environments

MDPI

Article

Speech Identification and Comprehension in the Urban Soundscape

Letizia Marchegiani [1,*], Xenofon Fafoutis [2] and Sahar Abbaspour [3]

1 Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK
2 Department of Electrical and Electronic Engineering, University of Bristol, Bristol BS8 1UB, UK; xenofon.fafoutis@bristol.ac.uk
3 Volvo Car Corporation, 223 63 Lund, Sweden; sahar.abbaspour@volvocars.com
* Correspondence: letizia.marchegiani@eng.ox.ac.uk

Received: 15 March 2018; Accepted: 2 May 2018; Published: 7 May 2018

Abstract: Urban environments are characterised by the presence of copious and unstructured noise. This noise continuously challenges speech intelligibility both in normal-hearing and hearing-impaired individuals. In this paper, we investigate the impact of urban noise, such as traffic, on speech identification and, more generally, speech understanding. With this purpose, we perform listening experiments to evaluate the ability of individuals with normal hearing to detect words and interpret conversational speech in the presence of urban noise (e.g., street drilling, traffic jams). Our experiments confirm previous findings in different acoustic environments and demonstrate that speech identification is influenced by the similarity between the target speech and the masking noise also in urban scenarios. More specifically, we propose the use of the structural similarity index to quantify this similarity. Our analysis confirms that speech identification is more successful in presence of noise with tempo-spectral characteristics different from speech. Moreover, our results show that speech comprehension is not as challenging as word identification in urban sound environments that are characterised by the presence of severe noise. Indeed, our experiments demonstrate that speech comprehension can be fairly successful even in acoustic scenes where the ability to identify speech is highly reduced.

Keywords: speech identification; speech comprehension; speech intelligibility; urban soundscape; urban environments; masking; auditory perception

1. Introduction

The soundscape of modern urban environments is busy and full of noisy events such as traffic and construction works. Indeed, many major cities suffer from noise pollution [1], and there is significant evidence that this noise pollution contributes to several health disorders, including heart decease, diabetes and hearing loss, in millions of individuals [2,3]. At the same time, the urban soundscape is a rich source of information, and there is growing research interest in leveraging such information for realising the vision of autonomous driving in intelligent transportation systems [4]. Indeed, urban sound signals contain important cues that are vital for navigation in urban environments, yet difficult to obtain with traditional sensing modalities (e.g., cameras, lasers, etc.), such as horns or sirens from emergency vehicles [4–6].

The presence of copious and unstructured noise constitutes a continuous speech intelligibility challenge both in healthy and hearing-impaired individuals. Noise interferes with a speech signal at different levels, which are traditionally grouped into energetic masking and informational masking [7]. Energetic masking refers to the noise that physically interferes with the speech signal in the acoustic environment. Informational masking, on the other hand, refers to noise that perceptually

interferes with the speech signal as part of the cognitive process of perception in listeners. Not surprisingly, the effect of noise on speech intelligibility, defined as the percentage of a message understood correctly [8], goes beyond the effect of energetic masking. There is supporting evidence in the literature that speech intelligibility is particularly challenged in the presence of sudden or random unstructured noise, i.e., noise able to draw the attention of the listeners [9,10], as well as noise that creates linguistic confusion and increases cognitive load [11–13]. This work focuses on urban sound environments and investigates the impact of specific sources of urban noise on speech identification (i.e., the ability to identify specific words) and, more generally, on speech comprehension (i.e., the ability to understand content). We believe that a better understanding of the effects of background noise on speech recognition and understanding in real-world urban scenarios is fundamental for the development of effective and robust assistive technologies for people with hearing impairments [14]. Indeed, approximately 25% of hearing aid users report that they do not wear their hearing aids because they did not work effectively in hard listening situations, occasionally amplifying the background noise to unacceptable or even painful levels [15]. Furthermore, the analysis of the impact of urban noise on speech perception could also serve as a crucial element in the design of intelligent in-car noise-cancelling solutions, able to adapt to the characteristics of different maskers and their effects on speech intelligibility, both for listeners with normal and impaired hearing.

In this direction, we explore the ability to correctly perceive and interpret conversational speech in listeners without any hearing impairment, in the presence of urban noise, such as street traffic and drilling. Upon analysing data collected from 20 individuals, we present experimental evidence that indicates that in the presence of excessive urban noise, understanding particular words is very challenging, yet urban noise does not necessarily impact the ability of the listeners to understand content as severely as identifying the words. These results confirm that hearing aid technology would benefit from investigating the potential of improving the understanding of the relevant content rather than words in environments that are characterised by excessive background noise, such as urban scenarios.

In line with previous works that focus on different acoustic scenarios (e.g., [10,16]), our experiments indicate that word identification in urban environments is more successful under the presence of noise that is different from speech, as opposed to noise that is similar to speech. In particular, we employ the Structural Similarity (SSIM) index [17] to calculate the similarity between time-frequency visual representations (i.e., gammatone-based spectrograms [18]) of the speech and the noise signals. Our experiments suggest that there is a negative correlation between the SSIM index and the performance of the participants. The SSIM index is designed and traditionally used in image processing to identify the structural similarity of images, and to the extent of our knowledge, this is the first time in the literature that SSIM is applied to audio signals.

The remainder of this article is structured as follows. Section 2 briefly summarises the related work. Section 3 presents the experimental setup. Section 4 focuses on the analysis of collected data and presents the results. Lastly, Section 5 discusses the results, and Section 6 provides conclusive remarks.

2. Related Work

Speech perception in healthy and hearing-impaired individuals has been widely analysed in the literature, under several perspectives and for different goals. Much attention has been devoted to speech intelligibility against multi-talker noise and to the various factors that might influence it (for a general review, see [16]). In [13,19,20], the authors analyse consonant and sentence recognition under babble noise, exploring the role played by the proficiency of the listeners in the languages of the target speech. The impact of the characteristics of the speakers' voices has been explored in [11]. The influence of age, hearing loss and cognition on intelligibility and cognitive load has been investigated in [21]. Furthermore, the effect of different kinds of background music on word spotting has been presented in [10].

