **2022**, *31*, 980–992. [CrossRef]

28. Gordon-Salant, S. Hearing Loss and Aging: New Research Findings and Clinical Implications. *J. Rehabil. Res. Dev.* **2005**, *42*, 9. [CrossRef] [PubMed]

29. Leek, M.R. Adaptive Procedures in Psychophysical Research. *Percept. Psychophys.* **2001**, *63*, 1279–1292. [CrossRef]

30. Dubno, J.R.; Dirks, D.D. Evaluation of Hearing-Impaired Listeners Using a Nonsense-Syllable Test I. Test Reliability. *J. Speech Lang. Hear. Res.* **1982**, *25*, 135–141. [CrossRef] [PubMed]

31. Mattys, S.L.; Brooks, J.; Cooke, M. Recognizing Speech under a Processing Load: Dissociating Energetic from Informational Factors. *Cogn. Psychol.* **2009**, *59*, 203–243. [CrossRef]

32. Cooke, M.; Lecumberri, M.L.G.; Scharenborg, O.; van Dommelen, W.A. Language-Independent Processing in Speech Perception: Identification of English Intervocalic Consonants by Speakers of Eight European Languages. *Speech Commun.* **2010**, *52*, 954–967. [CrossRef]

33. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 2125–2136. [CrossRef]

34. Polo, E.M.; Mollura, M.; Paglialonga, A.; Barbieri, R. Listening effort: Cardiovascular investigation through the point process framework. In Proceedings of the 2022 Computing in Cardiology Conference (CinC 2022), Tampere, Finland, 4–7 September 2022. [CrossRef]

35. Eberhard, D.M.; Simons, G.F.; Fennig, C.D. *Ethnologue: Languages of the World*, 26th ed.; SIL International: Dallas, TX, USA, 2023.

36. *ISO8253-3:2012*; Acoustics Audiometric Test Methods Part 3: Speech Audiometry. International Organization for Standardization (ISO): Geneva, Switzerland, 2012.

37. Byrne, D.; Dillon, H.; Tran, K.; Arlinger, S.; Wilbraham, K.; Cox, R.; Hagerman, B.; Hetu, R.; Kei, J.; Lui, C.; et al. An International Comparison of Long-term Average Speech Spectra. *J. Acoust. Soc. Am.* **1994**, *96*, 2108–2120. [CrossRef]

38. Jensen, J.; Taal, C.H. Speech Intelligibility Prediction Based on Mutual Information. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 430–440. [CrossRef]

39. Li, J.; Xia, R.; Ying, D.; Yan, Y.; Akagi, M. Investigation of Objective Measures for Intelligibility Prediction of Noise-Reduced Speech for Chinese, Japanese, and English. *J. Acoust. Soc. Am.* **2014**, *136*, 3301–3312. [CrossRef]

40. Munro, M.J.; Derwing, T.M. Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Lang. Learn.* **1995**, *45*, 73–97. [CrossRef]

41. Lyregaard, P. Towards a theory of speech audiometry tests. In *Speech Audiometry*, 2nd ed.; Martin, M., Ed.; John Wiley & Sons: Chichester, UK, 1997; pp. 34–62.

42. Strasburger, H. Converting between Measures of Slope of the Psychometric Function. *Percept. Psychophys.* **2001**, *63*, 1348–1355. [CrossRef] [PubMed]

43. Falk, T.H.; Parsa, V.; Santos, J.F.; Arehart, K.; Hazrati, O.; Huber, R.; Kates, J.M.; Scollie, S. Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and Limitations of Existing Tools. *IEEE Signal. Process. Mag.* **2015**, *32*, 114–124. [CrossRef]

44. Bernardi, G.; van Waterschoot, T.; Wouters, J.; Moonen, M. Subjective and Objective Sound-Quality Evaluation of Adaptive Feedback Cancellation Algorithms. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1010–1024. [CrossRef]

45. Kates, J.; Arehart, K. The Hearing-Aid Speech Quality Index (HASQI) Version 2. *J. Audio Eng. Soc.* **2014**, *62*, 99–117. [CrossRef]

46. Manabe, Y.; Tamagawa, K.; Sato, K. Speech Intelligibility in Environmental Sound Maskers and Prediction Based on Envelope-Power Based Models. In Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC) 2018, Miyazaki, Japan, 7–10 October 2018. [CrossRef]

47. Van Kuyk, S.; Kleijn, W.B.; Hendriks, R.C. An Evaluation of Intrusive Instrumental Intelligibility Metrics. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 2153–2166. [CrossRef]

48. Kandagatla, R.K.; Subbaiah, P.V. Speech Enhancement Using MMSE Estimation of Amplitude and Complex Speech Spectral Coefficients under Phase-Uncertainty. *Speech Commun.* **2018**, *96*, 10–27. [CrossRef]

49. Kluge, D.; Reis, M. Intelligibility of Brazilian Portuguese-accented English Realization of Nasals in Word-final Position by Brazilian and Dutch EFL Learners. *CROP-Rev. Área Estud. Linguíst. E Lit. Ingl.* **2008**, *13*, 215–229.

50. Beddor, P.S.; Krakow, R.A.; Goldstein, L.M. Perceptual Constraints and Phonological Change: A Study of Nasal Vowel Height. *Phonol. Yearb.* **1986**, *3*, 197–217.

51. Jensen, J.; Taal, C.H. An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 2009–2022. [CrossRef]

52. Sharma, B.; Wang, Y. Automatic Evaluation of Song Intelligibility Using Singing Adapted STOI and Vocal-Specific Features. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 319–331. [CrossRef]

53. Feng, Y.; Chen, F. Nonintrusive Objective Measurement of Speech Intelligibility: A Review of Methodology. *Biomed. Signal. Process. Control* **2022**, *71*, 103204. [CrossRef]

54. Kates, J.M.; Arehart, K.H. Coherence and the Speech Intelligibility Index. *J. Acoust. Soc. Am.* **2005**, *117*, 2224–2237. [CrossRef]

55. Goldsworthy, R.L.; Greenberg, J.E. Analysis of Speech-Based Speech Transmission Index Methods with Implications for Nonlinear Operations. *J. Acoust. Soc. Am.* **2004**, *116*, 3679–3689. [CrossRef]

56. Hu, Y.; Loizou, P.C. Evaluation of Objective Quality Measures for Speech Enhancement. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 229–238. [CrossRef]

57. Polo, E.M.; Mollura, M.; Barbieri, R.; Paglialonga, A. Multivariate Classification of Mild and Moderate Hearing Loss Using a Speech-in-Noise Test for Hearing Screening at a Distance. In *HealthyIoT 2022: IoT Technologies for HealthCare*; Spinsante, S., Iadarola, G., Paglialonga, A., Tramarin, F., Eds.; Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering; Springer: Cham, Switzerland, 2023; Volume 456, pp. 81–92. [CrossRef]

58. Watson, C.S.; Kidd, G.R.; Miller, J.D.; Smits, C.; Humes, L.E. Telephone Screening Tests for Functionally Impaired Hearing: Current Use in Seven Countries and Development of a US Version. *J. Am. Acad. Audiol.* **2012**, *23*, 757–767. [CrossRef]