In recent years, noise pollution has become an increasingly important concern, instigating the development of new research areas in the attempt to reduce the negative effects of environmental noise, as well as safeguarding the health and well-being of people and, more generally, of the ecosystem. In [22], the authors investigate the way urban noise shapes animal communication systems, while [23] explores the effects on the human population, such as physiological responses, health outcomes, annoyance and sleep disturbance. Acoustic comfort has been analysed in [24], while [25] provides an extensive survey on sound preference towards laying the foundation of a more sustainable and human-friendly soundscape design. While we share the aspirations of these works, our focus is on speech perception considering different types of common urban noise. In particular, we aim to quantitatively analyse the distracting and masking effect of urban noise on speech identification and comprehension. Holmes et al. [26] investigated the relation between speech intelligibility and semantic context for hearing aid users. They investigated the effect of keeping a consistent topic across sentences on the reduction of subjects' effort when listening to competing talkers in a reverberant background. Their results show improved speech intelligibility for same-topic compared to different-topic sentences. Previous studies (e.g., [27]) proved that context improves speech intelligibility and that there is not necessarily a correlation between speech comprehension and speech intelligibility. In [28], the authors demonstrated that, in the case of babble background noise, while speech comprehension scores increase linearly with the noise level, speech intelligibility scores are characterised by a much higher variability, even for moderate noise conditions. Building on those results, we analyse the relation between speech identification and comprehension in an urban case scenario, comparing the effects and the efficacy of different traffic noise maskers. The literature on speech perception in urban scenarios is, at this time, quite limited. In [29], which is one of the few examples in this direction, intelligibility is based on the correct identification of the final word in noisy sentences, where the word predictability is low. Different from [29], in this paper, we focus on the identification of randomly-selected words with contextual weight. In [30], the authors investigate speech intelligibility in primary schools, in the presence of various types of noise that are common in a classroom, such as traffic noise, room babble and fan coil noise. Different from [30], our work focuses on street noise, such as traffic, construction work and car engine noise. Furthermore, rather than relying on more traditional speech intelligibility indexes (e.g., [31,32]), we propose a new measure of speech intelligibility for word spotting, which is based on applying image processing techniques to visual representations of the tempo-spectral content of the acoustic signals. More specifically, we employ the SSIM index [17] to investigate how the difference between the target and the masker signals can affect the ability to hear specific words. Our results confirm previous findings in different noisy conditions (e.g., [10,16]) and show that speech identification is conditioned by the similarity between the target and the masker sound.

3. Experimental Setup

To analyse how different types of urban noise affect speech perception, we have created stimuli using three types of urban noise. For these stimuli, we designed a set of experiments using the PsyToolkit [33,34]. We asked 20 participants to perform the experiments after they confirmed they understood the experiments. The participants' ages were between 25 and 45, and they did not suffer from any form of hearing impairment. All participants were either native English speakers or proficient L2 (second language) English speakers. Moreover, the participants performed the experiments in a silent environment using headphones. The experiments were as follows.

The participants listened to three short stories (each story about one minute long) where we added heavy traffic, construction work and running car engine noise to each story, respectively, i.e., a monaural composition of noise and target speech, where both stimuli are diffused. We added the noise to each story from the beginning of the story. We recorded these stories in a silent environment, with no environmental noise. The stimuli used to generate the noisy samples were extracted from the urban sound dataset [35]. The relative statistical levels of the noises are provided in Table 1.

Table 1. Statistical noise levels in decibels relative to full scale (dBFS) .

Level	Construction Work	Car Engine	Heavy Traffic
L10	−16.7	−16.2	−16.5
L90	−40.7	−39.1	−41.7

Figure 1 plots the spectrum of the maskers in one-third octave bands. Figure 2 shows the behaviour of the spectrum of the maskers over time. The target-to-masker ratio (i.e., Signal-to-Noise Ratio (SNR)) of each test was −5 dB, and all the tests were normalised to have the same Root Mean Square (RMS) energy (i.e., the RMS energy of the urban noise was set 5 dB lower than the RMS energy of the narrative). We performed some pilots to empirically choose the most appropriate value of the SNR, to make sure that the task was neither too easy, nor too hard. The participants were not able to pause or rewind the stories while being played. The stories were fairly simple and used everyday language that was easy to understand. After listening to the stories, we asked the participants a set of questions, in two separate tasks. During this step, the participants had the option of taking breaks while answering the questions, in an attempt to diminish any effect on the performance due to tiredness. The first task was word spotting. In this task, we presented a list of 20 words to the participants; the presented word list being a combination of random words with some correlation to the story line and specific words selected from each story with a normal distribution. The words selected directly from each story had a contextual weight, i.e., we did not present words such as "the" or "and" that appear in spoken English with a high frequency. The main purpose of choosing the words with correlation to the story line was to determine whether the participants relied on speculation rather than actual hearing when answering the questions. The participants specified which of the words they could recognise from the story they had just heard. The average frequency of the selected words in the British National Corpus [36] is 11,732, 10,678 and 8098 for each story, respectively. ANOVA tests did not find any significant difference between the stories ($p = 0.70$), nor between the true and false words (i.e., words actually present in the listening and words not part of the listening, but presented to the subjects for the word spotting task) in each story ($p = 0.91$, $p = 0.72$, $p = 0.75$, respectively). Lastly, the selected words were not phonetically balanced. For the second task, we asked the participants 5 questions with contextual relevance to each story line, with either "yes", "no", or "I don't know" as available answers.

Aiming to emulate a natural environment, before the actual experiments, a trial story was presented to the subjects, allowing them to adjust the volume. The trial story, characterised by the same SNR level and presenting the same structure of the real experiments (i.e., a narrator's voice masked by urban noise of a different kind from the one utilised in the real experiments), had also the goal of letting the subjects get familiar with the experiments, avoiding any "surprise effect" at the beginning of the real experiments, which might have caused the subjects to miss part of the listening, due to the attentive mechanisms operating.

At the end of the experiment, we asked the participants a set of multiple answer questions in a post-questionnaire. These questions mainly aimed at collecting information about the difficulties they experienced while listening to the stories and the potential reasons behind such difficulties. Their answers confirmed that the reductions in speech identification were due to the noise, rather than the loss of concentration or memory issues in remembering the specific words heard.

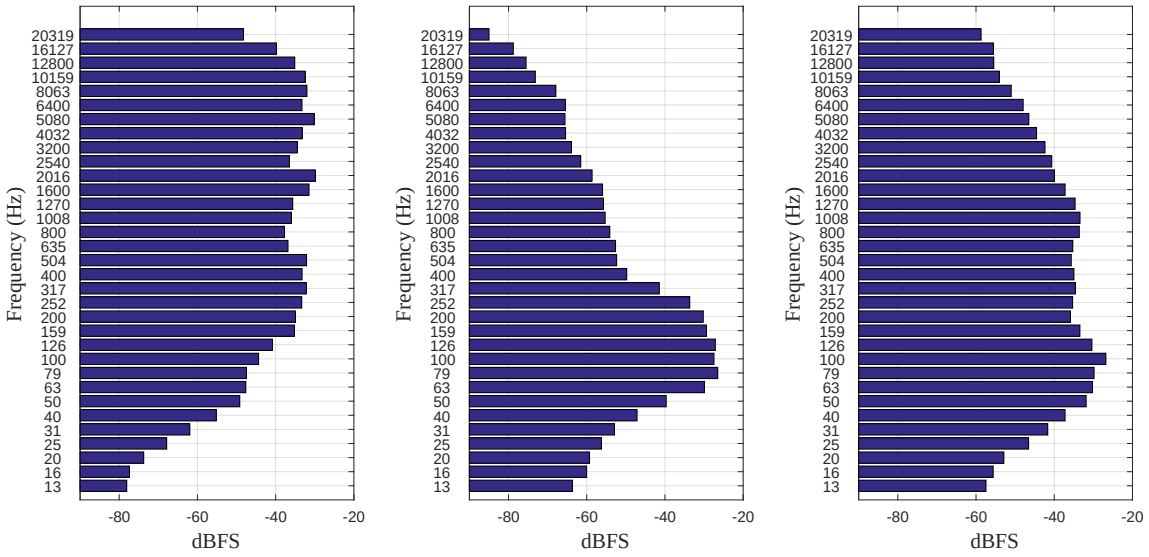


Figure 1. The spectrum of the three maskers in one-third octave bands: (**a**) Construction Work; (**b**) Car Engine; and (**c**) Heavy Traffic.

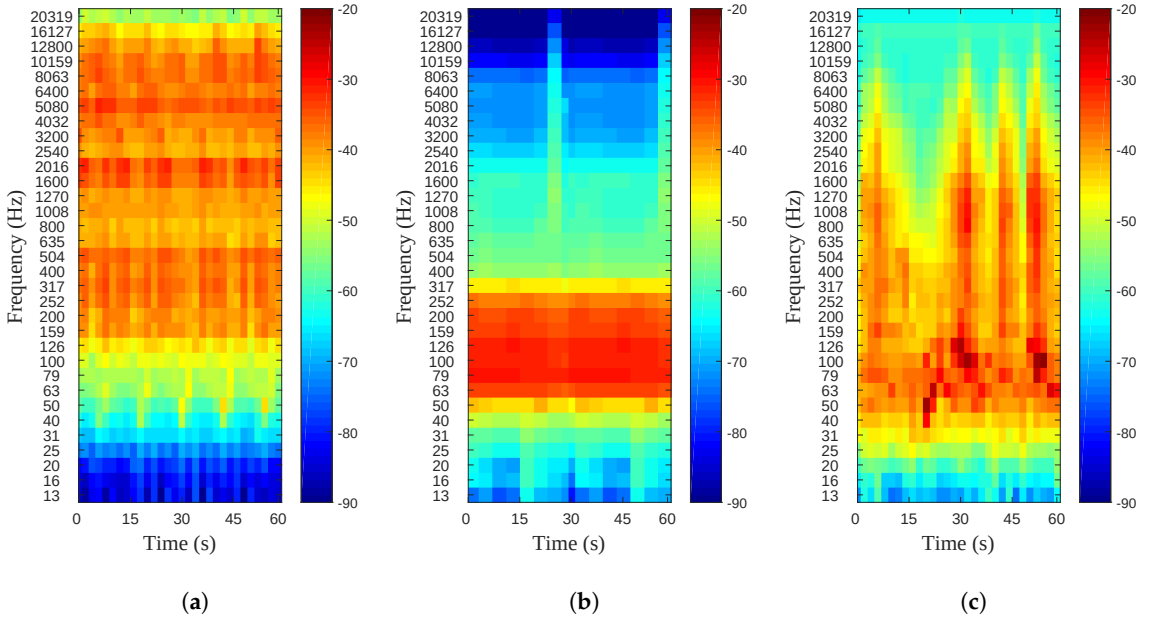


Figure 2. Spectrograms of the three maskers in one-third octave bands in dBFS: (**a**) Construction Work; (**b**) Car Engine; and (**c**) Heavy Traffic.

4. Results

In this section, we analyse the performance of our subjects in both tasks (i.e., word spotting and answering the questions). Figure 3 shows the true positive rate for all maskers. HT indicates the case of Heavy Traffic noise; CW refers to Construction Work noise; while CE indicates the case of Car Engine noise. Specifically, the figure reports the mean and standard deviation (on error bars) across all subjects. We observe that in all three cases, the subjects perform better in the question answering

task compared to the word spotting one. A repeated-measures Analysis of Variance (ANOVA) test with the task type as the factor confirmed the statistical significance of this difference ($p < 0.01$, degrees of freedom (DF) = 1). The Mauchly test verified that the sphericity condition was not violated. Furthermore, we note that the car engine noise results to be the most challenging one with regard to the word spotting task, while yielding the greatest performance in the context-related questions.

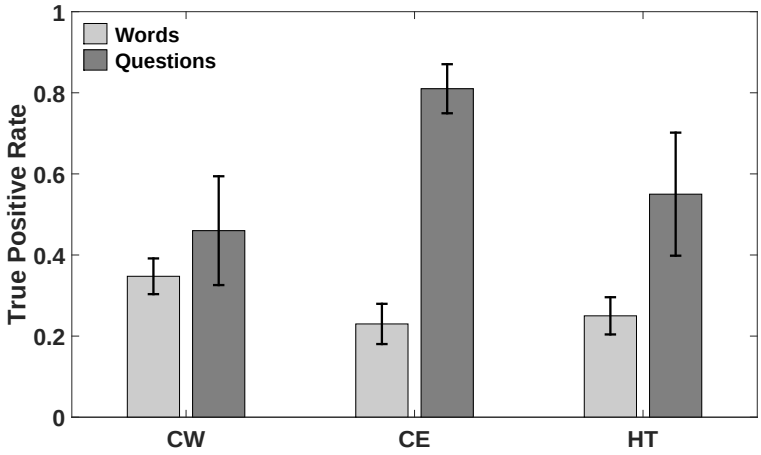
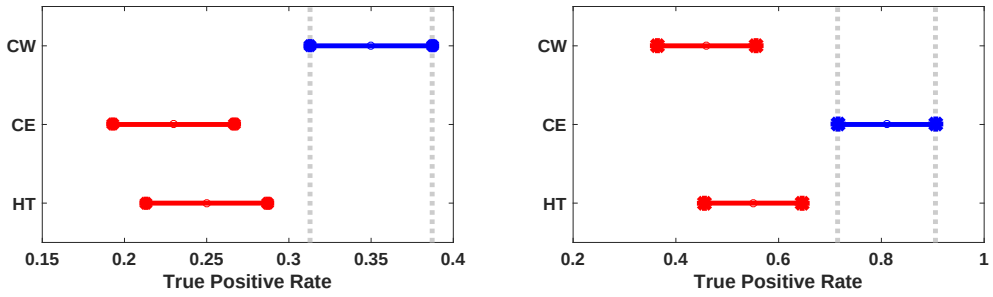


Figure 3. True positive rate for all maskers, both in the word spotting and question answering tasks. HT indicates the case of Heavy Traffic noise; CW refers to Construction Work noise; while CE indicates the case of Car Engine noise. The figure reports the mean and standard deviation (on error bars) across all subjects.

4.1. Word Spotting and Question Answering

Figure 3 suggests a significant difference in the performance of the subjects in the word spotting task in varying types of background noise. A repeated-measures ANOVA test with masker type as the factor confirmed the statistical significance of this difference ($p < 0.01$, DF = 2). The Mauchly test verified that the sphericity condition was not being violated. Car engine noise resulted in being the most challenging one for this task, while construction work noise was the one yielding the greatest performance. Post hoc tests (with Bonferroni adjustment for multiple comparisons) proved a significant difference ($p < 0.01$) in the performance of the subjects between the CW noise and the other cases, while no significance difference was observed between the HT and the CE cases. Figure 4a shows the result of the post hoc tests for this task.

Figure 3 also suggests a significant difference in the performance of the subjects in the question answering task with varying background noise. A repeated-measures ANOVA with masker type as the factor confirmed the statistical significance of this difference ($p < 0.01$, DF = 2). The Mauchly test verified that the sphericity condition was not being violated. The results from the story with construction work noise indicated it as the most challenging one for this task, while car engine noise was the one yielding the greatest performance. Post hoc tests (with Bonferroni adjustment for multiple comparisons) proved a significant difference ($p < 0.01$) in the performance of the subjects between the CE noise and the other cases, while no significant difference was observed between the CW and the HT cases. Figure 4b shows the result of the post hoc tests, providing a direct comparison of the performance of the subjects in the question answering task.



(**a**) Word Spotting Task (**b**) Question Answering Task

Figure 4. The figure shows the results of the post hoc analysis, both in the case of word spotting (**a**) and question answering (**b**) tasks, using the type of masker as the factor. The mean and standard deviation across all speakers for all three masking conditions are shown.

4.2. Masker Characterisation

To investigate the effect of the tempo-spectral characteristics of the maskers on the performance of the subjects in word spotting, we analysed the time-frequency content of the various stimuli. Specifically, we were interested in the analysis of all audio fragments in the stories containing each of the words presented to the subjects. Our tempo-spectral analysis relied on the use of gammatone filter banks [18]. The gammatone filter banks were first introduced in [37], as an approximation to the human cochlear frequency selectivity. These filter banks are often used in the literature in speech intelligibility and, more generally, psychoacoustic experiments (e.g., [38,39], among others). The impulse response of a gammatone filter centred at frequency f_c is:

$$g(t, f_c) = \begin{cases} t^{a-1} e^{-2\pi b t} \cos 2\pi f_c t & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where a indicates the order of the filter and b is the bandwidth, which increases as the centre frequency f_c increases. The frequency-dependent bandwidth yielded narrower filters at low frequencies and broader filters at high frequencies. Several investigations have been carried out to compute the values of the filters' parameters that best approximate the human auditory filter. In this work, following [40] and the implementation proposed by [41], we utilised fourth-order filters (i.e., $a = 4$) and approximated b as:

$$b = 1.09 \left(\frac{f_c}{9.26449} + 24.7 \right) \quad (2)$$

The centre frequencies f_c of the filters are distributed across the available spectrum in proportion to their bandwidth. The identification of those frequencies can be achieved by using the Equivalent Rectangular Bandwidth (ERB) scale [42]. In this work, we used gammatone filter banks to generate a visual representation of the Short-Time Fourier Transform (STFT) of the signals similar to common spectrograms, the gammatonegrams. Building on previous work [10,11], which suggested that stream segregation and, consequently, speech intelligibility were extremely challenged when the target sound and the masking noise had similar characteristics, we proposed a new method to quantify tempo-spectral similarities, by applying image processing techniques to the stimuli's gammatonegrams. In particular, we make use of the Structural Similarity (SSIM) index [17]. This method estimates the visual impact of shifts in image luminance and changes in contrast, as well as any other remaining errors, collectively identified as structural changes. Since its introduction, the SSIM index has been employed in a variety of image processing problems, which aimed to assess image quality or similarity, including image denoising [43], image restoration [44] and contrast enhancement [45]. In this work, we explore the possibility of measuring the difference in the time and frequency domain

of two audio signals, by comparing their gammatonegrams using the SSIM as the metric. Let **x** be the image signal related to the gammatonegram of the audio frame of a clean word uttered and **y** the image signal related to the gammatonegram of the correspondent audio frame of the masker overlapping with the word. We compute the SSIM between the two signals as:

$$SSIM(\mathbf{x}, \mathbf{y}) = l(\mathbf{x}, \mathbf{y}) + c(\mathbf{x}, \mathbf{y}) + s(\mathbf{x}, \mathbf{y}) \tag{3}$$

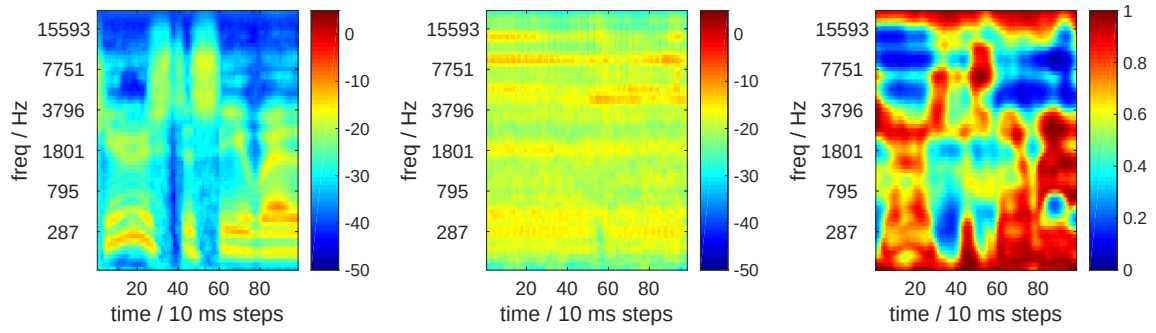
where $l(\mathbf{x}, \mathbf{y})$ refers to the luminance term, $c(\mathbf{x}, \mathbf{y})$ to the contrast term and $s(\mathbf{x}, \mathbf{y})$ to the structural term, and they are computed as:

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2\mu_y^2 + C1} \tag{4}$$

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{5}$$

$$s(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{6}$$

where μ_x and μ_y indicate the mean intensity of signal **x** and **y**, respectively. Moreover, σ_x, σ_y and σ_{xy} refer to the standard deviations and cross-covariances for the signal **x** and **y**. C_1 is a constant introduced to avoid instability when $\mu_x^2\mu_y^2$ is very close to zero. Similar reasoning is applied to the other constants C_2 and C_3. It is possible to generate a SSIM index map, by applying a sliding window over the two image signals and computing the SSIM index for each of these windows. It is also possible to employ different window structures to weight specific parts of the images differently. In this case, we follow the original approach proposed by [17]. Figures 5–7 show examples of gammatonegrams for the first, second and third story, respectively. These figures report both the clean utterance of one of the words presented to the subjects and the correspondent (overlapping) noise frame. The figures also illustrate the SSIM index map between the two gammatonegrams. In all three cases, local (frame related) $SNR = -5$ dB.

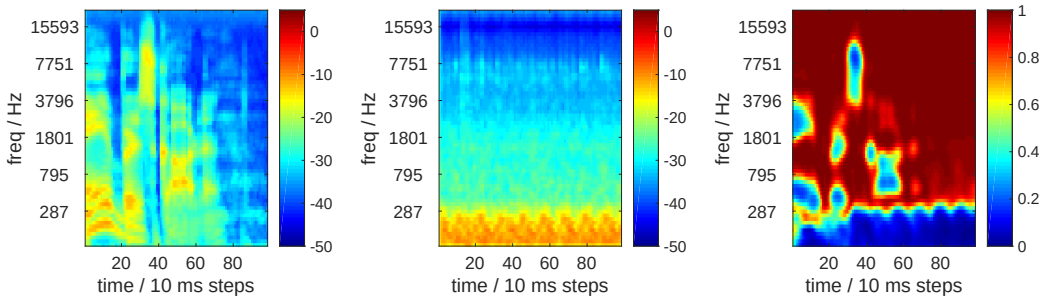


(**a**) First Story: Clean Speech (**b**) First Story: CW Noise (**c**) First Story: SSIM map

Figure 5. (**a**) The gammatonegram of one of the words uttered (and presented to the subjects in the experiments for the word spotting task) from the first story without any masker; (**b**) the gammatonegram of the masker (construction work noise) sound frame overlapping the word in (**a**); and (**c**) the SSIM map between the two gammatonegrams in (**a**,**b**). The Structural Similarity (SSIM) index is 0.59.

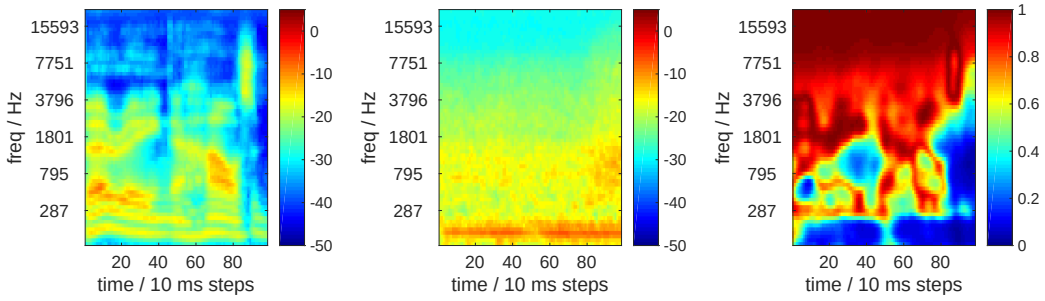
Figure 8 shows the SSIM indexes and the local SNR values for each of the words presented to the subjects across all of the stories. Figure 8a indicates a moderate negative correlation of -0.39 between the performance of the subjects (expressed as the average number of times each word was heard) and the SSIM indexes. A permutation test with 10,000 re-samples at the 5% significance level validated the results, proving the significance of the correlation. No significant correlation ($p = 0.13$)

was observed between the local SNR values and the performance of the subjects (expressed as the average number of times each word was heard).



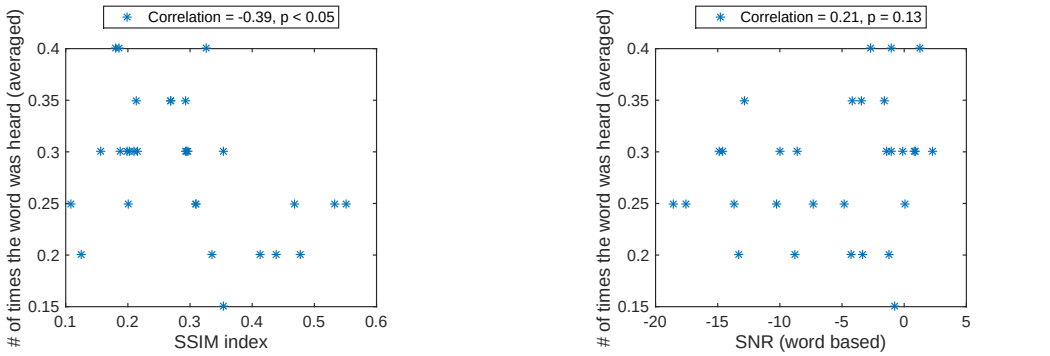
(**a**) Second Story: Clean Speech (**b**) Second Story: CE Noise (**c**) Second Story: SSIM map

Figure 6. (**a**) The gammatonegram of one of the words uttered (and presented to the subjects in the experiments for the word spotting task) from the second story without any masker; (**b**) the gammatonegram of the masker (car engine noise) sound frame overlapping the word in (**a**); and (**c**) the SSIM map between the two gammatonegrams in (**a**,**b**). The SSIM index is 0.78.



(**a**) Third Story: Clean Speech (**b**) Third Story: HT Noise (**c**) Third Story: SSIM map

Figure 7. (**a**) The gammatonegram of one of the words uttered (and presented to the subjects in the experiments for the word spotting task) from the third story without any masker; (**b**) the gammatonegram of the masker (heavy traffic) sound frame overlapping the word in (**a**); and (**c**) the SSIM map between the two gammatonegrams in (**a**,**b**). The SSIM index is 0.69.



(**a**) SSIM vs. Subjects' Performance (**b**) Local SNR values vs. Subjects' Performance

Figure 8. (**a**) The SSIM against the average number of times the words were identified for all stories; the correlation between them is shown on the legend; (**b**) local SNR values against the average number of times the words were identified for all stories; the correlation between them is shown on the legend.

5. Discussion

Our experiments aim to evaluate the relation between speech identification and comprehension in the presence of common urban noise maskers. The results confirm previous work in different acoustic scenarios that there is no correlation between the two aspects. Indeed, in the same noisy conditions, subjects perform significantly different in the two tasks (i.e., word spotting and question answering). In particular, in the word spotting task, construction work noise seems to be the least challenging masker, while in the question answering task, the listeners were most successful in the presence of car engine noise. It is interesting to notice that the subjects perform quite poorly in word spotting against car engine noise. Several factors could explain this behaviour. When noisy conditions are extremely challenging, the attentive effort and the cognitive load necessary to interpret the speech stimuli are greater. As there is evidence that human processing of spoken language is hierarchically organised [46], it is possible, then, that the subjects tend to focus on detecting and discriminating words, without being able to proceed at higher levels of the processing ladder. On the other hand, when the noisy condition is not as hard, the subjects are able to analyse the speech stimuli at a higher level of abstraction and improve speech comprehension.

In the attempt at quantifying the masking power of the different noisy conditions utilised, we use the SSIM index (a metric for comparing the structural similarity of images) on visual representations of the speech and the noise signals. The experiments suggest that the SSIM index is a more effective predictor of word intelligibility than the SNR. This result is in line with the previous work (e.g., [10,16,47], among others) and highlights the fact that the tempo-spectral similarity between the target and the noise signals is what best characterises the capability of discriminating words against a specific masker. In addition, we believe that precedent exposure to the maskers might have played a role. Indeed, previous investigations (e.g., [48]) showed that auditory perception in noise can be improved by training. Extensive exposure could result in involuntary training, as well, making the more common noise conditions, such as a car engine, less distracting and influential. Lastly, we observe that performance is generally higher in the question answering task. One explanation for this can be the natural redundancy of speech, which allows humans to understand the meaning of a sentence without necessarily identifying all of its words.

6. Conclusions

In this work, we explored speech identification and comprehension in the urban soundscape. More specifically, we carried out listening experiments to evaluate the masking effect of different types of common urban noise on everyday conversational speech. Our evaluation did not identify any statistically-significant relationship between speech understanding and word identification (i.e., higher speech identification does not necessarily translate to more accurate speech comprehension), suggesting that the nature of the noise and behavioural aspects play a major role in the listening performance. To further investigate the reasons behind those findings, we proposed a new masker characterisation metric, using a similarity index traditionally used in the image processing literature, the Structural Similarity index (SSIM). SSIM aims to quantify the difference between tempo-spectral representations of the target signal and the masker, analysing luminance, contrast and structure. Our experiments suggest that word identification is dependent on the similarity between the target signal and the masker. In particular, higher structural similarity decreases the identification rate. These encouraging results confirm a relationship that has been demonstrated in different scenarios (e.g., [10,16,47]) in the context of urban environments and acts to support evidence on the validity of the SSIM to characterise maskers. It is the authors' hope that future studies will validate this approach in different contexts. Future work could also investigate if SSIM can be used to predict attention bias. In addition, future work could explore the impact of linguistic related factors naturally operating in the same setting, encouraging research towards hearing assistance solutions (e.g., hearing aids) that do not only focus on improving the hearability of words, but also sentence-based meaning extrapolation and understanding. Furthermore, future work could explore

speech intelligibility in urban environments using rigorous intelligibility tests [49] and the potential of other similarity metrics (e.g., Kullback–Leibler divergence) to quantify the relationship between the masker and the target signal. Lastly, it would be interesting to evaluate the effect of other elements of the environment, such as the directional characteristics of the noise, in more natural settings [50,51].

Author Contributions: L.M. and S.A. conceived of and designed the experiments. All authors contributed to the implementation of the experiments. L.M. and X.F. analysed the data. All authors wrote the paper.

Funding: No funding sponsor had any role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

Acknowledgments: The authors wish to thank Fredrik Strömbergsson for volunteering to act as a narrator and Gijsbert Stoet for upgrading PhyToolkit to support our experiments. The authors would also like to thank all the participants of the experiments.

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

References

1. Hammer, M.S.; Swinburn, T.K.; Neitzel, R.L. Environmental Noise Pollution in the United States: Developing an Effective Public Health Response. *Environ. Health Perspect.* **2014**, *122*, 115–119. [CrossRef] [PubMed]
2. Sørensen, M.; Andersen, Z.J.; Nordsborg, R.B.; Becker, T.; Tjønneland, A.; Overvad, K.; Raaschou-Nielsen, O. Long-Term Exposure to Road Traffic Noise and Incident Diabetes: A Cohort Study. *Environ. Health Perspect.* **2013**, *121*, 217–222. [PubMed]
3. Passchier-Vermeer, W.; Passchier, W.F. Noise exposure and public health. *Environ. Health Perspect.* **2010**, *108*, 123–131. [CrossRef]
4. Marchegiani, L.; Posner, I. Leveraging the urban soundscape: Auditory perception for smart vehicles. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 6547–6554.
5. Meucci, F.; Pierucci, L.; Re, E.D.; Lastrucci, L.; Desii, P. A real-time siren detector to improve safety of guide in traffic environment. In Proceedings of the 16th European Signal Processing Conference, Lausanne, Switzerland, 25–29 August 2008; pp. 1–5.
6. Schröder, J.; Goetze, S.; Grützmacher, V.; Anemüller, J. Automatic acoustic siren detection in traffic noise by part-based models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 493–497.
7. Lidestam, B.; Holgersson, J.; Moradi, S. Comparison of informational vs. energetic masking effects on speechreading performance. *Front. Psychol.* **2014**, *5*, 639. [CrossRef] [PubMed]
8. International Organization for Standardization. Ergonomics—Assessment of Speech Communication. ISO9921, 2013. Available online: https://www.iso.org/standard/33589.html (accessed on 4 May 2018).
9. Stone, M.A.; Füllgrabe, C.; Mackinnon, R.C.; Moore, B.C.J. The importance for speech intelligibility of random fluctuations in "steady" background noise. *J. Acoust. Soc. Am.* **2011**, *130*, 2874–2881. [CrossRef] [PubMed]
10. Marchegiani, L.; Fafoutis, X. A Behavioral Study on the Effects of Rock Music on Auditory Attention. In Proceedings of the International Workshop on Human Behavior Understanding, Barcelona, Spain, 22 October 2013; pp. 15–26.
11. Moore, B.C.J.; Gockel, H. Factors Influencing Sequential Stream Segregation. *Acta Acust. United Acust.* **2002**, *88*, 320–333.
12. Cooke, M.; Lecumberri, M.L.G.; Barker, J. The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *J. Acoust. Soc. Am.* **2008**, *123*, 414–427. [CrossRef] [PubMed]
13. Marchegiani, L.; Fafoutis, X. On cross-language consonant identification in second language noise. *J. Acoust. Soc. Am.* **2015**, *138*, 2206–2209. [CrossRef] [PubMed]
14. Levitt, H. Noise reduction in hearing aids: A review. *J. Rehabilit. Res. Dev.* **2001**, *21*, 111–121.
15. Kochkin, S. MarkeTrak V: "Why my hearing aids are in the drawer" The consumers' perspective. *Hear. J.* **2000**, *52*, 34–41. [CrossRef]

16. Bronkhorst, A.W. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acust. United Acust.* **2000**, *86*, 117–128.
17. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
18. Lyon, R.F.; Katsiamis, A.G.; Drakakis, E.M. History and future of auditory filter models. In Proceedings of the IEEE International Symposium on Circuits and Systems, Paris, France, 30 May–2 June 2010; pp. 3809–3812.
19. Van Engen, K.J.; Bradlow, A.R. Sentence recognition in native-and foreign-language multi-talker background noise. *J. Acoust. Soc. Am.* **2007**, *121*, 519–526. [CrossRef] [PubMed]
20. Lecumberri, M.G.; Cooke, M. Effect of masker type on native and non-native consonant perception in noise. *J. Acoust. Soc. Am.* **2006**, *119*, 2445–2454. [CrossRef]
21. Zekveld, A.A.; Kramer, S.E.; Festen, J.M. Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear Hear.* **2011**, *32*, 498–510. [CrossRef] [PubMed]
22. Warren, P.S.; Katti, M.; Ermann, M.; Brazel, A. Urban bioacoustics: It's not just noise. *Anim. Behav.* **2006**, *71*, 491–502. [CrossRef]
23. Stansfeld, S.; Haines, M.; Brown, B. Noise and health in the urban environment. *Rev. Environ. Health* **2000**, *15*, 43–82. [CrossRef] [PubMed]
24. Yang, W.; Kang, J. Acoustic comfort evaluation in urban open public spaces. *Appl. Acoust.* **2005**, *66*, 211–229. [CrossRef]
25. Yang, W.; Kang, J. Soundscape and sound preferences in urban squares: A case study in Sheffield. *J. Urban Des.* **2005**, *10*, 61–80. [CrossRef]
26. Holmes, E.; Folkeard, P.; Johnsrude, I.S.; Scollie, S. Semantic context improves speech intelligibility and reduces listening effort for listeners with hearing impairment. *Int. J. Audiol.* **2018**, doi:10.1080/14992027.2018.1432901. [CrossRef] [PubMed]
27. Miller, G.A.; Heise, G.A.; Lichten, W. The intelligibility of speech as a function of the context of the test materials. *J. Exp. Psychol.* **1951**, *41*, 329. [CrossRef] [PubMed]
28. Fontan, L.; Tardieu, J.; Gaillard, P.; Woisard, V.; Ruiz, R. Relationship between speech intelligibility and speech comprehension in babble noise. *J. Speech Lang. Hear. Res.* **2015**, *58*, 977–986. [CrossRef] [PubMed]
29. Davies, W.; Mahnken, P.; Gamble, P.; Plack, C. Measuring and mapping soundscape speech intelligibility. In Proceedings of the Euronoise 2009, Edinburgh, UK, 26–28 October 2009.
30. Astolfi, A.; Bottalico, P.; Barbato, G. Subjective and objective speech intelligibility investigations in primary school classrooms. *J. Acoust. Soc. Am.* **2012**, *131*, 247–257. [CrossRef] [PubMed]
31. Cooke, M. A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.* **2006**, *119*, 1562–1573. [CrossRef] [PubMed]
32. 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]
33. Stoet, G. PsyToolkit: A software package for programming psychological experiments using Linux. *Behav. Res. Methods* **2010**, *42*, 1096–1104. [CrossRef] [PubMed]
34. Stoet, G. PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Teach. Psychol.* **2017**, *44*, 24–31. [CrossRef]
35. Salamon, J.; Jacoby, C.; Bello, J.P. A Dataset and Taxonomy for Urban Sound Research. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3 November 2014.
36. Davies, M. *BYU-BNC; Based on the British National Corpus from Oxford University Press*; Oxford University Press: Oxford, UK, 2004. Available online: https://corpus.byu.edu/bnc/ (accessed on 4 May 2018).
37. Holdsworth, J.; Nimmo-Smith, I.; Patterson, R.; Rice, P. Implementing a Gammatone Filter Bank. Available online: https://www.pdn.cam.ac.uk/other-pages/cnbh/files/publications/SVOSAnnexC1988.pdf (accessed on 15 March 2018).
38. Kjems, U.; Boldt, J.B.; Pedersen, M.S.; Lunner, T.; Wang, D. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *J. Acoust. Soc. Am.* **2009**, *126*, 1415–1426. [CrossRef] [PubMed]
39. Marchegiani, L.; Karadogan, S.G.; Andersen, T.; Larsen, J.; Hansen, L.K. The role of top-down attention in the cocktail party: Revisiting cherry's experiment after sixty years. In Proceedings of the 10th International Conference on Machine Learning and Applications and Workshops (ICMLA), Honolulu, HI, USA, 18–21 December 2011; Volume 1, pp. 183–188.

40. Toshio, I. An optimal auditory filter. In Proceedings of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 15–18 October 1995; pp. 198–201.
41. Ellis, D.P.W. "Gammatone-Like Spectrograms". 2009. Available online: http://www.ee.columbia.edu/dpwe/resources/matlab/gammatonegram/ (accessed on 15 March 2018).
42. Glasberg, B.R.; Moore, B.C. Derivation of auditory filter shapes from notched-noise data. *Hear. Res.* **1990**, *47*, 103–138. [CrossRef]
43. Rehman, A.; Wang, Z.; Brunet, D.; Vrscay, E.R. SSIM-inspired image denoising using sparse representations. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 1121–1124.
44. Channappayya, S.S.; Bovik, A.C.; Caramanis, C.; Heath, R.W. SSIM-optimal linear image restoration. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, NV, USA, 31 March–4 April 2008; pp. 765–768.
45. Avanaki, A.N. Exact global histogram specification optimized for structural similarity. *Opt. Rev.* **2009**, *16*, 613–621. [CrossRef]
46. Davis, M.H.; Johnsrude, I.S. Hierarchical processing in spoken language comprehension. *J. Neurosci.* **2003**, *23*, 3423–3431. [CrossRef] [PubMed]
47. Drullman, R.; Bronkhorst, A.W. Speech perception and talker segregation: Effects of level, pitch, and tactile support with multiple simultaneous talkers. *J. Acoust. Soc. Am.* **2004**, *116*, 3090–3098. [CrossRef] [PubMed]
48. Song, J.H.; Skoe, E.; Banai, K.; Kraus, N. Training to improve hearing speech in noise: Biological mechanisms. *Cerebral Cortex* **2011**, *22*, 1180–1190. [CrossRef] [PubMed]
49. Kollmeier, B.; Warzybok, A.; Hochmuth, S.; Zokoll, M.A.; Uslar, V.; Brand, T.; Wagener, K.C. The multilingual matrix test: Principles, applications, and comparison across languages: A review. *Int. J. Audiol.* **2015**, *54*, 3–16. [CrossRef] [PubMed]
50. Brungart, D.S.; Sheffield, B.M.; Kubli, L.R. Development of a test battery for evaluating speech perception in complex listening environments. *J. Acoust. Soc. Am.* **2014**, *136*, 777–790. [CrossRef] [PubMed]
51. Keidser, G. Introduction to Special Issue: Towards Ecologically Valid Protocols for the Assessment of Hearing and Hearing Devices. *J. Am. Acad. Audiol.* **2016**, *27*, 502–503. [CrossRef] [PubMed]